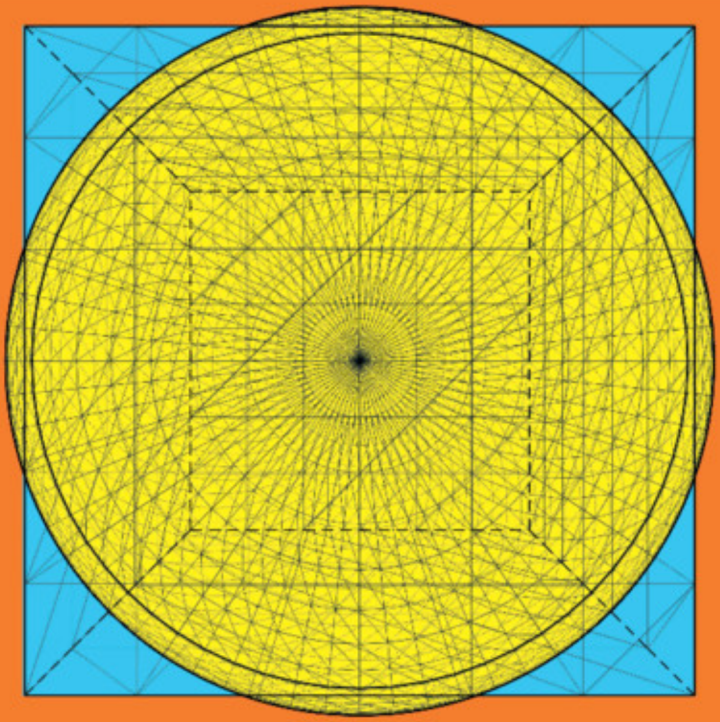
Foundations of Data Science

Avrim Blum
John Hopcroft
Ravindran Kannan



Foundations of Data Science

Avrim Blum

Toyota Technological Institute at Chicago

John Hopcroft

Cornell University, New York

Ravindran Kannan

Microsoft Research, India

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108485067
DOI: 10.1017/9781108755528

First published 2020

Printed in the United Kingdom by TJ International, Padstow Cornwall

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Names: Blum, Avrim, 1966– author. | Hopcroft, John E., 1939– author. | Kannan, Ravindran, 1953– author.
Title: Foundations of data science / Avrim Blum, Toyota Technological Institute at Chicago, John Hopcroft, Cornell University, New York, Ravindran Kannan, Microsoft Research, India.
Description: First edition. | New York, NY : Cambridge University Press, 2020. | Includes bibliographical references and index.
Identifiers: LCCN 2019038133 (print) | LCCN 2019038134 (ebook) | ISBN 9781108485067 (hardback) | ISBN 9781108755528 (epub)
Subjects: LCSH: Computer science. | Statistics. | Quantitative research.
Classification: LCC QA76 .B5675 2020 (print) | LCC QA76 (ebook) | DDC 004–dc23
LC record available at https://lccn.loc.gov/2019038133
LC ebook record available at https://lccn.loc.gov/2019038134

ISBN 978-1-108-48506-7 Hardback

Contents

CHAPTER ONE

Introduction

Computer science as an academic discipline began in the 1960s. Emphasis was on programming languages, compilers, operating systems, and the mathematical theory that supported these areas. Courses in theoretical computer science covered finite automata, regular expressions, context-free languages, and computability. In the 1970s, the study of algorithms was added as an important component of theory. The emphasis was on making computers useful. Today, a fundamental change is taking place and the focus is more on a wealth of applications. There are many reasons for this change. The merging of computing and communications has played an important role. The enhanced ability to observe, collect, and store data in the natural sciences, in commerce, and in other fields calls for a change in our understanding of data and how to handle it in the modern setting. The emergence of the web and social networks as central aspects of daily life presents both opportunities and challenges for theory.

While traditional areas of computer science remain highly important, increasingly researchers of the future will be involved with using computers to understand and extract usable information from massive data arising in applications, not just how to make computers useful on specific well-defined problems. With this in mind we have written this book to cover the theory we expect to be useful in the next 40 years, just as an understanding of automata theory, algorithms, and related topics gave students an advantage in the last 40 years. One of the major changes is an increase in emphasis on probability, statistics, and numerical methods.

Early drafts of the book have been used for both undergraduate and graduate courses. Background material needed for an undergraduate course has been put into a background chapter with associated homework problems.

Modern data in diverse fields such as information processing, search, and machine learning is often advantageously represented as vectors with a large number of components. The vector representation is not just a book-keeping device to store many fields of a record. Indeed, the two salient aspects of vectors – geometric (length, dot products, orthogonality, etc.) and linear algebraic (independence, rank, singular values, etc.) – turn out to be relevant and useful. Chapters 2 and 3 lay the foundations of geometry and linear algebra, respectively. More specifically, our intuition from two- or three-dimensional space can be surprisingly off the mark when it comes to high dimensions. Chapter 2 works out the fundamentals needed to understand the differences. The emphasis of the chapter, as well as the book in general, is to get across the intellectual ideas and the mathematical foundations rather than focus

on particular applications, some of which are briefly described. Chapter 3 focuses on singular value decomposition (SVD) a central tool to deal with matrix data. We give a from-first-principles description of the mathematics and algorithms for SVD. Applications of singular value decomposition include principal component analysis, a widely used technique we touch on, as well as modern applications to statistical mixtures of probability densities, discrete optimization, etc., which are described in more detail.

Exploring large structures like the web or the space of configurations of a large system with deterministic methods can be prohibitively expensive. Random walks (also called Markov chains) turn out often to be more efficient as well as illuminative. The stationary distributions of such walks are important for applications ranging from web search to the simulation of physical systems. The underlying mathematical theory of such random walks, as well as connections to electrical networks, forms the core of Chapter 4 on Markov chains.

One of the surprises of computer science over the last two decades is that some domain-independent methods have been immensely successful in tackling problems from diverse areas. Machine learning is a striking example. Chapter 5 describes the foundations of machine learning, both algorithms for optimizing over given training examples as well as the theory for understanding when such optimization can be expected to lead to good performance on new, unseen data. This includes important measures such as the Vapnik–Chervonenkis dimension; important algorithms such as the Perceptron Algorithm, stochastic gradient descent, boosting, and deep learning; and important notions such as regularization and overfitting.

The field of algorithms has traditionally assumed that the input data to a problem is presented in random access memory, which the algorithm can repeatedly access. This is not feasible for problems involving enormous amounts of data. The streaming model and other models have been formulated to reflect this. In this setting, sampling plays a crucial role and, indeed, we have to sample on the fly. In Chapter 6 we study how to draw good samples efficiently and how to estimate statistical and linear algebra quantities with such samples.

While Chapter 5 focuses on supervised learning, where one learns from labeled training data, the problem of unsupervised learning, or learning from unlabeled data, is equally important. A central topic in unsupervised learning is clustering, discussed in Chapter 7. Clustering refers to the problem of partitioning data into groups of similar objects. After describing some of the basic methods for clustering, such as the k-means algorithm, Chapter 7 focuses on modern developments in understanding these, as well as newer algorithms and general frameworks for analyzing different kinds of clustering problems.

Central to our understanding of large structures, like the web and social networks, is building models to capture essential properties of these structures. The simplest model is that of a random graph formulated by Erdös and Renyi, which we study in detail in Chapter 8, proving that certain global phenomena, like a giant connected component, arise in such structures with only local choices. We also describe other models of random graphs.

Chapter 9 focuses on linear-algebraic problems of making sense from data, in particular topic modeling and nonnegative matrix factorization. In addition to discussing well-known models, we also describe some current research on models and

probability that the difference will exceed ϵ and hence ϵ is in the denominator. Notice that squaring ϵ makes the fraction a dimensionless quantity.

We use two inequalities to prove the Law of Large Numbers. The first is Markov's inequality that states that the probability that a nonnegative random variable exceeds a is bounded by the expected value of the variable divided by a.

Theorem 2.1 (Markov's inequality) *Let x be a nonnegative random variable. Then for $a > 0$,*

$$Prob(x \geq a) \leq \frac{E(x)}{a}.$$

Proof For a continuous nonnegative random variable x with probability density p,

$$E(x) = \int_0^\infty xp(x)dx = \int_0^a xp(x)dx + \int_a^\infty xp(x)dx$$

$$\geq \int_a^\infty xp(x)dx \geq a\int_a^\infty p(x)dx = a\text{Prob}(x \geq a).$$

Thus, $\text{Prob}(x \geq a) \leq \frac{E(x)}{a}$. ■

The same proof works for discrete random variables with sums instead of integrals.

Corollary 2.2 $Prob(x \geq bE(x)) \leq \frac{1}{b}$

Markov's inequality bounds the tail of a distribution using only information about the mean. A tighter bound can be obtained by also using the variance of the random variable.

Theorem 2.3 (Chebyshev's inequality) *Let x be a random variable. Then for $c > 0$,*

$$Prob\Big(|x - E(x)| \geq c\Big) \leq \frac{Var(x)}{c^2}.$$

Proof $\text{Prob}(|x - E(x)| \geq c) = \text{Prob}(|x - E(x)|^2 \geq c^2)$. Note that $y = |x - E(x)|^2$ is a nonnegative random variable and $E(y) = Var(x)$, so Markov's inequality can be applied giving:

$$\text{Prob}(|x - E(x)| \geq c) = \text{Prob}\left(|x - E(x)|^2 \geq c^2\right) \leq \frac{E(|x - E(x)|^2)}{c^2} = \frac{Var(x)}{c^2}.$$

■

The Law of Large Numbers follows from Chebyshev's inequality together with facts about independent random variables. Recall that:

$$E(x + y) = E(x) + E(y),$$
$$Var(x - c) = Var(x),$$
$$Var(cx) = c^2 Var(x).$$

Also, if x and y are independent, then $E(xy) = E(x)E(y)$. These facts imply that if x and y are independent, then $Var(x+y) = Var(x) + Var(y)$, which is seen as follows:

$$\begin{aligned} Var(x+y) &= E(x+y)^2 - E^2(x+y) \\ &= E(x^2 + 2xy + y^2) - \left(E^2(x) + 2E(x)E(y) + E^2(y)\right) \\ &= E(x^2) - E^2(x) + E(y^2) - E^2(y) = Var(x) + Var(y), \end{aligned}$$

where we used independence to replace $E(2xy)$ with $2E(x)E(y)$.

Theorem 2.4 (Law of Large Numbers) *Let $x_1, x_2, \ldots, x_n$ be n independent samples of a random variable x. Then*

$$Prob\left(\left|\frac{x_1 + x_2 + \cdots + x_n}{n} - E(x)\right| \geq \epsilon\right) \leq \frac{Var(x)}{n\epsilon^2}$$

Proof $E(\frac{x_1+x_2+\cdots+x_n}{n}) = E(x)$ and thus

$$\text{Prob}\left(\left|\frac{x_1 + x_2 + \cdots + x_n}{n} - E(x)\right| \geq \epsilon\right) = \text{Prob}\left(\left|\frac{x_1 + x_2 + \cdots + x_n}{n} - E\left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right)\right| \geq \epsilon\right)$$

By Chebyshev's inequality,

$$\begin{aligned} \text{Prob}\left(\left|\frac{x_1 + x_2 + \cdots + x_n}{n} - E(x)\right| \geq \epsilon\right) &= \text{Prob}\left(\left|\frac{x_1 + x_2 + \cdots + x_n}{n} - E\left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right)\right| \geq \epsilon\right) \\ &\leq \frac{Var\left(\frac{x_1+x_2+\cdots+x_n}{n}\right)}{\epsilon^2} \\ &= \frac{1}{n^2\epsilon^2} Var(x_1 + x_2 + \cdots + x_n) \\ &= \frac{1}{n^2\epsilon^2}\left(Var(x_1) + Var(x_2) + \cdots + Var(x_n)\right) \\ &= \frac{Var(x)}{n\epsilon^2}. \end{aligned}$$

■

The Law of Large Numbers is quite general, applying to any random variable x of finite variance. Later we will look at tighter concentration bounds for spherical Gaussians and sums of 0–1 valued random variables.

One observation worth making about the Law of Large Numbers is that the size of the universe does not enter into the bound. For instance, if you want to know what fraction of the population of a country prefers tea to coffee, then the number n of people you need to sample in order to have at most a δ chance that your estimate is off by more than ϵ depends only on ϵ and δ and not on the population of the country.

As an application of the Law of Large Numbers, let $\mathbf{z}$ be a d-dimensional random point whose coordinates are each selected from a zero mean, $\frac{1}{2\pi}$ variance Gaussian.

Table 2.1: Table of tail bounds. The Higher Moments bound is obtained by applying Markov to x^r. The Chernoff, Gaussian Annulus, and Power Law bounds follow from Theorem 2.5 which is proved in Chapter 12.

	Condition	Tail bound
Markov	$x \geq 0$	$\text{Prob}(x \geq a) \leq \frac{E(x)}{a}$
Chebyshev	Any x	$\text{Prob}\big(\lvert x - E(x)\rvert \geq a\big) \leq \frac{\text{Var}(x)}{a^2}$
Chernoff	$x = x_1 + x_2 + \cdots + x_n$ $x_i \in [0,1]$ i.i.d. Bernoulli;	$\text{Prob}(\lvert x - E(x)\rvert \geq \varepsilon E(x))$ $\leq 3e^{-c\varepsilon^2 E(x)}$
Higher Moments	r positive even integer	$\text{Prob}(\lvert x\rvert \geq a) \leq E(x^r)/a^r$
Gaussian Annulus	$x = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$ $x_i \sim N(0,1); \beta \leq \sqrt{n}$ indep.	$\text{Prob}(\lvert x - \sqrt{n}\rvert \geq \beta) \leq 3e^{-c\beta^2}$
Power Law for x_i; order $k \geq 4$	$x = x_1 + x_2 + \cdots + x_n$ x_i i.i.d ; $\varepsilon \leq 1/k^2$	$\text{Prob}\big(\lvert x - E(x)\rvert \geq \varepsilon E(x)\big)$ $\leq (4/\varepsilon^2 kn)^{(k-3)/2}$

We set the variance to $\frac{1}{2\pi}$ so the Gaussian probability density equals one at the origin and is bounded below throughout the unit ball by a constant.[1] By the Law of Large Numbers, the square of the distance of $\mathbf{z}$ to the origin will be $\Theta(d)$ with high probability. In particular, there is vanishingly small probability that such a random point $\mathbf{z}$ would lie in the unit ball. This implies that the integral of the probability density over the unit ball must be vanishingly small. On the other hand, the probability density in the unit ball is bounded below by a constant. We thus conclude that the unit ball must have vanishingly small volume.

Similarly if we draw two points $\mathbf{y}$ and $\mathbf{z}$ from a d-dimensional Gaussian with unit variance in each direction, then $|\mathbf{y}|^2 \approx d$ and $|\mathbf{z}|^2 \approx d$. Since for all i,

$$E(y_i - z_i)^2 = E(y_i^2) + E(z_i^2) - 2E(y_i z_i) = Var(y_i) + Var(z_i) - 2E(y_i)E(z_i) = 2,$$

$|\mathbf{y} - \mathbf{z}|^2 = \sum_{i=1}^{d}(y_i - z_i)^2 \approx 2d$. Thus by the Pythagorean theorem, the random d-dimensional $\mathbf{y}$ and $\mathbf{z}$ must be approximately orthogonal. This implies that if we scale these random points to be unit length and call $\mathbf{y}$ the North Pole, much of the surface area of the unit ball must lie near the equator. We will formalize these and related arguments in subsequent sections.

We now state a general theorem on probability tail bounds for a sum of independent random variables. Tail bounds for sums of Bernoulli, squared Gaussian, and Power Law distributed random variables can all be derived from this. Table 2.1 summarizes some of the results.

[1] If we instead used variance 1, then the density at the origin would be a decreasing function of d, namely $(\frac{1}{2\pi})^{d/2}$, making this argument more complicated.

