TOPICS IN MATRIX APPROXIMATION

By

Srinivas Nambirajan

A Dissertation Submitted to the Graduate
Faculty of Rensselaer Polytechnic Institute
in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY
Major Subject: MATHEMATICS

Examining Committee:

_____
Peter R. Kramer, Dissertation Adviser

_____
Malik Magdon-Ismail, Member

_____
Joyce McLaughlin, Member

_____
John Mitchell, Member

Rensselaer Polytechnic Institute
Troy, New York

December 2015
(For Graduation December 2015)

CONTENTS

LIST OF TABLES v

LIST OF FIGURES vi

ACKNOWLEDGMENT vii

ABSTRACT ix

1. INTRODUCTION 1

1.1 Subspace Restricted Low Rank Approximation 2

1.1.1 Why Study This Problem? 2

1.1.2 Results 2

1.2 Existence of Exact Low Rank Approximations 3

1.2.1 Why Study This Problem? 3

1.2.2 Results 3

1.3 Element Sparsification in Finite Element Meshes 3

1.3.1 Why Study This Problem? 4

1.3.2 Results 4

2. PRELIMINARIES 5

2.1 Linear Algebra 5

2.1.1 Notation 5

2.1.2 Singular Value Decomposition 5

2.1.3 Norms 6

2.1.4 Rank 6

2.1.5 Low Rank Approximation 7

2.1.6 Inverse and The Moore-Penrose Pseudoinverse 7

2.1.7 Condition Number 8

2.1.8 Schur Complement 8

2.1.9 Symmetric Positive Semidefinite Matrices 9

2.1.10 Square Roots 9

2.1.11 Restrictions of Symmetric Positive Semidefinite Matrices . . . 10

2.1.12 Projections 10

2.2 Graph Theory 10

2.2.1 Matrices Associated with Graphs . . . 11
2.2.2 Effective Resistance . . . 12
2.3 Large Deviation and Concentration of Measure . . . 15
2.3.1 Johnson-Lindenstrauss Lemma . . . 15
2.3.2 Tail Bounds . . . 16
2.4 Sampling . . . 17
2.4.1 The ϵ-JLT . . . 17
2.4.2 The ϵ- FJLT . . . 17
2.5 Preconditioning Linear Systems . . . 18
2.6 Finite Element Matrices . . . 19
3. BACKGROUND AND RELATED WORK . . . 21
3.1 Low Rank Approximation . . . 21
3.2 Graph Sparsification . . . 22
3.3 Effective Resistance and Leverage Scores . . . 24
3.4 Sampling from Finite Element Matrices . . . 24
4. SUBSPACE RESTRICTED LOW RANK APPROXIMATION . . . 26
4.1 Introduction . . . 26
4.2 Main Result . . . 27
4.2.1 The Critical Rank . . . 28
4.3 Construction Of Subspace Restricted Low Rank Approximations . . . 31
4.3.1 Removing The Explicit Subspace Restriction . . . 31
4.3.2 Extracting The Subspace-Restricted Low Rank Approximation 33
4.3.3 The Algorithm . . . 34
4.4 Proofs . . . 37
4.4.1 Lemma 5 . . . 37
4.4.2 Proofs of Theorems 19, 20 . . . 38
4.4.2.1 Proof of Lemma 26 . . . 41
4.5 Conclusion . . . 44
4.6 Acknowledgement . . . 45

5. Fast Low Rank Approximations of Matrices . . . 46
5.1 Introduction . . . 46
5.2 Approximating The Range By A Single Action . . . 47
5.2.1 Understanding the Bounds for Expectation and Deviation . . . 48
5.3 Approximating The Range By Repeated Action . . . 50
5.4 The Existence of Exact Approximations . . . 52
5.5 Conclusion . . . 55
5.6 Acknowledgment . . . 56
6. ELEMENT SPARSIFICATION IN FINITE ELEMENT MESHES . . . 57
6.1 Introduction . . . 57
6.2 Mathematical Players . . . 57
6.3 The Sampling Approach . . . 58
6.3.1 Sampling Using Effective Stiffness . . . 60
6.3.1.1 Effective Stiffness . . . 60
6.3.1.2 Computation of Effective Stiffness . . . 61
6.4 Comparison of Sampling Techniques . . . 63
6.4.1 Frobenius Sampling . . . 64
6.4.2 Uniform Sampling . . . 66
6.4.3 Spectral Sampling . . . 67
6.5 Computing Effective Stiffness . . . 68
6.5.1 Randomized Pseudo-Inversion Using FJLT . . . 70
6.5.2 Estimating $\tilde{\lambda}_{\mathcal{I}}$ . . . 72
6.5.3 Sampling Using JLT Approximation . . . 73
6.5.4 Algorithm getEffectiveStiffness . . . 79
6.6 Conclusion . . . 80
6.7 Acknowledgments . . . 81
7. FUTURE DIRECTIONS . . . 82
7.1 Subspace Restricted Low Rank Approximation . . . 82
7.2 Exact Low Rank Approximations . . . 82
7.3 Sampling from Finite Element Matrices . . . 83
LITERATURE CITED . . . 84

LIST OF TABLES

5.1 Examples of the bounds stated in lemma 39. In all these examples, $q = 5$ and $p = k$, implying a probability of ≈ 1.0 that the bounds presented are true. 49

5.2 Examples of the bounds stated in lemmas 39 and 44. In all these examples, $q = 5$ and $p = k$, implying a probability of ≈ 1.0 that the bounds presented are true. 51

LIST OF FIGURES

5.1 The number of power iterations, q, of $\mathbf{CC}^\top$ required to guarantee exact rank k approximation, as a function of $\sigma_{k+1}/\sigma_{k+3}$. 54

6.1 Worst-case condition numbers of sampled preconditioners relative to the preconditioned matrix, $\mathbf{A}$ as a function of the number of elements sampled according to the three sampling probabilities considered: frobenius, spectral and uniform. 69

ACKNOWLEDGMENT

There are numerous people and factors that were critical to this dissertation. I would like to thank Malik Magdon-Ismail foremost, for his time and his persistent excitement through the varying degrees of my own motivation. I am greatly thankful to Peter Kramer for his patient counsel and for his time and diligence in tackling problems, both technical and otherwise, without the solution of which this dissertation would not have matured to a finish. I thank Petros Drineas, without the encouragement and support of whom this dissertation would not have begun at all. I would like to thank John Mitchell and Joyce McLaughlin for their feedback, and for the time and effort they invested in understanding and judging the work here.

I would like to thank my fellow 'problem-jammers': Dane Bush, Mithun Chakroborty, Jayanth Jagalur Mohan, John Postl, Maksim Tsikhanovich and other passing problem-solvers that filled the many whiteboards at Lally 03A with interesting scribbles and filled many evenings with intellectual event. I thank Bulent Yener for welcoming my ideas and consultation in a problem unrelated to any discussed in this dissertation. The amateur graduate student's idea of the doctorate life being one of general intellectual hunger and nourishment was kept alive into my veteran graduate student years due to these people. Thanks to Su Peng-Yu and Kenny Wu for being the best housemates a graduate student could ask for during the four years that we stayed together. Thanks to Ademola Akinlalu for making me a welcome guest at his basement for many days leading to my defense, long after I had moved out of my last rented apartment in Troy.

Troy has been home to many personal and academic experiences. I would like to thank Ricky and First Choice Carribbean American Cuisine for fueling much of the work here, to Spillin' The Beans for welcoming my caffeine habit with open arms and bottomless mugs and to Footsy Magoo for the lax last calls. Thanks to Troy Bike Rescue for allowing me to fix a ride for myself for the many months after internal combustion had failed me. Thanks to the folks at The Edge in Halfmoon for letting me spend many evening hours climbing on a membership that always

seemed to have another day before it expired.

Studying graduate mathematics is an intellectual luxury to most. I thank my parents for raising me with a healthy recklessness without which this intellectual luxury would have always seemed to be the impractical choice, and for always wishing for me the best of my own standards instead of imposing theirs. I especially thank my uncle, Srini Padmanabhan, for his friendship, counsel and support. Most of all, I thank my grandparents, Rukmani and Padmanabhan, for their quiet affection during the years when it mattered the most. I dedicate any idea that I am even remotely proud of in this dissertation to them, although they will likely not have understand a single word of it had they been around today.

ABSTRACT

A fundamental need in computational linear algebra is computing with matrices quickly but approximately. This is commonly achieved by approximating matrices, either deterministically or randomly such that the structure in these matrices essential to computation is preserved well. We study two useful and natural problems in this area, one involving deterministic, low-rank approximation of a matrix, and the other involving randomized approximation.

First, we study the low-rank approximation of a matrix, $\mathbf{C} \in \mathbb{R}^{m,n}$, using a matrix of rank at most $k < \min(m, n)$ under spectral (operator) norm with the additional constraint that the approximation contains columns belonging to a specified, r-dimensional subspace $\mathcal{B}$. We derive a closed form expression for the solution to this problem and present an algorithm to compute it. A similarly constrained approximation under the *Frobenius* norm allows a quick solution obtained in $O(T_{svd}(\mathbf{B}))$, where $T_{svd}(\mathbf{B})$ is the number of operations taken to compute the full singular value decomposition of a matrix $\mathbf{B} \in \mathbb{R}^{m,n}$ whose range is $\mathcal{B}$. However, there was no known algorithm for the problem in *spectral* norm. We provide the first closed form solution to the problem and an algorithm to compute it that runs in $O(T_{svd}(\mathbf{C}))$. We use this algorithm to then improve an existing result in low-rank approximation drastically: The best known result in computing a general low-rank approximation of a matrix guarantees only a *relative error* approximation; we guarantee the existence of *optimal* low-rank approximations.

Next, we study a randomized approximation of a matrix to obtain good preconditioners to it. A ubiquitous operation in computational linear algebra is the solution of a linear system $\mathbf{Ax} = \mathbf{b}$. The technique used to quickly obtain relative-error solutions to such systems with high probability is finding good randomized preconditioners to $\mathbf{A}$ for use in an appropriate iterative algorithm - Chebyshev or Conjugate Gradient, for instance. An established result for such preconditioning of symmetric, diagonally dominant (SDD) matrices has recently been extended to finite element matrices arising from finite element meshes for elliptic PDEs. The

computation of such preconditioners is expensive, requiring $O(rn^2 + n^3)$ operations for a matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ for an $r > n$, of the order of the number of elements in the finite element mesh. We provide a method that computes these preconditioners in $\tilde{O}(n^3 \log(rn))$ (where $\tilde{O}$ hides poly-logarithmic factors), which is a significant improvement for $r = \omega(n)$.

1. INTRODUCTION

This is a thesis, the bulk of which is devoted to the study of two important problems in the approximation of matrices. The first is the computation of a low rank approximation to a matrix and the second is the computation of a sparse finite element mesh that preserves the spectral characteristics of the original mesh. In both cases, the algorithms are randomized. Broadly, the research that this thesis presents makes three important contributions:

1. It guarantees the existence of *exact* low rank approximations to a matrix in computationally important subspaces.

2. In doing so, it solves a previously unsolved problem of obtaining the optimal low-rank approximation to a matrix, with respect to the spectral norm, with columns of the approximation constrained to belong to a specified subspace.

3. It studies and presents a quick algorithm to sample a small number of important elements from a finite element mesh while still preserving important structures in the mesh (such as the spectrum).

We introduce the document first, followed by the introduction of the problems themselves. This document is self-contained, for the most part. It is structured to offer the necessary preliminaries to the reader first, followed by a brief background to put the work presented here in context. We then present the problem of Subspace Restricted Low Rank Approximations, and solve a previously unsolved problem, the solution of which we use to improve the best existence results in computing low rank spectral approximations. The use of this solution toward this improvement features in the following chapter on Exact Low Rank Approximations of Matrices. Finally, we study the sparsification of elements from a finite element mesh, followed by a discussion of future directions that this work naturally guides us toward.

1.1 Subspace Restricted Low Rank Approximation

Here we are concerned with solving the following problem. Let $\mathbf{C} \in \mathbb{R}^{m,n}$ and let $\mathcal{B}$ be a subspace in $\mathbb{R}^m$ of dimension r.

$$\mathbf{C}_{\mathcal{B},k} = \arg\min_{\mathbf{Z}} \|\mathbf{C} - \mathbf{Z}\|_2 , \;\; \text{rank}(\mathbf{Z}) \leq k, \;\; \mathbf{Z} \text{ is made of columns from } \mathcal{B}.$$

1.1.1 Why Study This Problem?

We study this problem for two reasons. First, the solution to this problem implies the existence of exact low rank approximations in important subspaces, where only very crude approximations were guaranteed before. It opens up the possibility of obtaining the exact SVD of the matrix $\mathbf{C}$ in $O(mnk)$. Next, the problem is very natural to consider for those already familiar with the usual form of low-rank approximation, solved by Eckart and Young in [1]: an optimal rank k approximation to $\mathbf{C}$ in both the spectral and the frobenius norms, is given by its singular value decomposition, truncated at the first k singular values. Low-rank approximations to $\mathbf{C}$ with the added constraint of a subspace-membership on the columns is a subject of immediate curiosity.

1.1.2 Results

We obtain a closed form solution to $\mathbf{C}_{\mathcal{B},k}$, as

$$\mathbf{C}_{\mathcal{B},k} = \left(\mathbf{C}_{\mathcal{B}}\mathbf{T}\right)_k \mathbf{T}^{-1},$$

for a $\mathbf{T} \in \mathbb{R}^{n,n}, \mathbf{T} \succeq \mathbf{0}$ that is invertible almost always, and define the important notion of the *critical rank*, which is instrumental in improving the best known theoretical guarantees of low rank approximations available in certain subspaces. We provide an algorithm to compute $\mathbf{C}_{\mathcal{B},k}$ that runs in $O(T_{svd}(\mathbf{C}) + mnr)$ flops.
Our study is purely analytical, employing recent results in linear algebra to obtain the solution: we consider the orthogonal projection, $\mathbf{C}_{\mathcal{B}}$, of $\mathbf{C}$ onto $\mathcal{B}$ and make some elaborate arguments concerning the row-spaces of this projection and its residual, $\mathbf{C}_{\mathcal{N}} = \mathbf{C} - \mathbf{C}_{\mathcal{B}}$, to arrive at the solution.

1.2 Existence of Exact Low Rank Approximations

1.2.1 Why Study This Problem?

Low rank approximation of matrices is has important uses in PCA, image processing, compressed sensing, noise removal, Bayesian learning and many other fields [2]. The quick computation of a k rank approximation in spectral norm, in $O(mnk)$ flops, is a goal that remains elusive. Attempts to get nearer to this goal have been made by algorithms due to [3, 4, 5] that first obtain a good subspace and then obtain the best k rank approximation from within this subspace. While these attempts work well heuristically the theoretical guarantees are loose [3]. While our guarantee is probabilistic like the other guarantees in this field, it is considerably stronger. For further details, the reader is referred to 5.4.

1.2.2 Results

Our chief result is that, in the same computational time it takes to guarantee a subspace containing something 'roughly equivalent' to the optimal rank k approximation to a matrix, $\mathbf{C}$, we guarantee the existence of an *exact* rank k approximation. We use results in the study of Gaussian test matrices, made available for the purpose by [5, 3], and our results from the previous study, to make our high-probability guarantee.

1.3 Element Sparsification in Finite Element Meshes

Here, we study the sampling of elements from a finite element discretization of an elliptic partial differential equation. Specifically: given an elliptic PDE of the form $\mathbf{L}u = f$ over the domain Ω, let us denote the discretized problem as $\mathbf{Ax} = \mathbf{b}$, with $\mathbf{A} \in \mathbb{R}^{n,n}, \mathbf{x}, \mathbf{b} \in \mathbb{R}^n$ being the stiffness matrix, discretized solution and the discretized forcing function respectively (for further details, see 2.6). It is known that $\mathbf{A}$ can be written as the sum of *element stiffness matrices*, A_e, associated with every element in the discretization/partition of Ω. The matrix square-root, $\mathbf{F}_e$, of $\mathbf{A}_e$ can be assumed as being individually available [6] for computational purposes. Under this assumption, we develop and compute a probability distribution over the set of matrices $\{\mathbf{F}_e\}$, and therefore over the set of element-stiffness matrices, $\{\mathbf{A}_e\}$,

so that sampling and adding a small number of element-stiffness matrices according to this distribution provides a good preconditioner to $\mathbf{A}$ with high probability.

1.3.1 Why Study This Problem?

The quick solution of linear systems is a classic and ubiquitous problem in linear algebra, of which solving discretized PDEs quickly is of special importance. In solving these problems, speed is often obtained by *preconditioning* the original linear system (the reader is referred to 2.5 for further details), which are more easily obtained for matrices with a certain structure. The solution of a symmetric diagonally dominant (SDD) system of equations in nearly linear time [7, 8] and the work on sampling edges from graphs using effective resistance [9] has made it seem promising that similar methods can be employed to solve the large and commonly occurring linear systems in finite element analysis. In the approach above, the eventual quick solution of the system came after the initial investigation of graph-sparsification by Spielman et. al. Under the same belief, we use recent results on the generalization to finite element matrices [10, 6] of the method of sampling edges from graphs using effective resistances, to design a fast algorithm to 'sparsify' a finite element mesh i.e. we effectively provide a subset of the elements in a finite element mesh so that computationally relevant parameters such as element stiffness matrices may be computed only on these elements in order to obtain a good preconditioner to the original discretization, $\mathbf{Ax} = \mathbf{b}$, of the elliptic PDE $\mathbf{L}u = f$.

1.3.2 Results

Our chief result is that we may sample $O(n \log n)$ elements from a finite element mesh, where n is the number of nodes, in time $\tilde{O}(n^3 \log(rn))$ flops. This is considerably faster than naively sparsifying the mesh, which takes $O(rn^2)$ flops. We use a mix of linear algebraic analysis, special random projections and power iterations to achieve this run time with no asymptotic increase to the number of elements sampled, relative to the best theoretical guarantee by [6].
We proceed to the preliminaries.

2. PRELIMINARIES

2.1 Linear Algebra

2.1.1 Notation

Matrices will be denoted by bold, upper-case letters. We follow Parlett's convention [11] by denoting symmetric matrices by letters symmetric about their vertical axis, excepting two common practices. First, matrices with orthonormal rows/columns will be denoted by $\mathbf{U}, \mathbf{V}$ although these symbols are symmetric about their vertical axes. Next, $\mathbf{\Sigma_G}$, a symbol asymmetric about its vertical axis, will denote the diagonal matrix containing along its diagonal, the singular values of a matrix, $\mathbf{G}$, although this matrix is clearly symmetric. Scalars and vectors will be denoted in lower case with the latter being in bold font. Let $\mathbf{C} \in \mathbb{R}^{m,n}, m \geq n$ for the following exposition.

2.1.2 Singular Value Decomposition

The matrix, $\mathbf{C}$, has a description in which it can be thought of as 'essentially diagonal', up to rotations of its domain and range. This description, due to Beltrami and Jordan [12], is the singular value decomposition, widely used in numerical linear algebra since Gene Golub's presentation of a quick algorithm to compute it.
Specifically, the full Singular Value Decomposition (full SVD) of $\mathbf{C}$ is the factorization

$$\mathbf{C} = \mathbf{U_C}\mathbf{\Sigma_C}\mathbf{V_C}^\top, \quad \mathbf{U_C} \in \mathbb{R}^{m,n},\ \mathbf{\Sigma_C} \in \mathbb{R}^{n,n},\ \mathbf{V_C} \in \mathbb{R}^{n,n}$$

where $\mathbf{U_C}$ contains columns that form an orthonormal basis for $\mathcal{C}$ and $\mathbf{V_C}$ contains columns that form an orthonormal basis for $\mathcal{C}'$. The diagonal entries in $\mathbf{\Sigma_C}$ are non-negative and decreasingly ordered: $\mathbf{\Sigma_C}(i,i) \geq \mathbf{\Sigma_C}(j,j)$ for $i \leq j$; they are the singular values of $\mathbf{C}$, denoted, for concision, as $\sigma_i = \mathbf{\Sigma_C}(i,i)$. To resolve ambiguity where necessary, the matrix whose singular values are used in discussion will be made explicit as $\sigma_i(\mathbf{C})$.
The SVD is a particularly descriptive factorization of a matrix since many useful characteristics of a matrix may be defined in terms of the singular values and singular

vectors.

2.1.3 Norms

The matrix, $\mathbf{C}$, has two common families of norms: the induced vector norms and the Schatten norms, defined as follows.

Definition 1. *Given that a norm* $\|\mathbf{x}\|_\alpha$ *is defined on* $\mathbf{x} \in \mathbb{R}^n$*, the induced* α*-norm of* $\mathbf{C}$ *is defined as*

$$\|\mathbf{C}\|_\alpha = \max_{\mathbf{x}\in\mathbb{R}^n} \frac{\|\mathbf{C}\mathbf{x}\|_\alpha}{\|\mathbf{x}\|_\alpha}.$$

Specifically, for $\alpha = 2$, we get the induced 2-norm, also called as the operator/spectral norm of $\mathbf{C}$. It has the property that

$$\|\mathbf{C}\|_2 = \max_i \sigma_i.$$

As a convention, $\|\cdot\|$ will denote spectral norm henceforth unless otherwise specified.

Definition 2. *The p-Schatten norm of* $\mathbf{C}$*,* $\|\mathbf{C}\|_{(p)}$*, is the p-norm of the vector* $\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_n)$*:*

$$\|\mathbf{C}\|_{(p)} = \|\boldsymbol{\sigma}\|_p$$

Specifically, the Frobenius norm of $\mathbf{C}$, $\|\mathbf{C}\|_F$, is the 2-norm of $\boldsymbol{\sigma}$:

$$\|\mathbf{C}\|_F = \sqrt{\sum_i \sigma_i^2}.$$

2.1.4 Rank

The rank of $\mathbf{C}$, ρ, is as the number of singular values of $\mathbf{C}$ that are non-zero. If $\rho = \min(m, n)$, $\mathbf{C}$ is said to be *full rank*, and *rank deficient* otherwise. The *reduced SVD* of $\mathbf{C}$ is defined as

$$\mathbf{C} = \bar{\mathbf{U}}_\mathbf{C}\bar{\boldsymbol{\Sigma}}_\mathbf{C}\bar{\mathbf{V}}_\mathbf{C}^\top, \quad \bar{\mathbf{U}}_\mathbf{C} = \mathbf{U}_\rho \in \mathbb{R}^{m,\rho}, \quad \bar{\boldsymbol{\Sigma}}_\mathbf{C} = \boldsymbol{\Sigma}_\rho \in \mathbb{R}^{\rho,\rho}, \quad \bar{\mathbf{V}}_\mathbf{C} = \mathbf{V}_\rho \in \mathbb{R}^{n,\rho}.$$

This is a more compact representation of the SVD of $\mathbf{C}$ if $\rho < \min m, n$, where $\bar{\boldsymbol{\Sigma}}_\mathbf{C}$ is invertible as will be useful in upcoming discussions.

2.1.5 Low Rank Approximation

We may collect the first k columns of $\mathbf{U_C}$ to form $\mathbf{U}_k \in \mathbb{R}^{m,k}$, the first k diagonal entries of $\mathbf{\Sigma_C}$ to form $\mathbf{\Sigma}_k \in \mathbb{R}^{k,k}$, and the first k columns of $\mathbf{V_C}$ to form $\mathbf{V}_k \in \mathbb{R}^{n,k}$. The rank-$k$ truncation, $\mathbf{C}_k$, of $\mathbf{C}$ is then defined as

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

The best approximation to $\mathbf{C}$ of rank k is $\mathbf{C}_k$ under both the Frobenius and spectral norms:

$$\mathbf{C}_k = \arg \min_{\text{rank}(\mathbf{Z}) \leq k} \|\mathbf{C} - \mathbf{Z}\| = \arg \min_{\text{rank}(\mathbf{Z}) \leq k} \|\mathbf{C} - \mathbf{Z}\|_F .$$

The spans of columns of $\mathbf{U_C}, \mathbf{V_C}$ are called the *left* and *right singular subspaces* of $\mathbf{C}$. The ranges of $\mathbf{U}_k, \mathbf{V}_k$ are the *top k left and right singular subspaces*, notated as $\mathcal{L}_k, \mathcal{R}_k$ respectively

2.1.6 Inverse and The Moore-Penrose Pseudoinverse

Since a matrix can be viewed as a linear operator on $\mathbb{R}^n$, one can define the inverse of $\mathbf{C}$. If $\mathbf{C}$ is invertible, the inverse of $\mathbf{C}$, $\mathbf{C}^{-1}$ is simply defined as

$$\mathbf{C}^{-1} = \mathbf{V}_C \mathbf{\Sigma}_C^{-1} \mathbf{U}_C^\top,$$

where $\mathbf{\Sigma}^{-1}$ is the diagonal containing the inverse of the diagonal elements, σ_i, of $\mathbf{\Sigma}$. If $\mathbf{C}$ is not invertible, a *pseudoinverse* can be defined, which is analytically similar to the inverse. In fact, it is exactly the inverse of $\mathbf{C}$ in its range. Formally, a matrix, $\mathbf{C}^+$, is said to be the Moore-Penrose pseudoinverse of $\mathbf{C}$ if and only if it obeys the Penrose conditions:

1. $\mathbf{C}\mathbf{C}^+\mathbf{C} = \mathbf{C}$,
2. $\mathbf{C}^+\mathbf{C}\mathbf{C}^+ = \mathbf{C}^+$,
3. $(\mathbf{C}\mathbf{C}^+)^\top = \mathbf{C}\mathbf{C}^+$,
4. $(\mathbf{C}^+\mathbf{C})^\top = \mathbf{C}^+\mathbf{C}$.

The Moore-Penrose pseudoinverse is unique and the reduced SVD provides a very convenient way to construct it:

$$\mathbf{C}^{+} = \bar{\mathbf{V}}_{\mathbf{C}}\bar{\mathbf{\Sigma}}_{\mathbf{C}}^{-1}\bar{\mathbf{U}}_{\mathbf{C}}^{\top}.$$

The Moore-Penrose pseudoinverse has several useful properties, mainly arising from the fact that $\mathbf{C}\mathbf{C}^{+}$ is an orthogonal projector onto the range of $\mathbf{C}$. For instance, it is used to find the least squares solution, $\mathbf{x}^{*}$, to an overdetermined system of linear equations

$$\mathbf{C}\mathbf{x} = \mathbf{b},$$

which is simply

$$\mathbf{x}^{*} = \mathbf{C}\mathbf{C}^{+}\mathbf{b}.$$

2.1.7 Condition Number

In solving systems of linear equations, a useful characterization of the coefficient matrix, say $\mathbf{Q}$, for determining the computational stability of solutions is using its *condition number*, denoted as $\kappa(\mathbf{Q})$ and defined as

$$\kappa(\mathbf{Q}) = \frac{\sigma_{\max}(\mathbf{Q}}{\sigma_{\min}(\mathbf{Q})},$$

where $\sigma_{\max}(\cdot), \sigma_{\min}(\cdot)$ denote the largest and smallest non-zero singular values of their argument.

