Spectral Algorithms

**Foundations and Trends® in
Theoretical Computer Science**

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is R. Kannan and S. Vempala, Spectral Algorithms, Foundation and Trends® in Theoretical Computer Science, vol 4, nos 3–4, pp 157–288, 2008

ISBN: 978-1-60198-274-2

Foundations and Trends® in Theoretical Computer Science

Volume 4 Issues 3–4, 2008

Editorial Scope

Foundations and Trends® in Theoretical Computer Science will publish survey and tutorial articles in the following topics:

- Algorithmic game theory
- Computational algebra
- Computational aspects of combinatorics and graph theory
- Computational aspects of communication
- Computational biology
- Computational complexity
- Computational geometry
- Computational learning
- Computational Models and Complexity
- Computational Number Theory
- Cryptography and information security
- Data structures
- Database theory
- Design and analysis of algorithms
- Distributed computing
- Information retrieval
- Operations Research
- Parallel algorithms
- Quantum Computation
- Randomness in Computation

Information for Librarians

Foundations and Trends® in Theoretical Computer Science, 2008, Volume 4, 4 issues. ISSN paper version 1551-305X. ISSN online version 1551-3068. Also available as a combined paper and online subscription.

Foundations and Trends® in
Theoretical Computer Science
Vol. 4, Nos. 3–4 (2008) 157–288
© 2009 R. Kannan and S. Vempala
DOI: 10.1561/0400000025

Spectral Algorithms

Ravindran Kannan[1] and Santosh Vempala[2]

[1] *Microsoft Research, India, kannan@microsoft.com*
[2] *Georgia Institute of Technology, USA, vempala@cc.gatech.edu*

Abstract

Spectral methods refer to the use of eigenvalues, eigenvectors, singular values, and singular vectors. They are widely used in Engineering, Applied Mathematics, and Statistics. More recently, spectral methods have found numerous applications in Computer Science to "discrete" as well as "continuous" problems. This monograph describes modern applications of spectral methods and novel algorithms for estimating spectral parameters. In the first part of the monograph, we present applications of spectral methods to problems from a variety of topics including combinatorial optimization, learning, and clustering. The second part of the monograph is motivated by efficiency considerations. A feature of many modern applications is the massive amount of input data. While sophisticated algorithms for matrix computations have been developed over a century, a more recent development is algorithms based on "sampling on the fly" from massive matrices. Good estimates of singular values and low-rank approximations of the whole matrix can be provably derived from a sample. Our main emphasis in the second part of the monograph is to present these sampling methods with rigorous error bounds. We also present recent extensions of spectral methods from matrices to tensors and their applications to some combinatorial optimization problems.

Contents

1

The Best-Fit Subspace

Many computational problems have explicit matrices as their input (e.g., adjacency matrices of graphs, experimental observations, etc.) while others refer to some matrix implicitly (e.g., document-term matrices, hyperlink structure, object–feature representations, network traffic, etc.). We refer to algorithms which use the spectrum, i.e., eigenvalues and vectors, singular values, and vectors, of the input data or matrices derived from the input as *Spectral Algorithms*. Such algorithms are the focus of this monograph. In the first part of this monograph, we describe applications of spectral methods in algorithms for problems from combinatorial optimization, learning, clustering, etc. In the second part, we study efficient randomized algorithms for computing basic spectral quantities such as low-rank approximations.

The Singular Value Decomposition (SVD) from linear algebra and its close relative, Principal Component Analysis (PCA), are central tools in the design of spectral algorithms. If the rows of a matrix are viewed as points in a high-dimensional space, with the columns being the coordinates, then SVD/PCA are typically used to reduce the dimensionality of these points, and solve the target problem in the lower-dimensional space. The computational advantages of such a projection are apparent; in addition, these tools are often able to highlight hidden structure in the data. Section 1 provides an introduction to SVD via an application to a generalization of the least-squares fit problem. The next three chapters are motivated by one of the most popular applications of spectral methods, namely clustering. Section 2 tackles a classical problem from Statistics, learning a mixture of Gaussians from unlabeled samples; SVD leads to the current best guarantees. Section 3 studies spectral clustering for discrete random inputs, using classical results from random matrices, while Section 4 analyzes spectral clustering for arbitrary inputs to obtain approximation guarantees. In Section 5, we turn to optimization and see the application of tensors to solving maximum constraint satisfaction problems with a bounded number of literals in each constraint. This powerful application of low-rank tensor approximation substantially extends and generalizes a large body of work.

In the second part of this monograph, we begin with algorithms for matrix multiplication and low-rank matrix approximation. These algorithms (Section 6) are based on sampling rows and columns of the matrix from explicit, easy-to-compute probability distributions and lead to approximations additive error. In Section 7, the sampling methods are refined to obtain multiplicative error guarantees. Finally, in Section 8, we see an affine-invariant extension of standard PCA and a sampling-based algorithm for low-rank tensor approximation.