Theorem 2.5 (Master Tail Bounds Theorem) *Let $x = x_1 + x_2 + \cdots + x_n$, where $x_1, x_2, \ldots, x_n$ are mutually independent random variables with zero mean and variance at most σ^2. Let $0 \leq a \leq \sqrt{2}n\sigma^2$. Assume that $|E(x_i^s)| \leq \sigma^2 s!$ for $s = 3, 4, \ldots, \lfloor (a^2/4n\sigma^2) \rfloor$. Then,*

$$Prob\left(|x| \geq a\right) \leq 3e^{-a^2/(12n\sigma^2)}.$$

The proof of Theorem 2.5 is elementary. A slightly more general version, Theorem 12.5, is given in Chapter 12. For a brief intuition of the proof, consider applying Markov's inequality to the random variable x^r where r is a large even number. Since r is even, x^r is nonnegative, and thus $\text{Prob}(|x| \geq a) = \text{Prob}(x^r \geq a^r) \leq E(x^r)/a^r$. If $E(x^r)$ is not too large, we will get a good bound. To compute $E(x^r)$, write $E(x)$ as $E(x_1 + \cdots + x_n)^r$ and expand the polynomial into a sum of terms. Use the fact that by independence $E(x_i^{r_i} x_j^{r_j}) = E(x_i^{r_i})E(x_j^{r_j})$ to get a collection of simpler expectations that can be bounded using our assumption that $|E(x_i^s)| \leq \sigma^2 s!$. For the full proof, see Chapter 12.

2.3. The Geometry of High Dimensions

An important property of high-dimensional objects is that most of their volume is near the surface. Consider any object A in R^d. Now shrink A by a small amount ϵ to produce a new object $(1 - \epsilon)A = \{(1 - \epsilon)x | x \in A\}$. Then the following equality holds:

$$\text{volume}\big((1 - \epsilon)A\big) = (1 - \epsilon)^d \text{volume}(A).$$

To see that this is true, partition A into infinitesimal cubes. Then, $(1-\varepsilon)A$ is the union of a set of cubes obtained by shrinking the cubes in A by a factor of $1 - \varepsilon$. When we shrink each of the $2d$ sides of a d-dimensional cube by a factor f, its volume shrinks by a factor of f^d. Using the fact that $1 - x \leq e^{-x}$, for any object A in R^d we have:

$$\frac{\text{volume}\big((1 - \epsilon)A\big)}{\text{volume}(A)} = (1 - \epsilon)^d \leq e^{-\epsilon d}.$$

Fixing ϵ and letting $d \to \infty$, the above quantity rapidly approaches zero. This means that nearly all of the volume of A must be in the portion of A that does not belong to the region $(1 - \epsilon)A$.

Let S denote the unit ball in d-dimensions, that is, the set of points within distance one of the origin. An immediate implication of the above observation is that at least a $1 - e^{-\epsilon d}$ fraction of the volume of the unit ball is concentrated in $S \setminus (1 - \epsilon)S$, namely in a small annulus of width ϵ at the boundary. In particular, most of the volume of the d-dimensional unit ball is contained in an annulus of width $O(1/d)$ near the boundary. This is illustrated in Figure 2.1. If the ball is of radius r, then the annulus width is $O\left(\frac{r}{d}\right)$.

2.4. Properties of the Unit Ball

We now focus more specifically on properties of the unit ball in d-dimensional space. We just saw that most of its volume is concentrated in a small annulus of width $O(1/d)$ near the boundary. Next we will show that in the limit as d goes to infinity,

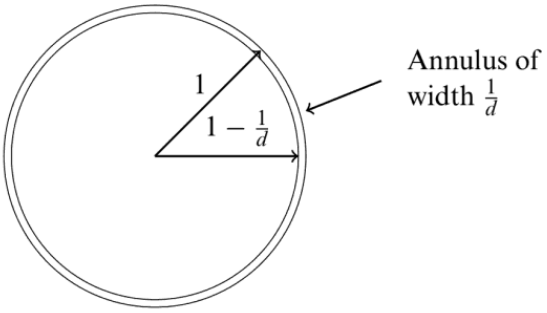


Figure 2.1: Most of the volume of the d-dimensional ball of radius r is contained in an annulus of width $O(r/d)$ near the boundary.

the volume of the ball goes to zero. This result can be proven in several ways. Here we use integration.

2.4.1. Volume of the Unit Ball

To calculate the volume $V(d)$ of the unit ball in R^d, one can integrate in either Cartesian or polar coordinates. In Cartesian coordinates the volume is given by

$$V(d) = \int_{x_1=-1}^{x_1=1} \int_{x_2=-\sqrt{1-x_1^2}}^{x_2=\sqrt{1-x_1^2}} \cdots \int_{x_d=-\sqrt{1-x_1^2-\cdots-x_{d-1}^2}}^{x_d=\sqrt{1-x_1^2-\cdots-x_{d-1}^2}} dx_d \cdots dx_2 dx_1.$$

Since the limits of the integrals are complicated, it is easier to integrate using polar coordinates. In polar coordinates, $V(d)$ is given by

$$V(d) = \int_{S^d} \int_{r=0}^{1} r^{d-1} dr d\Omega.$$

Since the variables Ω and r do not interact,

$$V(d) = \int_{S^d} d\Omega \int_{r=0}^{1} r^{d-1} dr = \frac{1}{d} \int_{S^d} d\Omega = \frac{A(d)}{d}$$

where $A(d)$ is the surface area of the d-dimensional unit ball. For instance, for $d = 3$ the surface area is 4π and the volume is $\frac{4}{3}\pi$. The question remains how to determine the surface area $A(d) = \int_{S^d} d\Omega$ for general d.

Consider a different integral,

$$I(d) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\left(x_1^2 + x_2^2 + \cdots x_d^2\right)} dx_d \cdots dx_2 dx_1.$$

Including the exponential allows integration to infinity rather than stopping at the surface of the sphere. Thus, $I(d)$ can be computed by integrating in both Cartesian

$$\text{volume}(A) \leq \int_{\frac{c}{\sqrt{d-1}}}^{\infty} \frac{x_1\sqrt{d-1}}{c} e^{-\frac{d-1}{2}x_1^2} V(d-1)dx_1$$
$$= V(d-1)\frac{\sqrt{d-1}}{c} \int_{\frac{c}{\sqrt{d-1}}}^{\infty} x_1 e^{-\frac{d-1}{2}x_1^2} dx_1$$

Now

$$\int_{\frac{c}{\sqrt{d-1}}}^{\infty} x_1 e^{-\frac{d-1}{2}x_1^2} dx_1 = -\frac{1}{d-1} e^{-\frac{d-1}{2}x_1^2} \Big|_{\frac{c}{\sqrt{(d-1)}}}^{\infty} = \frac{1}{d-1} e^{-\frac{c^2}{2}}$$

Thus, an upper bound on volume(A) is $\frac{V(d-1)}{c\sqrt{d-1}} e^{-\frac{c^2}{2}}$.

The volume of the hemisphere below the plane $x_1 = \frac{1}{\sqrt{d-1}}$ is a lower bound on the entire volume of the upper hemisphere, and this volume is at least that of a cylinder of height $\frac{1}{\sqrt{d-1}}$ and radius $\sqrt{1 - \frac{1}{d-1}}$. The volume of the cylinder is $V(d-1)(1 - \frac{1}{d-1})^{\frac{d-1}{2}} \frac{1}{\sqrt{d-1}}$. Using the fact that $(1-x)^a \geq 1 - ax$ for $a \geq 1$, the volume of the cylinder is at least $\frac{V(d-1)}{2\sqrt{d-1}}$ for $d \geq 3$.

Thus,

$$\text{ratio} \leq \frac{\text{upper bound above plane}}{\text{lower bound total hemisphere}} = \frac{\frac{V(d-1)}{c\sqrt{d-1}} e^{-\frac{c^2}{2}}}{\frac{V(d-1)}{2\sqrt{d-1}}} = \frac{2}{c} e^{-\frac{c^2}{2}}.$$

■

One might ask why we computed a lower bound on the total hemisphere, since it is one-half of the volume of the unit ball, which we already know. The reason is that the volume of the upper hemisphere is $\frac{1}{2}V(d)$, and we need a formula with $V(d-1)$ in it to cancel the $V(d-1)$ in the numerator.

Near Orthogonality

One immediate implication of the above analysis is that if we draw two points at random from the unit ball, with high probability their vectors will be nearly orthogonal to each other. Specifically, from our previous analysis in Section 2.3, with high probability both will be close to the surface and will have length $1 - O(1/d)$. From our analysis earlier, if we define the vector in the direction of the first point as "north," with high probability the second will have a projection of only $\pm O(1/\sqrt{d})$ in this direction, and thus their dot product will be $\pm O(1/\sqrt{d})$. This implies that with high probability the angle between the two vectors will be $\pi/2 \pm O(1/\sqrt{d})$. In particular, we have the following theorem that states that if we draw n points at random in the unit ball, with high probability all points will be close to unit length and each pair of points will be almost orthogonal.

Theorem 2.8 *Consider drawing n points $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$ at random from the unit ball. With probability $1 - O(1/n)$*

1. $|\mathbf{x_i}| \geq 1 - \frac{2\ln n}{d}$ *for all i, and*
2. $|\mathbf{x_i} \cdot \mathbf{x_j}| \leq \frac{\sqrt{6\ln n}}{\sqrt{d-1}}$ *for all $i \neq j$.*

Proof For the first part, for any fixed i by the analysis of Section 2.3, the probability that $|\mathbf{x_i}| < 1 - \epsilon$ is less than $e^{-\epsilon d}$. Thus

$$\text{Prob}\left(|\mathbf{x_i}| < 1 - \frac{2\ln n}{d}\right) \leq e^{-(\frac{2\ln n}{d})d} = 1/n^2.$$

By the union bound, the probability there exists an i such that $|\mathbf{x_i}| < 1 - \frac{2\ln n}{d}$ is at most $1/n$.

For the second part, Theorem 2.7 states that for a component of a Gaussian vector the probability $|x_i| > \frac{c}{\sqrt{d-1}}$ is at most $\frac{2}{c}e^{-\frac{c^2}{2}}$. There are $\binom{n}{2}$ pairs i and j, and for each such pair, if we define $\mathbf{x_i}$ as "north," the probability that the projection of $\mathbf{x_j}$ onto the "north" direction is more than $\frac{\sqrt{6\ln n}}{\sqrt{d-1}}$ is at most $O(e^{-\frac{6\ln n}{2}}) = O(n^{-3})$. Thus, the dot product condition is violated with probability at most $O\left(\binom{n}{2}n^{-3}\right) = O(1/n)$ as well. ■

Alternative Proof That Volume Goes to Zero

Another immediate implication of Theorem 2.7 is that as $d \to \infty$, the volume of the ball approaches zero. Specifically, consider a small box centered at the origin of side length $\frac{2c}{\sqrt{d-1}}$. Using Theorem 2.7, we show that for $c = 2\sqrt{\ln d}$, this box contains over half of the volume of the ball. On the other hand, the volume of this box clearly goes to zero as d goes to infinity, since its volume is $O((\frac{\ln d}{d-1})^{d/2})$. Thus the volume of the ball goes to zero as well.

By Theorem 2.7, with $c = 2\sqrt{\ln d}$, the fraction of the volume of the ball with $|x_1| \geq \frac{c}{\sqrt{d-1}}$ is at most:

$$\frac{2}{c}e^{-\frac{c^2}{2}} = \frac{1}{\sqrt{\ln d}}e^{-2\ln d} = \frac{1}{d^2\sqrt{\ln d}} < \frac{1}{d^2}.$$

Since this is true for each of the d dimensions, by a union bound, at most a $O(\frac{1}{d}) \leq \frac{1}{2}$ fraction of the volume of the ball lies outside the cube, completing the proof.

Discussion

One might wonder how it can be that nearly all the points in the unit ball are very close to the surface and yet at the same time nearly all points are in a box of side-length $O(\frac{\ln d}{d-1})$. The answer is to remember that points on the surface of the ball satisfy $x_1^2 + x_2^2 + \cdots + x_d^2 = 1$, so for each coordinate i, a typical value will be $\pm O(\frac{1}{\sqrt{d}})$. In fact, it is often helpful to think of picking a random point on the sphere as very similar to picking a random point of the form $(\pm\frac{1}{\sqrt{d}}, \pm\frac{1}{\sqrt{d}}, \pm\frac{1}{\sqrt{d}}, \ldots \pm\frac{1}{\sqrt{d}})$. A schematic illustration of the relationship between the unit-radius sphere and unit-volume cube is given in Figure 2.3.

2.5. Generating Points Uniformly at Random from a Ball

Consider generating points uniformly at random on the surface of the unit ball. For the two-dimensional version of generating points on the circumference of a unit-radius circle, independently generate each coordinate uniformly at random from the interval $[-1, 1]$. This produces points distributed over a square that is large enough

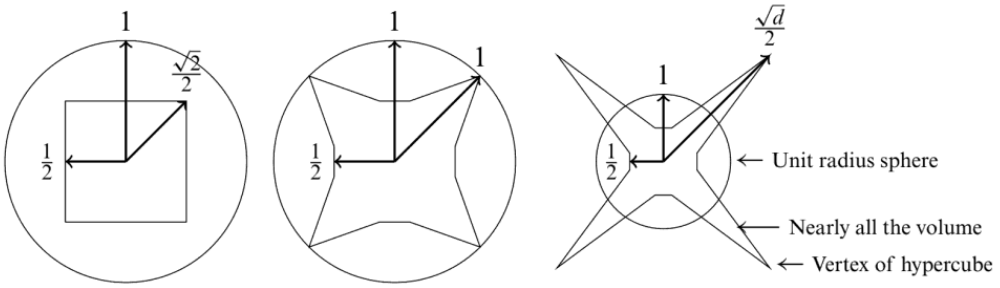


Figure 2.3: Illustration of the relationship between the sphere and the cube in 2, 4, and d-dimensions.

to completely contain the unit circle. Project each point onto the unit circle. The distribution is not uniform, since more points fall on a line from the origin to a vertex of the square than fall on a line from the origin to the midpoint of an edge of the square due to the difference in length. To solve this problem, discard all points outside the unit circle and project the remaining points onto the circle.

In higher dimensions, this method does not work, since the fraction of points that fall inside the ball drops to zero and all of the points would be thrown away. The solution is to generate a point each of whose coordinates is an independent Gaussian variable. Generate $x_1, x_2, \ldots, x_d$, using a zero mean, unit variance Gaussian, namely $\frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$ on the real line.[2] Thus, the probability density of $\mathbf{x}$ is

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{x_1^2+x_2^2+\cdots+x_d^2}{2}}$$

and is spherically symmetric. Normalizing the vector $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ to a unit vector, namely $\frac{\mathbf{x}}{|\mathbf{x}|}$, gives a distribution that is uniform over the surface of the sphere. Note that once the vector is normalized, its coordinates are no longer statistically independent.

To generate a point $\mathbf{y}$ uniformly over the ball (surface and interior), scale the point $\frac{\mathbf{x}}{|\mathbf{x}|}$ generated on the surface by a scalar $\rho \in [0, 1]$. What should the distribution of ρ be as a function of r? It is certainly not uniform, even in two dimensions. Indeed, the density of ρ at r is proportional to r for $d = 2$. For $d = 3$, it is proportional to r^2. By similar reasoning, the density of ρ at distance r is proportional to r^{d-1} in d dimensions. Solving $\int_{r=0}^{r=1} cr^{d-1} dr = 1$ (the integral of density must equal 1), one should set $c = d$. Another way to see this formally is that the volume of the radius r ball in d dimensions is $r^d V(d)$. The density at radius r is exactly $\frac{d}{dr}(r^d V_d) = dr^{d-1} V_d$. So, pick $\rho(r)$ with density equal to dr^{d-1} for r over $[0, 1]$.

[2]One might naturally ask: "How do you generate a random number from a 1-dimensional Gaussian?" To generate a number from any distribution given its cumulative distribution function P, first select a uniform random number $u \in [0, 1]$ and then choose $x = P^{-1}(u)$. For any $a < b$, the probability that x is between a and b is equal to the probability that u is between $P(a)$ and $P(b)$, which equals $P(b) - P(a)$ as desired. For the two-dimensional Gaussian, one can generate a point in polar coordinates by choosing angle θ uniform in $[0, 2\pi]$ and radius $r = \sqrt{-2\ln(u)}$ where u is uniform random in $[0, 1]$. This is called the Box-Muller transform.

We have succeeded in generating a point

$$\mathbf{y} = \rho \frac{\mathbf{x}}{|\mathbf{x}|}$$

uniformly at random from the unit ball by using the convenient spherical Gaussian distribution. In the next sections, we will analyze the spherical Gaussian in more detail.

2.6. Gaussians in High Dimension

A one-dimensional Gaussian has its mass close to the origin. However, as the dimension is increased, something different happens. The d-dimensional spherical Gaussian with zero mean and variance σ^2 in each coordinate has density function

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sigma^d} \exp\left(-\frac{|\mathbf{x}|^2}{2\sigma^2}\right).$$

The value of the density is maximum at the origin, but there is very little volume there. When $\sigma^2 = 1$, integrating the probability density over a unit ball centered at the origin yields almost zero mass, since the volume of such a ball is negligible. In fact, one needs to increase the radius of the ball to nearly $\sqrt{d}$ before there is a significant volume and hence significant probability mass. If one increases the radius much beyond $\sqrt{d}$, the integral barely increases even though the volume increases, since the probability density is dropping off at a much higher rate. The following theorem formally states that nearly all the probability is concentrated in a thin annulus of width $O(1)$ at radius $\sqrt{d}$.