2.1.8 Schur Complement

An important analytical tool for handling matrices, due to Issai Schur, results from the block-wise LU decomposition [13] of a matrix. This tool, called the *schur complement*, is defined as follows.

Definition 3. *For* $\mathbf{C} \in \mathbb{R}^{m,m}$ *let*

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix}, \mathbf{C}_{11} \in \mathbb{R}^{m_1,m_1}, \mathbf{C}_{22} \in \mathbb{R}^{m_2,m_2}, \mathbf{C}_{12} \in \mathbb{R}^{m_1,m_2}, \mathbf{C}_{21} \in \mathbb{R}^{m_2,m_1}.$$

Provided $\mathbf{C}_{11}$ *is invertible, the schur complement of* $\mathbf{C}_{11}$ *is defined as*

$$\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}.$$

The schur complement has many applications arising mainly from its power in describing certain properties of a matrix, $\mathbf{C}$, in terms of properties of smaller matrices,namely a diagonal block of $\mathbf{C}$ and its schur complement. For instance, one may check to see if $\mathbf{C}$ is positive definite by merely checking if $\mathbf{C}_{11}$ and its schur complement are positive definite owing to an equivalence between the two conditions. In this work, it will be used for a purely analytical reason presented in more detail in situ.

2.1.9 Symmetric Positive Semidefinite Matrices

A class of matrices are particularly useful due to the reason that quadratic forms using them result in seminorms in the appropriate spaces, and that they are the matrix-equivalent of the non-negative reals in that they have a square root. Matrices in this class are said to be *symmetric positive semidefinite* (SPSD). Formally,

Definition 4. *A matrix,* $\mathbf{A} \in \mathbb{R}^{m,m}$*, is SPSD if and only if*

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0 \; \forall \mathbf{x} \in \mathbb{R}^m.$$

In addition, if $\mathbf{x}^\top \mathbf{A}\mathbf{x} = 0$ only if $\mathbf{x} = \mathbf{0}$, then $\mathbf{A}$ is said to be *positive definite*. In terms of its singular values, $\mathbf{A}$ is SPSD if and only if $\sigma_i(\mathbf{A}) \geq 0 \; \forall i$, and $\mathbf{A}$ is positive definite if and only if $\sigma_i(\mathbf{A}) > 0 \; \forall i$.

There is a natural partial ordering on the class of SPSD matrices. The semidefinite partial ordering, $\preceq$, is defined by $\mathbf{X} \preceq \mathbf{Y}$ ($\mathbf{Y}$ dominates $\mathbf{X}$), if and only if $\mathbf{Y} - \mathbf{X}$ is positive semidefinite.

2.1.10 Square Roots

For a symmetric matrix, $\mathbf{A} \succeq \mathbf{0}$, we mean, by $\sqrt{\mathbf{A}}$, the unique symmetric square root of $\mathbf{A}$ given by $\mathbf{U}_\mathbf{A}\sqrt{\mathbf{\Sigma}}_\mathbf{A}\mathbf{U}_\mathbf{A}^\top$, where the square root of a diagonal matrix

is defined as the diagonal matrix containing its entry-wise square roots. Any real symmetric matrix that has a real square root is SPSD.

2.1.11 Restrictions of Symmetric Positive Semidefinite Matrices

That every restriction of $\mathbf{Y}$ must dominate the corresponding restriction of $\mathbf{X}$ is a useful observation, formally contained in the following lemma.

Lemma 5. *Let* $\mathbf{X} = \mathbf{F}^\top\mathbf{F}$ *and* $\mathbf{Y}$ *be symmetric positive semidefinite (SPSD) matrices. Let SVD of* $\mathbf{Y} = \mathbf{V}_\mathbf{Y}\mathbf{\Sigma}_\mathbf{Y}\mathbf{V}_\mathbf{Y}^\top$, $\mathbf{V}_\mathbf{Y} = (\mathbf{V}_+, \mathbf{V}_0)$ *where* $\mathbf{V}_0$ *is a basis for the null space of* $\mathbf{Y}$ *- so* $\mathbf{Y}\mathbf{V}_0 = \mathbf{0}$*. Then,*

$$\mathbf{X} \preceq \mathbf{Y} \quad \textit{if and only if} \quad \mathbf{X}\mathbf{V}_0 = \mathbf{0} \quad \textit{and} \quad \left\|\mathbf{F}\sqrt{\mathbf{Y}^\dagger}\right\| \leq 1.$$

This lemma is a generalization of a transformation shown in [14], and is proved in section 3.

2.1.12 Projections

A certain class of matrices generalize the notion of 'projection' onto a plane preserving the intuition that projecting an already projected object onto the same plane is does not make a difference. Formally,

Definition 6. *A matrix,* $\mathbf{\Pi} \in \mathbb{R}^{m,m}$*, is a projection if and only if it is idempotent:* $\mathbf{\Pi}^2 = \mathbf{\Pi}\mathbf{\Pi} = \mathbf{\Pi}$.

Furthermore, a projection, $\mathbf{\Pi}$, is an orthogonal projection if $\mathbf{\Pi} = \mathbf{\Pi}^\top$. The range of a projection is the space it projects on to, and is made explicit in the following manner. Let $\mathcal{B}$ be a subspace of $\mathbb{R}^m$. Then $\mathbf{\Pi}_\mathcal{B}$ denotes the orthogonal projector onto $\mathcal{B}$. An important and useful property of an orthogonal projector, $\mathbf{\Pi}$, is that $\|\mathbf{\Pi}\| = 1$.

2.2 Graph Theory

Some tools in graph theory are essential in understanding the intuition behind sampling algorithms we study in chapter 6 of this work. A graph, say G, is the

triplet $(V, E, w : E \to \mathbb{R})$, where V is a set of elements called *vertices*, $E \subset V \times V$, is a set of ordered pairs called *edges*, and w is a *weight* function from E to the reals that can be thought of as assigning weights to edges.

2.2.1 Matrices Associated with Graphs

We introduce some fundamental matrices associated with graphs here. The adjacency matrix, $\mathbf{A} \in \mathbb{R}^{V,V}$, of the graph G is the matrix such that

$$\mathbf{A}(i,j) = \begin{cases} w((i,j)) & (i,j) \in E \\ 0 & (i,j) \notin E \end{cases}$$

The degree, $d(v)$, of a vertex, $v \in V$, is simply the sum of the weights of all the edges 'incident' on v:

$$d(v) = \sum_i w((i,v)),$$

using which we define the degree matrix, $\mathbf{D} \in \mathbb{R}^{V,V}$, as a diagonal matrix with

$$\mathbf{D}(i,j) = \delta_{ij} d(i).$$

We now define a matrix that is critical to our work here and, in fact, to the vast field of *spectral graph theory* [15], the *Laplacian* matrix of $\mathbf{G}$.

Definition 7. *The Laplacian, $\mathbf{\Lambda} \in \mathbb{R}^{V,V}$, is defined as*

$$\mathbf{\Lambda} = \mathbf{D} - \mathbf{A}.$$

This simply defined object has many useful properties:

Lemma 8. *The Laplacian, $\mathbf{\Lambda}$, of the graph $\mathbf{G}$ is a matrix such that*

1. *Every graph has a unique Laplacian and every Laplacian has a unique graph associated with it,*
2. $\mathbf{\Lambda} \succeq \mathbf{0}$*,*
3. $\mathbf{1} \in \mathbb{R}^V$ *is in the nullspace of* $\mathbf{\Lambda}$*,*

4. $\mathbf{\Lambda}$ *is of rank* $|V|-1$ *if and only if* G *is connected.*

Property 1 states an isomorphism between graphs and Laplacians, which implies that linear algebraic tools may be employed to describe graphs and that graph theoretic ideas may be used to inform linear algebraic intuition. Properties 3, 4 are of special consequence to this work. To see why property 3 is true, we define yet another matrix associated with a graph, the *edge incidence* matrix.

Definition 9. $\mathbf{E} \in \mathbb{R}^{E\times V}$*, the edge incidence matrix of the graph,* $\mathbf{G}$*, is a matrix such that for all* $e=(u,v)\in E$*,*

$$|\mathbf{E}(e,v)| = \sqrt{w(e)}, \quad \mathbf{E}(e,u) = -\mathbf{E}(e,v), \quad \mathbf{E}(e,j) = 0 \ \forall j \neq u,v.$$

It follows immediately from the above definition that

$$\mathbf{\Lambda} = \mathbf{E}^\top \mathbf{E},$$

and hence, that $\mathbf{\Lambda} \geq \mathbf{0}$. The existence of an intuitive square root of $\mathbf{\Lambda}$ is crucial to the generalization to the context of finite element analysis, as is the mathematical player to be introduced next.

2.2.2 Effective Resistance

The concept of effective resistance of an edge in a graph, and its relation to the Laplacian of the graph, is both simple and beautiful. First, we view the graph, $G=(V,E,w)$ as a network of resistances where $w(e)$ represents the *conductance* of an edge $e\in E$. Clearly, the resistance, $\rho(e)$, of the edge is simply $1/w(e)$. One recalls Ohm's law from high school physics, which states that

$$v = i\rho, \tag{2.1}$$

where v, i, ρ are the potential at one end of a resistor assuming the other end is grounded, the current through the resistor and the resistance of the resistor respectively.

The *effective resistance* across a resistor with resistance ρ, in a network of resistances is simply defined as the net current flowing through the resistor having unit potential difference across its terminals. Alternatively, it is the net potential difference across a resistor when a unit current is injected into one of its terminals and extracted from the other.

We may 'batch-state' Ohm's law for an entire network simultaneously using matrix notation. In this linear algebraic description, the effective resistance is measured as a simple seminorm of an edge indicator vector, as is shown below. First, let us restate 2.1 using conductance, $w = 1/\rho$, as

$$wv = i.$$

More conveniently, let $e = (r, s)$ be an edge with conductance $w(e)$, and with current i_e flowing across it with potentials v_r, v_s at vertices r, s are respectively. Let the *edge incidence matrix*, $\mathbf{B}(e, r) \in \mathbb{R}^{|E|,|V|}$, be defined such that for $e = (i, j) \in E, l \in V$,

$$\mathbf{B}(e, l) = \begin{cases} 1, & l = i \\ -1, & l = j \\ 0, & l \notin \{i, j\} \end{cases}.$$

Then

$$w(e)(v_r - v_s) = i_e \quad \Rightarrow \quad \sqrt{w(e)}\,(\mathbf{B}(e, r)v_r + \mathbf{B}(e, s)v_s) = i_e.$$

For a graph, the above statement may be simultaneously expressed for all the edges in the graph as

$$\mathbf{W}^{1/2}\mathbf{B}\mathbf{v} = \mathbf{i}_e, \tag{2.2}$$

where $\mathbf{W} \in \mathbb{R}^{E,E}$ is the diagonal matrix containing $w(e)$ in $\mathbf{W}(e, e)$, and $\mathbf{v} \in \mathbb{R}^V$ and $\mathbf{i}_e \in \mathbb{R}^E$ are the potentials at the vertices in the graphs and currents on edges in the graph. Now we 'batch-state' Kirchoff's law, which is the physical law stating that the sum of the currents entering a vertex is the sum of the currents leaving the vertex. If $\mathbf{i}_s \in \mathbb{R}^V$ denotes the current leaving the vertex, s, Kirchoff's law can be

stated as

$$\mathbf{W}^{-1/2}\mathbf{B}^\top \mathbf{i}_e = \mathbf{i}_s, \tag{2.3}$$

Putting (2.2), (2.3) together,

$$\mathbf{\Lambda v} = \mathbf{i}$$

where $\mathbf{i} \in \mathbb{R}^V$ is the vector of current leaving the appropriate vertices and $\mathbf{v} \in \mathbb{R}^V$ is the vector of potentials on these vertices. Since ohm's law only depends upon potential differences and not actual potentials, replacing $\mathbf{v}$ with $\mathbf{v} + \delta\mathbf{1}$ should not change the result, and indeed it doesn't since $\delta\mathbf{1}$ is in the nullspace of $\mathbf{\Lambda}$ by lemma 8. If we are only concerned with the potential differences, we may limit ourselves to $\mathbf{v}$ such that the mean of entries is 0, whence we get

$$\mathbf{v} = \mathbf{\Lambda}^+\mathbf{i}, , \tag{2.4}$$

the collective statement of Ohm's law for the entire network of resistances, G. We now simply inject and extract a unit of current from vertices, $r, s \in V$, respectively and obtain the potential difference across $e = (r, s) \in E$ as

$$r_{eff}(e) = (\mathbf{e}_r - \mathbf{e}_s)\, \mathbf{\Lambda}^+ (\mathbf{e}_r - \mathbf{e}_s)\,.$$

This simple result has yet another underlying structure, namely that we may construct an orthogonal projector, $\mathbf{\Pi}$, onto the range of $\mathbf{B}$, such that its diagonals contain the effective resistances.

Lemma 10. *Let $\mathbf{\Pi} \in \mathbb{R}^{E,E}$ be defined as*

$$\mathbf{\Pi} = \mathbf{B}\mathbf{\Lambda}^+\mathbf{B}^\top.$$

Then $\mathbf{\Pi}$ is an orthogonal projector and $r_{eff}(e) = \mathbf{\Pi}(e, e)$.

A generalization of this fact will be used extensively in this work.

2.3 Large Deviation and Concentration of Measure

Randomized algorithms play a central role in the work here, and their existence and analysis is justified by the important idea from probability theory that a random variable that is 'nicely' dependent on numerous other random variables (Lipschitz-dependent, in particular) varies less with increase in the number of variables it depends upon [16]. The intuition is extended from that of the popular Central Limit Theorem, and concentration results for numerous other well-behaved functions of random variables exist, of the form

$$\mathsf{Pr}\left(\|f(X) - \mathsf{E}\left[X\right]\| \geq \epsilon\right) \leq \delta.$$

The theory of large deviations has a nearly identical motivation, and although concentration of measure and the theory of large deviations are considered to be different, they are presented as being effectively identical here.

2.3.1 Johnson-Lindenstrauss Lemma

This powerful lemma [17] states that m points in n dimensions may be embedded into a $O(\log(m)/\epsilon^2)$-dimensional space with minimal distortion. Specifically,

Lemma 11. *let $\mathbf{x}_i \in \mathbb{R}^n$ for $1 \leq i \leq m$. Then there exists an embedding function, $g : \mathbb{R}^n \to \mathbb{R}^k$, $k = O(\log(m)/\epsilon^2)$, such that*

$$(1-\epsilon)\left\|\mathbf{x}_i - \mathbf{x}_j\right\| \leq \left\|g(\mathbf{x}_i) - g(\mathbf{x}_j)\right\| \leq (1+\epsilon)\left\|\mathbf{x}_i - \mathbf{x}_j\right\|.$$

Owing to the phenomenon of measure concentration one may construct such a relative isometry, g, with high probability. Specifically,

Lemma 12. *Let $\mathcal{S}$ be a random, k-dimensional linear vector space, and let $\mathbf{\Pi}_{\mathcal{S}}$ be an orthogonal projector on to $\mathcal{S}$. Let $\mathbf{x}_i \in \mathbb{R}^n$ for $1 \leq i \leq m$ be m points. Define*

$$\tilde{g} = \sqrt{\frac{n}{k}}\mathbf{\Pi}_{\mathcal{S}}$$

Then, for m points $\mathbf{x}_i, 1 \leq i \leq m$ and $k = c\log(m)/\epsilon^2$, with probability at least

$1 - m^{2-c}$,

$$(1-\epsilon)\left\|\mathbf{x}_i - \mathbf{x}_j\right\| \leq \left\|\tilde{g}(\mathbf{x}_i) - \tilde{g}(\mathbf{x}_j)\right\| \leq (1+\epsilon)\left\|\mathbf{x}_i - \mathbf{x}_j\right\|.$$

2.3.2 Tail Bounds

Given some information about a random variable, such as its expectation, variance, one may bound the probability of deviation of a function of multiple, independent copies of this random variable from its mean. These bounds are appropriately named tail bounds, of which the most relevant to this work is the Bernstein bound for matrix random variables [18]:

Lemma 13. *Let $\mathbf{Z} \in \mathbb{R}^{n,n}$ be a matrix random variable defined such that*

$$\left\|\mathsf{E}\left[\mathbf{Z}\right]\right\|_2 = 0; \quad \left\|\mathbf{Z}\right\| \leq \gamma; \quad \left\|\mathsf{E}\left[\mathbf{Z}^T\mathbf{Z}\right]\right\| \leq s^2.$$

Further, let $\{\mathbf{Z}_n\}_1^m$ be m independent copies of $\mathbf{Z}$, with $\bar{\mathbf{Z}}_m = \sum_{n=1}^m \mathbf{Z}_n$. Then,

$$\mathsf{Pr}\left(\left\|\bar{\mathbf{Z}}_m\right\| \geq \epsilon\right) \leq 2n \exp\left(\frac{-m\epsilon^2}{2s^2 + 2\gamma\epsilon/3}\right).$$

Specifically, for $\gamma \leq 3s^2/\epsilon$,

$$\mathsf{Pr}\left(\left\|\bar{\mathbf{Z}}_m\right\| \geq \epsilon\right) \leq 2n \exp\left(\frac{-m\epsilon^2}{4s^2}\right).$$

In particular these tail bounds facilitate the randomized approximation of a finite sum by a smaller sum. For instance, suppose that

$$\mathbf{X} = \sum_{i=1}^{n} \mathbf{X}_i,$$

with all the quantities being bounded. We may design a random variable $\mathbf{Y}_j$ that takes values $\mathbf{X}_i/p_i$ with probability p_i and take k draws of this variable to form the sum

$$\mathbf{Y} = \sum_{j=1}^{k} \mathbf{Y}_j.$$

Clearly, $\mathbf{Y}$ is a random variable. To see how well $\mathbf{Y}$ approximates $\mathbf{X}$, we only need to compute bounding parameters for $\mathbf{Z} = \mathbf{X} - \mathbf{Y}$ that are required by lemma 13.

2.4 Sampling

Sampling may be done, in general, in a variety of contexts. In this work we are eventually concerned with sampling rows or columns from a matrix. In particular, if $\mathbf{B} \in \mathbb{R}^{m,n}$, we sample r rows from $\mathbf{B}$ as $\mathbf{SB}$, where $\mathbf{S} \in \mathbb{R}^{r,m}$ [19, 20, 21]. In analyzing and understanding algorithms employing sampling of this kind in this work, the following two matrices play a crucial role.

2.4.1 The ϵ-JLT

We define a matrix, $\tilde{\mathbf{\Pi}}_2 \in \mathbb{R}^{n,r_2}$, which is preserves the norms of the differences in $\{\mathbf{x}_i\}_1^{r_2}$ up to ϵ (as in 11) with a high probability [22, 23]

Lemma 14. *Let $\{\mathbf{x}_i \in \mathbb{R}^n\}_1^r$ be a set of r points in $\mathbb{R}^n$. Construct $\tilde{\mathbf{\Pi}}_2 \in \mathbb{R}^{n,r_2}$ such that*

$$\tilde{\mathbf{\Pi}}_2(i,j) = \begin{cases} \pm\sqrt{3/r_2} & \textit{with probability 1/6 each} \\ 0 & \textit{with probability 2/3} \end{cases}$$

Finally, let $0 < \epsilon \leq 1/2$. If

$$r_2 \geq \frac{1}{\epsilon^2}\left(12\ln n + 6\ln\frac{1}{\delta}\right),$$

then $\tilde{\mathbf{\Pi}}_2$ is an ϵ-JLT with probability at least $1-\delta$.

2.4.2 The ϵ- FJLT

We now define a random projection which preserves both the norms of differences and the inner products of r_1 vectors in $\{\mathbf{x}_i\}_1^{r_1}$ up to ϵ (we refer the reader to [22] for further details). We begin by introducing the Subsampled Randomized Hadamard Transform (SRHT): Suppose $\mathbf{H}_n \in \mathbb{R}^{n,n}$ is the hadamard matrix, defined as

$$\mathbf{H}_{2n} = \begin{pmatrix} \mathbf{H}_n & \mathbf{H}_n \\ \mathbf{H}_n & -\mathbf{H}_n \end{pmatrix}, \quad \mathbf{H}_1 = 1.$$

Let $\hat{\mathbf{H}}_n = \mathbf{H}/\sqrt{n}$ be the normalized hadamard matrix. The SRHT is the matrix formed by sampling rows from randomly signed columns of $\hat{H}_n$. Formally, if $\mathbf{D}_n \in \mathbb{R}^{n,n}$ contains $+1, -1$ on its diagonal with equal probabilities, and $\mathbf{S} \in \mathbb{R}^{r,n}$ is a row sampling matrix that samples r_1 rows from $\hat{\mathbf{H}}_n$, then $\mathbf{S}\hat{\mathbf{H}}_n\mathbf{D}_n$ is the SRHT. SRHTs have the useful property that they are, with high probability, Fast Johnson Lindenstrauss Transforms, as captured in the following lemma which appears as Lemma 3 in [22].

Lemma 15. *If* $\mathbf{U} \in \mathbb{R}^{r,n}, r >> n$ *contains orthonormal columns and if*

$$r_1 \geq \frac{14^2 n \ln(40rn)}{\epsilon^2} \ln\left(\frac{30^2 \ln(40rn)}{\epsilon^2}\right),$$

then $\mathbf{S}\hat{\mathbf{H}}_r\mathbf{D}_r \in \mathbb{R}^{r_1,r}$ *is an* ϵ*-FJLT with probability at least 0.9.*

2.5 Preconditioning Linear Systems

In obtaining a solution to the system $\mathbf{Ax} = \mathbf{b}$ quickly, a useful progression is obtaining a good *preconditioner* to the matrix $\mathbf{A}$. This objective of computing a good preconditioner to the matrix $\mathbf{A}$ may be thought of as a relaxation of the problem of exact inversion of $\mathbf{A}$: $\mathbf{X}$ is considered a good preconditioner to $\mathbf{A}$ if the condition number, $\kappa(\mathbf{A}, \mathbf{X}) = \kappa(\mathbf{X}^{-1}\mathbf{A})$, is a constant, C, close to unity. We note that this can be accomplished by finding an $\mathbf{X}$ such that

$$\left\|\mathbf{X}^{-1}\mathbf{A} - \mathbf{I}\right\|_2 \leq \epsilon$$

for an $\epsilon < 1$, since this would imply that

$$\kappa(\mathbf{A}, \mathbf{X}) \leq \frac{1+\epsilon}{1-\epsilon},$$

hence bounding C as an increasing function of ϵ. This 'approximate inverse' of $\mathbf{A}$ may be obtained by inverting a sub-sampled $\mathbf{A}$ that preserves enough of the structure of $\mathbf{A}$. This is the primary method of preconditioning a matrix involved in the study here, and its performance may be analyzed using the results in sampling presented above.

2.6 Finite Element Matrices

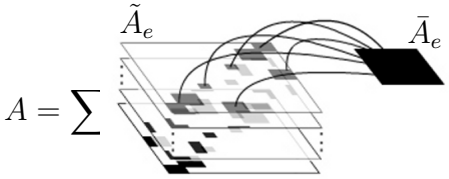
The matrix, $\mathbf{A}$, the preconditioning of which we study in this thesis, arises specifically from finite element analysis. Specifically, the *finite element matrix* (or stiffness matrix) is an SPSD matrix that is a matrix-representation of a discretized PDE. In purely linear algebraic terms, $\mathbf{A}$ is a finite element matrix if: $\mathbf{A}$ is SPSD; There is a natural covering of $\mathcal{A}$, the indices of $\mathbf{A}$, by subsets $\mathcal{A}_e$ of $\mathcal{A}$ - we call e an *element*; $\mathbf{A}$ has a natural decomposition $\mathbf{F}^\top\mathbf{F}$, where $\mathbf{F} \in \mathbb{R}^{m,n}, m \geq n$, and $\mathbf{F}$ has the following row-block structure, with the blocks indexed by elements:

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_1 \\ \vdots \\ \mathbf{F}_s \end{bmatrix}; \quad \mathbf{F}_e = \begin{bmatrix} \blacksquare & \square & \blacksquare & \blacksquare \end{bmatrix}$$

In such a structure, each $\mathbf{F}_e$ has non-zero columns only in column indices $\mathcal{A}_e$, corresponding to the element e, resulting in a vertically banded sparsity/density. As a result, $\mathbf{A}$ may be written as a sum of *element matrices*, $\tilde{A}_e = {\mathbf{F}_e}^T\mathbf{F}_e$:

$$\mathbf{A} = \sum_e \tilde{\mathbf{A}}_e.$$

The non-zero submatrix of $\tilde{\mathbf{A}}_e$ is denoted as $\bar{\mathbf{A}}_e$, as illustrated.



Finally, $\bar{\mathbf{A}}_e$ is SPSD and has a nullspace $\mathcal{N}_e$ that is compatible with the nullspace, $\mathcal{N}$, of $\mathbf{A}$ (see [6] for more details). In essence, a finite element matrix, $\mathbf{A}$, may be thought of as a generalization of the Laplacian matrix. The Laplacian matrix is the sum of edgewise-Laplacian matrices that are 2×2 matrices embedded in a larger matrix, and the finite element matrix, $\mathbf{A}$, is the sum of elementwise-stiffness matrices that are $|\mathcal{A}_e| \times |\mathcal{A}_e|$ embedded in a larger matrix as depicted above. Conversely, the Laplacian matrix is a special case of a finite element matrix with elements

representing edges in a graph.

This closes the presentation of the mathematical and conceptual preliminaries required in conducting the study presented in this thesis.

3. BACKGROUND AND RELATED WORK

The purpose of this chapter is to present a necessarily broad but brief overview of studies and results that have enabled the work presented in this thesis. Specifically, the work we overview here falls into three broad categories:

1. Results in low rank approximation of matrices
2. Results in sampling from graphs
3. Results in sampling from finite element matrices

3.1 Low Rank Approximation

Since the work of Eckart and Young [1] in showing that the k rank truncation, $\mathbf{C}_k$, of the singular value decomposition of a matrix, $\mathbf{C} \in \mathbb{R}^{m,n}, m < n$, was the best rank k approximation in both the spectral and Frobenius norms:

$$\mathbf{C}_k = \arg \min_{\text{rank}(\mathbf{Z}) \leq k} \|\mathbf{C} - \mathbf{Z}\|_\alpha , \quad \alpha \in \{2, F\}, \tag{3.1}$$

much of the field of low rank approximations has centered around the computation of some singular value decomposition. The full SVD of $\mathbf{C}$ takes $O(nm^2)$ flops, and is an infeasible computational method for obtaining low rank approximations for large matrices. It is usually sufficient to low-rank approximate $\mathbf{C}$ up to some relative or multiplicative error in exchange for the increase in the speed of computation of such an approximation. Gu et. at. [24], Woolfe et. al. [25] and Sarlos [26] produce algorithms running in $O(nmk)$ that produce low rank approximations that produce approximation errors that are a factor of $\sqrt{m}$ from the optimal, σ_{k+1}.
Tygert et. al. [3] produce a randomized algorithm that runs in time $O(mnki)$ that produces the approximation

$$\|\mathbf{C} - \mathbf{Z}\| \leq C m^{1/4i+2} \sigma_{k+1}$$

by employing a power iterative scheme (outlined in [5]) to obtain a subspace very close to $\mathcal{L}_k$, the top k left singular subspace of $\mathbf{C}$, and approximating $\mathbf{C}$ with a rank k projection onto this subspace.