To provide an in-depth and relatively quick introduction to SVD and its applicability, in this opening chapter, we consider the *best-fit subspace* problem. Finding the best-fit line for a set of data points is a classical problem. A natural measure of the quality of a line is the least-squares measure, the sum of squared (perpendicular) distances of the points to the line. A more general problem, for a set of data points in $\mathbf{R}^n$, is finding the best-fit k-dimensional subspace. SVD can be used to find a subspace that minimizes the sum of squared distances to the given set of points in polynomial time. In contrast, for other measures such as the sum of distances or the maximum distance, no polynomial-time algorithms are known.

A clustering problem widely studied in theoretical computer science is the k-median problem. In one variant, the goal is to find a set of k

points that minimize the sum of the squared distances of the data points to their nearest facilities. A natural relaxation of this problem is to find the k-dimensional subspace for which the sum of the squared distances of the data points to the subspace is minimized (we will see that this is a relaxation). We will apply SVD to solve this relaxed problem and use the solution to approximately solve the original problem.

1.1 Singular Value Decomposition

For an $n \times n$ matrix A, an eigenvalue λ and corresponding eigenvector v satisfy the equation

$$Av = \lambda v.$$

In general, i.e., if the matrix has nonzero determinant, it will have n nonzero eigenvalues (not necessarily distinct) and n corresponding eigenvectors.

Here we deal with an $m \times n$ rectangular matrix A, where the m rows denoted $A_{(1)}, A_{(2)}, \dots A_{(m)}$ are points in $\mathbf{R}^n$; $A_{(i)}$ will be a row vector.

If $m \neq n$, the notion of an eigenvalue or eigenvector does not make sense, since the vectors Av and λv have different dimensions. Instead, a *singular value* σ and corresponding *singular vectors* $u \in \boldsymbol{R}^m, v \in \boldsymbol{R}^n$ simultaneously satisfy the following two equations

1. $Av = \sigma u$
2. $u^T A = \sigma v^T$.

We can assume, without loss of generality, that u and v are unit vectors. To see this, note that a pair of singular vectors u and v must have equal length, since $u^T Av = \sigma \|u\|^2 = \sigma \|v\|^2$. If this length is not 1, we can rescale both by the same factor without violating the above equations.

Now we turn our attention to the value $\max_{\|v\|=1} \|Av\|^2$. Since the rows of A form a set of m vectors in R^n, the vector Av is a list of the projections of these vectors onto the line spanned by v, and $\|Av\|^2$ is simply the sum of the squares of those projections.

Instead of choosing v to maximize $\|Av\|^2$, the Pythagorean theorem allows us to equivalently choose v to minimize the sum of the squared distances of the points to the line through v. In this sense, v defines the line through the origin that best fits the points.

To argue this more formally, Let $d(A_{(i)}, v)$ denote the distance of the point $A_{(i)}$ to the line through v. Alternatively, we can write

$$d(A_{(i)}, v) = \|A_{(i)} - (A_{(i)}v)v^T\|.$$

For a unit vector v, the Pythagorean theorem tells us that

$$\|A_{(i)}\|^2 = \|(A_{(i)}v)v^T\|^2 + d(A_{(i)}, v)^2.$$

Thus we get the following proposition:

Proposition 1.1.

$$\begin{aligned} \max_{\|v\|=1} \|Av\|^2 &= ||A||_F^2 - \min_{\|v\|=1} \|A - (Av)v^T\|_F^2 \\ &= ||A||_F^2 - \min_{\|v\|=1} \sum_i \|A_{(i)} - (A_{(i)}v)v^T\|^2 \end{aligned}$$

Proof. We simply use the identity:

$$\|Av\|^2 = \sum_i \|(A_{(i)}v)v^T\|^2 = \sum_i \|A_{(i)}\|^2 - \sum_i \|A_{(i)} - (A_{(i)}v)v^T\|^2$$

□

The proposition says that the v which maximizes $\|Av\|^2$ is the "best-fit" vector which also minimizes $\sum_i d(A_{(i)}, v)^2$.

Next, we claim that v is in fact a singular vector.

Proposition 1.2. The vector $v_1 = \arg\max_{\|v\|=1} \|Av\|^2$ is a singular vector, and moreover $\|Av_1\|$ is the largest (or "top") singular value.

Proof. For any singular vector v,

$$(A^TA)v = \sigma A^T u = \sigma^2 v.$$

Thus, v is an eigenvector of A^TA with corresponding eigenvalue σ^2. Conversely, an eigenvector of A^TA is also a singular vector of A. To see this, let v be an eigenvector of A^TA with corresponding eigenvalue λ. Note that λ is positive, since

$$\|Av\|^2 = v^TA^TAv = \lambda v^Tv = \lambda\|v\|^2$$

and thus

$$\lambda = \frac{\|Av\|^2}{\|v\|^2}.$$

Now if we let $\sigma = \sqrt{\lambda}$ and $u = \frac{Av}{\sigma}$. it is easy to verify that u,v, and σ satisfy the singular value requirements.