Theorem 2.9 (Gaussian Annulus Theorem) *For a d-dimensional spherical Gaussian with unit variance in each direction, for any $\beta \leq \sqrt{d}$, all but at most $3e^{-c\beta^2}$ of the probability mass lies within the annulus $\sqrt{d} - \beta \leq |\mathbf{x}| \leq \sqrt{d} + \beta$, where c is a fixed positive constant.*

For a high-level intuition, note that $E(|\mathbf{x}|^2) = \sum_{i=1}^{d} E(x_i^2) = dE(x_1^2) = d$, so the mean squared distance of a point from the center is d. The Gaussian Annulus Theorem says that the points are tightly concentrated. We call the square root of the mean squared distance, namely $\sqrt{d}$, the radius of the Gaussian.

To prove the Gaussian Annulus Theorem, we make use of a tail inequality for sums of independent random variables of bounded moments (Theorem 12.5).

Proof (Gaussian Annulus Theorem) Let $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ be a point selected from a unit variance Gaussian centered at the origin, and let $r = |\mathbf{x}|$. $\sqrt{d} - \beta \leq |\mathbf{x}| \leq \sqrt{d} + \beta$ is equivalent to $|r - \sqrt{d}| \geq \beta$. If $|r - \sqrt{d}| \geq \beta$, then multiplying both sides by $r + \sqrt{d}$ gives $|r^2 - d| \geq \beta(r + \sqrt{d}) \geq \beta\sqrt{d}$. So, it suffices to bound the probability that $|r^2 - d| \geq \beta\sqrt{d}$.

Rewrite $r^2 - d = (x_1^2 + \cdots + x_d^2) - d = (x_1^2 - 1) + \cdots + (x_d^2 - 1)$ and perform a change of variables: $y_i = x_i^2 - 1$. We want to bound the probability that $|y_1 + \cdots + y_d| \geq \beta\sqrt{d}$. Notice that $E(y_i) = E(x_i^2) - 1 = 0$. To apply Theorem 12.5, we need to bound the s^{th} moments of y_i.

For $|x_i| \leq 1, |y_i|^s \leq 1$ and for $|x_i| \geq 1, |y_i|^s \leq |x_i|^{2s}$. Thus,

$$|E(y_i^s)| = E(|y_i|^s) \leq E(1 + x_i^{2s}) = 1 + E(x_i^{2s})$$
$$= 1 + \sqrt{\frac{2}{\pi}} \int_0^\infty x^{2s} e^{-x^2/2} dx.$$

Using the substitution $2z = x^2$,

$$|E(y_i^s)| = 1 + \frac{1}{\sqrt{\pi}} \int_0^\infty 2^s z^{s-(1/2)} e^{-z} dz$$
$$\leq 2^s s! .$$

The last inequality is from the Gamma integral.

Since $E(y_i) = 0$, $Var(y_i) = E(y_i^2) \leq 2^2 2 = 8$. Unfortunately, we do not have $|E(y_i^s)| \leq 8s!$ as required in Theorem 12.5. To fix this problem, perform one more change of variables, using $w_i = y_i/2$. Then, $Var(w_i) \leq 2$ and $|E(w_i^s)| \leq 2s!$, and our goal is now to bound the probability that $|w_1 + \cdots + w_d| \geq \frac{\beta\sqrt{d}}{2}$. Applying Theorem 12.5 where $\sigma^2 = 2$ and $n = d$, this occurs with probability less than or equal to $3e^{-\frac{\beta^2}{96}}$. ■

In the next sections we will see several uses of the Gaussian Annulus Theorem.

2.7. Random Projection and Johnson-Lindenstrauss Lemma

One of the most frequently used subroutines in tasks involving high-dimensional data is nearest neighbor search. In nearest neighbor search we are given a database of n points in $\mathbf{R}^d$ where n and d are usually large. The database can be preprocessed and stored in an efficient data structure. Thereafter, we are presented "query" points in $\mathbf{R}^d$ and are asked to find the nearest or approximately nearest database point to the query point. Since the number of queries is often large, the time to answer each query should be very small, ideally a small function of $\log n$ and $\log d$, whereas preprocessing time could be larger, namely a polynomial function of n and d. For this and other problems, dimension reduction, where one projects the database points to a k-dimensional space with $k \ll d$ (usually dependent on $\log d$), can be very useful so long as the relative distances between points are approximately preserved. We will see, using the Gaussian Annulus Theorem, that such a projection indeed exists and is simple.

The projection $f : \mathbf{R}^d \to \mathbf{R}^k$ that we will examine (many related projections are known to work as well) is the following. Pick k Gaussian vectors $\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_k}$ in $\mathbf{R}^d$ with unit-variance coordinates. For any vector $\mathbf{v}$, define the projection $f(\mathbf{v})$ by:

$$f(\mathbf{v}) = (\mathbf{u_1} \cdot \mathbf{v}, \mathbf{u_2} \cdot \mathbf{v}, \ldots, \mathbf{u_k} \cdot \mathbf{v}).$$

The projection $f(\mathbf{v})$ is the vector of dot products of $\mathbf{v}$ with the $\mathbf{u_i}$. We will show that with high probability, $|f(\mathbf{v})| \approx \sqrt{k}|\mathbf{v}|$. For any two vectors $\mathbf{v_1}$ and $\mathbf{v_2}$, $f(\mathbf{v_1} - \mathbf{v_2}) = f(\mathbf{v_1}) - f(\mathbf{v_2})$. Thus, to estimate the distance $|\mathbf{v_1} - \mathbf{v_2}|$ between two vectors $\mathbf{v_1}$ and $\mathbf{v_2}$ in $\mathbf{R}^d$, it suffices to compute $|f(\mathbf{v_1}) - f(\mathbf{v_2})| = |f(\mathbf{v_1} - \mathbf{v_2})|$ in the k-dimensional space, since the factor of $\sqrt{k}$ is known and one can divide by it. The reason distances increase when we project to a lower-dimensional space is that the vectors $\mathbf{u_i}$ are not

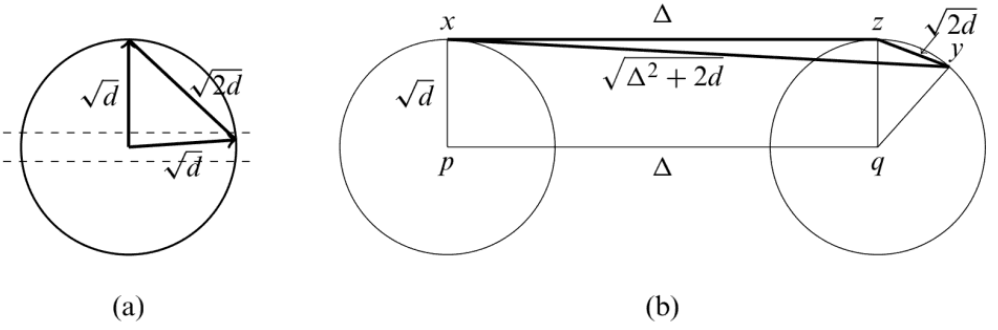


Figure 2.4: (a) indicates that two randomly chosen points in high dimension are surely almost nearly orthogonal. (b) indicates the distance between a pair of random points from two different unit balls approximating the annuli of two Gaussians.

at radius $\sqrt{d}$. Also $e^{-|\mathbf{x}|^2/2} = \prod_i e^{-x_i^2/2}$ and almost all of the mass is within the slab $\{\mathbf{x} \mid -c \leq x_1 \leq c\}$, for $c \in O(1)$. Pick a point $\mathbf{x}$ from this Gaussian. After picking $\mathbf{x}$, rotate the coordinate system to make the first axis align with $\mathbf{x}$. Independently pick a second point $\mathbf{y}$ from this Gaussian. The fact that almost all of the probability mass of the Gaussian is within the slab $\{\mathbf{x} \mid -c \leq x_1 \leq c, c \in O(1)\}$ at the equator implies that $\mathbf{y}$'s component along $\mathbf{x}$'s direction is $O(1)$ with high probability. Thus, $\mathbf{y}$ is nearly perpendicular to $\mathbf{x}$. So, $|\mathbf{x} - \mathbf{y}| \approx \sqrt{|\mathbf{x}|^2 + |\mathbf{y}|^2}$. See Figure 2.4(a). More precisely, since the coordinate system has been rotated so that $\mathbf{x}$ is at the North Pole, $\mathbf{x} = (\sqrt{d} \pm O(1), 0, \ldots, 0)$. Since $\mathbf{y}$ is almost on the equator, further rotate the coordinate system so that the component of $\mathbf{y}$ that is perpendicular to the axis of the North Pole is in the second coordinate. Then $\mathbf{y} = (O(1), \sqrt{d} \pm O(1), 0, \ldots, 0)$. Thus,

$$(\mathbf{x} - \mathbf{y})^2 = d \pm O(\sqrt{d}) + d \pm O(\sqrt{d}) = 2d \pm O(\sqrt{d})$$

and $|\mathbf{x} - \mathbf{y}| = \sqrt{2d} \pm O(1)$ with high probability.

Consider two spherical unit variance Gaussians with centers $\mathbf{p}$ and $\mathbf{q}$ separated by a distance Δ. The distance between a randomly chosen point $\mathbf{x}$ from the first Gaussian and a randomly chosen point $\mathbf{y}$ from the second is close to $\sqrt{\Delta^2 + 2d}$, since $\mathbf{x} - \mathbf{p}$, $\mathbf{p} - \mathbf{q}$, and $\mathbf{q} - \mathbf{y}$ are nearly mutually perpendicular. Pick $\mathbf{x}$ and rotate the coordinate system so that $\mathbf{x}$ is at the North Pole. Let $\mathbf{z}$ be the North Pole of the ball approximating the second Gaussian. Now pick $\mathbf{y}$. Most of the mass of the second Gaussian is within $O(1)$ of the equator perpendicular to $\mathbf{z} - \mathbf{q}$. Also, most of the mass of each Gaussian is within distance $O(1)$ of the respective equators perpendicular to the line $\mathbf{q} - \mathbf{p}$. See Figure 2.4 (b). Thus,

$$\begin{aligned} |\mathbf{x} - \mathbf{y}|^2 &\approx \Delta^2 + |\mathbf{z} - \mathbf{q}|^2 + |\mathbf{q} - \mathbf{y}|^2 \\ &= \Delta^2 + 2d \pm O(\sqrt{d}). \end{aligned}$$

Ensuring that two points picked from the same Gaussian are closer to each other than two points picked from different Gaussians requires that the upper limit of the distance between a pair of points from the same Gaussian is at most the lower limit of distance between points from different Gaussians. This requires that $\sqrt{2d} + O(1) \leq \sqrt{2d + \Delta^2} - O(1)$ or $2d + O(\sqrt{d}) \leq 2d + \Delta^2$, which holds when $\Delta \in \omega(d^{1/4})$. Thus,

mixtures of spherical Gaussians can be separated in this way, provided their centers are separated by $\omega(d^{1/4})$. If we have n points and want to correctly separate all of them with high probability, we need our individual high-probability statements to hold with probability $1-1/poly(n)$,[3] which means our $O(1)$ terms from Theorem 2.9 become $O(\sqrt{\log n})$. So we need to include an extra $O(\sqrt{\log n})$ term in the separation distance.

Algorithm for Separating Points from Two Gaussians Calculate all pairwise distances between points. The cluster of smallest pairwise distances must come from a single Gaussian. Remove these points. The remaining points come from the second Gaussian.

One can actually separate Gaussians where the centers are much closer. In the next chapter we will use singular value decomposition to separate points from a mixture of two Gaussians when their centers are separated by a distance $O(1)$.

2.9. Fitting a Spherical Gaussian to Data

Given a set of sample points, $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$, in a d-dimensional space, we wish to find the spherical Gaussian that best fits the points. Let f be the unknown Gaussian with mean $\boldsymbol{\mu}$ and variance σ^2 in each direction. The probability density for picking these points when sampling according to f is given by

$$c \exp\left(-\frac{(\mathbf{x_1}-\boldsymbol{\mu})^2+(\mathbf{x_2}-\boldsymbol{\mu})^2+\cdots+(\mathbf{x_n}-\boldsymbol{\mu})^2}{2\sigma^2}\right)$$

where the normalizing constant c is the reciprocal of $\left[\int e^{-\frac{|\mathbf{x}-\boldsymbol{\mu}|^2}{2\sigma^2}} dx\right]^n$. In integrating from $-\infty$ to ∞, one can shift the origin to $\boldsymbol{\mu}$ and thus c is $\left[\int e^{-\frac{|\mathbf{x}|^2}{2\sigma^2}} dx\right]^{-n} = \frac{1}{(2\pi)^{\frac{n}{2}}}$ and is independent of $\boldsymbol{\mu}$.

The *Maximum Likelihood Estimator* (MLE) of f, given the samples $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$, is the f that maximizes the above probability density.

Lemma 2.12 *Let $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be a set of n d-dimensional points. Then $(\mathbf{x}_1-\boldsymbol{\mu})^2+(\mathbf{x}_2-\boldsymbol{\mu})^2+\cdots+(\mathbf{x}_n-\boldsymbol{\mu})^2$ is minimized when $\boldsymbol{\mu}$ is the centroid of the points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, namely $\boldsymbol{\mu}=\frac{1}{n}(\mathbf{x}_1+\mathbf{x}_2+\cdots+\mathbf{x}_n)$.*

Proof Setting the gradient of $(\mathbf{x}_1-\boldsymbol{\mu})^2+(\mathbf{x}_2-\boldsymbol{\mu})^2+\cdots+(\mathbf{x}_n-\boldsymbol{\mu})^2$ with respect to $\boldsymbol{\mu}$ to zero yields

$$-2(\mathbf{x}_1-\boldsymbol{\mu})-2(\mathbf{x}_2-\boldsymbol{\mu})-\cdots-2(\mathbf{x}_n-\boldsymbol{\mu})=0.$$

Solving for $\boldsymbol{\mu}$ gives $\boldsymbol{\mu}=\frac{1}{n}(\mathbf{x}_1+\mathbf{x}_2+\cdots+\mathbf{x}_n)$. ■

To determine the maximum likelihood estimate of σ^2 for f, set $\boldsymbol{\mu}$ to the true centroid. Next, show that σ is set to the standard deviation of the sample. Substitute $v=\frac{1}{2\sigma^2}$ and $a=(\mathbf{x}_1-\boldsymbol{\mu})^2+(\mathbf{x}_2-\boldsymbol{\mu})^2+\cdots+(\mathbf{x}_n-\boldsymbol{\mu})^2$ into the formula for the probability of picking the points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. This gives

[3] poly(n) means bounded by a polynomial in n.

$$\frac{e^{-av}}{\left[\int_x e^{-x^2 v} dx\right]^n}.$$

Now, a is fixed and v is to be determined. Taking logs, the expression to maximize is

$$-av - n \ln \left[\int_x e^{-vx^2} dx\right].$$

To find the maximum, differentiate with respect to v, set the derivative to zero, and solve for σ. The derivative is

$$-a + n \frac{\int_x |x|^2 e^{-vx^2} dx}{\int_x e^{-vx^2} dx}.$$

Setting $y = \left|\sqrt{v}\mathbf{x}\right|$ in the derivative, yields

$$-a + \frac{n}{v} \frac{\int_y y^2 e^{-y^2} dy}{\int_y e^{-y^2} dy}.$$

Since the ratio of the two integrals is the expected distance squared of a d-dimensional spherical Gaussian of standard deviation $\frac{1}{\sqrt{2}}$ to its center, and this is known to be $\frac{d}{2}$, we get $-a + \frac{nd}{2v}$. Substituting σ^2 for $\frac{1}{2v}$ gives $-a + nd\sigma^2$. Setting $-a + nd\sigma^2 = 0$ shows that the maximum occurs when $\sigma = \frac{\sqrt{a}}{\sqrt{nd}}$. Note that this quantity is the square root of the average coordinate distance squared of the samples to their mean, which is the standard deviation of the sample. Thus, we get the following lemma.

Lemma 2.13 *The maximum likelihood spherical Gaussian for a set of samples is the Gaussian with center equal to the sample mean and standard deviation equal to the standard deviation of the sample from the true mean.*

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be a sample of points generated by a Gaussian probability distribution. Then $\boldsymbol{\mu} = \frac{1}{n}(\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_n)$ is an unbiased estimator of the expected value of the distribution. However, if in estimating the variance from the sample set we use the estimate of the expected value rather than the true expected value, we will not get an unbiased estimate of the variance, since the sample mean is not independent of the sample set. One should use $\tilde{\boldsymbol{\mu}} = \frac{1}{n-1}(\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_n)$ when estimating the variance. See Section 12.4.10 in Chapter 12.