This is a specific implementation of a more general randomized algorithm used in obtaining low rank approximations. In particular, the common proto-algorithm used to quickly find a low-rank approximation of $\mathbf{C}$ is

Algorithm 1. spectralLowRank($\mathbf{C}$, k)

1. Obtain a low-dimensional approximation, $\mathcal{B}$, to the range of $\mathbf{C}$,
2. Return $\mathbf{C}_{\mathcal{B},k}$ as the desired approximation.

For such an algorithm, Martinsson et. al. guarantee that the produced subspace contains a 'good enough' approximation [5]. Boutsidis et. al. guarantee even better [27]. Both these results are presented and used in Chapter 5.

A closely related problem to (3.1) is studied by Sou et. al. [14], who solve

$$\min \operatorname{rank}(\mathbf{X}) \;\; s.t. \;\; \|\mathbf{C} + \mathbf{BXR}\| < 1.$$

Their method uses basic linear algebraic arguments to produce what is in essence, the result that k is the minimum achievable rank to achieve an error, σ_{k+1}, if this error can, in fact be achieved. We recognize that this problem is, in spirit, the'dual' to the problem we wish to solve. Chapter 4 results from this recognition.

3.2 Graph Sparsification

Much of the results we use to sample elements from finite element matrices, due to Avron et. al. [6], are direct generalizations of those used in sampling edges from a graph. Specifically, Spielman, et. al. [9], show that a graph, $G = (V_G, E_G, w_G)$, may be *sparsified* by sampling only a subset of its edges, to obtain a graph $H = (V_H, E_H, w_H)$ with the property that the quadratic forms using their respective

Laplacian matrices is relatively intact i.e. preserved up to relative error ϵ. An edge, $e \in E_G$ is sampled with $r_{eff}(e)$, its effective resistance (see chapter 2.2.2) and added to H a certain number of times, as in the algorithm below.

Algorithm 2. sparsify(G, q)

1. Perform q times:
   (a) Choose edge, $e \in E_G$, with probability $p_e = r_{eff}/\left(|V_G| - 1\right)$
   (b) Add e to H with weight $(p_e q)^{-1}$.
2. Return H

They show in their work that

Lemma 16. *If $\mathbf{\Lambda}_G, \mathbf{\Lambda}_H$ are laplacians of G, H respectively, and H is obtained as $sparsify(G, q)$ with $q = O(n \log n/\epsilon^2)$, then for $\epsilon \in (1/\sqrt{n}, 1)$*

$$(1-\epsilon)\mathbf{x}^\top \mathbf{\Lambda}_H \mathbf{x} \leq \mathbf{x}^\top \mathbf{\Lambda}_G \mathbf{x} \leq (1+\epsilon)\mathbf{x}^\top \mathbf{\Lambda}_H \mathbf{x}$$

with probability at least $1/2$ for all $\mathbf{x} \in \mathbb{R}^n$.

The construction of such sparsifiers resulted in a swarm of results in solving systems of linear equations in Symmetric Diagonally Dominant (SDD) matrices since solving any such system was shown to be reducible [28] to approximating a Laplacian as in lemma 16. Such an approximation has the unique characteristic (by design) that $\mathbf{\Lambda}_H$ is a very good preconditioner to $\mathbf{\Lambda}_G$. Koutis et. al. [7, 8] developed a particularly fast way of approximating $p_e = r_{eff}/(|V_G| - 1)$ by using the notion of *stretch* of a edge in a tree to approximate r_{eff}, resulting in the quick computation of such preconditioners. Specifically, Koutis et. al. show that, if T is a spanning tree of G, and the stretch, $stretch(e)$ of an edge, $e = (u, v) \in E_G$ is defined as

$$stretch(e) = w_e l_T(u, v),$$

where $l_T(u, v)$ is the length of the unique path from u to v in the tree T, then

$$r_{eff}(e) \leq stretch(e),$$

and compute low stretch spanning trees to get better approximations to r_{eff}. Based on this and the construction of *incremental sparsifiers* with only a small fraction of $|E_G|$ more edges than a spanning tree, they produce a randomized preconditioner using which they obtain an approximate solution $\tilde{\mathbf{x}}$ to a SDD system, $\mathbf{Ax} = \mathbf{b}$ in expected time $\tilde{O}(m \log^n \log(1/\epsilon))$ such that

$$\|\mathbf{x} - \tilde{\mathbf{x}}\|_{\mathbf{A}} \leq \|\mathbf{x}\|_{\mathbf{A}} .$$

3.3 Effective Resistance and Leverage Scores

The important observation [29] is that the effective resistance, $r_{eff}(e)$, of an edge is identical to the statistical leverage score [30] of a row in an orthonormal matrix, $\mathbf{U} \in \mathbb{R}^{m,n}, m > n$, namely $l_i = \mathbf{U}\mathbf{U}^\top(i, i) = \|\mathbf{U}(i, :)\|^2$. We refer the reader to lemma 10 for a detailed statement. Existing linear algebraic techniques to approximate such leverage scores [22] depend on the dimension, m, which in the case of the edge incidence, $\mathbf{B}$, matrix of a Laplacian (see 2.2.2), is $|E|$, the number of edges in the Laplacian-generating graph, G, which is $O(n^2)$ in general.

3.4 Sampling from Finite Element Matrices

Avron, Toledo [6] suggest a generalization of effective resistance to finite element matrices. Effectively, they observe that the granularity of describing a graph does not have to be at the edge level. It could be at the subgraph level. For example, if $\mathbf{\Lambda}$, is the Laplacian of the graph G, we saw that $\mathbf{\Lambda} \succeq \mathbf{0}$ and had a square root. One such square root is $\mathbf{B}$ the edge incidence matrix introduced in the previous chapter. There are numerous other square roots of $\mathbf{\Lambda}$, however, including a symmetric square root. If the smallest unit we are willing to describe is a triangle, instead of an edge, then we get a square root, $\mathbf{B}_3 \in \mathbb{R}^{3m_3,n}$, where m_3 is the number of triangles in a triangle-decomposition of G, each such triangle containing a block of three rows in

$\mathbf{B}_3$. In general, an element is viewed in the work of Avron and Toledo as a subgraph of lowest resolution, as is the triangle in this example. Specifically, they show that

Lemma 17. *Let $\bar{p}_e = \Lambda(\bar{\mathbf{A}}_e, \mathbf{K}_e)$, and let $\tilde{p}_e$ be a relative error approximation to $\bar{p}_e$,*

$$|\bar{p}_e - \tilde{p}_e| \leq \delta \bar{p}_e.$$

Let probabilities p_e be defined relative to the set $\{\tilde{p}_e\}$:

$$p_e = \frac{\tilde{p_e}}{\sum_e \tilde{p}_e}.$$

If

$$\mathbf{Y} = \frac{\sum_1^m \mathbf{X}_i}{m},$$

with $\mathbf{X}_i = \mathbf{A}_e/p_e$ with probability p_e, and

$$m = m(\delta) = \Omega(\tilde{n} \log \tilde{n}); \quad \tilde{n} = n\left(\frac{1+\delta}{1-\delta}\right),$$

then

$$\mathsf{Pr}\left(\kappa(\mathbf{A}, \mathbf{Y}) > 2\right) \leq \frac{1}{poly(m)},$$

where $\mathsf{Pr}\left(\mathcal{E}\right)$ denotes the probability of the event $\mathcal{E}$, $\kappa(\mathbf{A}, \mathbf{Y})$ is the relative condition number of $\mathbf{A}$ with respect to $\mathbf{Y}$, defined as

$$\kappa(\mathbf{A}, \mathbf{Y}) = \kappa(\mathbf{Y}^{-1}\mathbf{A}),$$

where $\kappa(\cdot)$ is as defined in 2.1.7, and $poly(m)$ stands for a polynomial function of m.

This lemma is used as a fundamental analytical tool in our work. This closes the brief presentation of the background for the work discussed in this thesis.

4. SUBSPACE RESTRICTED LOW RANK APPROXIMATION

4.1 Introduction

Approximating a matrix using a low-rank matrix is a problem that has been widely studied and has a variety of applications in compression, signal processing and statistics, among other fields [2]. The two common forms of such a problem vary only in the norm used in the approximation, with the common goal being the construction of a $\tilde{\mathbf{Z}} \in \mathbb{R}^{m,n}$ of rank at most k, such that

$$\tilde{\mathbf{Z}}_{F,2} \in \arg \min_{\text{rank}(\mathbf{Z}) \leq k} \|\mathbf{C} - \mathbf{Z}\|_{F,2} \,,$$

where we note that the minimizing argument of $\|\mathbf{C} - \mathbf{Z}\|_{F,2}$ does not have to be unique in general, and arg returns a set of minimizers. Here, the subscripts $F, 2$ denote the frobenius and spectral norms respectively. In both cases, it is a well-known result, due to Eckart and Young [1], that $\tilde{\mathbf{Z}}_F = \tilde{\mathbf{Z}}_2 = \mathbf{C}_k$, the rank k truncation of $\mathbf{C}$ obtained from the singular value decomposition of $\mathbf{C}$. In the case where an additional, mathematically natural property of the desired approximation is known, namely that its range is contained in a subspace $\mathcal{B}$, we arrive at the following two variants: suppose $\mathbf{C} \in \mathbb{R}^{m,n}$ and let $\mathbf{B} \in \mathbb{R}^{m,r}$ be a basis for an r dimensional subspace $\mathcal{B}$ of $\mathbb{R}^m$; we seek to find $\tilde{\mathbf{Z}}^{(k)}_{F,2}$ such that

$$\tilde{\mathbf{Z}}^{(k)} \in \arg \min_{\text{rank}(\mathbf{Z}) \leq k} \|\mathbf{C} - \mathbf{Z}\|_{F,2} \,; \quad \mathbf{Z} = \mathbf{B}\boldsymbol{\Gamma},\ \boldsymbol{\Gamma} \in \mathbb{R}^{r,n},$$

if indeed such a $\tilde{\mathbf{Z}}^{(k)}$ exists. A natural approach to solve these variants is to first project $\mathbf{C}$ onto $\mathcal{B}$, and then take the best rank-k approximation of this projection. This simple approach gives the best rank-k approximation to $\mathbf{C}$ with columns in $\mathcal{B}$ under the Frobenius norm (lemma 7 in [27]), in running time $O(T_{svd(\mathbf{B})} + mnr)$; but this is only known to give a $\sqrt{2}$-approximation to the best subspace constrained rank-k approximation in spectral norm (lemma 7 in [27]).

Here, we show how to compute this optimal spectral approximation. Specifically, we provide a deterministic algorithm that constructs $\mathbf{C}_{\mathcal{B},k} \in \mathbb{R}^{m,n}$ with the property that

$$\mathbf{C}_{\mathcal{B},k} \in \arg\min_{\mathrm{rank}(\mathbf{Z})\leq k} \|\mathbf{C}-\mathbf{Z}\|_2, \quad \mathbf{Z}=\mathbf{B}\mathbf{\Gamma},\ \mathbf{\Gamma}\in\mathbb{R}^{r,n}, \tag{4.1}$$

when the set defined by the arg min is non-empty i.e. when $\mathbf{C}_{\mathcal{B},k}$ exists. Our main result is the following theorem.

Theorem 18. *There is a deterministic algorithm (Algorithm 3) running in time $O(T_{svd}(\mathbf{C})+mnr)$ to construct $\mathbf{C}_{\mathcal{B},k}$ which solves the optimization problem in (4.1).*

Henceforth, we will denote the spectral norm simply by $\|\cdot\|$. Other norms will be explicitly stated.

4.2 Main Result

Let $\mathbf{C}\in\mathbb{R}^{m\times n}$, $m\geq n$ and let $\mathbf{B}\in\mathbb{R}^{m\times r}$ be a basis for $\mathcal{B}$, an r-dimensional subspace of $\mathbb{R}^m$. The main observation, perhaps counter-intuitive, is that the best rank-k approximation to $\mathbf{C}$ with columns in $\mathcal{B}$, $\mathbf{C}_{\mathcal{B},k}$, is either

- as good as the best rank-k approximation, $\mathbf{C}_k$, without any subspace-restriction, or
- as good as the best subspace restricted approximation, $\mathbf{C}_{\mathcal{B}}$, without any rank restrictions.

A single parameter, the *critical rank* of $\mathbf{C}$ with respect to $\mathbf{B}$, controls which of the above two cases takes effect. Furthermore, $\mathbf{C}_{\mathcal{B},k}$ can be constructed for the most part, and a parameter called *rank slack* specifies what we mean by 'for the most part'. Formally: let us define the *critical rank*, k^*, as the number of singular values of $\mathbf{C}$ strictly larger than the orthogonal complement of $\mathbf{C}$ with respect to $\mathcal{B}$ in spectral norm:

$$\sigma_{k^*}(\mathbf{C}) > \|\mathbf{C}_{\mathcal{N}}\| \geq \sigma_{k^*+1}(\mathbf{C}). \tag{4.2}$$

Then

Theorem 19. *Given* $1 \leq k < r$,

$$\|\mathbf{C} - \mathbf{C}_{\mathcal{B},k}\| = \begin{cases} \|\mathbf{C} - \mathbf{C}_k\| = \sigma_{k+1}(\mathbf{C}), & k < k^*, \\ \|\mathbf{C}_{\mathcal{N}}\|, & k \geq k^*, \end{cases}$$

Furthermore,

Theorem 20. *In addition to the objects defined in Theorem 19, let* $h = \text{rank}(\mathbf{C}_{\mathcal{B}}\mathbf{V}_*\mathbf{V}_*{}^\top)$, *where* $\mathbf{V}_*$ *is an orthonormal basis for the rowspace of* $\mathbf{C}_{\mathcal{N}} = \mathbf{C} - \mathbf{C}_{\mathcal{B}}$ *corresponding to its top singular value,* $\|\mathbf{C}_{\mathcal{N}}\|$. *Then for* $1 \leq k < r$, $\mathbf{C}_{\mathcal{B},k}$ *can be constructed for* $k < k^*$ *and* $k \geq k^* + h$, *When* $k^* \leq k < k^* + h$, *one can construct* $\mathbf{Z}(\epsilon) = \mathbf{B}\mathbf{\Gamma}(\epsilon)$ *such that* $\text{rank}(\mathbf{Z}(\epsilon)) = \text{rank}(\mathbf{\Gamma}(\epsilon)) = k$ *and*

$$\|\mathbf{C} - \mathbf{Z}(\epsilon)\| \leq (1 + \epsilon) \|\mathbf{C}_{\mathcal{N}}\|.$$

The main results above state two main ideas: 1. That the effect of a subspace-restriction is in presenting a critical rank, k^* 2. That the errors one can achieve with low rank approximations in the presence of a subspace-restriction are either those that can be achieved using a rank-truncation of $\mathbf{C}$ or the projection of $\mathbf{C}$ onto $\mathcal{B}$, $\mathbf{C}_{\mathcal{B}}$. It is also evident from the result above that there are values of k, specifically for $k^* \leq k < k^* + h$, where we do not construct an approximation with a fixed approximation guarantee. We observe in our work that this region is a region of measure 0 under nice measures, and that a case where $h = 1$ can be obtained by perturbing the matrix $\mathbf{C}$ by a non-singular matrix that is arbitrarily small. We discuss the two approximation regimes presented above by illustrating the concept of critical rank below.

4.2.1 The Critical Rank

We provide a comparison between the usual low-rank approximation without subspace restrictions and the subspace-restricted low-rank approximation to provide the intuition for the concept of critical rank. Let $\mathbf{\Sigma}_{\mathbf{C}}$ contain the singular values of $\mathbf{C}$ in its diagonal in the usual, non-increasing order. By the Eckart-Young theorem,

all the optimal errors from low-rank approximations to $\mathbf{C}$ are "contained" in $\mathbf{\Sigma_C}$ as

$$\mathbf{\Sigma_C}(i,i) = \sigma_i = \min_{\operatorname{rank}\mathbf{Z}\leq(i-1)} \|\mathbf{C}-\mathbf{Z}\|\,.$$

This is illustrated schematically below where the optimal errors that can be achieved using low-rank approximations are denoted by $\bullet$.

$$\begin{pmatrix} \bullet\ \sigma_1 & & & \\ & \bullet\ \sigma_2 & & \\ & & \ddots & \\ & & & \bullet\ \sigma_n \end{pmatrix}$$

Now, let us restrict the columns of the low-rank approximation to $\mathbf{C}$ to come from a subspace $\mathcal{B}$. Let us first consider the case where $\mathcal{B}$ is k dimensional, and spanned by the first k left-singular vectors of $\mathbf{C}$. Note that obtaining an approximation to $\mathbf{C}$, $\operatorname{rank}\mathbf{C} \leq k$, with columns restricted to be in $\mathcal{B}$, is as good as not restricting the columns to be in $\mathcal{B}$ at all; the columns of low-rank approximations to $\mathbf{C}$ up to rank k are contained in $\mathcal{B}$ by design. Finally, since $\dim\mathcal{B} = k$, the best approximation to $\mathbf{C}$ we can obtain with columns in $\mathcal{B}$ is $\mathbf{C}_k$, irrespective of the rank we allow the approximation to have. The least error we can achieve using such an approximation, therefore, is σ_{k+1}. We illustrate this below, denoting achievable errors by $\bullet$ and non-achievable errors by $\circ$.

$$\begin{pmatrix} \bullet\ \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \bullet\ \sigma_{k+1} & & & \\ & & & \circ\ \sigma_{k+2} & & \\ & & & & \ddots & \\ & & & & & \circ\ \sigma_n \end{pmatrix}$$

We observe that σ_{k+1} is the smallest singular value of $\mathbf{C}$ that is no smaller than the error from the best possible approximation to $\mathbf{C}$ with columns in $\mathcal{B}$, and that $k+1$, the index of this singular value, plays an important role: all singular values at least as

large as σ_{k+1} are achievable errors by low-rank approximations to $\mathbf{C}$ with columns in $\mathcal{B}$, and no errors smaller than σ_{k+1} are achievable using any approximation to $\mathbf{C}$ with columns in $\mathcal{B}$. We may therefore think of $k+1$ as the *critical rank*. A fundamental contribution of this work is a surprising generalization of this observation to the case where $\mathcal{B}$ does not necessarily have to contain any singular vectors of $\mathbf{C}$: restricting the columns of approximations to $\mathbf{C}$ to belong to $\mathcal{B}$ is *almost always* equivalent to defining a *critical rank* on the spectrum of $\mathbf{C}$! To illustrate the general case, suppose that $\mathcal{B}$ is an arbitrary subspace with $\mathbf{C}_{\mathcal{B}} = \Pi_{\mathcal{B}}(\mathbf{C})$ and $\mathbf{C}_{\mathcal{N}} = \mathbf{C} - \mathbf{C}_{\mathcal{B}}$. Clearly, the least error that can possibly be achieved by an approximation to $\mathbf{C}$ with columns in $\mathcal{B}$ is $\|\mathbf{C}_{\mathcal{N}}\|$. Similar to the simpler case outlined previously, we define k^* such that σ_{k^*} is the smallest singular value that is no smaller than $\|\mathbf{C}_{\mathcal{N}}\|$. The result presented above is the statement that all singular values of $\mathbf{C}$ at least as large as σ_{k^*} are achievable optimal errors using low-rank approximations to $\mathbf{C}$ with columns in $\mathcal{B}$.

$$\begin{pmatrix} \bullet\ \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \bullet\ \sigma_{k^*} & & & \\ & & & \circ\ \sigma_{k^*+1} & & \\ & & & & \ddots & \\ & & & & & \circ\ \sigma_n \end{pmatrix}$$

The reader may note that the effect of the subspace restriction, $\mathcal{B}$, on the spectral approximation error is to place an upper bound, k^*, on the rank of the approximation up to which we may expect to do as well as a subspace-unrestricted rank-k approximation to $\mathbf{C}$. When the rank of the approximation meets or exceeds this critical rank, k^*, we construct an approximation that provides the same, or almost the same, error as $\mathbf{C}_{\mathcal{B}}$, the best approximation to $\mathbf{C}$ possible with columns in $\mathcal{B}$. Whether or not we construct an approximation to obtain exactly $\|\mathbf{C}_{\mathcal{N}}\|$ (the least error one can possibly achieve with an approximation having columns in $\mathcal{B}$) or almost the same, $(1+\epsilon)\|\mathbf{C}_{\mathcal{N}}\|$, is governed by the rank slack, h.

We close this discussion with a typical case: consider a full rank $\mathbf{C}$. If the critical rank, k^*, provided by $\mathcal{B}$ is very small ($k^* \ll r$) then $\mathbf{C}_{\mathcal{B},k^*}$ is as good an approxi-

mation to $\mathbf{C}$ as $\mathbf{C}_{\mathcal{B}}$ (which is of rank r), but with far lesser rank. If, on the other hand, $k^* = r$, we see that restricting the columns to belong to $\mathcal{B}$ doesn't affect the approximation at all, and $\mathbf{C}_{\mathcal{B},k\leq k^*}$ is as good as $\mathbf{C}_k$!

4.3 Construction Of Subspace Restricted Low Rank Approximations

We now turn to the construction of approximations having the properties promised in Theorem 20. The construction-strategy, following that of Sou and Rantzer in [14], will be to remove the explicit subspace-restriction, but retain the rank-restriction, by effectively constructing a series of transformations of problem (4.1) to a problem of approximating a modified matrix appropriately. We then extract a solution to (4.1) by inverse-transforming the solution to this equivalent problem.

4.3.1 Removing The Explicit Subspace Restriction

Suppose spectral error $\beta \geq \|\mathbf{C}_{\mathcal{N}}\|$ is achievable. So $\exists \mathbf{Z} = \mathbf{B}\mathbf{\Gamma}$, such that $\|\mathbf{C} - \mathbf{Z}\| \leq \beta$. Then,

$$\begin{aligned}\|\mathbf{C} - \mathbf{Z}\| \leq \beta &\Leftrightarrow (\mathbf{C} - \mathbf{Z})^\top (\mathbf{C} - \mathbf{Z}) \preceq \beta^2 \mathbf{I} \\ &\Leftrightarrow (\mathbf{C}_{\mathcal{N}} + \mathbf{C}_{\mathcal{B}} - \mathbf{Z})^\top (\mathbf{C}_{\mathcal{N}} + \mathbf{C}_{\mathcal{B}} - \mathbf{Z}) \preceq \beta^2 \mathbf{I} \\ &\Leftrightarrow (\mathbf{C}_{\mathcal{B}} - \mathbf{Z})^\top (\mathbf{C}_{\mathcal{B}} - \mathbf{Z}) \preceq \beta^2 \mathbf{I} - \mathbf{C}_{\mathcal{N}}^\top \mathbf{C}_{\mathcal{N}} \quad (\text{since } \mathbf{C}_{\mathcal{N}}^\top (\mathbf{C}_{\mathcal{B}} - \mathbf{Z}) = \mathbf{0}).\end{aligned}$$

Define $\mathbf{\Delta} = \beta^2 \mathbf{I} - \mathbf{C}_{\mathcal{N}}^\top \mathbf{C}_{\mathcal{N}}$, and let the full SVD of $\mathbf{C}_{\mathcal{N}}$ be $\mathbf{C}_{\mathcal{N}} = \mathbf{U}_N \mathbf{\Sigma}_N \mathbf{V}_N^\top$. Then

$$\begin{aligned}\mathbf{\Delta} = \beta^2 \mathbf{I} - \mathbf{C}_{\mathcal{N}}^\top \mathbf{C}_{\mathcal{N}} &= \mathbf{V}_N (\beta^2 \mathbf{I} - \Sigma_N^2) \mathbf{V}_N^\top \\ &= (\mathbf{V}_+ \quad \mathbf{V}_0) \begin{pmatrix} \Sigma_+ & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{V}_+^{\mathrm{T}} \\ \mathbf{V}_0^{\mathrm{T}} \end{pmatrix} \\ &= \mathbf{V}_+ \Sigma_+ \mathbf{V}_+^\top\end{aligned} \tag{4.3}$$

where the orthonormal bases $\mathbf{V}_+$ and $\mathbf{V}_0$ correspond to the positive and zero eigenvalues of $\mathbf{\Delta}$ respectively. We observe that $\mathbf{V}_0$ spans the subspace corresponding to

the top singular value of $\mathbf{C}_{\mathcal{N}}$ when $\beta = \|\mathbf{C}_{\mathcal{N}}\|$. We can now apply Lemma 5 with the null space of $\mathbf{\Delta}$ spanned by $\mathbf{V}_0$, the result of which we capture in the following lemma.