The right singular vectors $\{v_i\}$ are thus exactly equal to the eigenvectors of A^TA. Since A^TA is a real, symmetric matrix, it has n orthonormal eigenvectors, which we can label $v_1,\ldots,v_n$. Expressing a unit vector v in terms of $\{v_i\}$ (i.e., $v = \sum_i \alpha_i v_i$ where $\sum_i \alpha_i^2 = 1$), we see that $\|Av\|^2 = \sum_i \sigma_i^2\alpha_i^2$ which is maximized exactly when v corresponds to the top eigenvector of A^TA. If the top eigenvalue has multiplicity greater than 1, then v should belong to the space spanned by the top eigenvectors. □

More generally, we consider a k-dimensional subspace that best fits the data. It turns out that this space is specified by the top k singular vectors, as stated precisely in the following proposition.

Theorem 1.3. Define the k-dimensional subspace V_k as the span of the following k vectors:

$$\begin{aligned} v_1 &= \arg\max_{\|v\|=1} \|Av\| \\ v_2 &= \arg\max_{\|v\|=1, v\cdot v_1=0} \|Av\| \\ &\vdots \\ v_k &= \arg\max_{\|v\|=1, v\cdot v_i=0\ \forall i<k} \|Av\|, \end{aligned}$$

where ties for any $\arg\max$ are broken arbitrarily. Then V_k is *optimal* in the sense that

$$V_k = \arg \min_{dim(V)=k} \sum_i d(A_{(i)}, V)^2.$$

Further, $v_1, v_2, \ldots, v_n$ are all singular vectors, with corresponding singular values $\sigma_1, \sigma_2, \ldots, \sigma_n$ and

$$\sigma_1 = \|Av_1\| \geq \sigma_2 = \|Av_2\| \geq \cdots \geq \sigma_n = \|Av_n\|.$$

Finally, $A = \sum_{i=1}^n \sigma_i u_i v_i^T$.

Such a decomposition where,

1. The sequence of σ_is is nonincreasing
2. The sets $\{u_i\}, \{v_i\}$ are orthonormal

is called the *Singular Value Decomposition (SVD)* of A.

Proof. We first prove that V_k are optimal by induction on k. The case $k = 1$ is by definition. Assume that V_{k-1} is optimal.

Suppose V_k' is an optimal subspace of dimension k. Then we can choose an orthonormal basis for V_k', say $w_1, w_2, \ldots w_k$, such that w_k is orthogonal to V_{k-1}. By the definition of V_k', we have that

$$||Aw_1||^2 + ||Aw_2^2|| + \ldots ||Aw_k||^2$$

is maximized (among all sets of k orthonormal vectors.) If we replace w_i by v_i for $i = 1, 2, \ldots, k-1$, we have

$$\|Aw_1\|^2 + \|Aw_2^2\| + \ldots \|Aw_k\|^2 \leq \|Av_1\|^2 + \ldots + \|Av_{k-1}\|^2 + \|Aw_k\|^2.$$

Therefore we can assume that V_k' is the span of V_{k-1} and w_k. It then follows that $\|Aw_k\|^2$ maximizes $\|Ax\|^2$ over all unit vectors x orthogonal to V_{k-1}.

Proposition 1.2 can be extended to show that $v_1, v_2, \ldots, v_n$ are all singular vectors. The assertion that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$ follows from the definition of the v_is.

We can verify that the decomposition

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

is accurate. This is because the vectors $v_1, v_2, \ldots, v_n$ form an orthonormal basis for $\boldsymbol{R}^n$, and the action of A on any v_i is equivalent to the action of $\sum_{i=1}^n \sigma_i u_i v_i^T$ on v_i. □

Note that we could actually decompose A into the form $\sum_{i=1}^n \sigma_i u_i v_i^T$ by picking $\{v_i\}$ to be any orthogonal basis of $\boldsymbol{R}_n$, but the proposition actually states something stronger: that we can pick $\{v_i\}$ in such a way that $\{u_i\}$ is also an orthogonal set.

We state one more classical theorem. We have seen that the span of the top k singular vectors is the best-fit k-dimensional subspace for the rows of A. Along the same lines, the partial decomposition of A obtained by using only the top k singular vectors is the best rank-k matrix approximation to A.