2.10. Bibliographic Notes

The vector space model was introduced by Salton [SWY75]. There is vast literature on the Gaussian distribution, its properties, drawing samples according to it, etc. The reader can choose the level and depth according to his/her background. The Master Tail Bounds theorem and the derivation of Chernoff and other inequalities from it are from [Kan09]. The original proof of the Random Projection Theorem by Johnson

and Lindenstrauss was complicated. Several authors used Gaussians to simplify the proof. The proof here is due to Dasgupta and Gupta [DG99]. See [Vem04] for details and applications of the theorem. [MU05] and [MR95b] are textbooks covering much of the material touched upon here.

2.11. Exercises

Exercise 2.1

1. Let x and y be independent random variables with uniform distribution in $[0, 1]$. What is the expected value $E(x)$, $E(x^2)$, $E(x - y)$, $E(xy)$, and $E(x - y)^2$?
2. Let x and y be independent random variables with uniform distribution in $[-\frac{1}{2}, \frac{1}{2}]$. What is the expected value $E(x)$, $E(x^2)$, $E(x-y)$, $E(xy)$, and $E(x-y)^2$?
3. What is the expected squared distance between two points generated at random inside a unit d-dimensional cube?

Exercise 2.2 Randomly generate 30 points inside the cube $[-\frac{1}{2}, \frac{1}{2}]^{100}$ and plot distance between points and the angle between the vectors from the origin to the points for all pairs of points.

Exercise 2.3 Show that for any $a \geq 1$ there exist distributions for which Markov's inequality is tight by showing the following:

1. For each $a = 2, 3$, and 4 give a probability distribution $p(x)$ for a nonnegative random variable x where $\text{Prob}(x \geq a) = \frac{E(x)}{a}$.
2. For arbitrary $a \geq 1$ give a probability distribution for a nonnegative random variable x where $\text{Prob}(x \geq a) = \frac{E(x)}{a}$.

Exercise 2.4 Show that for any $c \geq 1$ there exist distributions for which Chebyshev's inequality is tight, in other words, $Prob(|x - E(x)| \geq c) = Var(x)/c^2$.

Exercise 2.5 Let x be a random variable with probability density $\frac{1}{4}$ for $0 \leq x \leq 4$ and zero elsewhere.

1. Use Markov's inequality to bound the probability that $x \geq 3$.
2. Make use of $\text{Prob}(|x| \geq a) = \text{Prob}(x^2 \geq a^2)$ to get a tighter bound.
3. What is the bound using $\text{Prob}(|x| \geq a) = \text{Prob}(x^r \geq a^r)$?

Exercise 2.6 Consider the probability distribution $p(x = 0) = 1 - \frac{1}{a}$ and $p(x = a) = \frac{1}{a}$. Plot the probability that x is greater than or equal to a as a function of a for the bound given by Markov's inequality and by Markov's inequality applied to x^2 and x^4.

Exercise 2.7 Consider the probability density function $p(x) = 0$ for $x < 1$ and $p(x) = c\frac{1}{x^4}$ for $x \geq 1$.

1. What should c be to make p a legal probability density function?
2. Generate 100 random samples from this distribution. How close is the average of the samples to the expected value of x?

Exercise 2.8 Let U be a set of integers and X and Y be subsets of U whose symmetric difference $X \Delta Y$ is 1/10 of U. Prove that the probability that none of the elements selected at random from U will be in $X \Delta Y$ is less than $e^{-0.1n}$.

Exercise 2.9 Let G be a d-dimensional spherical Gaussian with variance $\frac{1}{2}$ in each direction, centered at the origin. Derive the expected squared distance to the origin.

Exercise 2.10 Consider drawing a random point $\mathbf{x}$ on the surface of the unit sphere in R^d. What is the variance of x_1 (the first coordinate of $\mathbf{x}$)? See if you can give an argument without doing any integrals.

Exercise 2.11 How large must ε be for 99% of the volume of a 1000-dimensional unit-radius ball to lie in the shell of ε-thickness at the surface of the ball?

Exercise 2.12 Prove that $1 + x \leq e^x$ for all real x. For what values of x is the approximation $1 + x \approx e^x$ within 0.01?

Exercise 2.13 For what value of d does the volume, $V(d)$, of a d-dimensional unit ball take on its maximum? Hint: Consider the ratio $\frac{V(d)}{V(d-1)}$.

Exercise 2.14 A three-dimensional cube has vertices, edges, and faces. In a d-dimensional cube, these components are called faces. A vertex is a zero-dimensional face, an edge a one-dimensional face, etc.

1. For $0 \leq k \leq d$, how many k-dimensional faces does a d-dimensional cube have?
2. What is the total number of faces of all dimensions? The d-dimensional face is the cube itself, which you can include in your count.
3. What is the surface area of a unit cube in d-dimensions (a unit cube has a side-length of 1 in each dimension)?
4. What is the surface area of the cube if the length of each side is 2?
5. Prove that the volume of a unit cube is close to its surface.

Exercise 2.15 Consider the portion of the surface area of a unit radius, three-dimensional ball with center at the origin that lies within a circular cone whose vertex is at the origin. What is the formula for the incremental unit of area when using polar coordinates to integrate the portion of the surface area of the ball that is lying inside the circular cone? What is the formula for the integral? What is the value of the integral if the angle of the cone is 36°? The angle of the cone is measured from the axis of the cone to a ray on the surface of the cone.

Exercise 2.16 Consider a unit radius, circular cylinder in three-dimensions of height 1. The top of the cylinder could be a horizontal plane or half of a circular ball. Consider these two possibilities for a unit radius, circular cylinder in four dimensions. In four dimensions the horizontal plane is three-dimensional and the half circular ball is four-dimensional. In each of the two cases, what is the surface area of the top face of the cylinder? You can use $V(d)$ for the volume of a unit radius, d-dimension ball, and $A(d)$ for the surface area of a unit radius, d-dimensional ball. An infinite-length, unit radius, circular cylinder in 4-dimensions would be the set $\{(x_1, x_2, x_3, x_4) | x_2^2 + x_3^2 + x_4^2 \leq 1\}$ where the coordinate x_1 is the axis.

Exercise 2.17 Given a d-dimensional circular cylinder of radius r and height h,

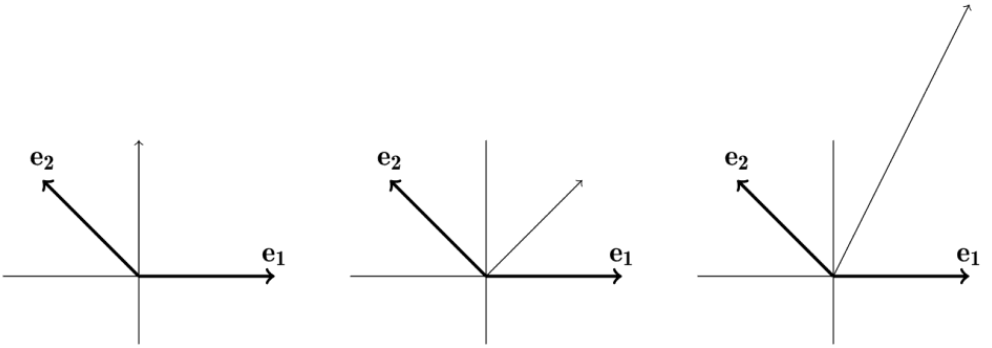
1. What is the surface area in terms of $V(d)$ and $A(d)$?
2. What is the volume?

Exercise 2.18 How does the volume of a ball of radius 2 behave as the dimension of the space increases? What if the radius was larger than 2 but a constant independent of d? What function of d would the radius need to be for a ball of radius r to

2. Is the volume of a unit cube concentrated close to the equator?
3. Is the surface area of a unit cube concentrated close to the equator?

Exercise 2.37 Consider a non-orthogonal basis $\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_d}$. The $\mathbf{e_i}$ are a set of linearly independent unit vectors that span the space.

1. Prove that the representation of any vector in this basis is unique.
2. Calculate the squared length of $\mathbf{z} = \left(\frac{\sqrt{2}}{2}, 1\right)_e$ where $\mathbf{z}$ is expressed in the basis $\mathbf{e_1} = (1, 0)$ and $\mathbf{e_2} = \left(-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$
3. If $\mathbf{y} = \sum_i a_i \mathbf{e_i}$ and $\mathbf{z} = \sum_i b_i \mathbf{e_i}$, with $0 < a_i < b_i$, is it necessarily true that the length of $\mathbf{z}$ is greater than the length of $\mathbf{y}$? Why or why not?
4. Consider the basis $\mathbf{e_1} = (1, 0)$ and $\mathbf{e_2} = \left(-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$.
   (a) What is the representation of the vector (0,1) in the basis $(\mathbf{e_1}, \mathbf{e_2})$?
   (b) What is the representation of the vector $\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$?
   (c) What is the representation of the vector $(1, 2)$?



Exercise 2.38 Generate 20 points uniformly at random on a 900-dimensional sphere of radius 30. Calculate the distance between each pair of points. Then, select a method of projection and project the data onto subspaces of dimension k = 100, 50, 10, 5, 4, 3, 2, 1 and calculate the difference between $\sqrt{k}$ times the original distances and the new pairwise distances. For each value of k what is the maximum difference as a percent of $\sqrt{k}$?

Exercise 2.39 What happens in high dimension to a lower-dimensional manifold? To see what happens, consider a sphere of dimension 100 in a 1,000-dimensional space when the 1,000-dimensional space is projected to a random 500-dimensional space. Will the sphere remain essentially spherical? Given an intuitive argument justifying your answer.

Exercise 2.40 In d-dimensions there are exactly d-unit vectors that are pairwise orthogonal. However, if you wanted a set of vectors that were almost orthogonal, you might squeeze in a few more. For example, in two dimensions, if almost orthogonal meant at least 45 degrees apart, you could fit in three almost orthogonal vectors. Suppose you wanted to find 1,000 almost orthogonal vectors in 100 dimensions. Here are two ways you could do it:

1. Begin with 1,000 orthonormal 1,000-dimensional vectors, and then project them to a random 100-dimensional space.

2. Generate 1,000 100-dimensional random Gaussian vectors.

Implement both ideas and compare them to see which does a better job.

Exercise 2.41 Suppose there is an object moving at constant velocity along a straight line. You receive the GPS coordinates corrupted by Gaussian noise every minute. How do you estimate the current position?

Exercise 2.42

1. What is the maximum-size rectangle that can be fitted under a unit-variance Gaussian?
2. What unit area rectangle best approximates a unit-variance Gaussian if one measures goodness of fit by the symmetric difference of the Gaussian and the rectangle?

Exercise 2.43 Let $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$ be independent samples of a random variable $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and variance σ^2. Let $\mathbf{m_s} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x_i}$ be the sample mean. Suppose one estimates the variance using the sample mean rather than the true mean, that is,

$$\sigma_s^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x_i} - \mathbf{m_s})^2.$$

Prove that $E(\sigma_s^2) = \frac{n-1}{n}\sigma^2$ and thus one should have divided by $n-1$ rather than n.

Hint. First calculate the variance of the sample mean and show that $\text{var}(\mathbf{m_s}) = \frac{1}{n}\text{var}(\mathbf{x})$. Then calculate $E(\sigma_s^2) = E[\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x_i} - \mathbf{m_s})^2]$ by replacing $\mathbf{x_i} - \mathbf{m_s}$ with $(\mathbf{x_i} - \mathbf{m}) - (\mathbf{m_s} - \mathbf{m})$.

Exercise 2.44 Generate 10 values by a Gaussian probability distribution with zero mean and variance 1. What is the center determined by averaging the points? What is the variance? In estimating the variance, use both the real center and the estimated center. When using the estimated center to estimate the variance, use both $n = 10$ and $n = 9$. How do the three estimates compare?

Exercise 2.45 Suppose you want to estimate the unknown center of a Gaussian in d-space that has variance 1 in each direction. Show that $O(\log d/\varepsilon^2)$ random samples from the Gaussian are sufficient to get an estimate $\mathbf{m}_s$ of the true center $\boldsymbol{\mu}$, so that with probability at least 99%,

$$\|\boldsymbol{\mu} - \mathbf{m}_s\|_\infty \leq \varepsilon.$$

How many samples are sufficient to ensure that with probability at least 99%

$$\|\boldsymbol{\mu} - \mathbf{m}_s\|_2 \leq \varepsilon?$$

Exercise 2.46 Use the probability distribution $\frac{1}{3\sqrt{2\pi}}e^{-\frac{1}{2}\frac{(x-5)^2}{9}}$ to generate 10 points.

(a) From the 10 points estimate μ. How close is the estimate of μ to the true mean of 5?
(b) Using the true mean of 5, estimate σ^2 by the formula $\sigma^2 = \frac{1}{10}\sum_{i=1}^{10}(x_i - 5)^2$. How close is the estimate of σ^2 to the true variance of 9?

(c) Using your estimate m of the mean, estimate σ^2 by the formula $\sigma^2 = \frac{1}{10}\sum_{i=1}^{10}(x_i - m)^2$. How close is the estimate of σ^2 to the true variance of 9?

(d) Using your estimate m of the mean, estimate σ^2 by the formula $\sigma^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - m)^2$. How close is the estimate of σ^2 to the true variance of 9?

Exercise 2.47 Create a list of the five most important things that you learned about high dimensions.

Exercise 2.48 Write a short essay whose purpose is to excite a college freshman to learn about high dimensions.

CHAPTER THREE

Best-Fit Subspaces and Singular Value Decomposition (SVD)

3.1. Introduction

In this chapter, we examine the *singular value decomposition* (SVD) of a matrix. Consider each row of an $n \times d$ matrix A as a point in d-dimensional space. The singular value decomposition finds the best-fitting k-dimensional subspace for $k = 1, 2, 3, \ldots$, for the set of n data points. Here, "best" means minimizing the sum of the squares of the perpendicular distances of the points to the subspace, or equivalently, maximizing the sum of squares of the lengths of the projections of the points onto this subspace.[1] We begin with a special case where the subspace is one-dimensional, namely a line through the origin. We then show that the best-fitting k-dimensional subspace can be found by k applications of the best-fitting line algorithm, where on the i^{th} iteration we find the best-fit line perpendicular to the previous $i - 1$ lines. When k reaches the rank of the matrix, from these operations we get an exact decomposition of the matrix called the *singular value decomposition*.

In matrix notation, the singular value decomposition of a matrix A with real entries (we assume all our matrices have real entries) is the factorization of A into the product of three matrices, $A = UDV^T$, where the columns of U and V are orthonormal[2] and the matrix D is diagonal with positive real entries. The columns of V are the unit length vectors defining the best-fitting lines described above (the i^{th} column being the unit length vector in the direction of the i^{th} line). The coordinates of a row of U will be the fractions of the corresponding row of A along the direction of each of the lines.

The SVD is useful in many tasks. Often a data matrix A is close to a low-rank matrix, and it is useful to find a good low-rank approximation to A. For any k, the singular value decomposition of A gives the best rank-k approximation to A in a well-defined sense.

[1] This equivalence is due to the Pythagorean Theorem. For each point, its squared length (its distance to the origin squared) is exactly equal to the squared length of its projection onto the subspace plus the squared distance of the point to its projection; therefore, maximizing the sum of the former is equivalent to minimizing the sum of the latter. For further discussion, see Section 3.2.

[2] A set of vectors is orthonormal if each is of length one and they are pairwise orthogonal.

If $\mathbf{u_i}$ and $\mathbf{v_i}$ are columns of U and V, respectively, then the matrix equation $A = UDV^T$ can be rewritten as

$$A = \sum_i d_{ii}\mathbf{u_i}\mathbf{v_i}^T.$$

Since $\mathbf{u_i}$ is a $n \times 1$ matrix and $\mathbf{v_i}$ is a $d \times 1$ matrix, $\mathbf{u_i}\mathbf{v_i}^T$ is an $n \times d$ matrix with the same dimensions as A. The i^{th} term in the above sum can be viewed as giving the components of the rows of A along direction $\mathbf{v_i}$. When the terms are summed, they reconstruct A.