Lemma 21.

$$(\mathbf{C}_{\mathcal{B}} - \mathbf{Z})^\top (\mathbf{C}_{\mathcal{B}} - \mathbf{Z}) \preceq \mathbf{\Delta} \;\; \textit{if and only if} \;\; (\mathbf{C}_{\mathcal{B}} - \mathbf{Z})\mathbf{V}_0 = \mathbf{0} \;\; \textit{and} \;\; \left\| (\mathbf{C}_{\mathcal{B}} - \mathbf{Z}) \sqrt{\mathbf{\Delta}^\dagger} \right\| \leq 1.$$

This is a generalization of the case when $\mathbf{\Delta}$ is invertible (in which case $\mathbf{V}_0 = \mathbf{0}$ and $\mathbf{\Delta}^\dagger = \mathbf{\Delta}^{-1}$). We begin the construction of $\mathbf{Z}$ by first orthogonally resolving the rows of $\mathbf{Z}$ as

$$\mathbf{Z} = \mathbf{Z}_+ + \mathbf{Z}_0, \quad \mathbf{Z}_+ = \mathbf{Z}\mathbf{V}_+\mathbf{V}_+^\top, \quad \mathbf{Z}_0 = \mathbf{Z}\mathbf{V}_0\mathbf{V}_0^\top.$$

From the first condition in lemma 21,

$$\begin{aligned}
(\mathbf{C}_{\mathcal{B}} - \mathbf{Z})\mathbf{V}_0 = \mathbf{0} &\Leftrightarrow (\mathbf{C}_{\mathcal{B}} - \mathbf{Z}_+ - \mathbf{Z}_0)\mathbf{V}_0 = \mathbf{0} \\
&\Leftrightarrow \mathbf{C}_{\mathcal{B}}\mathbf{V}_0 - \mathbf{Z}_0\mathbf{V}_0 = \mathbf{0}, \quad \text{since } \mathbf{V}_+ \perp \mathbf{V}_0 \\
&\Leftrightarrow \mathbf{Z}_0 = \mathbf{C}_{\mathcal{B}}\mathbf{V}_0\mathbf{V}_0^\top \\
&\Leftrightarrow \mathbf{Z} = \mathbf{Z}_+ + \mathbf{C}_{\mathcal{B}}\mathbf{V}_0\mathbf{V}_0^\top
\end{aligned} \tag{4.4}$$

It now follows from the second condition in lemma 21 that

$$\begin{aligned}
\left\| (\mathbf{C}_{\mathcal{B}} - \mathbf{Z})\, \mathbf{V}_+\Sigma_+^{-1/2}\mathbf{V}_+^\top \right\| \leq 1 \quad &\Leftrightarrow \quad \left\| \left(\mathbf{C}_{\mathcal{B}} - \mathbf{C}_{\mathcal{B}}\mathbf{V}_0\mathbf{V}_0^\top - \mathbf{Z}_+\right) \mathbf{V}_+\Sigma_+^{-1/2}\mathbf{V}_+^\top \right\| \leq 1, \\
&\Leftrightarrow \quad \left\| \left(\mathbf{C}_{\mathcal{B}}\mathbf{V}_+\mathbf{V}_+^\top - \mathbf{Z}_+\right) \mathbf{V}_+\Sigma_+^{-1/2}\mathbf{V}_+^\top \right\| \leq 1, \\
&\Leftrightarrow \quad \left\| \mathbf{C}_{\mathcal{B}}\mathbf{V}_+\Sigma_+^{-1/2}\mathbf{V}_+^\top - \mathbf{Z}_+\mathbf{V}_+^\top\Sigma_+^{-1/2}\mathbf{V}_+^\top \right\| \leq 1 \\
&\Leftrightarrow \quad \left\| \tilde{\mathbf{C}} - \tilde{\mathbf{Z}} \right\| \leq 1,
\end{aligned}$$

where

$$\tilde{\mathbf{C}} = \mathbf{C}_{\mathcal{B}}\mathbf{V}_+\Sigma_+^{-1/2}\mathbf{V}_+^\top = \mathbf{C}_{\mathcal{B}}\sqrt{\mathbf{\Delta}^\dagger}, \quad \tilde{\mathbf{Z}} = \mathbf{Z}_+\mathbf{V}_+\Sigma_+^{-1/2}\mathbf{V}_+^\top = \mathbf{Z}_+\sqrt{\mathbf{\Delta}^\dagger}. \tag{4.5}$$

It can be shown that the row, column spaces of $\tilde{\mathbf{Z}}$ are contained in the row, column spaces of $\tilde{\mathbf{C}}$ respectively. So we may obtain $\tilde{\mathbf{Z}}$ that satisfies the above condition by

gathering all the singular components of $\tilde{\mathbf{C}}$ with singular values strictly larger than 1.

4.3.2 Extracting The Subspace-Restricted Low Rank Approximation

Since $\mathbf{V}_+, \mathbf{V}_0$ are mutually orthogonal,

$$\operatorname{rank}(\mathbf{Z}) = \operatorname{rank}(\mathbf{Z}_+) + \operatorname{rank}(\mathbf{C}_{\mathcal{B}}\mathbf{V}_0\mathbf{V}_0^\top) = \operatorname{rank}(\mathbf{Z}_+) + \bar{h}.$$

So, if error β can be achieved with rank $k+\bar{h}$, we construct $\tilde{\mathbf{Z}}$ as $(\tilde{\mathbf{C}})_k = (\mathbf{C}_{\mathcal{B}}\sqrt{\boldsymbol{\Delta}^\dagger})_k$. Extracting $\mathbf{Z}$ from $\tilde{\mathbf{Z}}$ is simply done by noting that

$$\mathbf{Z}_+\sqrt{\boldsymbol{\Delta}^\dagger} = \tilde{\mathbf{Z}} \quad \Rightarrow \quad \mathbf{Z}_+\sqrt{\boldsymbol{\Delta}^\dagger}\sqrt{\boldsymbol{\Delta}} = \mathbf{Z}_+\mathbf{V}_+\mathbf{V}_+^\top = \mathbf{Z}_+ = \tilde{\mathbf{Z}}\sqrt{\boldsymbol{\Delta}},$$

where we have noted that $\sqrt{\boldsymbol{\Delta}^\dagger}\sqrt{\boldsymbol{\Delta}} = \mathbf{V}_+\mathbf{V}_+^\top$, the orthogonal projection onto the row-space of $\sqrt{\boldsymbol{\Delta}}$ (see 2.1.6). We now simply recall (4.4), and construct $\mathbf{Z}$ as

$$\mathbf{Z} = \tilde{\mathbf{Z}}\sqrt{\boldsymbol{\Delta}} + \mathbf{C}_{\mathcal{B}}\mathbf{V}_0\mathbf{V}_0^\top.$$

We summarize our discussion in the following lemma.

Lemma 22. *Let $\beta \geq \|\mathbf{C}_{\mathcal{N}}\|$ and let $\boldsymbol{\Delta} = \beta^2 I - \mathbf{C}_{\mathcal{N}}^\top\mathbf{C}_{\mathcal{N}}$ with $\mathbf{V}_0$ as the basis for its nullspace. Let $\bar{h} = \operatorname{rank}(\mathbf{C}_{\mathcal{B}}\mathbf{V}_0\mathbf{V}_0^\top)$ and suppose error β can be achieved with rank $k + \bar{h}$. Then*

$$\mathbf{Z} = (\mathbf{C}_{\mathcal{B}}\sqrt{\boldsymbol{\Delta}^\dagger})_k\sqrt{\boldsymbol{\Delta}} + \mathbf{C}_{\mathcal{B}}\mathbf{V}_0\mathbf{V}_0^\top$$

is a solution satisfying $\|\mathbf{C} - \mathbf{Z}\| \leq \beta$, $\operatorname{rank}(\mathbf{Z}) \leq k + h$ and $\mathbf{Z} = \mathbf{B}\boldsymbol{\Gamma}$ for some $\boldsymbol{\Gamma} \in \mathbb{R}^{r,n}$.

Corollary 23. *If $\beta > \|C_{\mathcal{N}}\|$, then $\boldsymbol{\Delta}$ is invertible, $\mathbf{V}_0 = \mathbf{0}, \bar{h} = 0$ and there exists a $k \leq k^*$ such that*

$$\mathbf{Z} = (\mathbf{C}_{\mathcal{B}}\sqrt{\boldsymbol{\Delta}^{-1}})_k\sqrt{\boldsymbol{\Delta}}$$

satisfies $\|\mathbf{C} - \mathbf{Z}\| \leq \beta$ with $\operatorname{rank}(\mathbf{Z}) \leq k$ and $\mathbf{Z} = \mathbf{B}\boldsymbol{\Gamma}$ for some $\boldsymbol{\Gamma} \in \mathbb{R}^{r,n}$ of rank k.

In particular, if $\beta = \beta(\epsilon) = (1+\epsilon)\|\mathbf{C}_{\mathcal{N}}\|$ for a variable $\epsilon > 0$, then

$$\mathbf{\Delta}(\epsilon) = (1+\epsilon)^2 \|\mathbf{C}_{\mathcal{N}}\|^2 \mathbf{I} - \mathbf{C}_{\mathcal{N}}^{\top}\mathbf{C}_{\mathcal{N}},$$

corollary 23 constructs $\mathbf{Z}$ as a function of ϵ in the following manner

$$\mathbf{Z}(\epsilon) = (\mathbf{C}_{\mathcal{B}}\sqrt{\mathbf{\Delta}(\epsilon)^{-1}})_k \sqrt{\mathbf{\Delta}(\epsilon)}.$$

Corollary 24. *Let $\mathbf{V}_*, h$ be as defined in theorem 20. Then for $k \geq k^*$,*

$$\mathbf{Z} = (\mathbf{C}_{\mathcal{B}}\sqrt{\mathbf{\Delta}^{\dagger}})_k \sqrt{\mathbf{\Delta}} + \mathbf{C}_{\mathcal{B}}\mathbf{V}_*{\mathbf{V}_*}^{\top}$$

satisfies $\|\mathbf{C} - \mathbf{Z}\| = \|\mathbf{C}_{\mathcal{N}}\|$ with $\mathrm{rank}(\mathbf{Z}) \leq k + h$ and $\mathbf{Z} = \mathbf{B}\mathbf{\Gamma}$ for some $\mathbf{\Gamma} \in \mathbb{R}^{r,n}$ of rank $k + h$.

4.3.3 The Algorithm

An algorithm that solves problem (4.1) results naturally from our analysis above. This algorithm is presented as **Su**bspace **Re**stricted **Spectral Approx**imation below. It is split into modules for ease of understanding and design; the running time of each module may be improved independently of the others. We also provide the leading running times - running times of operations that are $\Omega(T_{svd(\mathbf{B})})$ - next to the corresponding operations.

Algorithm 3. SuReSpectralApprox(C, B, k, ϵ)

Input: $\mathbf{C} \in \mathbb{R}^{m \times n}$, $\mathbf{B} \in \mathbb{R}^{m \times r}$, $k : 1 \leq k \leq n$, and $\epsilon > 0$.
Output: $\mathbf{Z}$ such that $\mathbf{Z} = \mathbf{B}\mathbf{\Gamma}$, $\mathrm{rank}(\mathbf{Z}) \leq k$, with error $\beta = \|\mathbf{C} - \mathbf{Z}\|$ achieved

according to Theorem 20.

1. $(\mathbf{C}_{\mathcal{B}}, \mathbf{C}_{\mathcal{N}}) = \mathbf{Orthosplit}(\mathbf{C}, \mathbf{B})$ $O(mnr + T_{svd(\mathbf{B})})$
2. $\left(\bar{\mathbf{U}}_N, \bar{\mathbf{\Sigma}}_N, \bar{\mathbf{V}}_N\right) = \mathbf{reducedSVD}(\mathbf{C}_{\mathcal{N}})$ $O(T_{svd(\mathbf{C})})$
3. $h = \mathbf{rank}(\mathbf{C}_{\mathcal{B}}\mathbf{V}_0\mathbf{V}_0^\top)$ $O\left(T_{svd(\mathbf{B})}\right)$
   where $\mathbf{V}_0$ is made of the columns of $\bar{\mathbf{V}}_N$
   corresponding to the top singular value in $\bar{\mathbf{\Sigma}}_N$.
4. $\left(\beta, \bar{h}\right) = \mathbf{SetError}(\mathbf{C}, \mathbf{C}_{\mathcal{N}}, k, h, \epsilon)$ $O\left(T_{svd(\mathbf{C})}\right)$
5. $\mathbf{\Delta} = \beta^2\mathbf{I} - \mathbf{C}_{\mathcal{N}}^\top\mathbf{C}_{\mathcal{N}}$
6. $\mathbf{Z} = \mathbf{ExtractApproximation}(\mathbf{C}_{\mathcal{B}}, \mathbf{\Delta}, k, \bar{h})$ $O\left(T_{svd(\mathbf{C})}\right)$

Here, $\mathbf{Orthosplit}(\mathbf{C}, \mathbf{B})$ obtains the orthogonal components of $\mathbf{C}$ with respect to $\mathcal{B}$. We note that the approximation error may be obtained prior to obtaining the approximation itself, as in theorem 20. The integer, h, is the rank-slack for obtaining an approximation, $\mathbf{C}_{\mathcal{B},k}$, that is as good as the best possible approximation, $\mathbf{C}_{\mathcal{B}}$, to $\mathbf{C}$, with columns in $\mathcal{B}$. The integer $\bar{h}$ indicates the need for rank-slack, equaling h when $\beta = \|\mathbf{C}_{\mathcal{N}}\|$ and 0 otherwise. The running time of this algorithm is $O\left(mnr + T_{svd(\mathbf{C})}\right)$.

Algorithm 4. SetError$(\mathbf{C}, \mathbf{C}_{\mathcal{N}}, k, h, \epsilon)$

Input: $\mathbf{C}, \mathbf{C}_{\mathcal{N}} \in \mathbb{R}^{m\times n}$, $k, h \in \mathbf{Z}^+$, and $\epsilon \in \mathbb{R}, \epsilon > 0$.

Output: $\beta, \bar{h}$, the approximation error and rank-slack

1. $\left(\bar{\mathbf{U}}_{\mathbf{C}}, \bar{\mathbf{\Sigma}}_{\mathbf{C}}, \bar{\mathbf{V}}_{\mathbf{C}}\right) = \mathbf{reducedSVD}(\mathbf{C})$, $O\left(T_{svd(\mathbf{C})}\right)$
   where $\sigma_i := \bar{\mathbf{\Sigma}}_{\mathbf{C}}(i,i)$
2. Compute $\|\mathbf{C}_{\mathcal{N}}\|$
3. Compute $k^* = \arg\max_i \sigma_i > \|\mathbf{C}_{\mathcal{N}}\|$
4. **if** $k < k^*$ **return** $(\sigma_{k+1}, 0)$
   if $k^* \leq k < k^* + h$ **return** $((1+\epsilon)\|\mathbf{C}_{\mathcal{N}}\|, 0)$
   if $k \geq k^* + h$ **return** $(\|\mathbf{C}_{\mathcal{N}}\|, h)$

Algorithm 5. ExtractApproximation($\mathbf{C}_{\mathcal{B}}, \mathbf{\Delta}, k, \bar{h}$)

Input: $\mathbf{C}_{\mathcal{B}} \in \mathbb{R}^{m \times n}, \mathbf{\Delta} \in \mathbb{R}^{n \times n}$ and $k, \bar{h} \in \mathbb{N}$
Output: $\mathbf{Z}$, such that $\mathbf{Z} = \mathbf{B\Gamma}$, $\text{rank}(\mathbf{Z}) \leq k$, and error $\beta = \|\mathbf{C} - \mathbf{Z}\|$ achieved according to Theorem 20.

1. $(\mathbf{V}_+, \mathbf{\Sigma}_+, \mathbf{V}_+) = \mathbf{reducedSVD}(\mathbf{\Delta})$, $O\left(T_{svd(\mathbf{C})}\right)$
   where $\mathbf{\Sigma}_+$ has a positive diagonal
2. $\mathbf{R} = \mathbf{SVtruncate}\left(\mathbf{C}_{\mathcal{B}}\, \mathbf{V}_+ \sqrt{\mathbf{\Sigma}_+}^{-1} \mathbf{V}_+^\top, 1\right)$ $O\left(T_{svd(\mathbf{B})}\right)$
3. $\mathbf{Z} = \mathbf{R}\, \mathbf{V}_+ \sqrt{\mathbf{\Sigma}_+} \mathbf{V}_+^\top + \mathbf{C}_{\mathcal{B}} \mathbf{V}_0 \mathbf{V}_0^\top$.

Here, $\mathbf{SVruncate}(G, \alpha)$ is the τ-rank truncation of $\mathbf{G}$ where $\sigma_\tau(\mathbf{G}) \geq \alpha$ and $\sigma_{\tau+1}(\mathbf{G}) < \alpha$. We note that, for $k < k^* + h$, the rank-h additive correction, $\mathbf{C}_{\mathcal{B}} \mathbf{V}_0 \mathbf{V}_0^\top$ is $\mathbf{0}$. This additive correction only features non-trivially when obtaining an approximation to $\mathbf{C}$ that is as good as the projection, $\mathbf{C}_{\mathcal{B}}$, but possibly of lower rank than $\mathbf{C}_{\mathcal{B}}$. This, for instance, happens when $\mathbf{C}$ is full rank and $k + h < r = rank(\mathbf{B}) = rank(\mathbf{C}_{\mathcal{B}})$.

4.4 Proofs

4.4.1 Lemma 5

The proof of lemma 5 is intuitive; we state the intuition first, followed by a formal proof. We first note that $\mathbf{X} \preceq \mathbf{Y}$ implies that the action of $\mathbf{Y}$ is at least as large as the action of $\mathbf{X}$ everywhere in $\mathbb{R}^n$. Clearly, this implies that, $\mathbf{X}$ cannot have a non-zero action in the space where $\mathbf{Y}$ has zero action, leading to $\mathbf{X}\mathbf{V}_0 = \mathbf{0}$. This also implies that, in the space where $\mathbf{Y}$ has non-zero action in general, its pseudo inverse must reduce a vector more than $\mathbf{X}$ magnifies it, which is essentially the intuition contained in $\left\|\mathbf{F}\sqrt{\mathbf{Y}^\dagger}\right\| \leq 1$. This observation follows immediately from the following well-known property of the semi-definite ordering on matrices (Proposition 2.1.1 in [31])

Lemma 25. *For* $\mathbf{A}, \mathbf{H} \in \mathbb{R}^{m,n}$ *and* $\mathbf{A}, \mathbf{H} \succeq \mathbf{0}$*,*

$$\mathbf{A} \preceq \mathbf{H} \Rightarrow \mathbf{L}^\top \mathbf{A}\mathbf{L} \preceq \mathbf{L}^\top \mathbf{H}\mathbf{L} \quad \forall \mathbf{L} \in \mathbb{R}^{n,k}.$$

We now proceed with the formal proof of lemma 5.

Proof. We recall that

$$\mathbf{X} = \mathbf{F}^\top \mathbf{F}, \quad \mathbf{Y} = (\mathbf{V}_+, \mathbf{V}_0)\, \mathbf{\Sigma}_\mathbf{Y}\, (\mathbf{V}_+, \mathbf{V}_0)^\top,$$

with $\mathbf{V}_+, \mathbf{V}_0$ being the orthonormal bases for the range and nullspace of $\mathbf{Y}$ respectively. Now using Lemma 25 with $\mathbf{A}, \mathbf{H}$ as $\mathbf{X}, \mathbf{Y}$ respectively and first casting $\mathbf{L}$ as $\sqrt{\mathbf{Y}^\dagger}$ we see that

$$\begin{aligned}\mathbf{X} \preceq \mathbf{Y} \Rightarrow \mathbf{F}^\top \mathbf{F} \preceq \mathbf{Y} \Rightarrow \sqrt{\mathbf{Y}^\dagger}\mathbf{F}^\top \mathbf{F}\sqrt{\mathbf{Y}^\dagger} \preceq \sqrt{\mathbf{Y}^\dagger}\mathbf{Y}\sqrt{\mathbf{Y}^\dagger} = \mathbf{V}_+\mathbf{V}_+^\top \\ \Rightarrow \left\|\mathbf{F}\sqrt{\mathbf{Y}^\dagger}\right\|^2 \leq \left\|\mathbf{V}_+\mathbf{V}_+^\top\right\| = 1.\end{aligned}$$

Next, casting $\mathbf{L}$ as $\mathbf{V}_0$:

$$\begin{aligned}\mathbf{X} \preceq \mathbf{Y} &\Rightarrow \mathbf{V}_0^\top \mathbf{X}\mathbf{V}_0 \preceq \mathbf{V}_0^\top \mathbf{Y}\mathbf{V}_0 = \mathbf{0} \\ &\Rightarrow \mathbf{V}_0^\top \mathbf{F}^\top \mathbf{F}\mathbf{V}_0 = (\mathbf{F}\mathbf{V}_0)^\top (\mathbf{F}\mathbf{V}_0) = \mathbf{0} \\ &\Rightarrow \mathbf{F}\mathbf{V}_0 = \mathbf{0} \\ &\Rightarrow \mathbf{F}^\top \mathbf{F}\mathbf{V}_0 = \mathbf{X}\mathbf{V}_0 = \mathbf{0}.\end{aligned}$$

This closes the forward implication. To see the backward implication, we first note that

$$\begin{aligned}\mathbf{X}\mathbf{V}_0 = \mathbf{0} &\Rightarrow \mathbf{X}\mathbf{V}_0\mathbf{V}_0^\top = \mathbf{X}\left(\mathbf{I} - \mathbf{V}_+\mathbf{V}_+^\top\right) = \mathbf{0} \\ &\Rightarrow \mathbf{X}\mathbf{V}_+\mathbf{V}_+^\top = \mathbf{X} \qquad (4.6) \\ &\Rightarrow \mathbf{V}_+\mathbf{V}_+^\top\mathbf{X} = \mathbf{X} \qquad (4.7)\end{aligned}$$

where (4.7) follows from transposing (4.6) and noting that $\mathbf{X}$ is symmetric. Finally,

$$\begin{aligned}\left\|\mathbf{F}\sqrt{\mathbf{Y}^\dagger}\right\| \leq 1 &\Rightarrow \sqrt{\mathbf{Y}^\dagger}\mathbf{F}^\top\mathbf{F}\sqrt{\mathbf{Y}^\dagger} = \sqrt{\mathbf{Y}^\dagger}\mathbf{X}\sqrt{\mathbf{Y}^\dagger} \preceq \mathbf{I} \\ &\Rightarrow \sqrt{\mathbf{Y}}\sqrt{\mathbf{Y}^\dagger}\mathbf{X}\sqrt{\mathbf{Y}^\dagger}\sqrt{\mathbf{Y}} \preceq \mathbf{Y} \\ &\Rightarrow \mathbf{V}_+\mathbf{V}_+^\top\mathbf{X}\mathbf{V}_+\mathbf{V}_+^\top \overset{(a)}{=} \mathbf{V}_+\mathbf{V}_+^\top\mathbf{X} \overset{(b)}{=} \mathbf{X} \preceq \mathbf{Y},\end{aligned}$$

where $(a), (b)$ follow from (4.6), (4.7) respectively. This completes the backward implication and the proof. □

4.4.2 Proofs of Theorems 19, 20

In section 2, we constructed a $\mathbf{Z}$ such that $\mathbf{Z} = \mathbf{B}\boldsymbol{\Gamma}$ for some $\boldsymbol{\Gamma}$, satisfying $\|\mathbf{C} - \mathbf{Z}\| \leq \beta$, where $\text{rank}(\mathbf{Z})$ was related to the number of singular values of $\tilde{\mathbf{C}} = \mathbf{C}\sqrt{\boldsymbol{\Delta}^\dagger}$ that are larger than unity. Theorem 19 is the statement of relationship between the rank, k, of a subspace restricted approximation to $\mathbf{C}$ and the least error achievable with this rank, and theorem 20 is the relationship between these two quantities conditioned upon our method of constructing subspace constrained approximations to $\mathbf{C}$. We prove this by first establishing a relationship between

the number of singular values of $\tilde{\mathbf{C}}$ larger than unity (and hence the rank of $\mathbf{Z}$), and the number of singular values of $\mathbf{C}$ larger than β. We state this relationship in lemma 26, from which theorem 20 follows by simple arguments. Theorem 19 follows from theorem 20 by construction in all cases except in one. This case is individually proven immediately after.

Lemma 26. *Let $\sigma^{(\alpha)}(\mathbf{K})$ be the number of singular values of $\mathbf{K}$ that are larger than α and $\bar{h} = rank(\mathbf{C}_{\mathcal{B}}\mathbf{V}_0\mathbf{V}_0^{\top})$ be as defined in lemma 22. Suppose $\|\mathbf{C} - \mathbf{Z}\| \leq \beta$ and let $\tilde{\mathbf{C}}$ be as defined in (4.5). Then*

$$\beta > \|\mathbf{C}_{\mathcal{N}}\| \;\Rightarrow\; \sigma^{(1)}(\tilde{\mathbf{C}}) = \sigma^{(\beta)}(\mathbf{C}), \tag{4.8}$$

$$\beta = \|\mathbf{C}_{\mathcal{N}}\| \;\Rightarrow\; \sigma^{(1)}(\tilde{\mathbf{C}}) \leq \sigma^{(\beta)}(\mathbf{C}). \tag{4.9}$$

We assume the truth of this lemma, proceed to establish that theorem 20 follows directly from it, and subsequently prove the lemma itself. We prove that theorem 20 results from lemma 26 as follows.

Lemma 27. $\|\mathbf{C} - \mathbf{C}_{\mathbf{B},k}\| = \sigma_{k+1}(\mathbf{C})$*, for $k < k^*$.*

Proof. For $k < k^*$, let $\beta = \sigma_{k+1}(\mathbf{C})$. So $\beta > \|\mathbf{C}_{\mathcal{N}}\|$ by definition of k^*. Clearly, $\sigma^{(\beta)}(\mathbf{C}) = k$, and by lemma 26, we have

$$\sigma^{(1)}(\tilde{\mathbf{C}}) = k.$$

By corollary 23, $\bar{h} = 0$ and there exists a $\mathbf{Z} = \mathbf{B}\mathbf{\Gamma}$ of rank k for some $\mathbf{\Gamma} \in \mathbb{R}^{r,n}$ such that $\|\mathbf{C} - \mathbf{Z}\| \leq \sigma_{k+1}(\mathbf{C})$. However, by Eckart-Young theorem, $\forall \mathbf{D}$ with $\text{rank}(\mathbf{D}) \leq k$, $\|\mathbf{C} - \mathbf{D}\| \geq \sigma_{k+1}(\mathbf{C})$, implying that

$$\|\mathbf{C} - \mathbf{Z}\| = \sigma_{k+1}(\mathbf{C}).$$

So $\mathbf{C}_{\mathcal{B},k} = \mathbf{Z}$ provides us the promised result. □

Lemma 28. $\|\mathbf{C} - \mathbf{C}_{\mathcal{B},k}\| = \|\mathbf{C}_{\mathcal{N}}\|$*, for $k \geq k^* + h$.*

Proof. Let $\beta = \|\mathbf{C}_{\mathcal{N}}\|$. We have, from corollary 24 that $\mathbf{Z} = (\mathbf{C}_{\mathcal{B}}\sqrt{\mathbf{\Delta}^{\dagger}})_k\sqrt{\mathbf{\Delta}} +$

$\mathbf{C}_{\mathcal{B}}\mathbf{V}_*{\mathbf{V}_*}^\top$ has the property that $\|\mathbf{C}-\mathbf{Z}\| \leq \beta = \|\mathbf{C}_{\mathcal{N}}\|$, and

$$rank(\mathbf{Z}) = \sigma^{(1)}(\tilde{\mathbf{C}}) + h.$$

From lemma 26: $\sigma^{(1)}(\tilde{\mathbf{C}}) \leq \sigma^{(\beta)}(\mathbf{C}) \leq k^*$. So it follows that $\forall k \geq k^* + h$, $rank(\mathbf{Z}) \leq k$. However, for any $\mathbf{D}$ with columns in $\mathcal{B}$,

$$\|\mathbf{C}-\mathbf{D}\| \geq \|\mathbf{C}-\mathbf{C}_{\mathcal{B}}\| = \|\mathbf{C}_{\mathcal{N}}\|.$$

Together with $\|\mathbf{C}-\mathbf{Z}\| \leq \|\mathbf{C}_{\mathcal{N}}\|$, setting $\mathbf{C}_{\mathcal{B},k} = \mathbf{Z}$ proves the promised result. □

Lemma 29. *For $k^* \leq k < k^* + h$, we can construct $\mathbf{Z}(\epsilon)$ of rank k, with $\mathbf{Z}(\epsilon) = \mathbf{B}\mathbf{\Gamma}(\epsilon), \mathrm{rank}(\mathbf{\Gamma}(\epsilon)) = k$, such that $\|\mathbf{C}-\mathbf{Z}(\epsilon)\| \leq (1+\epsilon)\|\mathbf{C}_{\mathcal{N}}\|$.*

Proof. This follows by setting $\beta = \beta(\epsilon) = (1+\epsilon)\|\mathbf{C}_{\mathcal{N}}\|$ and arguing as in the proof for lemma 27 without the use of the observation by Eckart and Young. □

Indeed, the approximation, $\mathbf{Z}$, constructed in this manner is a function of $\beta(\epsilon)$, since $\mathbf{\Delta} = \mathbf{\Delta}(\epsilon) = \beta(\epsilon)^2\mathbf{I} - \mathbf{C}_{\mathcal{N}}^\top\mathbf{C}_{\mathcal{N}}$. We note that proving the existence of an exact approximation for k, $k^* \leq k < k^* + h$ (as supposed in lemma 29) proves theorem 19. We do this by simply extending our argument for proving lemma 29, as follows.