Theorem 1.4. Among all rank-k matrices D, the matrix $A_k = \sum_{i=1}^k \sigma_i u_i v_i^T$ is the one which minimizes $\|A - D\|_F^2 = \sum_{i,j}(A_{ij} - D_{ij})^2$. Further,

$$\|A - A_k\|_F^2 = \sum_{i=k+1}^{n} \sigma_i^2.$$

Proof. We have

$$\|A - D\|_F^2 = \sum_{i=1}^{m} \|A_{(i)} - D_{(i)}\|^2.$$

Since D is of rank at most k, we can assume that all the $D_{(i)}$ are projections of $A_{(i)}$ to some rank-k subspace and therefore,

$$\begin{aligned}\sum_{i=1}^{m} \|A_{(i)} - D_{(i)}\|^2 &= \sum_{i=1}^{m} \|A_{(i)}\|^2 - \|D_{(i)}\|^2 \\ &= \|A\|_F^2 - \sum_{i=1}^{m} \|D_{(i)}\|^2.\end{aligned}$$

Thus the subspace is exactly the SVD subspace given by the span of the first k singular vectors of A. □

1.2 Algorithms for Computing the SVD

Computing the SVD is a major topic of numerical analysis [48, 64, 67]. Here we describe a basic algorithm called the power method.

Assume that A is symmetric.

1. Let x be a random unit vector.
2. Repeat:

$$x := \frac{Ax}{\|Ax\|}$$

For a nonsymmetric matrix A, we can simply apply the power iteration to A^TA.

Exercise 1.5. Show that the power iteration applied k times to a symmetric matrix A finds a vector x^k such that

$$\mathsf{E}\left(\|Ax^k\|^2\right) \geq \left(\frac{1}{n}\right)^{1/k} \sigma_1^2(A).$$

[Hint: First show that $\|Ax^k\| \geq (|x \cdot v|)^{1/k}\sigma_1(A)$ where x is the starting vector and v is the top eigenvector of A; then show that for a random unit vector x, $\mathsf{E}\left((x \cdot v)^2\right) = 1/n$].

The second part of this monograph deals with faster, sampling-based algorithms.

1.3 The k-Variance Problem

This section contains a description of a clustering problem which is often called k-means in the literature and can be solved approximately using SVD. This illustrates a typical use of SVD and has a provable bound.

We are given m points $\mathcal{A} = \{A_{(1)}, A_{(2)}, \ldots A_{(m)}\}$ in n-dimensional Euclidean space and a positive integer k. The problem is to find k

points $\mathcal{B} = \{B_{(1)}, B_{(2)}, \ldots, B_{(k)}\}$ such that

$$f_{\mathcal{A}}(\mathcal{B}) = \sum_{i=1}^{m} (\mathrm{dist}(A_{(i)}, \mathcal{B}))^2$$

is minimized. Here $\mathrm{dist}(A_{(i)}, \mathcal{B})$ is the Euclidean distance of $A_{(i)}$ to its nearest point in $\mathcal{B}$. Thus, in this problem we wish to minimize the sum of squared distances to the nearest "cluster center". We call this the k-variance problem. The problem is NP-hard even for $k = 2$.

Note that the solution is given by k clusters S_j, $j = 1, 2, \ldots k$. The cluster center $B_{(j)}$ will be the centroid of the points in S_j, $j = 1, 2, \ldots, k$. This is seen from the fact that for any set $\mathcal{S} = \{X^{(1)}, X^{(2)}, \ldots, X^{(r)}\}$ and any point B we have

$$\sum_{i=1}^{r} \|X^{(i)} - B\|^2 = \sum_{i=1}^{r} \|X^{(i)} - \bar{X}\|^2 + r\|B - \bar{X}\|^2, \tag{1.1}$$

where $\bar{X}$ is the centroid $(X^{(1)} + X^{(2)} + \cdots + X^{(r)})/r$ of $\mathcal{S}$. The next exercise makes this clear.

Exercise 1.6. Show that for a set of point $X^1, \ldots, X^k \in \boldsymbol{R}^n$, the point Y that minimizes $\sum_{i=1}^{k} |X^i - Y|^2$ is their centroid. Give an example when the centroid is not the optimal choice if we minimize sum of distances rather than squared distances.

The k-variance problem is thus the problem of partitioning a set of points into clusters so that the *sum of the variances of the clusters* is minimized.

We define a relaxation called the *Continuous Clustering Problem* (CCP), as the problem of finding the subspace V of $\mathbf{R}^n$ of dimension at most k which minimizes

$$g_{\mathcal{A}}(V) = \sum_{i=1}^{m} \mathrm{dist}(A_{(i)}, V)^2.$$

The reader will recognize that this is given by the SVD. It is easy to see that the optimal value of the k-variance problem is an upper bound for the optimal value of the CCP. Indeed for any set $\mathcal{B}$ of k points,

$$f_{\mathcal{A}}(\mathcal{B}) \geq g_{\mathcal{A}}(V_{\mathcal{B}}), \tag{1.2}$$

where $V_{\mathcal{B}}$ is the subspace generated by the points in $\mathcal{B}$.