This decomposition of A can be viewed as analogous to writing a vector $\mathbf{x}$ in some orthonormal basis $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_d}$. The coordinates of $\mathbf{x} = (\mathbf{x} \cdot \mathbf{v_1}, \mathbf{x} \cdot \mathbf{v_2} \ldots, \mathbf{x} \cdot \mathbf{v_d})$ are the projections of $\mathbf{x}$ onto the $\mathbf{v_i}$'s. For SVD, this basis has the property that for any k, the first k vectors of this basis produce the least possible total sum of squares error for that value of k.

In addition to the singular value decomposition, there is an eigenvalue decomposition. Let A be a square matrix. A vector $\mathbf{v}$ such that $A\mathbf{v} = \lambda\mathbf{v}$ is called an eigenvector and λ the eigenvalue. When A is symmetric, the eigenvectors are orthogonal and A can be expressed as $A = VDV^T$ where the eigenvectors are the columns of V and D is a diagonal matrix with the corresponding eigenvalues on its diagonal. For a symmetric matrix A, the singular values are the absolute values of the eigenvalues. Some eigenvalues may be negative, but all singular values are positive by definition. If the singular values are distinct, then A's right singular vectors and eigenvectors are identical up to scalar multiplication. The left singular vectors of A are identical with the right singular vectors of A when the corresponding eigenvalues are positive and are the negative of the right singular vectors when the corresponding eigenvalues are negative. If a singular value has multiplicity d greater than one, the corresponding singular vectors span a subspace of dimension d, and any orthogonal basis of the subspace can be used as the eigenvectors or singular vectors.[3]

The singular value decomposition is defined for all matrices, whereas the more familiar eigenvector decomposition requires that the matrix A be square and certain other conditions on the matrix to ensure orthogonality of the eigenvectors. In contrast, the columns of V in the singular value decomposition, called the *right-singular vectors* of A, always form an orthogonal set with no assumptions on A. The columns of U are called the *left-singular vectors* and they also form an orthogonal set (see Section 3.6). A simple consequence of the orthonormality is that for a square and invertible matrix A, the inverse of A is $VD^{-1}U^T$.

Eigenvalues and eignevectors satisfy $A\mathbf{v} = \lambda\mathbf{v}$. We will show that singular values and vectors satisfy a somewhat analogous relationship. Since $A\mathbf{v_i}$ is a $n \times 1$ matrix (vector), the matrix A cannot act on it from the left. But A^T, which is a $d \times n$ matrix, can act on this vector. Indeed, we will show that

$$A\mathbf{v_i} = d_{ii}\mathbf{u_i} \quad \text{and} \quad A^T\mathbf{u_i} = d_{ii}\mathbf{v_i}.$$

In words, A acting on $\mathbf{v_i}$ produces a scalar multiple of $\mathbf{u_i}$ and A^T acting on $\mathbf{u_i}$ produces the same scalar multiple of $\mathbf{v_i}$. Note that $A^TA\mathbf{v_i} = d_{ii}^2\mathbf{v_i}$. The i^{th} singular vector of A is the i^{th} eigenvector of the square symmetric matrix A^TA.

[3]When $d = 1$, there are actually two possible singular vectors, one the negative of the other. The subspace spanned is unique.

and so on. The process stops when we have found singular vectors $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$, singular values $\sigma_1, \sigma_2, \ldots, \sigma_r$, and

$$\max_{\substack{\mathbf{v} \perp \mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r} \\ |\mathbf{v}|=1}} |A\mathbf{v}| = 0.$$

The greedy algorithm found the $\mathbf{v_1}$ that maximized $|A\mathbf{v}|$ and then the best-fit two-dimensional subspace containing $\mathbf{v_1}$. Is this necessarily the best-fit two-dimensional subspace overall? The following theorem establishes that the greedy algorithm finds the best subspaces of every dimension.

Theorem 3.1 (The Greedy Algorithm Works) *Let A be an $n \times d$ matrix with singular vectors $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$. For $1 \leq k \leq r$, let V_k be the subspace spanned by $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_k}$. For each k, V_k is the best-fit k-dimensional subspace for A.*

Proof The statement is obviously true for $k = 1$. For $k = 2$, let W be a best-fit two-dimensional subspace for A. For any orthonormal basis $(\mathbf{w_1}, \mathbf{w_2})$ of W, $|A\mathbf{w_1}|^2 + |A\mathbf{w_2}|^2$ is the sum of squared lengths of the projections of the rows of A onto W. Choose an orthonormal basis $(\mathbf{w}_1, \mathbf{w_2})$ of W so that $\mathbf{w_2}$ is perpendicular to $\mathbf{v_1}$. If $\mathbf{v_1}$ is perpendicular to W, any unit vector in W will do as $\mathbf{w_2}$. If not, choose $\mathbf{w_2}$ to be the unit vector in W perpendicular to the projection of $\mathbf{v_1}$ onto W. This makes $\mathbf{w_2}$ perpendicular to $\mathbf{v_1}$.[5] Since $\mathbf{v_1}$ maximizes $|A\mathbf{v}|^2$, it follows that $|A\mathbf{w_1}|^2 \leq |A\mathbf{v_1}|^2$. Since $\mathbf{v_2}$ maximizes $|A\mathbf{v}|^2$ over all $\mathbf{v}$ perpendicular to $\mathbf{v_1}$, $|A\mathbf{w_2}|^2 \leq |A\mathbf{v_2}|^2$. Thus,

$$|A\mathbf{w_1}|^2 + |A\mathbf{w_2}|^2 \leq |A\mathbf{v_1}|^2 + |A\mathbf{v_2}|^2.$$

Hence, V_2 is at least as good as W and so is a best-fit two-dimensional subspace.

For general k, proceed by induction. By the induction hypothesis, V_{k-1} is a best-fit k-1–dimensional subspace. Suppose W is a best-fit k-dimensional subspace. Choose an orthonormal basis $\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_k}$ of W so that $\mathbf{w_k}$ is perpendicular to $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_{k-1}}$. Then

$$|A\mathbf{w_1}|^2 + |A\mathbf{w_2}|^2 + \cdots + |A\mathbf{w_{k-1}}|^2 \leq |A\mathbf{v_1}|^2 + |A\mathbf{v_2}|^2 + \cdots + |A\mathbf{v_{k-1}}|^2$$

since V_{k-1} is an optimal $k-1$ dimensional subspace. Since $\mathbf{w_k}$ is perpendicular to $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_{k-1}}$, by the definition of $\mathbf{v_k}$, $|A\mathbf{w_k}|^2 \leq |A\mathbf{v_k}|^2$. Thus,

$$\begin{aligned} &|A\mathbf{w_1}|^2 + |A\mathbf{w_2}|^2 + \cdots + |A\mathbf{w_{k-1}}|^2 + |A\mathbf{w_k}|^2 \\ &\leq |A\mathbf{v_1}|^2 + |A\mathbf{v_2}|^2 + \cdots + |A\mathbf{v_{k-1}}|^2 + |A\mathbf{v_k}|^2, \end{aligned}$$

proving that V_k is at least as good as W and hence is optimal. ■

Note that the n-dimensional vector $A\mathbf{v_i}$ is a list of lengths (with signs) of the projections of the rows of A onto $\mathbf{v_i}$. Think of $|A\mathbf{v_i}| = \sigma_i(A)$ as the *component* of the matrix A along $\mathbf{v_i}$. For this interpretation to make sense, it should be true that adding up the squares of the components of A along each of the $\mathbf{v_i}$ gives the square of the "whole content of A". This is indeed the case and is the matrix analogy of decomposing a vector into its components along orthogonal directions.

[5] This can be seen by noting that $\mathbf{v_1}$ is the sum of two vectors that each are individually perpendicular to $\mathbf{w_2}$, namely the projection of $\mathbf{v_1}$ to W and the portion of $\mathbf{v_1}$ orthogonal to W.

Consider one row, say $\mathbf{a_j}$, of A. Since $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$ span the space of all rows of A, $\mathbf{a_j} \cdot \mathbf{v} = 0$ for all $\mathbf{v}$ perpendicular to $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$. Thus, for each row $\mathbf{a_j}$, $\sum_{i=1}^{r} (\mathbf{a_j} \cdot \mathbf{v_i})^2 = |\mathbf{a_j}|^2$. Summing over all rows j,

$$\sum_{j=1}^{n} |\mathbf{a_j}|^2 = \sum_{j=1}^{n} \sum_{i=1}^{r} (\mathbf{a_j} \cdot \mathbf{v_i})^2 = \sum_{i=1}^{r} \sum_{j=1}^{n} (\mathbf{a_j} \cdot \mathbf{v_i})^2 = \sum_{i=1}^{r} |A\mathbf{v_i}|^2 = \sum_{i=1}^{r} \sigma_i^2(A).$$

But $\sum_{j=1}^{n} |\mathbf{a_j}|^2 = \sum_{j=1}^{n} \sum_{k=1}^{d} a_{jk}^2$, the sum of squares of all the entries of A. Thus, the sum of squares of the singular values of A is indeed the square of the "whole content of A," i.e., the sum of squares of all the entries. There is an important norm associated with this quantity, the Frobenius norm of A, denoted $||A||_F$ defined as

$$||A||_F = \sqrt{\sum_{j,k} a_{jk}^2}.$$

Lemma 3.2 *For any matrix A, the sum of squares of the singular values equals the square of the Frobenius norm. That is, $\sum \sigma_i^2(A) = ||A||_F^2$.*

Proof By the preceding discussion. ■

The vectors $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$ are called the *right-singular vectors*. The vectors $A\mathbf{v_i}$ form a fundamental set of vectors, and we normalize them to length 1 by

$$\mathbf{u_i} = \frac{1}{\sigma_i(A)} A\mathbf{v_i}.$$

Later we will show that $\mathbf{u}_i$ similarly maximizes $|\mathbf{u}^T A|$ over all $\mathbf{u}$ perpendicular to $\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}$. These $\mathbf{u}_i$ are called the *left-singular vectors*. Clearly, the right-singular vectors are orthogonal by definition. We will show later that the left-singular vectors are also orthogonal.

3.4. Singular Value Decomposition (SVD)

Let A be an $n \times d$ matrix with singular vectors $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$ and corresponding singular values $\sigma_1, \sigma_2, \ldots, \sigma_r$. The left-singular vectors of A are $\mathbf{u_i} = \frac{1}{\sigma_i} A\mathbf{v_i}$ where $\sigma_i \mathbf{u_i}$ is a vector whose coordinates correspond to the projections of the rows of A onto $\mathbf{v_i}$. Each $\sigma_i \mathbf{u_i} \mathbf{v_i^T}$ is a rank one matrix whose rows are the "$\mathbf{v_i}$ components" of the rows of A, i.e., the projections of the rows of A in the $\mathbf{v_i}$ direction. We will prove that A can be decomposed into a sum of rank one matrices as

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T.$$

Geometrically, each point is decomposed in A into its components along each of the r orthogonal directions given by the $\mathbf{v_i}$. We will also prove this algebraically. We begin with a simple lemma that two matrices A and B are identical if $A\mathbf{v} = B\mathbf{v}$ for all $\mathbf{v}$.

Lemma 3.3 *Matrices A and B are identical if and only if for all vectors $\mathbf{v}$, $A\mathbf{v} = B\mathbf{v}$.*

Proof Clearly, if $A = B$, then $A\mathbf{v} = B\mathbf{v}$ for all $\mathbf{v}$. For the converse, suppose that $A\mathbf{v} = B\mathbf{v}$ for all $\mathbf{v}$. Let $\mathbf{e_i}$ be the vector that is all zeros except for the i^{th} component, which has value 1. Now $A\mathbf{e_i}$ is the i^{th} column of A, and thus $A = B$ if for each i, $A\mathbf{e_i} = B\mathbf{e_i}$. ■

Theorem 3.4 *Let A be an $n \times d$ matrix with right-singular vectors $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$, left-singular vectors $\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_r}$, and corresponding singular values $\sigma_1, \sigma_2, \ldots, \sigma_r$. Then*

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T.$$

Proof We first show that multiplying both A and $\sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T$ by $\mathbf{v_j}$ results in equality.

$$\sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T \mathbf{v_j} = \sigma_j \mathbf{u_j} = A\mathbf{v_j}.$$

Since any vector $\mathbf{v}$ can be expressed as a linear combination of the singular vectors plus a vector perpendicular to the $\mathbf{v_i}$, $A\mathbf{v} = \sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T \mathbf{v}$ for all $\mathbf{v}$ and by Lemma 3.3, $A = \sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T$. ■

The decomposition $A = \sum_i \sigma_i \mathbf{u_i} \mathbf{v_i^T}$ is called the *singular value decomposition*, *SVD*, of A. We can rewrite this equation in matrix notation as $A = UDV^T$ where $\mathbf{u_i}$ is the i^{th} column of U, $\mathbf{v_i^T}$ is the i^{th} row of V^T, and D is a diagonal matrix with σ_i as the i^{th} entry on its diagonal (see Figure 3.2). For any matrix A, the sequence of singular values is unique, and if the singular values are all distinct, then the sequence of singular vectors is unique up to signs. However, when some set of singular values are equal, the corresponding singular vectors span some subspace. Any set of orthonormal vectors spanning this subspace can be used as the singular vectors.

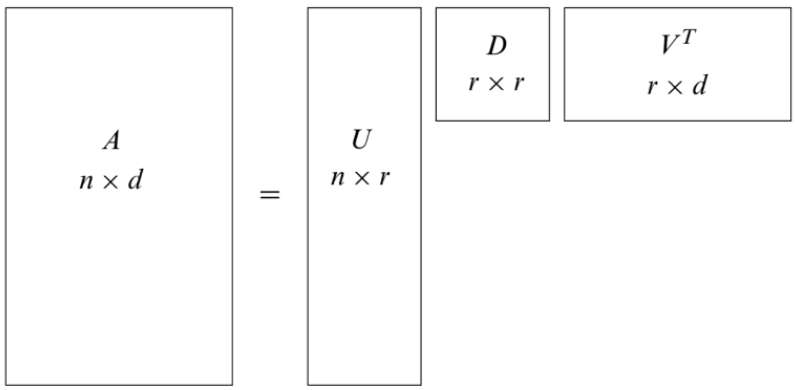


Figure 3.2: The SVD decomposition of an $n \times d$ matrix.

3.5. Best Rank-k Approximations

Let A be an $n \times d$ matrix and think of the rows of A as n points in d-dimensional space. Let

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T$$

be the SVD of A. For $k \in \{1, 2, \ldots, r\}$, let

$$A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u_i} \mathbf{v_i}^T$$

be the sum truncated after k terms. It is clear that A_k has rank k. We show that A_k is the best rank-k approximation to A, where error is measured in the Frobenius norm. Geometrically, this says that $\mathbf{v_1}, \ldots, \mathbf{v_k}$ define the k-dimensional space minimizing the sum of squared distances of the points to the space. To see why, we need the following lemma.

Lemma 3.5 *The rows of A_k are the projections of the rows of A onto the subspace V_k spanned by the first k singular vectors of A.*

Proof Let $\mathbf{a}$ be an arbitrary row vector. Since the $\mathbf{v_i}$ are orthonormal, the projection of the vector $\mathbf{a}$ onto V_k is given by $\sum_{i=1}^{k} (\mathbf{a} \cdot \mathbf{v_i}) \mathbf{v_i}^T$. Thus, the matrix whose rows are the projections of the rows of A onto V_k is given by $\sum_{i=1}^{k} A \mathbf{v_i} \mathbf{v_i}^T$. This last expression simplifies to

$$\sum_{i=1}^{k} A \mathbf{v_i} \mathbf{v_i}^T = \sum_{i=1}^{k} \sigma_i \mathbf{u_i} \mathbf{v_i}^T = A_k.$$

■

Theorem 3.6 *For any matrix B of rank at most k*

$$\|A - A_k\|_F \leq \|A - B\|_F$$

Proof Let B minimize $\|A - B\|_F^2$ among all rank k or less matrices. Let V be the space spanned by the rows of B. The dimension of V is at most k. Since B minimizes $\|A - B\|_F^2$, it must be that each row of B is the projection of the corresponding row of A onto V: Otherwise, replace the row of B with the projection of the corresponding row of A onto V. This still keeps the row space of B contained in V, and hence the rank of B is still at most k. But it reduces $\|A - B\|_F^2$, contradicting the minimality of $||A - B||_F$.