Lemma 30. *For $k^* \leq k < k^* + h$, $\exists \mathbf{Z}^* = \mathbf{B}\mathbf{\Gamma}^*$ with $\mathrm{rank}(\mathbf{Z}^*) = \mathrm{rank}(\mathbf{\Gamma}^*) \leq k$ such that $\|\mathbf{C}-\mathbf{Z}^*\| = \|\mathbf{C}_{\mathcal{N}}\|$.*

Proof. Our proof is in two parts: first, we construct a limiting sequence of $\mathbf{Z}(\epsilon)$ with the property guaranteed in lemma 29. Then we argue that the limit, $\mathbf{Z}^*$, of this sequence has the desired properties, namely that $\mathbf{Z}^* = \mathbf{B}\mathbf{\Gamma}^*$ for some $\mathbf{\Gamma}^*$ and that $\mathrm{rank}(Z^*) \leq k$.

First, let $\{\epsilon_l\}$ be an infinite sequence such that $1 > \epsilon_l \geq 0 \forall l$ and $\lim_{l\to\infty}\epsilon_l = 0$. One can let $\epsilon = 1/l^2$ for $l \geq 1$, for instance. Lemma 29 guarantees a corresponding sequence $\mathcal{Z} = \{\mathbf{Z}(\epsilon_l)\}, \mathbf{Z}(\epsilon_l) \in \mathcal{B}^n$ with $\mathcal{B}^n \sim \mathbb{R}^{r\times n}$ being a closed subspace of $\mathbb{R}^{m\times n}$. We note that $\mathcal{Z}$ is uniformly bounded, since

$$\|\mathbf{Z}(\epsilon_l)\| \leq \|\mathbf{C}-\mathbf{Z}(\epsilon_l)\| + \|\mathbf{C}\| \leq (1+\epsilon_l)\|\mathbf{C}_{\mathcal{N}}\| + \|\mathbf{C}\| < 2\|\mathbf{C}\| + \|\mathbf{C}\| = 3\|\mathbf{C}\|.$$

By virtue of m, n being finite and by the Bolzano-Weierstrass theorem, we have that $\mathcal{Z}$ must have a convergent sub-sequence converging in $\mathcal{B}^n$, say $\tilde{\mathcal{Z}} = \{\mathbf{Z}_{l_t} = \mathbf{Z}(1+\epsilon_{l_t})\}$. Let $\mathbf{Z}^*$ be the limit of $\tilde{\mathcal{Z}}$.
By virtue of $\mathbf{Z}^*$ being in $\mathcal{B}^n$ (i.e. the n columns of $\mathbf{Z}^*$ are each in $\mathcal{B}$), $\mathbf{Z}^* = \mathbf{B}\mathbf{\Gamma}^*$ for a $\mathbf{\Gamma}^* \in \mathbb{R}^{r,n}$. One such $\mathbf{\Gamma}^*$ is simply $\mathbf{\Gamma}^* = \mathbf{B}^\dagger\mathbf{Z}^*$. Next, by the lower semi-continuity of rank$(\cdot)$ on matrices [32], we have that

$$\text{rank}(\mathbf{Z}^*) \leq \text{rank}(\mathbf{Z}_{l_t}) \leq k,$$

which proves that $\mathbf{Z}^*$ obeys both the subspace and rank constraints in theorem 19. Finally, we have from the triangle inequality that

$$\|\mathbf{C} - \mathbf{Z}^*\| \leq \|\mathbf{C} - \mathbf{Z}_{l_t}\| + \|\mathbf{Z}^* - \mathbf{Z}_{l_t}\| \leq (1 + \epsilon_{l_t}) \|\mathbf{C}_\mathcal{N}\| + \|\mathbf{Z}^* - \mathbf{Z}_{l_t}\|,$$

which, at the limit $t \to \infty$, gives us

$$\|\mathbf{C} - \mathbf{Z}^*\| \leq \|\mathbf{C}_\mathcal{N}\|.$$

Since for any $\mathbf{D}$ with columns in $\mathcal{B}$,

$$\|\mathbf{C} - \mathbf{D}\| \geq \|\mathbf{C} - \mathbf{C}_\mathcal{B}\| = \|\mathbf{C}_\mathcal{N}\|,$$

we have that

$$\|\mathbf{C} - \mathbf{Z}^*\| = \|\mathbf{C}_\mathcal{N}\|,$$

closing the proof of theorem 19. □

4.4.2.1 Proof of Lemma 26

At the crux of the proof of lemma 26 is a law concerning an invariant property of the eigenvalues of a matrix $\mathbf{D} \in \mathbb{R}^{m,m}$ under transformations of the form $\mathbf{S}^\top\mathbf{D}\mathbf{S}$, due to Sylvester, presented below.

Theorem 31 (Sylvester's Law Of Inertia)**.** *Let $\mathbf{D} \in \mathbb{R}^{m,m}$, and define the signature of $\mathbf{D}$, $[\mathbf{D}] = (\lambda_+(\mathbf{D}), \lambda_0(\mathbf{D}), \lambda_-(\mathbf{D}))$, as the 3-tuple containing the number of positive,*

zero and negative eigenvalues of $\mathbf{D}$ *in that order. Then for all non-singular* $\mathbf{S} \in \mathbb{R}^{m,m}$,

$$[\mathbf{S}^\top \mathbf{D}\mathbf{S}] = [\mathbf{D}].$$

Let $\sigma^{(\alpha)}(\mathbf{C})$ be as defined in lemma 26, and let $\sigma_0^{(\alpha)}(\mathbf{C}), \sigma_-^{(\alpha)}(\mathbf{C})$ denote the number of singular values of $\mathbf{C}$ that are equal to and lesser than α respectively. To see the relevance of Sylvester's law, we simply observe that

$$[\mathbf{C}^\top \mathbf{C} - \alpha^2 \mathbf{I}] = \left(\sigma^{(\alpha)}(\mathbf{C}), \sigma_0^{(\alpha)}(\mathbf{C}), \sigma_-^{(\alpha)}(\mathbf{C})\right). \tag{4.10}$$

We prove lemma 26 by relating $[\mathbf{C}^\top \mathbf{C} - \alpha^2 \mathbf{I}]$ and $[\tilde{\mathbf{C}}^\top \tilde{\mathbf{C}} - \mathbf{I}]$. First, we recall that

$$\tilde{\mathbf{C}}^\top \tilde{\mathbf{C}} - \mathbf{I} = \sqrt{\boldsymbol{\Delta}^\dagger} \left(\mathbf{C}^\top \mathbf{C} - \beta^2 \mathbf{I}\right) \sqrt{\boldsymbol{\Delta}^\dagger}.$$

Clearly, $\sqrt{\boldsymbol{\Delta}^\dagger}$ is symmetric and expressing $\tilde{\mathbf{C}}^\top \tilde{\mathbf{C}} - \mathbf{I}$ as

$$\tilde{\mathbf{C}}^\top \tilde{\mathbf{C}} - \mathbf{I} = \mathbf{S}^\top \left(\mathbf{C}^\top \mathbf{C} - \beta^2 \mathbf{I}\right) \mathbf{S}$$

for a non-singular $\mathbf{S}$ follows immediately if $\boldsymbol{\Delta}$ is invertible. Certainly, this is the case when $\beta > \|\mathbf{C}_{\mathcal{N}}\|$, which gives (4.8), but a different approach is necessary when $\beta = \|\mathbf{C}_{\mathcal{N}}\|$. In general, suppose the full SVD of $\boldsymbol{\Delta}$ is $(\mathbf{V}_+, \mathbf{V}_0)\, \boldsymbol{\Sigma}_{\boldsymbol{\Delta}}\, (\mathbf{V}_+, \mathbf{V}_0)^\top$, as described in (4.3) and define

$$\boldsymbol{\Delta}_1 = \boldsymbol{\Delta} + \mathbf{V}_0 \mathbf{V}_0^\top,$$

effectively replacing the zeros in $\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}$ by ones. By construction, $\boldsymbol{\Delta}_1$ is invertible, commutes with $\boldsymbol{\Delta}$, and

$$\boldsymbol{\Delta}_1^{-1} \boldsymbol{\Delta} = \boldsymbol{\Delta}^\dagger \boldsymbol{\Delta} = \mathbf{V}_+ \mathbf{V}_+^\top = \mathbf{I} - \mathbf{V}_0 \mathbf{V}_0^\top.$$

Furthermore,

$$\sqrt{\boldsymbol{\Delta}_1^{-1}}^\top \left(\mathbf{C}^\top \mathbf{C} - \beta^2 \mathbf{I}\right) \sqrt{\boldsymbol{\Delta}_1^{-1}} \succeq \sqrt{\boldsymbol{\Delta}^\dagger}^\top \left(\mathbf{C}^\top \mathbf{C} - \beta^2 \mathbf{I}\right) \sqrt{\boldsymbol{\Delta}^\dagger} \tag{4.11}$$

resulting from the following lemma.

Lemma 32. *Suppose that* $\mathbf{M}, \mathbf{A}_1, \mathbf{A}_2 \succeq \mathbf{0}$ *are symmetric,* $\mathbf{A}_1 \succeq \mathbf{A}_2$ *and that* $\mathbf{A}_1, \mathbf{A}_2$ *commute. Then*

$$\mathbf{A}_1^\top \mathbf{M} \mathbf{A}_1 \succeq \mathbf{A}_2^\top \mathbf{M} \mathbf{A}_2.$$

Now we simply observe that $\sqrt{\boldsymbol{\Delta}_1^{-1}} = \sqrt{\boldsymbol{\Delta}^\dagger} + \mathbf{V}_0 \mathbf{V}_0^\top \succeq \sqrt{\boldsymbol{\Delta}^\dagger}$, and cast $\mathbf{M}, \mathbf{A}_1, \mathbf{A}_2$ in lemma 32 as $\left(\mathbf{C}^\top \mathbf{C} - \beta^2 \mathbf{I}\right), \sqrt{\boldsymbol{\Delta}_1^{-1}}, \sqrt{\boldsymbol{\Delta}^\dagger}$ respectively to obtain (4.11), whence there exists an $\mathbf{H} \succeq \mathbf{0}$ such that

$$\sqrt{\boldsymbol{\Delta}_1^{-1}}^\top \left(\mathbf{C}^\top \mathbf{C} - \beta^2 \mathbf{I}\right) \sqrt{\boldsymbol{\Delta}_1^{-1}} = \sqrt{\boldsymbol{\Delta}^\dagger}^\top \left(\mathbf{C}^\top \mathbf{C} - \beta^2 \mathbf{I}\right) \sqrt{\boldsymbol{\Delta}^\dagger} + \mathbf{H}. \tag{4.12}$$

By Sylvester's law of inertia then,

$$\begin{aligned}
[\mathbf{C}^\top \mathbf{C} - \beta^2 \mathbf{I}] &= [\mathbf{C}_\mathcal{B}^\top \mathbf{C}_\mathcal{B} - \boldsymbol{\Delta}] \\
&= \left[\sqrt{\boldsymbol{\Delta}_1^{-1}}^\top \left(\mathbf{C}_\mathcal{B}^\top \mathbf{C}_\mathcal{B} - \boldsymbol{\Delta}\right) \sqrt{\boldsymbol{\Delta}_1^{-1}}\right] \\
&= \left[\sqrt{\boldsymbol{\Delta}_1^{-1}}^\top \mathbf{C}_\mathcal{B}^\top \mathbf{C}_\mathcal{B} \sqrt{\boldsymbol{\Delta}_1^{-1}} - \mathbf{I} + \mathbf{V}_0 \mathbf{V}_0^\top\right] \\
&= \left[\left(\sqrt{\boldsymbol{\Delta}^\dagger}^\top \mathbf{C}_\mathcal{B}^\top \mathbf{C}_\mathcal{B} \sqrt{\boldsymbol{\Delta}^\dagger} + \mathbf{H}\right) - \mathbf{I} + \mathbf{V}_0 \mathbf{V}_0^\top\right] \\
&= \left[\tilde{\mathbf{C}}^\top \tilde{\mathbf{C}} - \mathbf{I} + \left(\mathbf{H} + \mathbf{V}_0 \mathbf{V}_0^\top\right)\right], \quad \text{where } \mathbf{H} + \mathbf{V}_0 \mathbf{V}_0^\top \succeq \mathbf{0}
\end{aligned} \tag{4.13}$$

Here on, we will proceed by considering only the first (and relevant) component, $\sigma^{(\beta)}(\mathbf{C})$, of the signature of $\mathbf{C}^\top \mathbf{C} - \beta^2 \mathbf{I}$. We briefly pause to recall a standard result in the following lemma:

Lemma 33. *Let* $\mathbf{W} \succeq \mathbf{0}$*. Then for all* $\mathbf{K}$ *where* $\mathbf{K} + \mathbf{W}$ *is defined,*

$$\lambda(\mathbf{K} + \mathbf{W}) \geq \lambda(\mathbf{K}).$$

The following corollary results immediately.

Corollary 34. *If* $\lambda_+(\mathbf{K})$ *denotes the number of positive eigenvalues of* $\mathbf{K}$*,*

$$\lambda_+(\mathbf{K} + \mathbf{W}) \geq \lambda_+(\mathbf{K}).$$

Now, recalling (4.10) and (4.13)

$$\begin{aligned}\sigma^{(\beta)}(\mathbf{C}) &= \lambda_+\left(\tilde{\mathbf{C}}^\top\tilde{\mathbf{C}} - \mathbf{I} + \left(\mathbf{H} + \mathbf{V}_0\mathbf{V}_0^\top\right)\right) \\ &\geq \lambda_+\left(\tilde{\mathbf{C}}^\top\tilde{\mathbf{C}} - \mathbf{I}\right) \\ &\geq \sigma^{(1)}(\tilde{\mathbf{C}}).\end{aligned}$$

The above course of reasoning, while designed to establish (4.9), is general and provides both statements contained in lemma 26 if we observe that $\mathbf{H} + \mathbf{V}_0\mathbf{V}_0^\top = \mathbf{0}$ when $\beta > \|\mathbf{C}_\mathcal{N}\|$ whence $\mathbf{\Delta}$ is invertible and $\mathbf{V}_0 = \mathbf{0}$. In this case the above statements hold with equality, providing (4.8). This closes the proof of lemma 26. □

4.5 Conclusion

In this chapter we derived a closed form solution to the problem of subspace restricted low rank approximation, and devised an efficient algorithm to compute such an approximation exactly. The running time of this algorithm, however, is dominated by the computation of the SVD of $\mathbf{C}_\mathcal{N}$. In the case where $\text{rank}(\mathbf{B}) << \text{rank}(\mathbf{C}_\mathcal{N})$, it would be particularly effective if one can obtain $\mathbf{C}_{\mathcal{B},k}$ in time comparable to performing an SVD on just $\mathbf{C}_\mathcal{B}$. Such an algorithm cannot be reported at the moment of this writing. However, there is sufficient reason to believe that, under a very intuitive assumption, one may be able to obtain a relative error approximation to $\mathbf{C}_{\mathcal{B},k}$ in time comparable to performing the SVD on $\mathbf{C}_\mathcal{B}$. Formally, one can find compelling reasons from both numerical experiments and analysis for the following conjecture to hold

Conjecture 35. *Let $\mathcal{L}_k$ be the range of $\mathbf{C}_k$. If*

$$\|\Pi_{\mathcal{L}_k}\Pi_\mathcal{B}\| \geq 1 - \epsilon,$$

$$\|\mathbf{C} - (\mathbf{C}_\mathcal{B})_k\| \leq (1 + \epsilon')\|\mathbf{C} - \mathbf{C}_{\mathcal{B},k}\|,$$

where ϵ, ϵ' are both small in magnitude.

Since $(\mathbf{C}_{\mathcal{B}})_k$ only requires an SVD of $\mathbf{C}_{\mathcal{B}}$, the truth of the above conjecture would imply precisely the sort of fast algorithm we desire. This closes our study of subspace restricted low rank approximations.

4.6 Acknowledgement

The author would like to acknowledge Malik Magdon-Ismail for his discussions concerning the rank slack, and Jeffrey Banks for his time and discussion, which informed conjecture 35, and Peter Kramer for the many expository and technical contributions to the chapter.

5. Fast Low Rank Approximations of Matrices

5.1 Introduction

In the previous chapter, we saw that a matrix $\mathbf{C}$ may be approximated by a rank-k matrix $\mathbf{C}_{\mathcal{B},k}$ containing columns from a specified subspace, $\mathcal{B}$, such that the approximation is the best one can do with rank k without additional subspace constraints. Whether or not this is possible is determined by whether k is smaller than a certain critical rank. In this chapter, we discuss how the results of our study of subspace restricted low rank approximation ensure that exact low rank approximations, unrestricted by any subspace-constraints, can be obtained using existing algorithms. This is an improvement to existing results which guarantee only that a relative error low-rank approximation may be obtained by existing algorithms. Furthermore, we provide a natural and intuitive conjecture the truth of which implies the existence of very good relative error approximations that can be computed very quickly.

Specifically, a common proto-algorithm used to quickly compute low rank approximations (II.4 in [5]) proceeds as follows

Algorithm 6. spectralLowRank($\mathbf{C}$, k)

1. Compute a low-dimensional approximation, $\mathcal{B}$, to the range of $\mathbf{C}$,
2. Return $\mathbf{C}_{\mathcal{B},k}$ as the desired approximation.

The approximation of the range of $\mathbf{C}$ may be done in several ways [5], two of which are in common use and discussed in this chapter.

1. Approximating the range of $\mathbf{C}$ as $\mathbf{im}\,(\mathbf{CG})$ for a matrix $\mathbf{G} \in \mathbb{R}^{n\times l}$ with entries coming from a standard normal.
2. Approximating the range of $\mathbf{C}$ as $\mathbf{im}\,(\mathbf{C}^r\mathbf{G})$ for $\mathbf{G} \in \mathbb{R}^{n\times l}$ with entries coming

from a standard normal. This can be viewed as letting $\mathbf{C}$ "pull" $\mathbf{im}(\mathbf{G})$ r times, closer to its top l singular subspace after every such pull.

Both of these ways is accompanied by a guarantee of how "good" a low-rank approximation they contain, where the "goodness" is usually measured as how close the approximation error is to the optimal approximation error in both frobenius and spectral norms.

The probabilistic guarantee of exact low rank approximations that we provide borrows its functional form and analytical tools used in providing guarantees provided by the above methods for approximating the range of $\mathbf{C}$. Due to this reason, this chapter first outlines these two methods and the probabilistic guarantees of spectral error that accompany them, setting the tools in place which are then used to state, prove and discuss the guarantee of exact approximations.

5.2 Approximating The Range By A Single Action

Here, the top l left singular subspace of $\mathbf{C}$ is approximated as $\mathbf{im}(\mathbf{CG})$ for a $\mathbf{G}$ such that $\mathbf{G}(i,j) \sim \mathcal{N}(0,1)$. It can be inferred immediately that this method approximates $\mathbf{im}(\mathbf{C})$ exactly if $\text{rank}(\mathbf{C}) \leq l$, where the least $m-l$ singular values are simply 0. Guiding the intuition using this example, one expects such a method to approximate $\mathbf{im}(\mathbf{C})$ relatively well if the top l singular values are well separated from the remaining singular values. Indeed, if every pair of singular values is well separated, the spectrum is said to decay quickly and approximating $\mathbf{im}(\mathbf{C})$ by a single action of $\mathbf{C}$ on $\mathbf{G}$ works particularly well. Studies by [5, 3, 4] reveal results we restate here. The mathematical objects that play a role in these results are introduced in the following hypothesis.

Hypothesis 36. *1. Let $\mathbf{C} \in \mathbb{R}^{m,n}$, $m < n$, with singular values σ_i indexed in non-increasing order.*

2. Let k, p, l be integers such that $k > 2, p > 2, l = k + p \leq \min\{m, n\} = m$.

3. Suppose that $\mathbf{G} \in \mathbb{R}^{n,l}$ such that $\mathbf{G}(i,j) \sim \mathcal{N}(0,1)$.

4. Let $\mathcal{B}'$ be the range of $\mathbf{CG}$ and let $\mathcal{N}'$, its orthogonal complement.

5. *Let* $\mathbf{C}_{\mathcal{B}'} = \Pi_{\mathcal{B}'}, \mathbf{C}_{\mathcal{N}'} = \mathbf{C} - \mathbf{C}_{\mathcal{B}'}$.

The expected size of the error incurred in approximating $\mathbf{C}$ as $\mathbf{C}_{\mathcal{B}'}$, and the deviation of the size of this error from its mean, are bounded as follows (Theorem 10.6, Corollary 10.9 in [5]).

Lemma 37. *Under hypothesis 36,*

$$\mathsf{E}\left[\|\mathbf{C}_{\mathcal{N}'}\|\right] \leq \left(1 + f_1(m,n,k,p)\right)\sigma_{k+1}, \quad f_1(m,n,k,p) = \sqrt{\frac{k}{p-1}} + \frac{e\sqrt{l\left(\min\{m,n\}-k\right)}}{p}.$$

Lemma 38. *Under hypothesis 36,*

$$\|\mathbf{C}_{\mathcal{N}'}\| \leq \left(1 + g_1(m,n,k,p)\right)\sigma_{k+1}, \quad g_1(m,n,k,p) = 17\sqrt{1+k/p} + \frac{8\sqrt{l\left(\min\{m,n\}-k\right)}}{p+1}$$

with probability at least $1 - 6p^{-p}$.

5.2.1 Understanding the Bounds for Expectation and Deviation

The bounds that appear above are partially simplified from their most general forms. To better intuit the content of the bounds that appear above and bounds that will be used henceforth, they may be simplified further. Specifically, it is clear that $l = O(k)$ and that this fact remains unchanged if $p = k$. While setting the oversampling parameter, p, to the target rank, k, is not a computationally necessary or prudent choice, it changes little in the asymptotic analysis of the resulting scheme. Effecting this change, we see that

$$\begin{aligned} f_1(m,n,k,k) &= \sqrt{\frac{k}{k-1}} + \frac{e\sqrt{2k\left(m-k\right)}}{k} \leq 1 + e\sqrt{\frac{2(m-k)}{k}} \\ &= O\left(\sqrt{\frac{m}{k}}\right), \end{aligned} \tag{5.1}$$

and that

$$\begin{aligned} g_1(m,n,k,k) &= 17\sqrt{2} + \frac{8\sqrt{l\left(m-k\right)}}{k+1} \leq 17\sqrt{2} + 8\sqrt{\frac{2(m-k)}{k+1}} \\ &= O\left(\sqrt{\frac{m}{k}}\right). \end{aligned} \tag{5.2}$$

We now recast lemmas 37, 38 as

Corollary 39. *Under hypothesis 36,*

$$\mathsf{E}\left[\|\mathbf{C}_{\mathcal{N}'}\|\right] \leq \left(2 + \sqrt{\frac{2(m-k)}{k}}\right)\sigma_{k+1}.$$

Furthermore,

$$\|\mathbf{C}_{\mathcal{N}'}\| \leq \left(26 + 8\sqrt{\frac{2(m-k)}{k+1}}\right)\sigma_{k+1} \tag{5.3}$$

with probability at least $1 - 6k^{-k}$.

Since k is often $o(m)$, these bounds are useful for good approximations as such only if σ_{k+1} is well separated from larger singular values. For instance, in (5.3), the bound is useful only if

$$\left(26 + 8\sqrt{\frac{2(m-k)}{k+1}}\right)\sigma_{k+1} \leq \sigma_j$$

for a j that is not too small. If $j = 1$ the approximation obtained is only slightly better than $\mathbf{0}$, since $\|\mathbf{C} - \mathbf{0}\| = \sigma_1$. From (38), it is clear that the spectrum of $\mathbf{C}$ must decay as $\Theta(m)$ if the bound is to be non-trivial. Mainly due to this imposition on the decay of the spectrum, the expectation and deviation bounds obtained by a single action of $\mathbf{C}$ on $\mathbf{G}$ are not immediately encouraging, as illustrated in the tables below.

Table 5.1: Examples of the bounds stated in lemma 39. In all these examples, $q = 5$ and $p = k$, implying a probability of ≈ 1.0 that the bounds presented are true.

m	n	k	$\|\mathbf{C}_{\mathcal{N}'}\| / \sigma_{k+1} \leq$
512	1024	10	102.43
512	1024	100	48.85
1024	2048	100	60.22
2048	4096	100	75.69

One can provide better guarantees, however, by simply allowing $\mathbf{C}$ to act on $\mathbf{G}$ repeatedly [5, 4]. A treatment of this idea is provided here.

5.3 Approximating The Range By Repeated Action

In addition to hypothesis 36, suppose the following

Hypothesis 40. *Suppose $\mathcal{B}$ is the range of $\left(\mathbf{C}\mathbf{C}^\top\right)^q \mathbf{C}\mathbf{G}$, that $\mathcal{N}$ is the orthogonal complement of $\mathcal{B}$ and that*

$$\mathbf{C}_\mathcal{B} = \Pi_\mathcal{B}\mathbf{C}, \quad \mathbf{C}_\mathcal{N} = \mathbf{C} - \mathbf{C}_\mathcal{B}.$$

The essential observation due to [5] that leads to $\mathbf{C}_\mathcal{B}$ approximating $\mathbf{C}$ better than $\mathbf{C}_{\mathcal{B}'}$ is that

Lemma 41. *Let $\mathbf{C}, \mathcal{B}', \mathcal{B}$ be as in hypothesis 40, then*

$$\|\mathbf{C} - \mathbf{C}_{\mathcal{B}'}\| \leq \|\mathbf{C} - \mathbf{C}_\mathcal{B}\|^{1/2q+1}.$$

Lemma 41 may now be used to immediately produce the following results that bound the expected size of the error incurred in approximating $\mathbf{C}$ as $\mathbf{C}_\mathcal{B}$, and the deviation in norm from its expected value.