We now present a factor-2 approximation algorithm for the k-variance problem using the relaxation to the best-fit subspace. The algorithm has two parts. First we project to the k-dimensional SVD subspace. Then we solve the problem in the smaller-dimensional space using a brute-force algorithm with the following guarantee.

Theorem 1.7. The k-variance problem can be solved in $O(m^{k^2 d/2})$ time when the input $\mathcal{A} \subseteq \mathbf{R}^d$.

We describe the algorithm for the low-dimensional setting. Each set $\mathcal{B}$ of "cluster centers" defines a Voronoi diagram where cell $\mathcal{C}_i = \{X \in \mathbf{R}^d : |X - B_{(i)}| \leq |X - B_{(j)}| \text{ for } j \neq i\}$ consists of those points whose closest point in $\mathcal{B}$ is $B_{(i)}$. Each cell is a polyhedron and the total number of faces in $C_1, C_2, \ldots, C_k$ is no more than $\binom{k}{2}$ since each face is the set of points equidistant from two points of $\mathcal{B}$.

We have seen in Equation (1.1) that it is the partition of $\mathcal{A}$ that determines the best $\mathcal{B}$ (via computation of centroids) and so we can move the boundary hyperplanes of the optimal Voronoi diagram, without any face passing through a point of $\mathcal{A}$, so that each face contains at least d points of $\mathcal{A}$.

Assume that the points of $\mathcal{A}$ are in general position and $0 \notin \mathcal{A}$ (a simple perturbation argument deals with the general case). This means that each face now contains d affinely independent points of $\mathcal{A}$. We ignore the information about which side of each face to place these points and so we must try all possibilities for each face. This leads to the following enumerative procedure for solving the k- variance problem:

Algorithm: k-variance

```
1. Enumerate all sets of $t$ hyperplanes, such that
   $k \leq t \leq k(k-1)/2$ hyperplanes, and each hyperplane
   contains $d$ affinely independent points of $\mathcal{A}$. The
   number of sets is at most

   $$\sum_{t=k}^{\binom{k}{2}} \binom{\binom{m}{d}}{t} = O(m^{dk^2/2}).$$
```

```
2. Check that the arrangement defined by these
   hyperplanes has exactly $k$ cells.
3. Make one of $2^{td}$ choices as to which cell to assign
   each point of $\mathcal{A}$ which lies on a hyperplane
4. This defines a unique partition of $\mathcal{A}$. Find
   the centroid of each set in the partition and
   compute $f_{\mathcal{A}}$.
```

Now we are ready for the complete algorithm. As remarked previously, CCP can be solved by Linear Algebra. Indeed, let V be a k-dimensional subspace of $\mathbf{R}^n$ and $\bar{A}_{(1)}, \bar{A}_{(2)}, \ldots, \bar{A}_{(m)}$ be the orthogonal projections of $A_{(1)}, A_{(2)}, \ldots, A_{(m)}$ onto V. Let $\bar{A}$ be the $m \times n$ matrix with rows $\bar{A}_{(1)}, \bar{A}_{(2)}, \ldots, \bar{A}_{(m)}$. Thus $\bar{A}$ has rank at most k and

$$\|A - \bar{A}\|_F^2 = \sum_{i=1}^{m} |A_{(i)} - \bar{A}_{(i)}|^2 = \sum_{i=1}^{m} (\text{dist}(A_{(i)}, V))^2.$$

Thus to solve CCP, all we have to do is find the first k vectors of the SVD of A (since by Theorem 1.4, these minimize $\|A - \bar{\mathbf{A}}\|_F^2$ over all rank-k matrices $\bar{A}$) and take the space V_{SVD} spanned by the first k singular vectors in the row space of A.

We now show that combining SVD with the above algorithm gives a 2-approximation to the k-variance problem in arbitrary dimension. Let $\bar{\mathcal{A}} = \{\bar{A}_{(1)}, \bar{A}_{(2)}, \ldots, \bar{A}_{(m)}\}$ be the projection of $\mathcal{A}$ onto the subspace V_k. Let $\bar{\mathcal{B}} = \{\bar{B}_{(1)}, \bar{B}_{(2)}, \ldots, \bar{B}_{(k)}\}$ be the optimal solution to k-variance problem with input $\bar{\mathcal{A}}$.

Algorithm for the k-variance problem

- Compute V_k.
- Solve the k-variance problem with input $\bar{\mathcal{A}}$ to obtain $\bar{\mathcal{B}}$.
- Output $\bar{\mathcal{B}}$.