Since each row of B is the projection of the corresponding row of A, it follows that $\|A - B\|_F^2$ is the sum of squared distances of rows of A to V. Since A_k minimizes the sum of squared distance of rows of A to any k-dimensional subspace, from Theorem 3.1, it follows that $\|A - A_k\|_F \leq \|A - B\|_F$. ■

In addition to the Frobenius norm, there is another matrix norm of interest. Consider an $n \times d$ matrix A and a large number of vectors where for each vector

$\mathbf{x}$ we wish to compute $A\mathbf{x}$. It takes time $O(nd)$ to compute each product $A\mathbf{x}$, but if we approximate A by $A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u_i} \mathbf{v_i}^T$ and approximate $A\mathbf{x}$ by $A_k\mathbf{x}$, it requires only k dot products of d-dimensional vectors, followed by a sum of k n-dimensional vectors, and takes time $O(kd + kn)$, which is a win provided $k \ll \min(d, n)$. How is the error measured? Since $\mathbf{x}$ is unknown, the approximation needs to be good for every $\mathbf{x}$. So we take the maximum over all $\mathbf{x}$ of $|(A_k - A)\mathbf{x}|$. Since this would be infinite if $|\mathbf{x}|$ could grow without bound, we restrict the maximum to $|\mathbf{x}| \leq 1$. Formally, we define a new norm of a matrix A by

$$||A||_2 = \max_{|\mathbf{x}|\leq 1} |A\mathbf{x}|.$$

This is called the 2-norm or the spectral norm. Note that it equals $\sigma_1(A)$.

As an application consider a large database of documents that form rows of an $n \times d$ matrix A. There are d terms, and each document is a d-dimensional vector with one component for each term, which is the number of occurrences of the term in the document. We are allowed to "preprocess" A. After the preprocessing, we receive queries. Each query $\mathbf{x}$ is an d-dimensional vector that specifies how important each term is to the query. The desired answer is an n-dimensional vector that gives the similarity (dot product) of the query to each document in the database, namely $A\mathbf{x}$, the "matrix-vector" product. Query time is to be much less than preprocessing time, since the idea is that we need to answer many queries for the same database. There are many other applications where one performs many matrix vector products with the same matrix. This technique is applicable to these situations as well.

3.6. Left Singular Vectors

The left singular vectors are also pairwise orthogonal. Intuitively if $\mathbf{u_i}$ and $\mathbf{u_j}$, $i < j$, were not orthogonal, one would suspect that the right singular vector $\mathbf{v_j}$ had a component of $\mathbf{v_i}$, which would contradict that $\mathbf{v_i}$ and $\mathbf{v_j}$ were orthogonal. Let i be the smallest integer such that $\mathbf{u_i}$ is not orthogonal to all other $\mathbf{u_j}$. Then to prove that $\mathbf{u_i}$ and $\mathbf{u_j}$ are orthogonal, we add a small component of $\mathbf{v_j}$ to $\mathbf{v_i}$, normalize the result to be a unit vector

$$\mathbf{v_i'} = \frac{\mathbf{v_i} + \epsilon \mathbf{v_j}}{|\mathbf{v_i} + \epsilon \mathbf{v_j}|}$$

and show that $|A\mathbf{v_i'}| > |A\mathbf{v_i}|$, a contradiction.

Theorem 3.7 *The left singular vectors are pairwise orthogonal.*

Proof Let i be the smallest integer such that $\mathbf{u_i}$ is not orthogonal to some other $\mathbf{u_j}$. Without loss of generality, assume that $\mathbf{u_i}^T \mathbf{u_j} = \delta > 0$. If $\mathbf{u_i}^T \mathbf{u_j} < 0$, then just replace $\mathbf{u_i}$ with $-\mathbf{u_i}$. Clearly $j > i$, since i was selected to be the smallest such index. For $\varepsilon > 0$, let

$$\mathbf{v_i'} = \frac{\mathbf{v_i} + \epsilon \mathbf{v_j}}{|\mathbf{v_i} + \epsilon \mathbf{v_j}|}.$$

Notice that $\mathbf{v_i'}$ is a unit length vector.

$$A\mathbf{v_i'} = \frac{\sigma_i \mathbf{u_i} + \varepsilon \sigma_j \mathbf{u_j}}{\sqrt{1 + \varepsilon^2}}$$

Now consider computing B^2.

$$B^2 = \left(\sum_i \sigma_i^2 \mathbf{v_i}\mathbf{v_i}^T\right)\left(\sum_j \sigma_j^2 \mathbf{v_j}\mathbf{v_j}^T\right) = \sum_{ij} \sigma_i^2\sigma_j^2 \mathbf{v_i}(\mathbf{v_i}^T\mathbf{v_j})\mathbf{v_j}^T.$$

When $i \neq j$, the dot product $\mathbf{v_i}^T\mathbf{v_j}$ is zero by orthogonality.[6] Thus, $B^2 = \sum_{i=1}^r \sigma_i^4 \mathbf{v_i}\mathbf{v_i}^T$. In computing the k^{th} power of B, all the cross-product terms are zero and

$$B^k = \sum_{i=1}^r \sigma_i^{2k} \mathbf{v_i}\mathbf{v_i}^T.$$

If $\sigma_1 > \sigma_2$, then the first term in the summation dominates, so $B^k \to \sigma_1^{2k}\mathbf{v_1}\mathbf{v_1}^T$. This means a close estimate to $\mathbf{v_1}$ can be computed by simply taking the first column of B^k and normalizing it to a unit vector.

3.7.1. A Faster Method

A problem with the above method is that A may be a very large, sparse matrix, say a $10^8 \times 10^8$ matrix with 10^9 non-zero entries. Sparse matrices are often represented by just a list of non-zero entries, say a list of triples of the form (i, j, a_{ij}). Though A is sparse, B need not be and in the worse case may have all 10^{16} entries non-zero[7] and it is then impossible to even write down B, let alone compute the product B^2. Even if A is moderate in size, computing matrix products is costly in time. Thus, a more efficient method is needed.

Instead of computing B^k, select a random vector $\mathbf{x}$ and compute the product $B^k\mathbf{x}$. The vector $\mathbf{x}$ can be expressed in terms of the singular vectors of B augmented to a full orthonormal basis as $\mathbf{x} = \sum_{i=1}^d c_i\mathbf{v_i}$. Then

$$B^k\mathbf{x} \approx (\sigma_1^{2k}\mathbf{v_1}\mathbf{v_1}^T)\left(\sum_{i=1}^d c_i\mathbf{v_i}\right) = \sigma_1^{2k}c_1\mathbf{v_1}.$$

Normalizing the resulting vector yields $\mathbf{v_1}$, the first singular vector of A. The way $B^k\mathbf{x}$ is computed is by a series of matrix vector products, instead of matrix products. $B^k\mathbf{x} = A^TA\ldots A^TA\mathbf{x}$, which can be computed right-to-left. This consists of $2k$ vector times sparse matrix multiplications.

To compute k singular vectors, one selects a random vector $\mathbf{r}$ and finds an orthonormal basis for the space spanned by $\mathbf{r}, A\mathbf{r}, \ldots, A^{k-1}\mathbf{r}$. Then compute A times each of the basis vectors, and find an orthonormal basis for the space spanned by the resulting vectors. Intuitively, one has applied A to a subspace rather than a single vector. One repeatedly applies A to the subspace, calculating an orthonormal basis after each application to prevent the subspace collapsing to the one-dimensional subspace spanned by the first singular vector. The process quickly converges to the first k singular vectors.

An issue occurs if there is no significant gap between the first and second singular values of a matrix. Take for example the case when there is a tie for the first singular vector and $\sigma_1 = \sigma_2$. Then, the above argument fails. We will overcome this hurdle.

[6] The "outer product" $\mathbf{v_i}\mathbf{v_j}^T$ is a matrix and is not zero even for $i \neq j$.
[7] E.g., suppose each entry in the first row of A is non-zero and the rest of A is zero.

Theorem 3.11 below states that even with ties, the power method converges to some vector in the span of those singular vectors corresponding to the "nearly highest" singular values. The theorem assumes it is given a vector $\mathbf{x}$, which has a component of magnitude at least δ along the first right singular vector $\mathbf{v_1}$ of A. We will see in Lemma 3.12 that a random vector satisfies this condition with fairly high probability.

Theorem 3.11 *Let A be an $n \times d$ matrix and $\mathbf{x}$ a unit length vector in $\mathbf{R}^d$ with $|\mathbf{x}^T\mathbf{v_1}| \geq \delta$, where $\delta > 0$. Let V be the space spanned by the right singular vectors of A corresponding to singular values greater than $(1-\varepsilon)\,\sigma_1$. Let $\mathbf{w}$ be the unit vector after $k = \frac{\ln(1/\varepsilon\delta)}{2\varepsilon}$ iterations of the power method, namely*

$$\mathbf{w} = \frac{\left(A^TA\right)^k \mathbf{x}}{\left|\left(A^TA\right)^k \mathbf{x}\right|}.$$

Then $\mathbf{w}$ has a component of at most ε perpendicular to V.

Proof Let

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T$$

be the SVD of A. If the rank of A is less than d, then for convenience complete $\{\mathbf{v_1}, \mathbf{v_2}, \ldots \mathbf{v}_r\}$ into an orthonormal basis $\{\mathbf{v_1}, \mathbf{v_2}, \ldots \mathbf{v}_d\}$ of d-space. Write $\mathbf{x}$ in the basis of the $\mathbf{v_i}$'s as

$$\mathbf{x} = \sum_{i=1}^{d} c_i \mathbf{v_i}.$$

Since $(A^TA)^k = \sum_{i=1}^{d} \sigma_i^{2k} \mathbf{v_i}\mathbf{v_i}^T$, it follows that $(A^TA)^k\mathbf{x} = \sum_{i=1}^{d} \sigma_i^{2k} c_i \mathbf{v_i}$. By hypothesis, $|c_1| \geq \delta$.

Suppose that $\sigma_1, \sigma_2, \ldots, \sigma_m$ are the singular values of A that are greater than or equal to $(1-\varepsilon)\,\sigma_1$ and that $\sigma_{m+1}, \ldots, \sigma_d$ are the singular values that are less than $(1-\varepsilon)\,\sigma_1$. Now,

$$|(A^TA)^k\mathbf{x}|^2 = \left|\sum_{i=1}^{d} \sigma_i^{2k} c_i \mathbf{v_i}\right|^2 = \sum_{i=1}^{d} \sigma_i^{4k} c_i^2 \geq \sigma_1^{4k} c_1^2 \geq \sigma_1^{4k} \delta^2.$$

The component of $|(A^TA)^k\mathbf{x}|^2$ perpendicular to the space V is

$$\sum_{i=m+1}^{d} \sigma_i^{4k} c_i^2 \leq (1-\varepsilon)^{4k}\, \sigma_1^{4k} \sum_{i=m+1}^{d} c_i^2 \leq (1-\varepsilon)^{4k}\, \sigma_1^{4k},$$

since $\sum_{i=1}^{d} c_i^2 = |\mathbf{x}| = 1$. Thus, the component of $\mathbf{w}$ perpendicular to V has squared length at most $\frac{(1-\varepsilon)^{4k}\sigma_1^{4k}}{\sigma_1^{4k}\delta^2}$ and so its length is at most

$$\frac{(1-\varepsilon)^{2k}\sigma_1^{2k}}{\delta\sigma_1^{2k}} = \frac{(1-\varepsilon)^{2k}}{\delta} \leq \frac{e^{-2k\varepsilon}}{\delta} = \varepsilon,$$

since $k = \frac{\ln(1/\epsilon\delta)}{2\epsilon}$. ■

Lemma 3.12 *Let* $\mathbf{y} \in \mathbf{R}^n$ *be a random vector with the unit-variance spherical Gaussian as its probability density. Normalize y to be a unit length vector by setting* $\mathbf{x} = \mathbf{y}/|\mathbf{y}|$. *Let* $\mathbf{v}$ *be any unit length vector. Then*

$$Prob\left(|\mathbf{x}^T\mathbf{v}| \leq \frac{1}{20\sqrt{d}}\right) \leq \frac{1}{10} + 3e^{-d/96}.$$

Proof Proving for the unit length vector $\mathbf{x}$ that $\text{Prob}\left(|\mathbf{x}^T\mathbf{v}| \leq \frac{1}{20\sqrt{d}}\right) \leq \frac{1}{10} + 3e^{-d/96}$ is equivalent to proving for the unnormalized vector $\mathbf{y}$ that $\text{Prob}(|\mathbf{y}| \geq 2\sqrt{d}) \leq 3e^{-d/96}$ and $\text{Prob}(|\mathbf{y}^T\mathbf{v}| \leq \frac{1}{10}) \leq 1/10$. That $\text{Prob}(|\mathbf{y}| \geq 2\sqrt{d})$ is at most $3e^{-d/96}$ follows from Theorem 2.9 with $\sqrt{d}$ substituted for β. The probability that $|\mathbf{y}^T\mathbf{v}| \leq \frac{1}{10}$ is at most $1/10$ because $\mathbf{y}^T\mathbf{v}$ is a random, zero mean, unit-variance Gaussian with density at most $1/\sqrt{2\pi} \leq 1/2$ in the interval $[-1/10, 1/10]$, so the integral of the Gaussian over the interval is at most 1/10. ■

3.8. Singular Vectors and Eigenvectors

For a square matrix B, if $B\mathbf{x} = \lambda\mathbf{x}$, then $\mathbf{x}$ is an *eigenvector* of B and λ is the corresponding *eigenvalue*. We saw in Section 3.7, if $B = A^TA$, then the right singular vectors $\mathbf{v_j}$ of A are eigenvectors of B with eigenvalues σ_j^2. The same argument shows that the left singular vectors $\mathbf{u_j}$ of A are eigenvectors of AA^T with eigenvalues σ_j^2.

The matrix $B = A^TA$ has the property that for any vector $\mathbf{x}, \mathbf{x}^TB\mathbf{x} \geq 0$. This is because $B = \sum_i \sigma_i^2\mathbf{v_i}\mathbf{v_i}^T$ and for any $\mathbf{x}$, $\mathbf{x}^T\mathbf{v_i}\mathbf{v_i}^T\mathbf{x} = (\mathbf{x}^T\mathbf{v_i})^2 \geq 0$. A matrix B with the property that $\mathbf{x}^TB\mathbf{x} \geq 0$ for all $\mathbf{x}$ is called *positive semi-definite*. Every matrix of the form A^TA is positive semi-definite. In the other direction, any positive semi-definite matrix B can be decomposed into a product A^TA, and so its eigenvalue decomposition can be obtained from the singular value decomposition of A. The interested reader should consult a linear algebra book.

3.9. Applications of Singular Value Decomposition

3.9.1. Centering Data

Singular value decomposition is used in many applications, and for some of these applications it is essential to first center the data by subtracting the centroid of the data from each data point.[8] If you are interested in the statistics of the data and how it varies in relationship to its mean, then you would center the data. On the other hand, if you are interested in finding the best low-rank approximation to a matrix, then you do not center the data. The issue is whether you are finding the best-fitting subspace or the best-fitting affine space. In the latter case you first center the data and then find the best-fitting subspace. See Figure 3.3.

We first show that the line minimizing the sum of squared distances to a set of points, if not restricted to go through the origin, must pass through the centroid of the points. This implies that if the centroid is subtracted from each data point, such a

[8] The centroid of a set of points is the coordinate-wise average of the points.

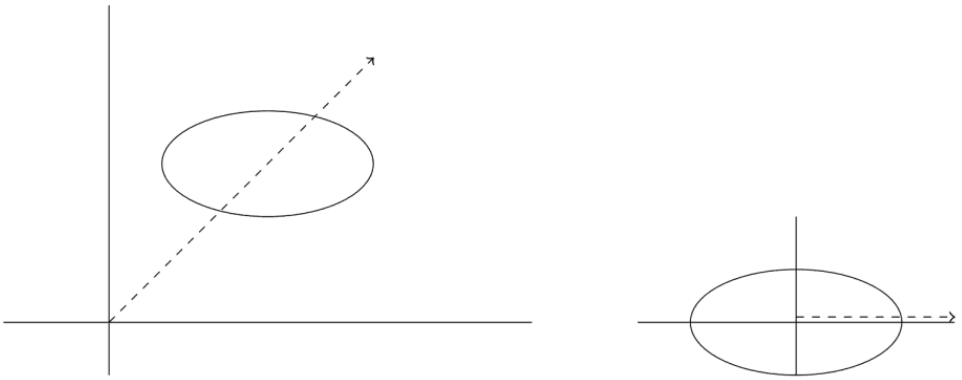


Figure 3.3: If one wants statistical information relative to the mean of the data, one needs to center the data. If one wants the best low-rank approximation, one would not center the data.

line will pass through the origin. The best-fit line can be generalized to k dimensional "planes." The operation of subtracting the centroid from all data points is useful in other contexts as well. We give it the name *centering data*.