Lemma 42. *Under hypothesis 40, with f_1 as in lemma 37,*

$$\mathsf{E}\left[\|\mathbf{C}_\mathcal{N}\|\right] \leq \left(1 + f_1(m,n,k,p)\right)^{\frac{1}{2q-1}} \sigma_{k+1}.$$

Lemma 43. *Under hypothesis 40, with g_1 as in lemma 38,*

$$\|\mathbf{C}_\mathcal{N}\| \leq \left(1 + g_1(m,n,k,p)\right)^{\frac{1}{2q-1}} \sigma_{k+1}$$

with probability at least $1 - 6p^{-p}$.

Setting $p = k$ to provide an intuitive bound, we obtain

Corollary 44. *Under hypothesis 40,*

$$\mathsf{E}\left[\|\mathbf{C}_\mathcal{N}\|\right] \leq \left(2 + \sqrt{\frac{2(m-k)}{k}}\right)^{\frac{1}{2q-1}} \sigma_{k+1}.$$

Furthermore,

$$\|\mathbf{C}_{\mathcal{N}}\| \leq \left(26 + 8\sqrt{\frac{2(m-k)}{k+1}}\right)^{\frac{1}{2q-1}} \sigma_{k+1}$$

with probability at least $1 - 6k^{-k}$.

The above guarantees are significantly lower upper bounds than the ones in lemmas 37, 38, as can be seen in table 5.3.

Table 5.2: Examples of the bounds stated in lemmas 39 and 44. In all these examples, $q = 5$ **and** $p = k$**, implying a probability of** ≈ 1.0 **that the bounds presented are true.**

m	n	k	$\|\mathbf{C}_{\mathcal{N}'}\| / \sigma_{k+1} \leq$	$\|\mathbf{C}_{\mathcal{N}}\| / \sigma_{k+1} \leq$
512	1024	10	102.43	1.67
512	1024	100	48.85	1.54
1024	2048	100	60.22	1.58
2048	4096	100	75.69	1.62

In summary, the best of the above results guarantees a $k+p$ rank matrix, $\mathbf{C}_{\mathcal{B}}$, that has a high probability of approximating $\mathbf{C}$ only slightly worse than $\mathbf{C}_k$. An even stronger result, due to [27], is that

Lemma 45. *Under hypothesis 40,*

$$\mathsf{E}\left[\|\mathbf{C} - \mathbf{C}_{\mathcal{B},k}\|\right] \leq \left(2 + \sqrt{\frac{2(m-k)}{k}}\right)^{\frac{1}{2q-1}} \sigma_{k+1}.$$

Furthermore,

$$\|\mathbf{C} - \mathbf{C}_{\mathcal{B},k}\| \leq \left(26 + 8\sqrt{\frac{2(m-k)}{k+1}}\right)^{\frac{1}{2q-1}} \sigma_{k+1}$$

with probability at least $1 - 6k^{-k}$.

Even this result only guarantees an approximation *close* to the best possible rank k approximation. A fundamental result arising from the work presented in the previous chapter is that one can guarantee the existence of approximations with columns in $\mathcal{B}$ that are *exactly* as good as the best possible rank k approximation to $\mathbf{C}$ in spectral norm.

5.4 The Existence of Exact Approximations

The central result in this chapter is a probabilistic guarantee on the existence of low rank approximations to a matrix $\mathbf{C} \in \mathbb{R}^{(m \times n)}, m \leq n$. Let the logarithmic gap, γ_{k+1}, be as defined in the following hypothesis.

Hypothesis 46. *In addition to hypothesis 40, let*

$$\gamma_{k+1} = \log(\sigma_{k+1}) - \log(\sigma_{k+3}) > 0. \tag{5.4}$$

Then

Theorem 47. *Under hypothesis 46, if*

$$q \geq \frac{1}{2}\left(\frac{\log\left(26 + 8\sqrt{\frac{2m}{k+3}}\right)}{\gamma_{k+1}} + 1\right), \tag{5.5}$$

then with probability at least $1 - 6(k+3)^{-(k+3)}$,

$$\|\mathbf{C} - \mathbf{C}_{\mathcal{B},k}\| = \sigma_{k+1}.$$

The salient features of the result above are that

1. the approximation is exact,
2. the number of power iterations required to guarantee the existence of exact rank k approximations to $\mathbf{C}$ in $\mathcal{B}$ is logarithmic in a low power of m/k,
3. the number of power iterations depends upon the gap between $\sigma_{k+1}, \sigma_{k+3}$,
4. the result is probabilistic with a quickly (exponentially) decaying failure probability, $6(k+3)^{-(k+3)}$.

We briefly discuss these features and present the intuition for the proof of the result, followed by the formal proof itself. The exactness of the approximation is the main contribution of this work to the study of obtaining low rank approximations quickly, and follows from the observation that the optimal spectral error from the rank-k

approximation of $\mathbf{C}$ can be achieved even if the columns of the usual solution, $\mathbf{C}_k$, are not available from within $\mathcal{B}$. The prerequisite to an exact rank k approximation, as discussed in the previous chapter, is that the target rank, k, be larger than the critical rank, k^* where $\sigma_{k^*} > \|\mathbf{C}_{\mathcal{N}}\|$. Clearly, if $k^* \geq k+1$, then $\mathbf{C}_{sB,k}$ is as good an approximation to $\mathbf{C}$ as $\mathbf{C}_k$, incurring the error σ_{k+1}. Since the algorithm in 6 is randomized, it is not possible to deterministically control $\|\mathbf{C}_{\mathcal{N}}\|$ and one settles for a high-probability upper bound of $\|\mathbf{C}_{\mathcal{N}}\|$ instead. The imposition that $k^* \geq k+1$ is present here as the imposition that there be a sufficient gap between σ_{k+1} and the controlling upper bound of $\|\mathbf{C}_{\mathcal{N}}\|$. The essential idea behind proving the result is that we construct a subspace $\mathcal{B} = \mathbf{im}\left(\left(\mathbf{C}\mathbf{C}^\top\right)^q \mathbf{C}\mathbf{G}\right)$ (as in hypothesis 40) such that $\mathbf{C}_{\mathcal{B}}$ is roughly as good as a rank $k+3$ approximation, yielding an approximation error close to σ_{k+3}. Now, provided we have that σ_{k+1} is strictly larger than this approximation error, we have $k^* \geq k+1$ and the result follows.

In this construction, how close the approximation error $\|\mathbf{C} - \mathbf{C}_{\mathcal{B},k}\|$ is to σ_{k+3} intuitively depends upon how close $\mathcal{B}$ is to the top l left singular subspace of $\mathbf{C}$, which is boosted by power iterations. The larger the number of power iterations, q, the closer $\|\mathbf{C} - \mathbf{C}_{\mathcal{B}}\|$ is to σ_{k+3}. Since it is only necessary to obtain an approximation error close enough to σ_{k+3}, so that this error is strictly less than σ_{k+1}, the number of power iterations depends inversely on this gap, as can be observed in figure 5.4.

The proof is a straightforward application of the intuition and the lemmas presented above:

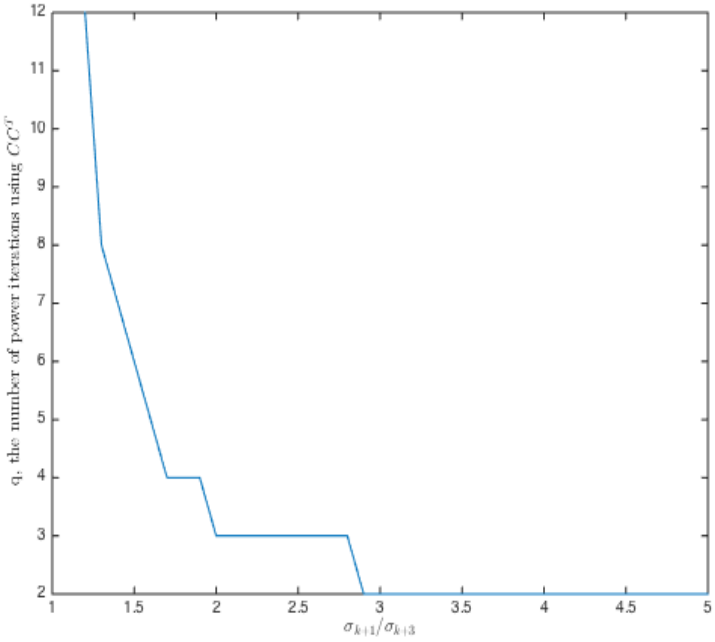


Figure 5.1: The number of power iterations, q, of $\mathbf{CC}^\top$ required to guarantee exact rank k approximation, as a function of $\sigma_{k+1}/\sigma_{k+3}$.

Proof. From (5.5), we have

$$
\begin{aligned}
& q \geq \frac{1}{2}\left(\frac{\log\left(26+8\sqrt{\frac{2m}{k+3}}\right)}{\gamma_{k+1}}+1\right) \\
\Rightarrow\ & 2q-1 \geq \frac{\log\left(26+8\sqrt{\frac{2m}{k+3}}\right)}{\gamma_{k+1}} \\
\Rightarrow\ & \log\left(\frac{\sigma_{k+1}}{\sigma_{k+3}}\right) \geq \frac{\log\left(26+8\sqrt{\frac{2m}{k+3}}\right)}{2q-1} \quad \text{from (5.4)} \\
\Rightarrow\ & \sigma_{k+1} \geq \left(26+8\sqrt{\frac{2m}{k+3}}\right)^{1/2q-1} \sigma_{k+3} \quad \text{from exponentiating both sides} \\
\Rightarrow\ & \sigma_{k+1} \geq \left(26+8\sqrt{\frac{2m}{k+3}}\right)^{1/2q-1} \sigma_{k+3}. \qquad (5.6)
\end{aligned}
$$

Clearly,

$$\sqrt{\frac{2m}{k+3}} > \sqrt{\frac{2m}{k+3} - 2} = \sqrt{\frac{2m - 2(k+3)}{k+3}} > \sqrt{\frac{2(m-k-3)}{k+4}},$$

whence, from (5.6), we have

$$\sigma_{k+1} \geq \left(26 + 8\sqrt{\frac{2m}{k+3}}\right)^{1/2q-1} \sigma_{k+3} > \left(26 + 8\sqrt{\frac{2(m-k-3)}{k+4}}\right) \sigma_{k+3} \tag{5.7}$$

We know from corollary 44 that with probability at least $1 - 6(k+3)^{-(k+3)}$, and with $\mathbf{G} \in \mathbb{R}^{n,l}, l = 2(k+3)$,

$$\|\mathbf{C}_{\mathcal{N}}\| \leq \left(26 + 8\sqrt{\frac{2(m-k-3)}{k+4}}\right) \sigma_{k+3}. \tag{5.8}$$

Combining (5.8), (5.7), we have that, with probability at least $1 - 6(k+3)^{-(k+3)}$,

$$\sigma_{k+1} > \|\mathbf{C}_{\mathcal{N}}\| \tag{5.9}$$

We now simply invoke the result of the study in the previous chapter, theorem 20, and have that

$$\|\mathbf{C} - \mathbf{C}_{\mathcal{B},k}\| = \sigma_{k+1}.$$

□

5.5 Conclusion

The results above provide guarantees on the goodness of the best possible rank k approximations we may obtain using power iterations on a random Gaussian matrix. The computation of the this approximation requires an SVD of $\mathbf{C}_{\mathcal{N}}$, defeating the purpose of using power iterations to reduce the dimension of the matrix we wish to compute the SVD of, as seen in the previous chapter. Due to the nature of power iterations, however, there seem to exist enough reasons to conjecture the following.

Conjecture 48. *Let $\mathcal{L}_k$ be the range of $\mathbf{C}_k$, and let $\mathcal{B}, q$ be as defined in hypothesis*

40. For q that is not too large and a small $\epsilon > 0$,

$$\|\Pi_{\mathcal{L}_k}\Pi_{\mathcal{B}}\| \geq 1 - \epsilon.$$

This, along with conjecture 35, implies that we may compute rank-k approximations to $\mathbf{C}$ in relative error and in running time that is dominated by computing the SVD of $\mathbf{C}_{\mathcal{B}}$. Summarily:

Conjecture 49. *Under hypothesis 40 and the conditions of theorem 47,*

$$\|\mathbf{C} - (\mathbf{C}_{\mathcal{B}})_k\| \leq (1 + \epsilon)\sigma_{k+1}.$$

This closes the presentation of probabilistic guarantees of exact low rank approximations.

5.6 Acknowledgment

The author would like to acknowledge Peter Kramer and Jeffrey Banks for their time and involvement in the discussion of ideas which led to conjecture 48.

6. ELEMENT SPARSIFICATION IN FINITE ELEMENT MESHES

6.1 Introduction

The common approach to solve the system $\mathbf{Ax} = \mathbf{b}$ is to obtain good preconditioners to $\mathbf{A}$. In this chapter, $\mathbf{A}$ arises from the finite element discretization of an elliptic PDE. These elements are readily available, and it is in our best interest to understand the relative importance of each of these elements. So we study, in particular, the quick construction of preconditioners to $\mathbf{A}$ that arise from reducing the number of elements in the mesh. In essence, this chapter may be broadly split into the study of natural probability distributions over the elements, the choice of the most appropriate probability distribution, and then the construction and analysis of a randomized algorithm to compute an approximation to this distribution quickly and with high probability. Specifically, we first provide an overview of the sampling methodology employed, and invoke some technical tools that will aid us in the process of designing and comparing probability distributions over the elements. We then design a probability distribution and provide a comparison to the benchmark and the state of the art. Finally, we design and analyze randomized algorithms to compute a relative error approximation to the probability distribution obtained. Currently, the best known manner of deterministically computing the best probability distribution we are currently aware of, takes $O(n^3N)$ operations, where N is the number of elements. We begin our presentation of ideas that result in the computation of a relative error approximation to these probabilities in $\tilde{O}(n^3 \log(rn))$ time.

6.2 Mathematical Players

We introduce the main players in this chapter here.

1. Let $\mathbf{A} \in \mathbb{R}^{n,n}$ be a finite element matrix (as defined in chapter 2.6), with the element-describing square root, $\mathbf{F} \in \mathbb{R}^{m,n}, m > n$, such that $\mathbf{A} = \mathbf{F}^\top \mathbf{F}$.

2. $\mathbf{F}$ has the QR factorization $\mathbf{UR}$, with $\mathbf{U}$ being an orthonormal basis to the range of $\mathbf{F}$.

3. Let $\tilde{A}_e$, $\bar{A}_e$ be the element matrix and the non-zero principal submatrix of the element matrix of the element e [see section 2.6 for more details].

4. Let $\mathcal{A}$ denote the set of all row indices of $\mathbf{F}$, and let $\mathcal{I}$, or any script letters, denote a subset of $\mathcal{A}$.

5. Let $\tilde{\mathbf{\Pi}}$ denote a randomized Johnson Lindenstrauss Transform (JLT). We do not make a distinction in symbols between the various kinds of JLTs. For instance, a Fast JLT (FJLT) will still be denoted by an appropriately subscripted $\tilde{\mathbf{\Pi}}$.

6.3 The Sampling Approach

We recall from chapter 2 that if $\mathbf{X} = \sum_{i=1}^{n} \mathbf{X}_i$, then $\mathbf{X}$ may be approximated by the independent, identically distributed random variable, $\mathbf{Y}$, where

$$\mathbf{Y} = \sum_{j=1}^{s} \mathbf{Y}_j, \quad \mathsf{Pr}\left(\mathbf{Y}_j = \frac{\mathbf{X}_i}{p_i}\right) = p_i \ \forall j, \ p_i > 0 \ \forall i.$$

Here, $\mathbf{Y}$ is $\mathbf{X}$ in expectation, but the likelihood that $\mathbf{Y}$ is close to $\mathbf{X}$ in any given realization is controlled by the choice of the probabilities, $p_i, 1 \leq i \leq n$. In general, we seek probabilities that minimize the variance of $\mathbf{Y}$.

We notice that the finite element matrix, $\mathbf{A}$, can be expressed as a sum of matrices associated with the elements in the mesh:

$$\mathbf{A} = \sum_{e} \tilde{\mathbf{A}}_e. \tag{6.1}$$

Following the intuition outlined above, we seek to associate with each element, e, a probability, p_e, such that the sum of a small number of draws of the random variable, $\mathbf{Y}_e = \mathbf{A}_e/p_e$, is 'close to' $\mathbf{A}$ with high probability. One may think of the probability, p_e, as a measure of how 'important' an element is to the sum in (6.1). The measure of closeness of $\mathbf{Y}$ to $\mathbf{A}$ determines this importance.

To design distributions that take the importance of an element to the mesh into account, we may look to generalize the results by Spielman et. al. and Miller et. al. [9, 7, 8] for sampling edges from a graph to provide a good preconditioner for the graph-laplacian, as in [6]. As shown in [33], this is equivalent to sampling edges according to the squared row-norms or *leverage scores* of the $|E| \times |V|$ left singular matrix of the edge-incidence matrix of the graph [see 2.2.2]. We recall that the effective resistance or leverage score of an edge was simply the squared row-norm of the appropriate row of an orthonormal basis for the edge incidence matrix. We consider two generalizations of the leverage score to row-blocks of $\mathbf{U}$, an orthonormal basis for $\mathbf{F}$: the first, where the leverage score, $l_1(\mathbf{U}_e)$, of the row-block $\mathbf{U}_e$ of $\mathbf{U}$ is given by $\left\|\mathbf{U}_e\mathbf{U}_e^\top\right\|_F$, and the second, where the leverage score, $l_2(\mathbf{U}_e)$, is given by $\left\|\mathbf{U}_e\mathbf{U}_e^\top\right\|$ as in [6].

The main tool for designing and analyzing the efficiency of our sampling methods will be the following corollary of the bernstein bound in lemma 13

Corollary 50. *For $\mathbf{Z}, \gamma, s^2$ and ϵ defined in Lemma 13, the number of copies, m, of $\mathbf{Z}$ needed such that*

$$\mathsf{Pr}\left(\left\|\bar{\mathbf{Z}}_m\right\|_2 \geq \epsilon\right) \leq \delta$$

is

$$\frac{4s^2}{\epsilon^2}\ln\left(\frac{2n}{\delta}\right) = O\left(\frac{s^2}{\epsilon^2}\log\left(\frac{n}{\delta}\right)\right).$$

The goal of designing an appropriate distribution, $P = \{p_i\}_{i=1}^N$, for sampling elements from the underlying finite element mesh of $\mathbf{A}$ will be to provide a randomized approximation, $\mathbf{Y}$, to $\mathbf{A}$, such that $\mathbf{Z} = \mathbf{A} - \mathbf{X}$ is a matrix random variable with the property that $\mathsf{E}[\mathbf{Z}] = \mathbf{0}, \|\mathbf{Z}\|_2 \leq \gamma$ and $\left\|\mathsf{E}\left[\mathbf{Z}^\top\mathbf{Z}\right]\right\|_2 \leq s^2$ with respect to P. The number of samples that will be required to provide a certain, fixed error, ϵ, is monotonically increasing with s^2, and this is used to compare and qualify sampling methods.

6.3.1 Sampling Using Effective Stiffness

6.3.1.1 Effective Stiffness

The chief goal of this discussion is to introduce to the reader a useful notion of importance of an element to the FE mesh, show that sampling according to this notion of importance leads to an economy of samples required to approximate $\mathbf{A}$ well, and show that relative error approximations to this measure of importance are nearly as good. By doing so, we summarize the work by Avron and Toledo in [6]. Intuitively, the chief observation by Avron and Toledo is that the significance, say $\bar{p}_e$, of an element - also called as *effective stiffness* or *leverage* of the element - can be measured by first letting the element 'absorb' all the connectivity it provides to the finite element mesh it is a part of, and measuring how 'close' this connectivity-concentrated element is to the original element, where 'closeness' is measured as their generalized top eigenvalue. Formally, let $\mathbf{A}$ be written such that $\mathbf{A}_e = \mathbf{A}(\mathcal{A}_e)$ appears as the last principal submatrix. Let us call this representation of $\mathbf{A}$ as $\mathbf{A}^{(e)}$. So

$$\mathbf{A}^{(e)} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_e \end{pmatrix}.$$

The *connectivity-concentrated* element corresponding to $\mathcal{A}_e$ is given by the schur complement, $\mathbf{K}_e$, of $\mathbf{A}_e$ in $\mathbf{A}$,

$$\mathbf{K}_e = \begin{cases} \mathbf{A}_e - \mathbf{A}_{21}{\mathbf{A}_{11}}^{-1}\mathbf{A}_{12}^T, & \text{if } \mathbf{A}_{11} \text{ is invertible,} \\ \bar{\mathbf{A}}_e, & \text{otherwise.} \end{cases}$$

Finally, the *relative leverage*, p_e, of an element (relative to the set of all the elements) is used as the probability with which e is sampled for addition to a randomized preconditioner for $\mathbf{A}$, denoted by $\mathbf{Y}$.

The following lemma (simplified from Theorem 6.1 in [6]) states that sampling an element according to its effective stiffness leads to an economy of samples, and that one does not lose too much of this economy if a relative error approximation to effective stiffness is used instead.

Lemma 51. *Let* $\bar{p}_e = \Lambda(\bar{\mathbf{A}}_e, \mathbf{K}_e)$*, and let* $\tilde{p}_e$ *be a relative error approximation to* $\bar{p}_e$*,*

$$|\bar{p}_e - \tilde{p}_e| \leq \delta \bar{p}_e.$$

Let probabilities p_e *be defined relative to the set* $\{\tilde{p}_e\}$*:*

$$p_e = \frac{\tilde{p_e}}{\sum_e \tilde{p}_e}.$$

If

$$\mathbf{Y} = \frac{\sum_1^m \mathbf{Y}_i}{m},$$

with $\mathbf{Y}_i = \mathbf{A}_e/p_e$ *with probability* p_e*, and*

$$m = m(\delta) = O(\tilde{n} \log \tilde{n}); \quad \tilde{n} = n\left(\frac{1+\delta}{1-\delta}\right),$$

then

$$\mathsf{Pr}\left(\kappa(\mathbf{A}, \mathbf{Y}) > 2\right) \leq \frac{1}{poly(m)}.$$

We note that if $\bar{p}_e$ are known exactly, then $\delta = 0$ and $m = O(n \log n)$. In other words, the only difference that a relative error approximation to $\bar{p}_e$ makes is a small increase in the number of elements picked to guarantee a good preconditioner.

6.3.1.2 Computation of Effective Stiffness

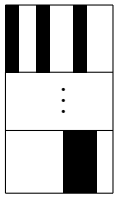
In order to compute p_e the following lemma, proved as Lemma 5.1 in Avron, Toledo, is of importance.

Lemma 52. *Let* $\mathbf{A} = \mathbf{F}^\top \mathbf{F}$ *and* $\mathbf{U}$*, an orthonormal basis for* Range($\mathbf{F}$)*. Let* $\mathbf{F}_e$ *be the set of rows corresponding to the element* $\mathcal{A}_e$*, and let* $\mathbf{U}_e$ *be the set of rows of* $\mathbf{U}$ *such that* $\mathbf{F}_e = \mathbf{U}_e\mathbf{R}$*. Then,*

$$\Lambda(\bar{\mathbf{A}}_e, \mathbf{K}_e) = \Lambda(\mathbf{U}_e \mathbf{U}_e^\top).$$

While the proof of this lemma is available in Avron and Toledo, it is presented here for the sake of completeness.

Proof. We begin by arguing with respect to a certain generic element, e. For convenience, let $\mathbf{F}^{(e)}$ denote $\mathbf{F}$ with its columns corresponding to indices $\mathcal{A}_e$ rearranged to the end, resulting in

$F^{(e)} =$ 

We obtain an orthonormal basis by QR factorization: $\mathbf{F}^{(e)} = \mathbf{UR}$, implying that $\mathbf{F}_e^{(e)} = \mathbf{U}_e\mathbf{R}_e$. As a result,

$$\mathbf{A}^{(e)} = {\mathbf{F}^{(e)}}^\top \mathbf{F}^{(e)} = \mathbf{R}^\top \mathbf{R},$$

and since $\mathbf{R}$ is a cholesky factor of $\mathbf{A}^{(e)}$,

$$\mathbf{K}_e = {\mathbf{R}_e}^\top \mathbf{R}_e.$$

Since ${\mathbf{F}_e}^\top \mathbf{F}_e = \bar{\mathbf{A}}_e$, we have

$$\begin{aligned}
\Lambda(\bar{\mathbf{A}}_e, \mathbf{K}_e) &= \Lambda({\mathbf{F}_e}^\top \mathbf{F}_e, {\mathbf{R}_e}^\top \mathbf{R}_e) \\
&= \Sigma(\left({\mathbf{R}_e}^\top \mathbf{R}_e\right)^\dagger {\mathbf{F}_e}^\top \mathbf{F}_e) \\
&= \Sigma(\mathbf{F}_e {\mathbf{R}_e}^\dagger {\mathbf{R}_e}^{\top\dagger} {\mathbf{F}_e}^\top) \quad \text{due to similarity} \\
&= \Sigma^2({\mathbf{R}_e}^{\top\dagger} {\mathbf{F}_e}^\top) \\
&= \Sigma^2({\mathbf{R}_e}^{\top\dagger} {\mathbf{R}_e}^\top \mathbf{U}_e^\top) \\
&= \Sigma^2({\mathbf{U}_e}^\top) = \Lambda(\mathbf{U}_e {\mathbf{U}_e}^\top).
\end{aligned}$$

This closes the proof. □

Additionally, the approximation

$$\mathbf{Y} = \sum_i \frac{X_i}{m}$$

can be written using matrix operations as $(\mathbf{SF})^\top(\mathbf{SF})$, where $\mathbf{S}$ is the random,

sampling matrix containing scaling coefficients that correspond to sampling $\mathbf{U}_e$ - effectively sampling $\mathbf{F}_e$ - with probability p_e and scaling with p_e^{-1}. Through a reasoning almost identical to the one provided above, it can be shown (see Lemma 5.1 in [6]) that

$$\kappa(\mathbf{A}, \mathbf{P}) = \Lambda(\mathbf{P}^\dagger \mathbf{A}) = \Lambda((\mathbf{SU})^\top (\mathbf{SU})).$$

As a result, in order to bound $\kappa(\mathbf{A}, \mathbf{P})$, it suffices to show that

$$\left\| (\mathbf{SU})^\top (\mathbf{SU}) - \mathbf{I} \right\| \geq \epsilon$$

with an appropriately small probability, δ, whence

$$\Lambda((\mathbf{SU})^\top (\mathbf{SU})) \leq \frac{1+\epsilon}{1-\epsilon}$$

with probability $1 - \delta$.