It follows from Equation (1.2) that the optimal value $Z_{\mathcal{A}}$ of the k-variance problem satisfies

$$Z_{\mathcal{A}} \geq \sum_{i=1}^{m} |A_{(i)} - \bar{A}_{(i)}|^2. \tag{1.3}$$

Note also that if $\hat{\mathcal{B}} = \{\hat{B}_{(1)}, \hat{B}_{(2)}, \ldots, \hat{B}_{(k)}\}$ is an optimal solution to the k-variance problem and $\tilde{\mathcal{B}}$ consists of the projection of the points in $\hat{\mathcal{B}}$ onto V, then

$$Z_{\mathcal{A}} = \sum_{i=1}^{m} \mathrm{dist}(A_{(i)}, \hat{\mathcal{B}})^2 \geq \sum_{i=1}^{m} \mathrm{dist}(\bar{A}_{(i)}, \tilde{\mathcal{B}})^2 \geq \sum_{i=1}^{m} \mathrm{dist}(\bar{A}_{(i)}, \bar{\mathcal{B}})^2.$$

Combining this with Equation (1.3) we get

$$2Z_{\mathcal{A}} \geq \sum_{i=1}^{m} (|A_{(i)} - \bar{A}_{(i)}|^2 + \mathrm{dist}(\bar{A}_{(i)}, \bar{\mathcal{B}})^2) = \sum_{i=1}^{m} \mathrm{dist}(A_{(i)}, \bar{\mathcal{B}})^2 = f_{\mathcal{A}}(\bar{\mathcal{B}})$$

proving that we do indeed get a 2-approximation.

Theorem 1.8. Algorithm *k-variance* finds a factor-2 approximation for the k-variance problem for m points in $\boldsymbol{R}^n$ in $O(mn^2 + m^{k^3/2})$ time.

1.4 Discussion

In this chapter, we reviewed basic concepts in linear algebra from a geometric perspective. The k-variance problem is a typical example of how SVD is used: project to the SVD subspace, then solve the original problem. In many application areas, the method known as "Principal Component Analysis" (PCA) uses the projection of a data matrix to the span of the largest singular vectors. There are several general references on SVD/PCA, e.g., [12, 48].

The application of SVD to the k-variance problem is from [33] and its hardness is from [3]. The following complexity questions are open: (1) Given a matrix A, is it NP-hard to find a rank-k matrix D that minimizes the error with respect to the L_1 norm, i.e., $\sum_{i,j} |A_{ij} - D_{ij}|$? (more generally for L_p norm for $p \neq 2$)? (2) Given a set of m points in $\boldsymbol{R}^n$, is it NP-hard to find a subspace of dimension at most k that minimizes the sum of distances of the points to the subspace? It is known that finding a subspace that minimizes the maximum distance is NP-hard [58]; see also [49].

References

[1] D. Achlioptas and F. McSherry, "On Spectral Learning of Mixtures of Distributions," in *Proceedings of COLT*, 2005.

[2] D. Achlioptas and F. McSherry, "Fast computation of low-rank matrix approximations," *Journal of the ACM*, vol. 54, no. 2, 2007.

[3] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "NP-hardness of Euclidean sum-of-squares clustering," *Machine Learning*, vol. 75, no. 2, pp. 245–248, 2009.

[4] N. Alon, W. DeLaVega, R. Kannan, and M. Karpinski, "Random sub-problems of Max-SNP problems," *Proceedings of the 34th Annual ACM Symposium on Theory on Computing*, pp. 668–677, 2002.

[5] N. Alon, M. Krivelevich, and B. Sudakov, "Finding a large hidden clique in a random graph," *Random Structures and Algorithms*, vol. 13, pp. 457–466, 1998.

[6] S. Arora and R. Kannan, "Learning mixtures of arbitrary Gaussians," *Annals of Applied Probability*, vol. 15, no. 1A, pp. 69–92, 2005.

[7] S. Arora, D. Karger, and M. Karpinski, "Polynomial time approximation schemes for dense instances of NP-hard problems," *Proceedings of the 27th Annual ACM Symposium on Theory of Computing*, pp. 284–293, 1995.

[8] S. Arora, S. Rao, and U. Vazirani, "Expander flows, geometric embeddings and graph partitioning," in *STOC '04: Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pp. 222–231, 2004.

[9] D. Arthur and S. Vassilvitskii, "k-means++: The Advantages of Careful Seeding," in *Proceedings of SODA*, 2007.

[10] Y. Azar, A. Fiat, A. Karlin, and F. McSherry, "Spectral analysis of data," in *Proceedings of STOC*, pp. 619–626, 2001.

[11] R. Bhatia, "Matrix factorizations and their perturbations," *Linear Algebra and its applications*, vol. 197, 198, pp. 245–276, 1994.

[12] R. Bhatia, *Matrix Analysis*. Springer, 1997.