Lemma 3.13 *The best-fit line (minimizing the sum of perpendicular distances squared) of a set of data points must pass through the centroid of the points.*

Proof Subtract the centroid from each data point so that the centroid is $\mathbf{0}$. After centering the data, let ℓ be the best-fit line and assume for contradiction that ℓ does not pass through the origin. The line ℓ can be written as $\{\mathbf{a} + \lambda\mathbf{v} | \lambda \in \mathbf{R}\}$, where $\mathbf{a}$ is the closest point to $\mathbf{0}$ on ℓ and $\mathbf{v}$ is a unit length vector in the direction of ℓ, which is perpendicular to $\mathbf{a}$. For a data point $\mathbf{a_i}$, let $dist(\mathbf{a_i}, \ell)$ denote its perpendicular distance to ℓ. By the Pythagorean theorem, we have $|\mathbf{a_i} - \mathbf{a}|^2 = dist(\mathbf{a_i}, \ell)^2 + (\mathbf{v} \cdot \mathbf{a_i})^2$, or equivalently, $dist(\mathbf{a_i}, \ell)^2 = |\mathbf{a_i} - \mathbf{a}|^2 - (\mathbf{v} \cdot \mathbf{a_i})^2$. Summing over all data points:

$$\begin{aligned}\sum_{i=1}^{n} dist(\mathbf{a_i}, \ell)^2 &= \sum_{i=1}^{n} \left(|\mathbf{a_i} - \mathbf{a}|^2 - (\mathbf{v} \cdot \mathbf{a_i})^2 \right) = \sum_{i=1}^{n} \left(|\mathbf{a_i}|^2 + |\mathbf{a}|^2 - 2\mathbf{a_i} \cdot \mathbf{a} - (\mathbf{v} \cdot \mathbf{a_i})^2 \right) \\ &= \sum_{i=1}^{n} |\mathbf{a_i}|^2 + n|\mathbf{a}|^2 - 2\mathbf{a} \cdot \left(\sum_i \mathbf{a_i} \right) - \sum_{i=1}^{n} (\mathbf{v} \cdot \mathbf{a_i})^2 \\ &= \sum_i |\mathbf{a_i}|^2 + n|\mathbf{a}|^2 - \sum_i (\mathbf{v} \cdot \mathbf{a_i})^2,\end{aligned}$$

where we used the fact that since the centroid is $\mathbf{0}$, $\sum_i \mathbf{a_i} = 0$. The above expression is minimized when $\mathbf{a} = \mathbf{0}$, so the line $\ell' = \{\lambda\mathbf{v} : \lambda \in \mathbf{R}\}$ through the origin is a better fit than ℓ, contradicting ℓ being the best-fit line. ■

A statement analogous to Lemma 3.13 holds for higher-dimensional objects. Define an *affine space* as a subspace translated by a vector. So an affine space is a set of the form

$$\left\{ \mathbf{v_0} + \sum_{i=1}^{k} c_i \mathbf{v_i} | c_1, c_2, \ldots, c_k \in \mathbf{R} \right\}.$$

Here, $\mathbf{v_0}$ is the translation and $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_k}$ form an orthonormal basis for the subspace.

Lemma 3.14 *The k-dimensional affine space that minimizes the sum of squared perpendicular distances to the data points must pass through the centroid of the points.*

Proof We only give a brief idea of the proof, which is similar to the previous lemma. Instead of $(\mathbf{v} \cdot \mathbf{a_i})^2$, we will now have $\sum_{j=1}^{k} (\mathbf{v_j} \cdot \mathbf{a_i})^2$, where the $\mathbf{v_j}, j = 1, 2, \ldots, k$ are an orthonormal basis of the subspace through the origin parallel to the affine space. ■

3.9.2. Principal Component Analysis

The traditional use of SVD is in principal component analysis (PCA). PCA is illustrated by a movie recommendation setting where there are n customers and d movies. Let matrix A with elements a_{ij} represent the amount that customer i likes movie j. One hypothesizes that there are only k underlying basic factors that determine how much a given customer will like a given movie, where k is much smaller than n or d. For example, these could be the amount of comedy, drama, and action, the novelty of the story, etc. Each movie can be described as a k-dimensional vector indicating how much of these basic factors the movie has, and each customer can be described as a k-dimensional vector indicating how important each of these basic factors is to that customer. The dot product of these two vectors is hypothesized to determine how much that customer will like that movie. In particular, this means that the $n \times d$ matrix A can be expressed as the product of an $n \times k$ matrix U describing the customers and a $k \times d$ matrix V describing the movies (see Figure 3.4). Finding the best rank-k approximation A_k by SVD gives such a U and V. One twist is that A may not be exactly equal to UV, in which case $A - UV$ is treated as noise. Another issue is that SVD gives a factorization with negative entries. Nonnegative matrix factorization (NMF) is more appropriate in some contexts where we want to keep entries nonnegative. NMF is discussed in Chapter 9.

In the above setting, A was available fully, and we wished to find U and V to identify the basic factors. However, in a case such as movie recommendations, each

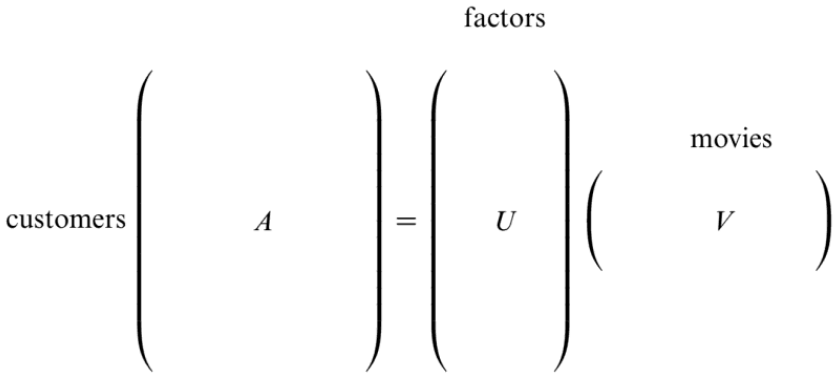


Figure 3.4: Customer-movie data

for $k \ll d$. Interestingly, we will see that the subspace spanned by the k-centers is essentially the best-fit k-dimensional subspace that can be found by singular value decomposition.

Lemma 3.15 *Suppose p is a d-dimensional spherical Gaussian with center $\boldsymbol{\mu}$ and standard deviation σ. The density of p projected onto a k-dimensional subspace V is a spherical Gaussian with the same standard deviation.*

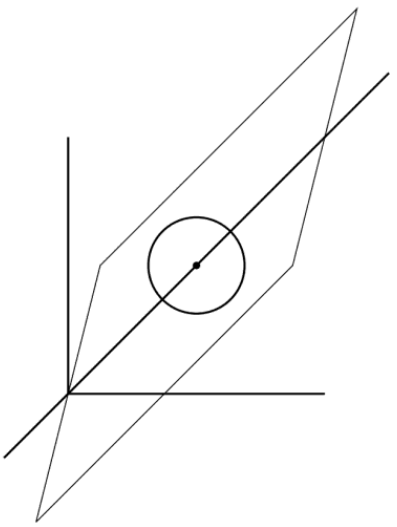
Proof Rotate the coordinate system so V is spanned by the first k coordinate vectors. The Gaussian remains spherical with standard deviation σ, although the coordinates of its center have changed. For a point $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, we will use the notation $\mathbf{x}' = (x_1, x_2, \ldots x_k)$ and $\mathbf{x}'' = (x_{k+1}, x_{k+2}, \ldots, x_n)$. The density of the projected Gaussian at the point $(x_1, x_2, \ldots, x_k)$ is

$$ce^{-\frac{|\mathbf{x}'-\boldsymbol{\mu}'|^2}{2\sigma^2}} \int_{\mathbf{x}''} e^{-\frac{|\mathbf{x}''-\boldsymbol{\mu}''|^2}{2\sigma^2}} d\mathbf{x}'' = c'e^{-\frac{|\mathbf{x}'-\boldsymbol{\mu}'|^2}{2\sigma^2}}.$$

This implies the lemma. ■

We now show that the top k singular vectors produced by the SVD span the space of the k centers. First, we extend the notion of best fit to probability distributions. Then we show that for a single spherical Gaussian whose center is not the origin, the best-fit one-dimensional subspace is the line though the center of the Gaussian and the origin. Next, we show that the best-fit k-dimensional subspace for a single Gaussian whose center is not the origin is any k-dimensional subspace containing the line through the Gaussian's center and the origin. Finally, for k spherical Gaussians, the best-fit k-dimensional subspace is the subspace containing their centers. Thus, the SVD finds the subspace that contains the centers (see Figure 3.5).

Recall that for a set of points, the best-fit line is the line passing through the origin that maximizes the sum of squared lengths of the projections of the points onto the line. We extend this definition to probability densities instead of a set of points.



1. The best fit 1-dimension subspace to a spherical Gaussian is the line through its center and the origin.
2. Any k-dimensional subspace containing the line is a best fit k-dimensional subspace for the Gaussian.
3. The best fit k-dimensional subspace for k spherical Gaussians is the subspace containing their centers.

Figure 3.5: Best fit subspace to a spherical Gaussian.

Definition 3.1 *If p is a probability density in d space, the best-fit line for p is the line in the $\mathbf{v}_1$ direction where*

$$\mathbf{v}_1 = \arg\max_{|\mathbf{v}|=1} \underset{\mathbf{x}\sim p}{E}\left[(\mathbf{v}^T\mathbf{x})^2\right].$$

For a spherical Gaussian centered at the origin, it is easy to see that any line passing through the origin is a best-fit line. Our next lemma shows that the best-fit line for a spherical Gaussian centered at $\boldsymbol{\mu} \neq 0$ is the line passing through $\boldsymbol{\mu}$ and the origin.

Lemma 3.16 *Let the probability density p be a spherical Gaussian with center $\boldsymbol{\mu} \neq 0$. The unique best-fit one-dimensional subspace is the line passing through $\boldsymbol{\mu}$ and the origin. If $\boldsymbol{\mu} = 0$, then any line through the origin is a best-fit line.*

Proof For a randomly chosen $\boldsymbol{x}$ (according to p) and a fixed unit length vector $\mathbf{v}$,

$$\begin{aligned}
\underset{\mathbf{x}\sim p}{E}\left[(\mathbf{v}^T\mathbf{x})^2\right] &= \underset{\mathbf{x}\sim p}{E}\left[v\mathbf{v}^T(\mathbf{x}-\boldsymbol{\mu}) + \mathbf{v}^T\boldsymbol{\mu}\right)^2\right] \\
&= \underset{\mathbf{x}\sim p}{E}\left[\left(\mathbf{v}^T(\mathbf{x}-\boldsymbol{\mu})\right)^2 + 2\left(\mathbf{v}^T\boldsymbol{\mu}\right)\left(\mathbf{v}^T(\mathbf{x}-\boldsymbol{\mu})\right) + \left(\mathbf{v}^T\boldsymbol{\mu}\right)^2\right] \\
&= \underset{\mathbf{x}\sim p}{E}\left[\left(\mathbf{v}^T(\mathbf{x}-\boldsymbol{\mu})\right)^2\right] + 2\left(\mathbf{v}^T\boldsymbol{\mu}\right)E\left[\mathbf{v}^T(\mathbf{x}-\boldsymbol{\mu})\right] + \left(\mathbf{v}^T\boldsymbol{\mu}\right)^2 \\
&= \underset{\mathbf{x}\sim p}{E}\left[\left(\mathbf{v}^T(\mathbf{x}-\boldsymbol{\mu})\right)^2\right] + \left(\mathbf{v}^T\boldsymbol{\mu}\right)^2 \\
&= \sigma^2 + \left(\mathbf{v}^T\boldsymbol{\mu}\right)^2
\end{aligned}$$

where the fourth line follows from the fact that $E[\mathbf{v}^T(\mathbf{x}-\boldsymbol{\mu})] = 0$, and the fifth line follows from the fact that $E[(\mathbf{v}^T(\mathbf{x}-\boldsymbol{\mu}))^2]$ is the variance in the direction $\boldsymbol{v}$. The best-fit line $\mathbf{v}$ maximizes $E_{\mathbf{x}\sim p}[(\mathbf{v}^T\mathbf{x})^2]$ and therefore maximizes $\left(\mathbf{v}^T\boldsymbol{\mu}\right)^2$. This is maximized when $\boldsymbol{v}$ is aligned with the center $\boldsymbol{\mu}$. To see uniqueness, just note that if $\boldsymbol{\mu} \neq 0$, then $\mathbf{v}^T\boldsymbol{\mu}$ is strictly less when $\mathbf{v}$ is not aligned with the center. ■

We now extend Definition 3.1 to k-dimensional subspaces.

Definition 3.2 *If p is a probability density in d-space then the best-fit k-dimensional subspace V_k is*

$$V_k = \underset{\substack{V \\ dim(V)=k}}{\operatorname{argmax}} \underset{\mathbf{x}\sim p}{E}\left(|\operatorname{proj}(\mathbf{x}, V)|^2\right),$$

where proj$(\mathbf{x}, V)$ is the orthogonal projection of $\mathbf{x}$ onto V.

Lemma 3.17 *For a spherical Gaussian with center $\boldsymbol{\mu}$, a k-dimensional subspace is a best-fit subspace if and only if it contains $\boldsymbol{\mu}$.*

Proof If $\boldsymbol{\mu} = \mathbf{0}$, then by symmetry any k-dimensional subspace is a best-fit subspace. If $\boldsymbol{\mu} \neq \mathbf{0}$, then the best-fit line must pass through $\boldsymbol{\mu}$ by Lemma 3.16. Now, as in the greedy algorithm for finding subsequent singular vectors, we would project perpendicular to the first singular vector. But after the projection,

the mean of the Gaussian becomes $\mathbf{0}$, and any vectors will do as subsequent best-fit directions. ■

This leads to the following theorem.

Theorem 3.18 *If p is a mixture of k spherical Gaussians, then the best-fit k-dimensional subspace contains the centers. In particular, if the means of the Gaussians are linearly independent, the space spanned by them is the unique best-fit k-dimensional subspace.*

Proof Let p be the mixture $w_1p_1 + w_2p_2 + \cdots + w_kp_k$. Let V be any subspace of dimension k or less. Then,

$$\underset{\mathbf{x}\sim p}{E}\left(|\text{proj}(\mathbf{x}, V)|^2\right) = \sum_{i=1}^{k} w_i \underset{\mathbf{x}\sim p_i}{E}\left(|\text{proj}(\mathbf{x}, V)|^2\right).$$

If V contains the centers of the densities p_i, by Lemma 3.17, each term in the summation is individually maximized, which implies the entire summation is maximized, proving the theorem. ■

For an infinite set of points drawn according to the mixture, the k-dimensional SVD subspace gives exactly the space of the centers. In reality, we have only a large number of samples drawn according to the mixture. However, it is intuitively clear that as the number of samples increases, the set of sample points will approximate the probability density, and so the SVD subspace of the sample will be close to the space spanned by the centers. The details of how close it gets as a function of the number of samples are technical, and we do not carry this out here.

3.9.4. Ranking Documents and Web Pages

An important task for a document collection is to rank the documents according to their intrinsic relevance to the collection. A good candidate definition of "intrinsic relevance" is a document's projection onto the best-fit direction for that collection, namely the top left-singular vector of the term-document matrix. An intuitive reason for this is that this direction has the maximum sum of squared projections of the collection and so can be thought of as a synthetic term-document vector best representing the document collection.

Ranking in order of the projection of each document's term vector along the best-fit direction has a nice interpretation in terms of the power method. For this, we consider a different example, that of the web with hypertext links. The World Wide Web can be represented by a directed graph whose nodes correspond to web pages and directed edges to hypertext links between pages. Some web pages, called *authorities*, are the most prominent sources for information on a given topic. Other pages, called *hubs*, are ones that identify the authorities on a topic. Authority pages are pointed to by many hub pages and hub pages point to many authorities. One is led to what seems like a circular definition: a hub is a page that points to many authorities and an authority is a page that is pointed to by many hubs.

One would like to assign hub weights and authority weights to each node of the web. If there are n nodes, the hub weights form an n-dimensional vector $\mathbf{u}$ and the

authority weights form an n-dimensional vector $\mathbf{v}$. Suppose A is the adjacency matrix representing the directed graph. Here a_{ij} is 1 if there is a hypertext link from page i to page j and 0 otherwise. Given hub vector $\mathbf{u}$, the authority vector $\mathbf{v}$ could be computed by the formula

$$v_j \propto \sum_{i=1}^{d} u_i a_{ij},$$

since the right hand side is the sum of the hub weights of all the nodes that point to node j. In matrix terms,

$$\mathbf{v} = A^T\mathbf{u}/|A^T\mathbf{u}|.$$

Similarly, given an authority vector $\mathbf{v}$, the hub vector $\mathbf{u}$ could be computed by $\mathbf{u} = A\mathbf{v}/|A\mathbf{v}|$. Of course, at the start, we have neither vector. But the above discussion suggests a power iteration. Start with any $\mathbf{v}$. Set $\mathbf{u} = A\mathbf{v}$, then set $\mathbf{v} = A^T\mathbf{u}$, then renormalize and repeat the process. We know from the power method that this converges to the left-singular and right-singular vectors. So after sufficiently many iterations, we may use the left vector $\mathbf{u}$ as the hub weights vector and project each column of A onto this direction and rank columns (authorities) in order of this projection. But the projections just form the vector $A^T\mathbf{u}$ that equals a multiple of $\mathbf{v}$. So we can just rank by order of the v_j. This is the basis of an algorithm called the HITS algorithm, which was one of the early proposals for ranking web pages.