6.4 Comparison of Sampling Techniques

In order to qualify the design and analysis of the algorithms presented here, we compare three sets of probabilities that may be used to sample elements from a given finite element mesh. Specifically, we present a comparison of sampling using *effective stiffness*, with probabilities, p_e,

$$p_e = \frac{\left\| \mathbf{U}_e \mathbf{U}_e{}^\top \right\|}{\sum_t \left\| \mathbf{U}_t \mathbf{U}_t{}^\top \right\|},$$

to the alternative generalization of leverage scores with probabilities ϕ_e,

$$\phi_e = \frac{\left\| \mathbf{U}_e \mathbf{U}_e{}^\top \right\|_F}{\sum_t \left\| \mathbf{U}_t \mathbf{U}_t{}^\top \right\|_F},$$

as well as sampling uniformly, with probabilities ν_e,

$$\nu_e = 1/N,$$

where N is the number of elements that make up $\mathbf{A}$. For the sake of brevity, we choose to call these probabilies as *spectral probabilities, frobenius probabilities* and *uniform probabilities*, respectively, and sampling using these probabilities as *spectral sampling, frobenius sampling* and *uniform sampling.*

The sampling procedures using the three distributions provided above differ only in the probabilities used. In all cases, the random variable, $\mathbf{Y}_i$ is defined as

$$\mathsf{Pr}\left(\mathbf{Y}_i = \frac{1}{w_e}\mathbf{U}_e\mathbf{U}_e^\top\right) = w_e, \quad w_e \in \{p_e, \phi_e, \nu_e\}.$$

We look to approximate $\mathbf{A}$ using $\mathbf{Y}$ as

$$\mathbf{Y} = \sum_{i=1}^{k} \mathbf{Y}_i,$$

with the residual

$$\mathbf{Z} = \mathbf{A} - \mathbf{Y},$$

which we wish to bound in norm with high probability. Our yardstick for comparison will be the number of samples required to approximate $\mathbf{A}$ with error less than ϵ with probability at least $1-\delta$. We will use corollary 50 to indicate the number of samples required in each case, for which three steps are necessary:

1. Bounding $\|\mathbf{Z}\|$ with γ,
2. Bounding $\left\|\mathsf{E}\left[\mathbf{Z}^\top\mathbf{Z}\right]\right\|$ with s^2,
3. Verifying that $\gamma \leq 3s^2/\epsilon$.

We proceed to do this for each of the sampling probabilities.

6.4.1 Frobenius Sampling

Bounding the Random Variable in Norm: We first observe that

$$\|\mathbf{Z}\| \leq \max_e \left\|\mathbf{U}_e{\mathbf{U}_e}^\top\right\| \frac{1}{\phi_e} = \max_e \left\|\mathbf{U}_e{\mathbf{U}_e}^\top\right\| \frac{\sum_f \left\|\mathbf{U}_f{\mathbf{U}_f}^\top\right\|_F}{\left\|\mathbf{U}_e{\mathbf{U}_e}^\top\right\|_F}.$$

Since $\frac{\left\|\mathbf{U}_e\mathbf{U}_e^\top\right\|_F}{\|\mathbf{U}_e\mathbf{U}_e^\top\|} \geq 1$, we have

$$\|\mathbf{Z}\| \leq \sum_f \left\|\mathbf{U}_f\mathbf{U}_f^\top\right\|_F \leq \sum_f \|\mathbf{U}_f\|_F^2 \leq \|\mathbf{U}\|_F^2 .$$

Since $\mathbf{U}$ is an orthonormal matrix, it follows that $\|\mathbf{U}\|_F^2 = \operatorname{rank}\mathbf{U} \leq n-d$, implying that

$$\|\mathbf{Z}\| = \gamma \leq n - d.$$

The inequality shown above is theoretically tight, in the sense that there exist matrices such that $\left\|\mathbf{U}_f\mathbf{U}_f^\top\right\|_F = \left\|\mathbf{U}_f\mathbf{U}_f^\top\right\|$, saturating the bound.

Bounding the Second Moment in Norm: We bound $\left\|\mathsf{E}\left[\mathbf{Z}^\top\mathbf{Z}\right]\right\|$ as follows. First, we note that

$$\left\|\mathsf{E}\left[\mathbf{Z}^\top\mathbf{Z}\right]\right\| = \left\|\sum_e \phi_e \frac{\mathbf{U}_e\mathbf{U}_e^\top\mathbf{U}_e\mathbf{U}_e^\top}{\phi_e^2}\right\| = \left\|\sum_e \frac{\mathbf{U}_e\mathbf{U}_e^\top\mathbf{U}_e\mathbf{U}_e^\top}{\phi_e}\right\|. \tag{6.2}$$

Now, using the definition of ϕ_e and positive homogeneity of a norm, we have

$$\begin{aligned}\left\|\sum_e \frac{\mathbf{U}_e\mathbf{U}_e^\top\mathbf{U}_e\mathbf{U}_e^\top}{\phi_e}\right\| &= \left\|\sum_e \frac{\mathbf{U}_e\mathbf{U}_e^\top\mathbf{U}_e\mathbf{U}_e^\top}{\|\mathbf{U}_e\mathbf{U}_e^\top\|_F}\left(\sum_e \left\|\mathbf{U}_e\mathbf{U}_e^\top\right\|_F\right)\right\| \\ &= \left\|\sum_e \frac{\mathbf{U}_e\mathbf{U}_e^\top\mathbf{U}_e\mathbf{U}_e^\top}{\|\mathbf{U}_e\mathbf{U}_e^\top\|_F}\right\|\left(\sum_e \left\|\mathbf{U}_e\mathbf{U}_e^\top\right\|_F\right). \end{aligned} \tag{6.3}$$

We observe that

$$\sum_e \left\|\mathbf{U}_e\mathbf{U}_e^\top\right\|_F \leq \sum_e \|\mathbf{U}_e\|_F^2 = \|\mathbf{U}\|_F^2 = n - d,, \tag{6.4}$$

$$\frac{\mathbf{U}_e\mathbf{U}_e^\top}{\|\mathbf{U}_e\mathbf{U}_e^\top\|_F} \preceq \frac{\mathbf{U}_e\mathbf{U}_e^\top}{\|\mathbf{U}_e\mathbf{U}_e^\top\|} \preceq \mathbf{I} \Rightarrow \frac{\mathbf{U}_e\mathbf{U}_e^\top\mathbf{U}_e\mathbf{U}_e^\top}{\|\mathbf{U}_e\mathbf{U}_e^\top\|_F} \preceq \mathbf{U}_e\mathbf{U}_e^\top. \tag{6.5}$$

Now, since

$$\sum_e \mathbf{U}_e\mathbf{U}_e^\top = \mathbf{U}^\top\mathbf{U} = \mathbf{I},$$

we have from (6.2), (6.3), (6.4) and (6.5) that

$$\left\|\mathsf{E}\left[\mathbf{Z}^\top\mathbf{Z}\right]\right\| = s^2 \leq n - d.$$

Since $\gamma \leq 3s^2/\epsilon$, we have from corollary 50 that the number of sampled elements to provide a $1-\delta$ probable guarantee on a good preconditioner to $\mathbf{A}$ is

$$m_\phi = O\left(\frac{(n-d)}{\epsilon^2}\log\frac{2n}{\delta}\right). \tag{6.6}$$

We adopt a line of reasoning identical to the one above, to provide bounds for the two following cases.

6.4.2 Uniform Sampling

Bounding the Random Variable in Norm: Here,

$$\|\mathbf{Z}\| \leq \max_e \frac{\left\|\mathbf{U}_e\mathbf{U}_e^\top\right\|}{1/N} \leq N.$$

Since, theoretically, there are matrices such that $\left\|\mathbf{U}_e\mathbf{U}_e^\top\right\| = 1$, the above bound is tight as well.

Bounding the Second Moment: We first note that

$$\left\|\mathbf{U}_e\mathbf{U}_e^\top\right\| \leq \left\|\mathbf{U}\mathbf{U}^\top\right\| = 1$$

since, otherwise, we can find a unit vector $\hat{\mathbf{x}} \in \mathbb{R}^m$ with support in the indices corresponding to the element e, such that

$$\hat{\mathbf{x}}^\top\mathbf{U}\mathbf{U}^\top\hat{\mathbf{x}} \geq 1,$$

which contradicts the fact that $\left\|\mathbf{U}\mathbf{U}^\top\right\| = 1$.
Next, we observe that $\mathbf{U}_e\mathbf{U}_e^\top$ is SPSD by definition. So

$$0 \preceq \mathbf{U}_e\mathbf{U}_e^\top\mathbf{U}_e\mathbf{U}_e^\top \preceq \mathbf{U}_e\mathbf{U}_e^\top \forall e$$

and we have

$$\left\|\mathsf{E}\left[\mathbf{Z}^\top\mathbf{Z}\right]\right\| = \left\|\sum_e \frac{\mathbf{U}_e\mathbf{U}_e^\top\mathbf{U}_e\mathbf{U}_e^\top}{1/N}\right\| \leq N\left\|\sum_e \mathbf{U}_e\mathbf{U}_e^\top\right\| \leq N.$$

Clearly, the conditions of lemma 13 are satisfied, and according to corollary 50, guaranteeing a good preconditioner for $\mathbf{A}$ with probability $1-\delta$ requires a sample set of size

$$m_\nu = O\left(\frac{N}{\epsilon^2}\log\frac{2n}{\delta}\right). \tag{6.7}$$

6.4.3 Spectral Sampling

Bounding the Random Variable in Norm: In this case,

$$\|\mathbf{Z}\| \leq \max_e \left\|\mathbf{U}_e{\mathbf{U}_e}^\top\right\| \frac{1}{p_e} \leq \max_e \left\|\mathbf{U}_e{\mathbf{U}_e}^\top\right\| \frac{\sum_f \left\|\mathbf{U}_f{\mathbf{U}_f}^\top\right\|}{\left\|\mathbf{U}_e{\mathbf{U}_e}^\top\right\|} \leq \sum_f \left\|\mathbf{U}_f\mathbf{U}_f^\top\right\|. \tag{6.8}$$

Here, we note that

$$\sum_f \left\|\mathbf{U}_f\mathbf{U}_f^\top\right\| \leq \sum_f \left\|\mathbf{U}_f\mathbf{U}_f^\top\right\|_F \leq n-d, \tag{6.9}$$

and that

$$\sum_f \left\|\mathbf{U}_f\mathbf{U}_f^\top\right\| \leq \sum_f 1 = N. \tag{6.10}$$

Now, combining (6.8) with (6.9), (6.10), we have

$$\|\mathbf{Z}\| \leq \min\{N, n-d\}.$$

Bounding the Second Moment: We argue similar to the previous arguments in bounding second moments, noting that

$$\mathsf{E}\left[\mathbf{Z}^\top\mathbf{Z}\right] = \sum_e \frac{\mathbf{U}_e\mathbf{U}_e^\top\mathbf{U}_e\mathbf{U}_e^\top}{\|\mathbf{U}_e\mathbf{U}_e^\top\|}\left(\sum_e \left\|\mathbf{U}_e\mathbf{U}_e^\top\right\|\right).$$

Since

$$\sum_e \left\|\mathbf{U}_e\mathbf{U}_e^\top\right\| \leq \min\{N, n-d\}$$

and

$$\frac{\mathbf{U}_e\mathbf{U}_e^\top}{\|\mathbf{U}_e\mathbf{U}_e^\top\|} \preceq \mathbf{I},$$

we immediately have that

$$\begin{aligned}\left\|\mathsf{E}\left[\mathbf{Z}^\top\mathbf{Z}\right]\right\| &\leq \left\|\sum_e \mathbf{U}_e\mathbf{U}_e^\top\right\| \min\{N, n_d\} \\ &\leq \min\{N, n-d\}.\end{aligned}$$

This results in requiring

$$m_p = O\left(\frac{\min\{N,(n-d)\}}{\epsilon^2}\log\frac{2n}{\delta}\right) \tag{6.11}$$

scaled samples taken from the set of elements that form $\mathbf{A}$, to provide a good preconditioner for it with probability $1-\delta$. In this sense, the spectral sampling probabilities used in this paper are 'the right' generalization of leverage scores for the context. This is observed in practice as well. The relative efficiency of spectral sampling, as well as the usefulness of the Bernstein bound in analyzing it are depicted in Figure 6.1.

Having determined that $\left\|\mathbf{U}_e\mathbf{U}_e^\top\right\|$ is the appropriate quantity to design sampling probabilities with, we next proceed to design a fast randomized algorithm to compute it.

6.5 Computing Effective Stiffness

We begin by assuming that $\mathbf{F}$ may be computed from the existing description of elements quickly, while $\mathbf{U}$ requires the SVD of $\mathbf{F}$, an expensive operation that is $O(r\hat{n}^2)$, which we would like to avoid. However, since $\mathbf{U}$ is an orthonormal basis of the range of $\mathbf{F}$, we have

$$\mathbf{F}\mathbf{F}^\dagger = \mathbf{U}\mathbf{U}^\top,$$

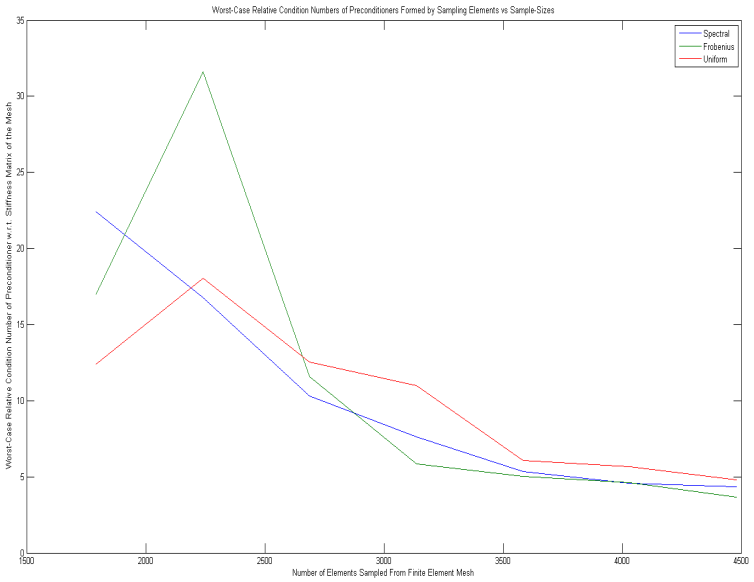


Figure 6.1: Worst-case condition numbers of sampled preconditioners relative to the preconditioned matrix, A as a function of the number of elements sampled according to the three sampling probabilities considered: frobenius, spectral and uniform.

$$\begin{aligned}\left\|\mathbf{U}_e\mathbf{U}_e^\top\right\| &= \max_{\|\mathbf{z}\|=1} \mathbf{z}^\top\mathbf{U}\mathbf{U}^\top\mathbf{z}, \quad \mathbf{z}\in\mathbb{R}^{\mathcal{A}_e},\\ &= \max_{\|\mathbf{z}\|=1} \mathbf{z}^\top\mathbf{F}\mathbf{F}^\dagger\mathbf{z},\end{aligned}$$

where $\mathbb{R}^{\mathcal{A}_e}$ is just the set of vectors in $\mathbb{R}^n$ with support in $\mathcal{A}_e$. Given a quick method of obtaining, $\mathbf{F}^\dagger$, therefore, we may effectively operate on $\mathbf{U}\mathbf{U}^\top$ by operating on $\mathbf{F}\mathbf{F}^\dagger$. This is precisely what we do.

We first get around the computation of $\mathbf{F}^\dagger$, which is once again as expensive as computing the SVD of $\mathbf{F}$, by a randomized pseudo-inversion: we sample rows from $\mathbf{F}$ using an ϵ-FJLT operator, $\tilde{\mathbf{\Pi}}_1$, and invert the resulting $\tilde{\mathbf{\Pi}}_1\mathbf{F}$. Next, we take one of two courses of action, depending upon the number of partitions, $\mathbf{U}_\mathcal{I}$, of $\mathbf{U}$ (equivalently, the number of blocks in $\mathbf{F}$):

1. When the number of partitions is small and the number of rows in each partition is large: This is equivalent to a small number of large elements forming the mesh. In this case we obtain a relative error approximation to $\left\|\mathbf{U}_e\mathbf{U}_e^\top\right\|$ by

carrying out power iterations on $\mathbf{F}\tilde{\mathbf{\Pi}}_1\mathbf{F}^\dagger$ restricting the support of the vector at each iterate to $\mathcal{A}_e$ (see [34]).

2. When the number of partitions of $\mathbf{F}$ is large, with a small number of rows in every partition: This corresponds to the typical case when the finite element mesh is made of numerous small elements. We note that

$$\|\mathbf{U}_e\mathbf{U}_e\|_F \leq |\mathcal{A}_e| \, \|\mathbf{U}_e\mathbf{U}_e\|,$$

and that, for a small $|\mathcal{A}_e|$, as is the case here, $\left\|\mathbf{U}_e\mathbf{U}_e^\top\right\|_F$ forms a good relative error approximation to $\left\|\mathbf{U}_e\mathbf{U}_e^\top\right\|$. Since relative error approximation is transitive, we compute relative error approximations to $\left\|\mathbf{U}_e\mathbf{U}_e^\top\right\|_F$ as follows. We first project $\mathbf{F}\tilde{\mathbf{\Pi}}_1\mathbf{F}^\dagger$ to a lower dimension using a second JLT, $\tilde{\mathbf{\Pi}}_2$, and use the squared frobenius norms of the resulting low-column-dimensional approximation to $\mathbf{U}_\mathcal{I}$, normalized to sum to unity, as sampling probabilities.

The analysis of both these procedures is presented in detail below along with the resulting running times.

6.5.1 Randomized Pseudo-Inversion Using FJLT

We first show that $\tilde{\mathbf{\Pi}}$, an ϵ-FJLT row-operator, preserves the spectra of all principal submatrices of $\mathbf{F}\mathbf{F}^\dagger$, using the following lemma, proven in [20, 35].

Lemma 53. *Let $\mathbf{F} \in \mathbb{R}^{r,n}$, with rank $\hat{n}$. Let $\mathbf{F} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be the SVD of $\mathbf{F}$, let $\tilde{\mathbf{\Pi}}$ be an ϵ-FJLT and let $\tilde{\mathbf{U}} = \tilde{\mathbf{\Pi}}\mathbf{U}$. Then*

$$\left\|\mathbf{I} - \left(\tilde{\mathbf{U}}^\top\tilde{\mathbf{U}}\right)^\dagger\right\| \leq \frac{\epsilon}{1-\epsilon} = O(\epsilon).$$

We say that $\left(\left(\tilde{\mathbf{U}}^\top\tilde{\mathbf{U}}\right)\right)^\dagger = \tilde{\mathbf{U}}^\dagger\left(\tilde{\mathbf{U}}\right)^{\dagger_\top} = \tilde{\mathbf{I}} \approx \mathbf{I}$ in this case.
We effectively approximate $\lambda_1\left(\mathbf{U}_\mathcal{I}\mathbf{U}_\mathcal{I}^\top\right)$ by $\lambda_1\left(\mathbf{U}_\mathcal{I}\tilde{\mathbf{I}}\mathbf{U}_\mathcal{I}^\top\right)$. However, since $\mathbf{U}$ is never computed, such an operation isn't explicitly performed. Instead, we operate upon $\mathbf{F}$ in a way that implies the above computation: we note that

$$\mathbf{F}\left(\tilde{\mathbf{\Pi}}\mathbf{F}\right)^\dagger = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}^\dagger\left(\tilde{\mathbf{\Pi}}\mathbf{U}\right)^\dagger$$

as a result of which

$$\begin{aligned}\mathbf{F}\Big(\tilde{\mathbf{\Pi}}\mathbf{F}\Big)^{\dagger}\Big(\tilde{\mathbf{\Pi}}\mathbf{F}\Big)^{\dagger^{\top}}\mathbf{F}^{\top} &= \mathbf{U}\Big(\tilde{\mathbf{\Pi}}\mathbf{U}\Big)^{\dagger}\Big(\tilde{\mathbf{\Pi}}\mathbf{U}\Big)^{\dagger^{\top}}\mathbf{U} \\ &= \mathbf{U}\tilde{\mathbf{I}}\mathbf{U}^{\top}.\end{aligned}$$

The preservation of $\lambda_1\left(\mathbf{U}_{\mathcal{I}}\mathbf{U}_{\mathcal{I}}^{\top}\right)$ is shown as follows. For expository concision, we use $\hat{\mathbf{z}}$ to denote a generic member of the set of all unit vectors in $\mathbb{R}^n$ with support in $\mathcal{I}$. For the same reason, we denote $\lambda_1\left(\mathbf{U}_{\mathcal{I}}\mathbf{U}_{\mathcal{I}}^{\top}\right)$ by $\lambda_{\mathcal{I}}$ and $\lambda_1\left(\mathbf{U}_{\mathcal{I}}\tilde{\mathbf{I}}\mathbf{U}_{\mathcal{I}}^{\top}\right)$ as $\tilde{\lambda}_{\mathcal{I}}$. Now,

$$\begin{aligned}\lambda_{\mathcal{I}} - \tilde{\lambda}_{\mathcal{I}} &= \max \hat{z}^{\top}\mathbf{U}\mathbf{U}^{\top}\hat{z} - \max \hat{z}^{\top}\mathbf{U}\tilde{I}\mathbf{U}^{\top}\hat{z} \\ &\leq \max \hat{z}^{\top}(\mathbf{U}\mathbf{U}^{\top} - \mathbf{U}\tilde{I}\mathbf{U}^{\top})\hat{z} \\ &= \max \hat{z}^{\top}\mathbf{U}(I - \tilde{I})\mathbf{U}^{\top}\hat{z} \\ &\leq \left\|I - \tilde{I}\right\| \max \hat{z}^{\top}\mathbf{U}\mathbf{U}^{\top}\hat{z} \\ &\leq \frac{\epsilon}{1-\epsilon}\lambda_{\mathcal{I}} \\ &= O(\epsilon)\lambda_{\mathcal{I}}.\end{aligned}$$

While this result, as stated, is sufficiently tight for the our purposes here, it is noted that the whole spectrum of $\mathbf{U}_{\mathcal{I}}\mathbf{U}_{\mathcal{I}}^{\top}$ is preserved up to $O(\epsilon)$ and relatively. In order to construct an algorithm to approximate the pseudo-inverse of $\mathbf{F}$, we simply design an SRHT in accordance with lemma 15. The resulting algorithm is presented as Algorithm 7. We note that the right singular vectors of $(\mathbf{\Pi}_2\mathbf{F})^{\dagger}$ aren't returned, as this makes no difference to our calculations.

Algorithm 7. getPseudoInverse(F)

Input: The element-incidence matrix $\mathbf{F} \in \mathbb{R}^{r,n}, r \geq n$
Output: Effectively an approximate pseudo-inverse, $\tilde{\mathbf{F}}^{\dagger} = \left(\tilde{\mathbf{\Pi}}\mathbf{F}\right)^{\dagger}$, of $\mathbf{F}$.

1. Compute $\tilde{\mathbf{\Pi}}$ as in lemma 15.
2. Compute $\tilde{\mathbf{\Pi}}\mathbf{F}$.

3. $\bar{\mathbf{U}}, \bar{\mathbf{\Sigma}}, \bar{\mathbf{V}} = \text{SVD}(\tilde{\mathbf{\Pi}}\mathbf{F})$.

4. Return $\bar{\mathbf{V}}\bar{\mathbf{\Sigma}}^{\dagger}$

Operation Count: To compute $\tilde{\mathbf{\Pi}}\mathbf{F}$ takes $O(rn \log r_1)$ operations, where

$$r_1 = \Theta\left(\frac{n\log(rn)}{\epsilon^2}\log\left(\frac{n\log(rn)}{\epsilon^2}\right)\right),$$

as per lemma 15 providing an operation count that is

$$\tilde{O}(rn\log\frac{n}{\epsilon^2}).$$

Next, the full pseudo-inversion of $\tilde{\mathbf{\Pi}}F$ is done in

$$O(r_1 n^2) = \tilde{O}\left(\frac{n^3}{\epsilon^2}\log(rn)\right).$$

The net operation-count of the process thus far is dominated by the latter step, resulting in

$$\tilde{O}\left(\frac{n^3}{\epsilon^2}\log(rn)\right)$$

operations.

6.5.2 Estimating $\tilde{\lambda}_{\mathcal{I}}$

We observe that we may have one of two cases at hand. Recall that N is the size of the partition of the rows of $\mathbf{F}$ (or number of elements / row-blocks), and suppose that μ_e is the lowest upper bound for the number of rows in each row-block. We may either have a large number of row-blocks, with a small number of rows in each block (N is large and μ_e is small) or the other way around, where N is small and μ_e is large. We provide separate algorithms to approximate $\tilde{\lambda}_{\mathcal{I}}$ in each of these cases. First we tackle the case when N is large and μ_e is small.

6.5.3 Sampling Using JLT Approximation

Thus far we have shown that the element-wise spectrum, $\Lambda\left(\mathbf{U}_{\mathcal{I}}\mathbf{U}_{\mathcal{I}}\right)$ was relatively preserved in $\Lambda\left(\tilde{\mathbf{U}}_{\mathcal{I}}\tilde{\mathbf{U}}_{\mathcal{I}}^{\top\dagger}\right)$. However, computing the top eigenvalue of $\tilde{\mathbf{U}}_{\mathcal{I}}\tilde{\mathbf{U}}_{\mathcal{I}}^{\top}$ takes $O(\mu_e^2 r_1)$, which we would like to avoid. While this beats computing the exact top eigenvalue of $\mathbf{U}_{\mathcal{I}}\mathbf{U}_{\mathcal{I}}$, we may do significantly better if we are willing to suffer a small, multiplicative cost. Specifically, we first compute an approximation, α_e, to $\tilde{\beta}_e$,

$$\tilde{\beta}_e = \frac{\lambda_1\left(\tilde{\mathbf{U}}_e\tilde{\mathbf{U}}_e^{\top}\right)}{\sum_e \lambda_1\left(\tilde{\mathbf{U}}_e\tilde{\mathbf{U}}_e^{\top}\right)}, \tag{6.12}$$

such that

$$|\alpha_e - \tilde{\beta}_e| \leq (\mu_e - 1)\tilde{\beta}_e, , \tag{6.13}$$

and then proceed to find a relative error approximation, $\tilde{\alpha}_e$ to α_e with high probability.

We begin by claiming that

Lemma 54. *Let μ_e be the upper bound on the number of rows of $\tilde{\mathbf{U}}_e$. Suppose further that μ_e is small: $n - d >> \mu_e$, but non-trivial: $\mu_e > 1 + \sqrt{2}$. If α_e are defined as*

$$\alpha_e = \frac{\text{trace}\left(\tilde{\mathbf{U}}_e\tilde{\mathbf{U}}_e^{\top}\right)}{\sum_e \text{trace}\left(\tilde{\mathbf{U}}_e\tilde{\mathbf{U}}_e^{\top}\right)}$$

and $\tilde{\beta}_e$is defined as in (6.12), then (6.13) is satisfied.