[13] R. Boppana, "Eigenvalues and graph bisection: An average-case analysis," in *Proceedings of the 28th IEEE Symposium on Foundations of Computer Science*, pp. 280–285, 1987.

[14] S. C. Brubaker, "Robust PCA and Clustering on Noisy Mixtures," in *Proceedings of SODA*, 2009.

[15] S. C. Brubaker and S. Vempala, "Isotropic PCA and affine-invariant clustering," in *Building Bridges Between Mathematics and Computer Science*, 19, (M. Grötschel and G. Katona, eds.), Bolyao Society Mathematical Studies, 2008.

[16] M. Charikar, S. Guha, Éva Tardos, and D. B. Shmoys, "A constant-factor approximation algorithm for the k-median problem," in *Proceedings of the 31st Annual ACM Symposium on Theory of Computing*, pp. 1–10, 1999.

[17] K. Chaudhuri and S. Rao, "Beyond Gaussians: Spectral Methods for Learning Mixtures of Heavy-Tailed Distributions," in *Proceedings of COLT*, 2008.

[18] K. Chaudhuri and S. Rao, "Learning mixtures of product distributions using correlations and independence," in *Proceedings of COLT*, 2008.

[19] D. Cheng, R. Kannan, S. Vempala, and G. Wang, "A divide-and-merge methodology for clustering," *ACM Transactions on Database Systems*, vol. 31, no. 4, pp. 1499–1525, 2006.

[20] A. Dasgupta, J. Hopcroft, R. Kannan, and P. Mitra, "Spectral clustering with limited independence," in *Proceedings of SODA*, pp. 1036–1045, Philadelphia, PA, USA, Society for Industrial and Applied Mathematics, 2007.

[21] A. Dasgupta, J. Hopcroft, J. Kleinberg, and M. Sandler, "On learning mixtures of heavy-tailed distributions," in *Proceedings of FOCS*, 2005.

[22] S. DasGupta, "Learning mixtures of Gaussians," in *Proceedings of FOCS*, 1999.

[23] S. DasGupta and L. Schulman, "A two-round variant of EM for Gaussian mixtures," in *Proceedings of UAI*, 2000.

[24] W. F. de-la Vega, "MAX-CUT has a randomized approximation scheme in dense graphs," *Random Structures and Algorithms*, vol. 8, pp. 187–199, 1996.

[25] W. F. de la Vega, M. Karpinski, R. Kannan, and S. Vempala, "Tensor decomposition and approximation schemes for constraint satisfaction problems," in *STOC '05: Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pp. 747–754, 2005.

[26] W. F. de la Vega, M. Karpinski, and C. Kenyon, "Approximation schemes for metric bisection and partitioning," in *Proceedings of 15th ACM-SIAM SODA*, pp. 499–508, 2004.

[27] W. F. de la Vega, M. Karpinski, C. Kenyon, and Y. Rabani, "Approximation schemes for clustering problems," in *Proceedings of 35th ACM STOC*, pp. 50–58, 2003.

[28] W. F. de la Vega and C. Kenyon, "A randomized approximation scheme for metric MAX-CUT," *Journal of Computer and System Sciences*, vol. 63, pp. 531–541, 2001.

[29] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang, "Matrix approximation and projective clustering via volume sampling," *Theory of Computing*, vol. 2, no. 1, pp. 225–247, 2006.

[30] A. Deshpande and S. Vempala, "Adaptive sampling and fast low-rank matrix approximation," in *APPROX-RANDOM*, pp. 292–303, 2006.

[31] P. Drineas and R. Kannan, "Fast Monte-Carlo algorithms for approximate matrix multiplication," in *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science*, pp. 452–459, 2001.

[32] P. Drineas and R. Kannan, "Pass efficient algorithms for approximating large matrices," in *SODA '03: Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 223–232, 2003.

[33] P. Drineas, R. Kannan, A. Frieze, S. Vempala, and V. Vinay, "Clustering large graphs via the singular value decomposition," *Machine Learning*, vol. 56, pp. 9–33, 2004.

[34] P. Drineas, R. Kannan, and M. Mahoney, "Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix," *SIAM Journal on Computing*, vol. 36, pp. 132–157, 2006.

[35] P. Drineas, R. Kannan, and M. Mahoney, "Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix," *SIAM Journal on Computing*, vol. 36, pp. 158–183, 2006.

[36] P. Drineas, R. Kannan, and M. Mahoney, "Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix," *SIAM Journal on Computing*, vol. 36, pp. 184–206, 2006.

[37] P. Drineas, I. Kerenidis, and P. Raghavan, "Competitive recommendation systems," *Proceedings of the 34th Annual ACM Symposium on Theory of Computing*, pp. 82–90, 2002.

[38] R. O. Duda, P. Hart, and D. Stork, *Pattern Classification.* John Wiley & Sons, 2001.