A different ranking called *pagerank* is widely used. It is based on a random walk on the graph described above. We will study random walks in detail in Chapter 4.

3.9.5. An Illustrative Application of SVD

A deep neural network in which inputs images are classified by category such as cat, dog, or car maps an image to an activation space. The dimension of the activation space might be 4,000, but the set of cat images might be mapped to a much lower-dimensional manifold. To determine the dimension of the cat manifold, we could construct a tangent subspace at an activation vector for a cat image. However, we only have 1,000 cat images and they are spread far apart in the activation space. We need a large number of cat activation vectors close to each original cat activation vector to determine the dimension of the tangent subspace. To do this we want to slightly modify each cat image to get many images that are close to the original. One way to do this is to do a singular value decomposition of an image and zero out a few very small singular values. If the image is 1,000 by 1,000, there will be 1,000 singular values. The smallest 100 will be essentially zero, and zeroing out a subset of them should not change the image much and produce images whose activation vectors are very close. Since there are $\binom{100}{10}$ subsets of 10 singular values, we can generate, say, 10,000 such images by zeroing out 10 singular values. Given the corresponding activation vectors, we can form a matrix of activation vectors and determine the rank of the matrix, which should give the dimension of the tangent subspace to the original cat activation vector.

To determine the rank of the matrix of 10,000 activation vectors, we again do a singular value decomposition. To determine the actual rank, we need to determine a cutoff point below which we conclude the remaining singular values are noise.

We might consider a sufficient number of the largest singular values so that their sum of squares is 95% of the square of the Frobenius norm of the matrix or look to see where there is a sharp drop in the singular values.

3.9.6. An Application of SVD to a Discrete Optimization Problem

In clustering a mixture of Gaussians, SVD was used as a dimension reduction technique. It found a k-dimensional subspace (the space of centers) of a d-dimensional space and made the Gaussian clustering problem easier by projecting the data to the subspace. Here, instead of fitting a model to data, we consider an optimization problem where applying dimension reduction makes the problem easier. The use of SVD to solve discrete optimization problems is a relatively new subject with many applications. We start with an important NP-hard problem, the maximum cut problem for a directed graph $G(V, E)$.

The maximum cut problem is to partition the nodes of an n-node directed graph into two subsets S and $\bar{S}$ so that the number of edges from S to $\bar{S}$ is maximized. Let A be the adjacency matrix of the graph. With each vertex i, associate an indicator variable x_i. The variable x_i will be set to 1 for $i \in S$ and 0 for $i \in \bar{S}$. The vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is unknown, and we are trying to find it or, equivalently, the cut, so as to maximize the number of edges across the cut. The number of edges across the cut is precisely

$$\sum_{i,j} x_i(1 - x_j)a_{ij}.$$

Thus, the maximum cut problem can be posed as the optimization problem:

$$\text{Maximize} \sum_{i,j} x_i(1 - x_j)a_{ij} \quad \text{subject to } x_i \in \{0, 1\}.$$

In matrix notation,

$$\sum_{i,j} x_i(1 - x_j)a_{ij} = \mathbf{x}^T A(\mathbf{1} - \mathbf{x}),$$

where $\mathbf{1}$ denotes the vector of all 1's. So, the problem can be restated as

$$\text{Maximize } \mathbf{x}^T A(\mathbf{1} - \mathbf{x}) \quad \text{subject to } x_i \in \{0, 1\}. \tag{3.1}$$

This problem is NP-hard. However we will see that for dense graphs – that is, graphs with $\Omega(n^2)$ edges and therefore whose optimal solution has size $\Omega(n^2)$[10] – we can use the SVD to find a near-optimal solution in polynomial time. To do so we will begin by computing the SVD of A and replacing A by $A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u_i} \mathbf{v_i}^T$ in (3.1) to get

$$\text{Maximize } \mathbf{x}^T A_k(\mathbf{1} - \mathbf{x}) \quad \text{subject to } x_i \in \{0, 1\}. \tag{3.2}$$

Note that the matrix A_k is no longer a 0-1 adjacency matrix.

We will show that:

1. For each 0-1 vector $\mathbf{x}$, $\mathbf{x}^T A_k(\mathbf{1} - \mathbf{x})$ and $\mathbf{x}^T A(\mathbf{1} - \mathbf{x})$ differ by at most $\frac{n^2}{\sqrt{k+1}}$. Thus, the maxima in (3.1) and (3.2) differ by at most this amount.

[10] Any graph of m edges has a cut of size at least $m/2$. This can be seen by noting that the expected size of the cut for a *random* $\mathbf{x} \in \{0, 1\}^n$ is exactly $m/2$.

3.11. Exercises

Exercise 3.1 (Least squares vertical error) In many experiments one collects the value of a parameter at various instances of time. Let y_i be the value of the parameter y at time x_i. Suppose we wish to construct the best linear approximation to the data in the sense that we wish to minimize the mean square error. Here error is measured vertically rather than perpendicular to the line. Develop formulas for m and b to minimize the mean square error of the points $\{(x_i, y_i) | 1 \leq i \leq n\}$ to the line $y = mx + b$.

Exercise 3.2 Given five observed variables – height, weight, age, income, and blood pressure of n people – how would one find the best least squares fit affine subspace of the form

$$a_1 \text{ (height)} + a_2 \text{ (weight)} + a_3 \text{ (age)} + a_4 \text{ (income)} + a_5 \text{ (blood pressure)} = a_6?$$

Here $a_1, a_2, \ldots, a_6$ are the unknown parameters. If there is a good best-fit four-dimensional affine subspace, then one can think of the points as lying close to a four-dimensional sheet rather than points lying in five dimensions. Why might it be better to use the perpendicular distance to the affine subspace rather than vertical distance where vertical distance is measured along the coordinate axis corresponding to one of the variables?

Exercise 3.3 Manually find the best-fit lines (not subspaces, which must contain the origin) through the points in the sets below. Best fit means minimize the perpendicular distance. Subtract the center of gravity of the points in the set from each of the points in the set and find the best-fit line for the resulting points. Does the best-fit line for the original data go through the origin?

1. (4,4) (6,2)
2. (4,2) (4,4) (6,2) (6,4)
3. (3,2.5) (3,5) (5,1) (5,3.5)

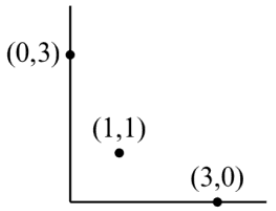
Exercise 3.4 Manually determine the best-fit line through the origin for each of the following sets of points. Is the best-fit line unique? Justify your answer for each of the subproblems.

1. $\{(0, 1), (1, 0)\}$
2. $\{(0, 1), (2, 0)\}$

Exercise 3.5 Manually find the left-singular and right-singular vectors, the singular values, and the SVD decomposition of the matrices in Figure 3.6.

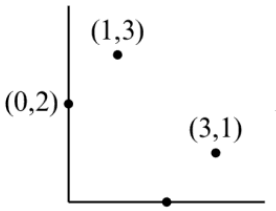
Exercise 3.6 Let A be a square $n \times n$ matrix whose rows are orthonormal. Prove that the columns of A are orthonormal.

Exercise 3.7 Suppose A is a $n \times n$ matrix with block diagonal structure with k equal size blocks where all entries of the ith block are a_i with $a_1 > a_2 > \cdots > a_k > 0$. Show that A has exactly k non-zero singular vectors $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_k}$ where $\mathbf{v_i}$ has the value $(\frac{k}{n})^{1/2}$ in the coordinates corresponding to the ith block and 0 elsewhere. In other words, the singular vectors exactly identify the blocks of the diagonal. What happens if $a_1 = a_2 = \cdots = a_k$? In the case where the a_i are equal, what is the structure of the set of all possible singular vectors?



$$M = \begin{pmatrix} 1 & 1 \\ 0 & 3 \\ 3 & 0 \end{pmatrix}$$

Figure 3.6 a



$$M = \begin{pmatrix} 0 & 2 \\ 2 & 0 \\ 1 & 3 \\ 3 & 1 \end{pmatrix}$$

Figure 3.6 b

Figure 3.6: SVD problem

Hint. By symmetry, the top singular vector's components must be constant in each block.

Exercise 3.8 Interpret the first right and left-singular vectors for the document term matrix.

Exercise 3.9 Verify that the sum of r-rank one matrices $\sum_{i=1}^{r} c_i \mathbf{x_i} \mathbf{y_i}^T$ can be written as XCY^T, where the $\mathbf{x_i}$ are the columns of X, the $\mathbf{y_i}$ are the columns of Y, and C is a diagonal matrix with the constants c_i on the diagonal.

Exercise 3.10 Let $\sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T$ be the SVD of A. Show that $\left|\mathbf{u}_1^T A\right| = \sigma_1$ and that $\left|\mathbf{u}_1^T A\right| = \max_{|\mathbf{u}|=1} \left|\mathbf{u}^T A\right|$.

Exercise 3.11 If $\sigma_1, \sigma_2, \ldots, \sigma_r$ are the singular values of A and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ are the corresponding right-singular vectors, show that

1. $A^T A = \sum_{i=1}^{r} \sigma_i^2 \mathbf{v_i} \mathbf{v_i}^T$
2. $\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_r$ are eigenvectors of $A^T A$.
3. Assuming that the eigenvectors of $A^T A$ are unique up to multiplicative constants, conclude that the singular vectors of A (which by definition must be unit length) are unique up to sign.

Exercise 3.12 Let $\sum_i \sigma_i u_i v_i^T$ be the singular value decomposition of a rank r matrix A. Let $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ be a rank k approximation to A for some $k < r$. Express the following quantities in terms of the singular values $\{\sigma_i, 1 \leq i \leq r\}$.

1. $||A_k||_F^2$
2. $||A_k||_2^2$
3. $||A - A_k||_F^2$
4. $||A - A_k||_2^2$

Exercise 3.13 If A is a symmetric matrix with distinct singular values, show that the left and right singular vectors are the same and that $A = VDV^T$.

Exercise 3.14 Use the power method to compute the singular value decomposition of the matrix

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

Exercise 3.15 Consider the matrix

$$A = \begin{pmatrix} 1 & 2 \\ -1 & 2 \\ 1 & -2 \\ -1 & -2 \end{pmatrix}$$

1. Run the power method starting from $x = \binom{1}{1}$ for $k = 3$ steps. What does this give as an estimate of v_1?
2. What actually are the v_i's, σ_i's, and u_i's? It may be easiest to do this by computing the eigenvectors of $B = A^T A$.
3. Suppose matrix A is a database of restaurant ratings: each row is a person, each column is a restaurant, and a_{ij} represents how much person i likes restaurant j. What might v_1 represent? What about u_1? How about the gap $\sigma_1 - \sigma_2$?

Exercise 3.16

1. Write a program to implement the power method for computing the first singular vector of a matrix. Apply your program to the matrix

$$A = \begin{pmatrix} 1 & 2 & 3 & \cdots & 9 & 10 \\ 2 & 3 & 4 & \cdots & 10 & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ 9 & 10 & 0 & \cdots & 0 & 0 \\ 10 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

2. Modify the power method to find the first four singular vectors of a matrix A as follows. Randomly select four vectors and find an orthonormal basis for the space spanned by the four vectors. Then multiply each of the basis vectors times A and find a new orthonormal basis for the space spanned by the resulting four vectors. Apply your method to find the first four singular vectors of matrix A from part 1. In Matlab the command orth finds an orthonormal basis for the space spanned by a set of vectors.

Exercise 3.17

1. For $n = 5, 10, \ldots, 25$ create random graphs by generating random vectors $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$. Create edges $(x_i, y_i) - (x_{i+1}, y_{i+1})$ for $i = 1 : n$ and an edge $(x_n, y_n) - (x_1, y_1)$.
2. For each graph create a new graph by selecting the midpoint of each edge for the coordinates of the vertices and add edges between vertices corresponding to the midpoints of two adjacent edges of the original graph. What happens when you iterate this process? It is best to draw the graphs.
3. Repeat the above step but normalize the vectors x and y to have unit length after each iteration. What happens?
4. One could implement the process by matrix multiplication where $x(t)$ and $y(t)$ are the vectors at the t iteration. What is the matrix A such that $x(t+1) = Ax(t)$?
5. What is the first singular vector of A and the first two singular values of A? Does this explain what happens and how long the process takes to converge?
6. If A is invertible, what happens when you run the process backwards?

Exercise 3.18 A matrix A is positive semi-definite if for all $\mathbf{x}, \mathbf{x}^T A\mathbf{x} \geq 0$.

1. Let A be a real valued matrix. Prove that $B = AA^T$ is positive semi-definite.
2. Let A be the adjacency matrix of a graph. The Laplacian of A is $L = D - A$ where D is a diagonal matrix whose diagonal entries are the row sums of A. Prove that L is positive semi-definite by showing that $L = B^T B$ where B is an m-by-n matrix with a row for each edge in the graph, a column for each vertex, and we define

$$b_{ei} = \begin{cases} -1 & \text{if } i \text{ is the endpoint of } e \text{ with lesser index} \\ 1 & \text{if } i \text{ is the endpoint of } e \text{ with greater index} \\ 0 & \text{if } i \text{ is not an endpoint of } e \end{cases}$$

Exercise 3.19 Prove that the eigenvalues of a symmetric real valued matrix are real.

Exercise 3.20 Suppose A is a square invertible matrix and the SVD of A is $A = \sum_i \sigma_i u_i v_i^T$. Prove that the inverse of A is $\sum_i \frac{1}{\sigma_i} v_i u_i^T$.

Exercise 3.21 Suppose A is square but not necessarily invertible and has SVD $A = \sum_{i=1}^r \sigma_i u_i v_i^T$. Let $B = \sum_{i=1}^r \frac{1}{\sigma_i} v_i u_i^T$. Show that $BA\mathbf{x} = \mathbf{x}$ for all $\mathbf{x}$ in the span of the right-singular vectors of A. For this reason B is sometimes called the pseudo-inverse of A and can play the role of A^{-1} in many applications.

Exercise 3.22

1. For any matrix A, show that $\sigma_k \leq \frac{||A||_F}{\sqrt{k}}$.
2. Prove that there exists a matrix B of rank at most k such that $||A - B||_2 \leq \frac{||A||_F}{\sqrt{k}}$.
3. Can the 2-norm on the left-hand side in (2) be replaced by Frobenius norm?

Exercise 3.23 Suppose an $n \times d$ matrix A is given and you are allowed to preprocess A. Then you are given a number of d-dimensional vectors $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_m}$ and for each of these vectors you must find the vector $A\mathbf{x_j}$ approximately, in the sense that you must find a vector $\mathbf{y_j}$ satisfying $|\mathbf{y_j} - A\mathbf{x_j}| \leq \varepsilon ||A||_F |\mathbf{x_j}|$. Here $\varepsilon > 0$ is a given error bound. Describe an algorithm that accomplishes this in time $O\left(\frac{d+n}{\varepsilon^2}\right)$ per $\mathbf{x_j}$ not counting the preprocessing time. Hint: use Exercise 3.22.

Exercise 3.24 Find the values of c_i to maximize $\sum_{i=1}^r c_i^2 \sigma_i^2$ where $\sigma_1^2 \geq \sigma_2^2 \geq \ldots$ and $\sum_{i=1}^r c_i^2 = 1$.

Exercise 3.25 (Document-Term Matrices) Suppose we have an $m \times n$ document-term matrix A where each row corresponds to a document and has been normalized to length 1. Define the "similarity" between two such documents by their dot product.

1. Consider a "synthetic" document whose sum of squared similarities with all documents in the matrix is as high as possible. What is this synthetic document and how would you find it?
2. How does the synthetic document in (1) differ from the center of gravity?
3. Building on (1), given a positive integer k, find a set of k synthetic documents such that the sum of squares of the mk similarities between each document in the matrix and each synthetic document is maximized. To avoid the trivial solution of selecting k copies of the document in (1), require the k synthetic