Proof. First note that

$$\text{trace}\left(\tilde{\mathbf{U}}_e\tilde{\mathbf{U}}_e^{\top}\right) = \left\|\tilde{\mathbf{U}}_e\right\|_F^2 \leq \mu_e \left\|\tilde{\mathbf{U}}_e\tilde{\mathbf{U}}_e^{\top}\right\|.$$

Since $\text{rank}(\tilde{\mathbf{U}}_e) \leq \mu_e < n - d$ by assumption that μ_e is an upper bound on the number of rows of $\tilde{\mathbf{U}}_e$, it follows that

$$\frac{\left\|\tilde{\mathbf{U}}_e\tilde{\mathbf{U}}_e^{\top}\right\|}{\mu_e\sum_e\left\|\tilde{\mathbf{U}}_e\tilde{\mathbf{U}}_e^{\top}\right\|} \leq \frac{\left\|\tilde{\mathbf{U}}_e\right\|_F^2}{\sum_e\left\|\tilde{\mathbf{U}}_e\right\|_F^2} \leq \frac{\mu_e\left\|\tilde{\mathbf{U}}_e\tilde{\mathbf{U}}_e^{\top}\right\|}{\sum_e\left\|\tilde{\mathbf{U}}_e\tilde{\mathbf{U}}_e^{\top}\right\|}.$$

Now, we observe that $\mu_e > 1 + \sqrt{2}$, whence

$$1 + \frac{1}{\mu_e} < m_e - 1,$$

and we have (6.13). □

We note that it is sufficient to approximate $\left\|\tilde{\mathbf{U}}_e\right\|_F^2$ up to relative error in order to approximate α_e up to relative error. We do the former with high probability, using the JLT.

Using the JLT: We choose a JLT operator $\tilde{\mathbf{\Pi}}_2$ that is a relative isometry for a fixed collection of n vectors in $\mathbb{R}^{r_1}$. Consequently, for the matrix $\tilde{\mathbf{U}} \in \mathbb{R}^{n,r_1}$ with the set of rows $\{\tilde{\mathbf{u}}_i\}_1^n$,

$$(1-\epsilon)\left\|\tilde{\mathbf{u}}_i\right\|^2 \leq \left\|\tilde{\mathbf{\Pi}}_2\tilde{\mathbf{u}}_i\right\|^2 \leq (1+\epsilon)\left\|\tilde{\mathbf{u}}_i\right\|^2,$$

whence

$$(1-\epsilon)\left\|\tilde{\mathbf{U}}_e\right\|_F^2 \leq \left\|\tilde{\mathbf{\Pi}}_2\tilde{\mathbf{U}}_e\right\|_F^2 \leq (1+\epsilon)\left\|\tilde{\mathbf{U}}_e\right\|_F^2, \tag{6.14}$$

where $\mathbf{U}_e$ is the row-block in $\mathbf{U}$ corresponding to element e. This immediately provides a way of approximating α_e:

Lemma 55. *Let $\tilde{\alpha}_e$ be defined as*

$$\tilde{\alpha}_e = \frac{\left\|\tilde{\mathbf{\Pi}}_2\tilde{\mathbf{U}}_e\right\|_F^2}{\sum_e \left\|\tilde{\mathbf{\Pi}}_2\tilde{\mathbf{U}}_e\right\|_F^2}.$$

Then for $\epsilon < 1/3$,

$$|\tilde{\alpha}_e - \tilde{\beta}_e| \leq (\mu_e)^2\,(1+3\epsilon)\,\tilde{\beta}_e$$

Proof. First, we prove that $|\tilde{\alpha}_e - \alpha_e| \leq (1+3\epsilon)\alpha_e$, from which the lemma follows by (6.13). Note that (6.14) and the definition of $\tilde{\alpha}_e, \alpha_e$ imply that

$$\frac{1-\epsilon}{1+\epsilon}\alpha_e \leq \tilde{\alpha}_e \leq \frac{1+\epsilon}{1-\epsilon}\alpha_e,$$

$$\Rightarrow \ |\tilde{\alpha}_e - \alpha_e| \leq 1 + \frac{2\epsilon}{1-\epsilon}\alpha_e,$$

whence, by $\epsilon < 1/3$, we have

$$|\tilde{\alpha}_e - \alpha_e| \leq (1+3\epsilon)\alpha_e$$

and hence the proof. □

Having shown that we may fruitfully use an ϵ-JLT, we now invoke lemma 14 for the quick, high-probability construction of this ϵ-JLT. As a result we have the algorithm 8.

Algorithm 8. getElementImportanceJLT(F)

Input: The element-incidence matrix $\mathbf{F} \in \mathbb{R}^{r,n}, r \geq n$
Output: Effectively an approximate pseudo-inverse, $\tilde{\mathbf{F}}^{\dagger} = \left(\tilde{\mathbf{\Pi}}\mathbf{F}\right)^{\dagger}$, of $\mathbf{F}$.

1. $\bar{\mathbf{V}}\bar{\mathbf{\Sigma}}^{\dagger} = \mathbf{getPseudoInverse}(\mathbf{F})$
2. Form $\tilde{\mathbf{\Pi}}_2 \in \mathbb{R}^{n,r_2}$ as in lemma 14
3. Compute $\bar{\mathbf{V}}\bar{\mathbf{\Sigma}}^{\dagger}\tilde{\mathbf{\Pi}}_2$
4. Compute α_e, the squared frobenius norms of blocks of $\mathbf{F}\bar{\mathbf{V}}\bar{\mathbf{\Sigma}}^{\dagger}\tilde{\mathbf{\Pi}}_2$ corresponding to element e

The runtime of this algorithm, note that step 1 takes $O(rn\log(r_1)+r_1n^2)$, step 2 and 3 together take $O(r_2n^2)$ followed by premultiplication with $\mathbf{F}$, which takes $O(rnr_2)$, followed by the computation of the trace of element-blocks, step 4, which takes an added $O(rr_2)$, resulting in a total operation count which is $\tilde{O}(rn\log r + n^3\log(rn))$.

Lemma 56. *Let μ_e be the upper bound on the number of rows of $\mathbf{F}_e$, and let*

$$1+\sqrt{2} < \mu_e < \sqrt{\frac{1-\epsilon}{4\epsilon}}, \ \ \mu_e << \sqrt{n-d}.$$

Then, under the hypothesis of lemma 51, with $\tilde{p}_e = \tilde{\alpha}_e$,

$$m = \Omega\left(n \log n\right)$$

samples suffice for

$$\mathsf{Pr}\left(\kappa(\mathbf{A}, \mathbf{Y}) > 2\right) \leq \frac{1}{poly(m)}.$$

Next we turn to the case where N, the number of row-blocks/elements is small and μ_e is large. It is immediate that one cannot approximate $\lambda_{\mathcal{I}}$ as $\left\|\mathbf{U}_{\mathcal{I}}\mathbf{U}_{\mathbf{I}}^{\top}\right\|_F$ since, by assumption, the factor, μ_e^2, by which the latter approximates the former is large, requiring a large number of samples according to lemma 51. We tackle this case by performing an operation more expensive than approximating the frobenius norms of row-blocks, as done above, but still much cheaper than an exhaustive computation.

Using Power Iteration: In the case where μ_e is large, it is no longer prudent to approximate $\tilde{\beta}_e$ using $\tilde{\alpha}_e$. In this case, we employ a time-tested tool to compute the top eigenvalue. We estimate $\tilde{\lambda}_{\mathcal{I}}$ by power iteration on a random Gaussian vector with support restricted to $\mathcal{I}$. Given a matrix, $\mathbf{B} \in \mathbb{R}^r$, the *power iteration* is an iterative algorithm cheifly used to approximate the eigenvalue of $\mathbf{B}$ having the largest modulus; for $\mathbf{M} \succeq \mathbf{0}$, this is the largest eigenvalue/singular value, $\lambda_1(\mathbf{M})$. In general, this method produces a sequence, $\{\rho_i\}$, of successively better approximates to $\lambda_1(\mathbf{B})$, starting by operating on a unit vector, $\mathbf{y}_0$, where

$$\mathbf{y}_i = \frac{\mathbf{B}\mathbf{y}_{i-1}}{||\mathbf{B}y_{i-1}||_2}; \quad \rho_i = \mathbf{y}_{i-1}^{\top}\mathbf{B}\mathbf{y}_{i-1},$$

called the *rayleigh quotient* at step i. For details, see [36, 37].

Given a $\mathbf{B} \in \mathbb{R}^{r,n}$, it can be shown that the rayleigh quotient, ρ_k, obtained at step k of the power iteration provides a constant factor approximation for the top eigenvalue of an $r_e \times r_e$ submatrix of $\mathbf{B}\mathbf{B}^{\top}$ if $k = \Omega(\log(r_e/\delta^3))$. Formally,

Lemma 57. *Let* $\mathbf{B} \in \mathbb{R}^{r,n}$*, and* $\mathbf{x}$*, a vector chosen isotropically with non-zero components only in* $\mathcal{I}_e$*, and iid in* $\mathcal{N}(0,1)$*, with* $r_e = |\mathcal{I}_e|$*. Let* $(\mathbf{B}\mathbf{B}^{\top})_e$ *be the* $\mathcal{I}_e \times \mathcal{I}_e$ *principal submatrix of* $\mathbf{B}\mathbf{B}^{\top}$*. Let* ρ_k *be the rayleigh quotient after step* k *of the power*

iteration method initiated with $\mathbf{x}$*. Then, for a constant,*

$$c \geq \left(\frac{2}{\pi} + 2\right)^3,$$

if $k \geq c \log \frac{r_e}{\delta^3}$*, then with probability at least* $1 - \delta$*,*

$$\rho_k \geq \frac{\lambda_1((\mathbf{B}\mathbf{B}^\top)_e)}{\sqrt{5}}.$$

The proof of this lemma is nearly identical to that in [34], with the only difference being that we restrict the support of the vectors being operated upon, to $\mathcal{I}_e$.

Since we are interested in approximating $\lambda_1(\mathbf{U}_e\mathbf{U}_e^\top)$, we effectively carry out these power iterations on the $\mathcal{I}_e \times \mathcal{I}_e$ principal submatrix of $\mathbf{F}\tilde{\mathbf{\Pi}}\mathbf{F}^\dagger\tilde{\mathbf{\Pi}}\mathbf{F}^{\dagger\top}\mathbf{F}^\top$. Operationally, this is done by restricting the vectors used in the power iteration on $\mathbf{F}\tilde{\mathbf{\Pi}}\mathbf{F}^\dagger\tilde{\mathbf{\Pi}}\mathbf{F}^{\dagger\top}\mathbf{F}^\top$ to have support in $\mathcal{I}_e$. The general form of power iteration used here is presented in in Algorithm 9, below.

Algorithm 9. POWERITERATE$(\mathbf{C}, \mathcal{I}_e)$

Input: An SPSD $\mathbf{C} \in \mathbb{R}^{r,n}$ and a subset of its row-indices $\mathcal{I}_e$.

Output: A constant factor approximation, ρ_k, to the top eigenvalue of the $|\mathcal{I}_e| \times |\mathcal{I}_e|$ submatrix of $\mathbf{C}\mathbf{C}^\top$.

1. $k \leftarrow |\mathcal{I}_e|$.
2. Pick vector $\mathbf{x} \in \mathbb{R}^r$ such that $\mathbf{x}(i) \in \mathcal{N}(0, 1)$ if $i \in \mathcal{I}_e$ and 0 otherwise.
3. $\mathbf{x} \leftarrow \mathbf{x}/||\mathbf{x}||$.
4. For $j = 1 \rightarrow \log(n/\delta^3)$
   (a) $\mathbf{w}^\top \leftarrow (\mathbf{x}^\top\mathbf{C})\mathbf{C}^\top$,
   (b) $\rho_j \leftarrow \sqrt{\mathbf{w}^\top\mathbf{x}}$.
   (c) $\mathbf{w}(i) \leftarrow 0 \forall i \notin \mathcal{I}_e$.

(d) $\mathbf{x} \leftarrow \mathbf{w}/r_j$.

5. Return ρ_j.

Remark: Algorithm 9 seems to require the matrix $\mathbf{C} = \mathbf{F}(\mathbf{\Pi F})^\dagger$, which is very inefficient to compute. However, this is only for expository convenience. In actuality, the matrix vector multiplications are preferred to any matrix matrix multiplications.

Operation Count: Let the number of rows in element $\mathbf{F}_e$ be r_e, with the average number of rows of $\mathbf{F}_e$ across all e being $\bar{r}_e$. Since $\mathbf{C} = \mathbf{F}_e\tilde{\mathbf{\Pi}}\mathbf{F}^\dagger \in \mathbb{R}^{r_e, r_1}$ here, it follows that $\Theta(\log(r_e/\delta^3))$ iterations are required. We note that the dominating step in powerIterate is step 5, requiring two multiplications of a vector with $\mathbf{C}$, or equivalently, two matrix-vector multiplications with each of $\mathbf{F}_e, \tilde{\mathbf{\Pi}}\mathbf{F}$. This requires $O((r_e + r_1)n)$ operations per iteration, providing the operation count

$$O((r_e + r_1)n \log(r_e/\delta^3))$$

for the computation of $(1-\delta)$-probable approximation to the spectral norm of the principal submatrix of $\mathbf{F}\tilde{\mathbf{\Pi}}\mathbf{F}^\dagger\tilde{\mathbf{\Pi}}\mathbf{F}^{\dagger\top}\mathbf{F}^\top$, corresponding to a single element. Since powerIterate will have to be carried out once for each element, we obtain a net operation count of

$$\sum_e O((r_e + r_1)n \log(r_e/\delta^3)) = O(rn \log(r/\delta^3) + nrr_1),$$

where we have used the observation that

$$\sum_e ((r_e + r_1)n \log(r_e/\delta^3))$$

$$\leq rn \log(r/\delta^3) + nrr_1 \log(1/\delta^3),$$

the detailed derivation of which is presented in the appendix.

6.5.4 Algorithm getEffectiveStiffness

To avoid the computation of $\mathbf{F}^\dagger$, which takes $O(rn^2)$ operations, we have performed a randomized-pseudoinversion of $\mathbf{F}$, which is shown to preserve the spectrum of the principal submatrices of $\mathbf{F}\mathbf{F}^\dagger$ up to relative error. Next, we have shown that a randomized-power iteration on an appropriately chosen initial vector provides a constant factor approximation to what is already a relative-error approximation to $\lambda_{\mathbf{I}}$. We now observe that if $a \approx b$ is defined as

$$(1 - O(\epsilon_1))a \leq b \leq (1 + O(e_2))a,$$

then $\approx$ is an equivalence relation. Since the randomized pseudo-inversion approximates $\lambda_{\mathcal{I}}$ in this sense of relative-error, and the randomized power iteration approximates this approximation in the sense of relative error, the transitivity of $\approx$ implies that we may chain these methods to obtain a relative-error algorithm for $\bar{p}_e = \lambda_{\mathcal{I}_e}$. We present this algorithm, getEffectiveStiffness, below.

Algorithm 10. getEffectiveStiffness($\mathbf{F}$)

Input: The element-incidence matrix $\mathbf{F} \in \mathbb{R}^{r,n}, r \geq n$ and indices $\mathcal{I}_e$ for each element, e. **Output:** A sampling probability, p_e, for each element.

1. $\mathbf{K} \leftarrow$ getPseudoInverse($\mathbf{F}$)
2. For each element, e,
   (a) $\tilde{p}_e \leftarrow$ powerIterate($\mathbf{F}\mathbf{K}, \mathcal{I}_e$)
3. Return $\{\tilde{p}_e\}$

We capture the study of this implementation of power iteration in the following lemma.

Lemma 58. *A relative error approximation,* $\tilde{p}_e = \tilde{\lambda}_{\mathcal{I}_e} \approx p_e = \lambda_{\mathcal{I}_e}$*, such that*

$$\frac{9}{10\sqrt{5}} \leq \tilde{p}_e \leq \frac{11}{10} p_e$$

can be computed in $\tilde{O}(n^3 \log(rn))$*.*

Proof. From sections 3.1.1 and 3.2.1, we have that the operation-count of is

$$\tilde{O}\left(\frac{n^3}{\epsilon^2}\log rn)\right) + O((r+r_1)n\log(r/\delta^3))$$

$$= \tilde{O}\left(\frac{n^3}{\epsilon^2}\log(rn)\right).$$

By setting $\epsilon = 1/11$, we have the result. □

6.6 Conclusion

The study presented above provides a quicker, albeit randomized, algorithm to obtain the importances of elements in a finite element mesh, to the solution of the corresponding discretized PDE. The promised running time of $O(n^3 \log(rn))$ is a considerable improvement over the otherwise current running time of $O(n^3 N)$, and it establishes the use of generalized leverage scores as a promising way of sampling from finite element meshes. That said, the author directs the reader to note that this work is eventually inspired by quickly obtaining a solution to the original system of equations $\mathbf{Ax} = \mathbf{b}$ arising from a finite element model of a PDE. Even the naive manner of obtaining a solution to this system takes merely $O(n^3)$ operations, a factor of $O(\log(rn))$ lesser than the algorithm to merely compute the importance of the elements in the mesh. It is clear, therefore, that more work is needed to attain the eventual goal of quickly solving the original linear matrix-equation. However, the author notes that incremental results in this field are sought after, and that this incremental work establishes the quickest known algorithm for a sub-routine in the overall process of solving $\mathbf{Ax} = \mathbf{b}$ quickly, that is also of independent interest.

6.7 Acknowledgments

The author thanks Haim Avron for providing the meshes on which the reported numerical experiments were performed, and Petros Drineas for numerous helpful suggestions.

7. FUTURE DIRECTIONS

The work presented in this thesis has consistently generated a lot of potential for furtherance. From amidst this potential, the author considers the following tasks to be the most immediately achievable.

7.1 Subspace Restricted Low Rank Approximation

The chief result in this chapter was that, given a matrix, $\mathbf{C} \in \mathbb{R}^{m,n}$, and a subspace $\mathcal{B} \subset \mathbb{R}^m$ of dimension r, we may find $\mathbf{C}_{\mathcal{B},k}$, the best k rank spectral approximation to $\mathbf{C}$ with columns in $\mathcal{B}$, that is equivalent to $\mathbf{C}_k$. This arises from the fact that there is no unique 'best' rank k approximation to $\mathbf{C}$, which immediately raises the following question

> Are there other solutions to $\mathbf{C}_{\mathcal{B},k}$ within $\mathcal{B}$? If so, are any of them computationally less expensive to obtain?

We already know some useful features of this equivalence class of the best rank k approximations to $\mathbf{C}$ from within $\mathcal{B}$, namely that

1. There is always a $\mathbf{C}_{\mathcal{B},k}$ that is an orthogonal projection of $\mathbf{C}$ onto some subspace of $\mathcal{B}$.

2. There is an orthogonal decomposition of $\mathcal{B} = \mathcal{B}_1 \oplus \mathcal{B}_2 \oplus \cdots \oplus \mathcal{B}_{k^*}$, such that orthogonally projecting $\mathbf{C}$ on to $\sum_{i=1}^{k} \mathcal{B}_i$ gives $\mathbf{C}_{\mathcal{B},k}$.

It remains to characterize these subspaces in a way such that we can compute them quickly. This seems to be the most pressing question the study of this problem generates, apart from the resolution of conjecture 35.

7.2 Exact Low Rank Approximations

The immediate goal in this chapter is to move beyond the existence of exact approximations to constructing them quickly. This is clearly a goal shared by the

study of subspace restricted low rank approximations as well. However, the study here poses the new challenge, namely

> If and under what conditions is $(\mathbf{C}_{\mathcal{B}})_k$ a good approximation to $\mathbf{C}_{\mathcal{B},k}$?

We state conjecture 48 to answer this question, but we note that the resolution of conjecture 48 only does so partially, in that it provides *sufficient* conditions for when $(\mathbf{C}_{\mathcal{B}})_k$ is nearly as good as $\mathbf{C}_{\mathcal{B},k}$ and not *necessary* ones, in the absence of which it is clear that we may be overlooking a good guarantee that is immediately available. We close with the posing of one last question our study encourages us to ask: Is there an algorithm that is a small perturbation of algorithm 6 in the sense of adding an easily computed correction to $(\mathbf{C}_{\mathcal{B}})_k$ that greatly improves the computed error and/or its analytical guarantee?

7.3 Sampling from Finite Element Matrices

We sample $O(n \log n)$ elements from a finite element mesh, but in order to do so, we spend $\tilde{O}(n^3 \log(rn))$ operations. While the subsampling of elements from a finite element mesh is interesting and useful in its own right, the runtime achieved in this document does not aid in solving a linear system in a finite element matrix. Whether techniques used in obtaining quick preconditioners to Laplacian matrices, such as low-stretch spanning trees, can be generalized to obtain preconditioners to a finite element mesh is the most immediate question that is posed by our study. The larger question, in essence, is

> Is there an algorithmic routine that runs in $O(n^3 \operatorname{plog}(n))$ that computes the importance of elements in the finite element mesh?

The author closes this thesis with the hope of being able to resolve one of these nagging curiosities.

LITERATURE CITED

[1] C. Eckart and G. Young, *The approximation of one matrix by another of lower rank*, Psychometrika, 1 (1936), pp. 211–218.

[2] I. Markovsky, *Low Rank Approximation - Algorithms, Implementation, Applications*, Springer, London, 2011.

[3] V. Rokhlin, A. Szlam, and M. Tygert, *A randomized algorithm for principal component analysis*, SIAM J. Matrix Anal. and Appl., 31 (2009), pp. 1100–1124.

[4] E. Liberty, F. Woolfe, P.G. Martinsson, V. Rokhlin, and M. Tygert, *Randomized algorithms for the low-rank approximation of matrices*, Proc. Natl. Acad. Sci. USA, 104 (2007), pp. 20167–20172.

[5] N Halko, P.G. Martinsson, and J.A. Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Rev., 53 (2011), pp. 217–288.

[6] H. Avron and S. Toledo, *Effective stiffness: Generalizing effective resistance sampling to finite element matrices.* arXiv preprint arXiv:1110.4437, 2011.

[7] I. Koutis, G. L. Miller, and R. Peng, *Approaching optimality for solving SDD linear systems*, in Proceedings Of The Fifty-First Annual Symposium On Foundations Of Computer Science, FOCS, IEEE, October 2010, pp. 235 – 244.

[8] I. Koutis, G. L. Miller, and R. Peng, *A nearly (m log n) time solver for SDD linear systems*, in Proceedings Of The Fifty-Second Annual Symposium On Foundations Of Computer Science, FOCS, IEEE, 2011, pp. 590–598.

[9] D. A. Spielman and N. Srivastava, *Graph sparsification by effective resistances*, SIAM J. Comput., 40 (2011), pp. 1913 – 1926.

[10] H. Avron, D. Chen, G. Shklarski, and S. Toledo, *Combinatorial preconditioners for scalar elliptic finite-element problems*, SIAM J. Matrix Anal. Appl., 31 (2009), pp. 694 – 720.

[11] B. N. Parlett, *The Symmetric Eigenvalue Problem*, vol. 7 of Classics Appl. Math., Prentice-Hall, Englewood Cliffs, NJ, 1980.

[12] G. W. Stewart, *On the early history of the singular value decomposition*, SIAM Rev., 35 (1993), pp. 551–566.

[13] A. Turing, *Rounding-off errors in matrix processes*, Quart. J. Mech. Appl. Math., 1 (1948), pp. 287–308.

[14] K. C. Sou and A. Rantzer, *On a generalized matrix approximation problem in the spectral norm*, Linear Algebra Appl., 436 (2012), pp. 2331–2341.

[15] F. R. K. Chung, *Spectral Graph Theory*, no. 92 in Regional Conference Series in Mathematics, American Mathematical Society, Providence, RI, 1997.

[16] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*, Springer, Berlin, 2013.

[17] W. B. Johnson and J. Lindenstrauss, *Extensions of lipschitz mappings into a hilbert space*, Contemp. Math., 26 (1984), p. 1.

[18] B. Recht, *A simpler approach to matrix completion*, J. Mach. Learn. Res., 12 (2011), pp. 3413–3430.

[19] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, *Fast monte carlo algorithms for matrices I: Approximating matrix multiplication*, SIAM J. Comput., 36 (2006), pp. 132–157.

[20] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, *Sampling algorithms for l_2 regression and applications*, in Proceedings Of The Seventeenth Annual ACM-SIAM Symposium On Discrete Algorithms, Society for Industrial and Applied Mathematics, January 2006, pp. 1127–1136.

[21] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, *Relative-error CUR matrix decompositions*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 844–881.

[22] P. Drineas, M. Magdon-Ismail, M. Mahoney, and D. P. Woodruff, *Fast approximation of matrix coherence and statistical leverage*, J. Mach. Learn. Res., 13 (2012), pp. 3475–3506.

[23] D. Achlioptas, *Database-friendly random projections: Johnson-lindenstrauss with binary coins*, J. Comput. System Sci., 66 (2003), pp. 671–687.

[24] M. Gu and S.C. Eisenstat, *Efficient algorithms for computing a strong rank-revealing QR factorization*, SIAM J. Sci. Comput., 17 (1996), pp. 848–869.

[25] F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert, *A fast randomized algorithm for the approximation of matrices*, Appl. Comput. Harmon. Anal., 25 (2008), pp. 335–366.

[26] T. Sarlos, *Improved approximation algorithms for large matrices via random projections*, in Proceedings Of The 47th Annual IEEE Symposium On Foundations Of Computer Science, FOCS'06, October 2006, pp. 143–152.

[27] C. Boutsidis, P. Drineas, and M. Magdon-Ismail, *Near-optimal column-based matrix reconstruction*, SIAM J. Comput., 43 (2014), pp. 687–717.

[28] K. Gremban, *Combinatorial Preconditioners for Sparse, Symmetric, Diagonally Dominant Linear Systems*, PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 1996.

[29] P. Drineas and M. W. Mahoney, *Effective resistances, statistical leverage, and applications to linear equation solving.* arXiv preprint arXiv:1005.3097, 2010.

[30] S. Chatterjee and A.S. Hadi, *Influential observations, high leverage points, and outliers in linear regression*, Stat. Sci., (1986), pp. 379–393.

[31] J. A. Tropp, *User friendly tail bounds for sums of random matrices*, Found. Comput. Math., 12 (2012), pp. 389–434.

[32] K. Lange, *Optimization*, Springer, New York, 2013.

[33] P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff, *Fast approximation of matrix coherence and statistical leverage.* arXiv preprint arXiv:1109.3843, 2011.

[34] M. Magdon-Ismail, *Using a non-commutative bernstein bound to approximate some matrix algorithms in spectral norm.* arXiv preprint arXiv:1103.5453, 2011.

[35] P. Drineas, M.W. Mahoney, S. Muthukrishnan, and T. Sarlos, *Faster least squares approximation*, Numer. Math., 117 (2011), pp. 219–249.

[36] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 2012.

[37] L. N. Trefethen and D. Bau III, *Numerical Linear Algebra*, vol. 50, SIAM, Philadelphia, 1997.