[39] J. Feldman, R. A. Servedio, and R. O'Donnell, "PAC learning axis-aligned mixtures of Gaussians with no separation assumption," in *Proceedings of COLT*, pp. 20–34, 2006.

[40] A. Frieze and R. Kannan, "The regularity lemma and approximation schemes for dense problems," *Proceedings of the 37th Annual IEEE Symposium on Foundations of Computing*, pp. 12–20, 1996.

[41] A. Frieze and R. Kannan, "MAX-CUT has a randomized approximation scheme in dense graphs," *Quick Approximation to matrices and applications*, vol. 19, no. 2, pp. 175–200, 1999.

[42] A. Frieze, R. Kannan, and S. Vempala, "Fast Monte-Carlo algorithms for finding low-rank approximations," in *Proceedings of FOCS*, pp. 370–378, 1998.

[43] A. Frieze, R. Kannan, and S. Vempala, "Fast Monte-Carlo algorithms for finding low-rank approximations," *Journal of the ACM*, vol. 51, no. 6, pp. 1025–1041, 2004.

[44] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* Academic Press, 1990.

[45] Z. Füredi and J. Komlós, "The eigenvalues of random symmetric matrices," *Combinatorica*, vol. 1, no. 3, pp. 233–241, 1981.

[46] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman, 1979.

[47] O. Goldreich, S. Goldwasser, and D. Ron, “Property testing and its connection to learning and approximation,” *Journal of the ACM*, vol. 5, no. 4, pp. 653–750, 1998.

[48] G. H. Golub and C. F. van Loan, *Matrix Computations.* Johns Hopkins University Press, 3rd ed., 1996.

[49] S. Har-Peled and K. R. Varadarajan, “Projective clustering in high dimensions using core-sets,” in *Symposium on Computational Geometry*, pp. 312–318, 2002.

[50] P. Indyk, “A sublinear time approximation scheme for clustering in metric spaces,” in *Proceedings of 40th IEEE FOCS*, pp. 154–159, 1999.

[51] R. Kannan, H. Salmasian, and S. Vempala, “The spectral method for general mixture models,” *SIAM Journal on Computing*, vol. 38, no. 3, pp. 1141–1156, 2008.

[52] R. Kannan, S. Vempala, and A. Vetta, “On clusterings: Good, bad and spectral,” *Journal of ACM*, vol. 51, no. 3, pp. 497–515, 2004.

[53] J. A. Kelner, “Spectral partitioning, eigenvalue bounds, and circle packings for graphs of bounded genus,” *SIAM Journal on Computing*, vol. 35, no. 4, pp. 882–902, 2006.

[54] F. T. Leighton and S. Rao, “Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms,” *Journal of the ACM*, vol. 46, no. 6, pp. 787–832, 1999.

[55] L. Lovász and S. Vempala, “The geometry of logconcave functions and sampling algorithms,” *Random Structures and Algorithms*, vol. 30, no. 3, pp. 307–358, 2007.

[56] F. Lust-Piquard, “Inégalites de Khinchin dans $C_p(1 < p < \infty)$,” *Comptes Rendus de l’Académie des sciences, Paris*, vol. 303, pp. 289–292, 1986.

[57] F. McSherry, “Spectral partitioning of random graphs,” in *FOCS*, pp. 529–537, 2001.

[58] N. Megiddo and A. Tamir, “On the complexity of locating facilities in the plane,” *Operations Research Letters*, vol. I, pp. 194–197, 1982.

[59] C. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala, “Latent semantic indexing: A probabilistic analysis,” in *Proceedings of PODS*, 1998.

[60] M. Rudelson, “Random vectors in the isotropic position,” *Journal of Functional Analysis*, vol. 164, pp. 60–72, 1999.

[61] T. Sarlós, “Improved approximation algorithms for large matrices via random projections,” in *FOCS*, pp. 143–152, 2006.

[62] A. Sinclair and M. Jerrum, “Approximate counting, uniform generation and rapidly mixing Markov chains,” *Information and Computation*, vol. 82, pp. 93–133, 1989.

[63] D. A. Spielman and S.-H. Teng, “Spectral partitioning works: Planar graphs and finite element meshes,” *Linear Algebra and its Applications*, vol. 421, no. 2–3, pp. 284–305, 2007.

[64] G. Strang, *Linear Algebra and Its Applications.* Brooks Cole, 1988.

[65] S. Vempala and G. Wang, “A spectral algorithm for learning mixtures of distributions,” *Journal of Computer and System Sciences*, vol. 68, no. 4, pp. 841–860, 2004.

[66] V. H. Vu, “Spectral norm of random matrices,” in *Proceedings of STOC*, pp. 423–430, 2005.

[67] J. Wilkinson, *The algebraic eigenvalue problem (paperback ed.)*. Oxford: Clarendon Press, 1988.