Thèse n° 10 967

Randomized low-rank approximation and its applications

Présentée le 28 juin 2024

Faculté des sciences de base
Algorithmes numériques et calcul haute performance - Chaire CADMOS
Programme doctoral en mathématiques

pour l'obtention du grade de Docteur ès Sciences

par

Ulf David PERSSON

Acceptée sur proposition du jury

Prof. F. Eisenbrand, président du jury
Prof. D. Kressner, directeur de thèse
Prof. A. Townsend, rapporteur
Prof. E. Jarlebring, rapporteur
Prof. L. Grigori, rapporteuse

École
polytechnique
fédérale
de Lausanne

2024

To Magna, my grandmother.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor Daniel Kressner. I thank him for helping me to navigate in the academic world and for giving me the freedom work on the problems that I found interesting. I am sincerely grateful for all the opportunities that he gave me; they allowed me to shape my future.

I thank Alex Townsend, Elias Jarlebring and Laura Grigori for being part of my jury. I thank you for taking the time to read my thesis and for the interesting questions asked during my private defence. I would also like to thank Friedrich Eisenbrand for agreeing to act as president of the jury for my private defense.

I thank the members of the ANCHP group for all the lunches, discussions, jokes and, most importantly, your kindness. Thank you Alice, Axel, Christoph, Fabio, Gianluca, Haoze, Hysan, Margherita, Nian, and Peter. I have been very fortunate to have you as colleagues.

In spring 2023 I had the opportunity to go to NYU to work together with Chris Musco. I am tremendously grateful to Chris for agreeing to host me. I benefited a lot from his vast expertise and his encouragement. I thank Raphael and Tyler for making me feel welcomed and our dinners, nights out in New York City, and discussions about mathematics and life.

I thank my Swedish friends Filip, Joel, John, and Wilhelm for all the good times and the support through the bad times. I am incredibly lucky to have such good friends.

I thank my girlfriend Ekaterina for your patience, pancakes, mathematical knowledge, and for distracting me when I needed it. Your support has truly been invaluable to me.

I devote a special thanks to Axel and Paride. You have been amazing friends during my time as a PhD-student. Thank you for all the discussions, jokes, laughs, and beers.

Finally, I send the warmest thanks to my family for your patience, support, the opportunities that I have been given, and for being my support system.

I gratefully acknowledge the support by SNSF: this PhD was supported by the SNSF research project "Fast algorithms from low-rank updates", with grant number 200020_178806.

Lausanne, June 6, 2024 David Persson

Abstract

In this thesis we will present and analyze randomized algorithms for numerical linear algebra problems. An important theme in this thesis is randomized low-rank approximation. In particular, we will study randomized low-rank approximation of matrix functions, the use of randomized low-rank approximation for trace estimation, and randomized low-rank approximation of self-adjoint non-negative trace class operators.

Chapters 3 to 5 will be concerned with low-rank approximations of matrix functions. We will present two methods to compute low-rank approximations of matrix functions. In Chapter 4 we will analyze a method called funNyström, which uses a low-rank approximation to $\boldsymbol{A}$ to obtain a low-rank approximation to $f(\boldsymbol{A})$, where f is a non-negative operator monotone function. In particular, we will show that a near-optimal Nyström low-rank approximation can be used to construct a near-optimal funNyström low-rank approximation. In Chapter 5 we will consider a block-Krylov subspace method to compute randomized low-rank approximations of general matrix-functions. We will provide probabilistic error bounds for the method.

Chapters 6 to 8 will be concerned with trace estimation. In Chapter 7 we will present an adaptive version of the Hutch++ algorithm. This algorithm takes an error tolerance as input, and returns an estimate of the trace within the error tolerance with a controllable failure probability, while minimizing the number of matrix-vector products with the matrix. In Chapter 8 we present a single pass version of the Hutch++ algorithm. This algorithm uses the Nyström approximation instead of the randomized SVD in the low-rank approximation phase of Hutch++, and we prove that it satisfies a similar complexity guarantee as Hutch++.

Chapter 9 will be concerned with an infinite-dimensional generalization of the Nyström approximation to compute randomized low-rank approximations to self-adjoint non-negative trace class operators. We will provide an error bound for the finite-dimensional Nyström approximation when it is implemented with non-standard Gaussian random vectors. We then use these bounds to prove an error bound for an infinite-dimensional generalization of the Nyström approximation.

Key words: Low-rank approximation, randomized numerical linear algebra, matrix functions, trace estimation, Hilbert-Schmidt operators, trace class operators.

Zusammenfassung

In dieser Arbeit werden wir randomisierte Algorithmen für Probleme der numerischen linearen Algebra vorstellen und analysieren. Ein wichtiges Thema in dieser Arbeit ist die randomisierte Approximation mit niedrigem Rang. Insbesondere werden wir randomisierte Niedrigrangapproximationen von Matrixfunktionen, die Verwendung randomisierter Low-Rank-Approximation für die Schätzung der Spur einer Matrix und randomisierte Niedrigrangapproximationen von selbstadjungierte nichtnegative Spurklassenoperatoren untersuchen.

Kapiteln 3 bis 5 werden sich mit Niedrigrangapproximationen von Matrixfunktionen beschäftigen. Wir werden zwei Methoden zur Berechnung von Low-Rank-Approximationen von Matrixfunktionen vorstellen. In Kapitel 4 werden wir eine Methode namens funNyström analysieren, die eine Niedrigrangapproximation von $\boldsymbol{A}$ verwendet, um eine Niedrigrangapproximation von $f(\boldsymbol{A})$ zu erhalten, wobei f eine nicht negative operatormonotone Matrixfunktion ist. Insbesondere wird gezeigt, dass eine nahezu optimale Nyström-Approximation verwendet werden kann, um eine nahezu optimale funNyström-Approximation zu konstruieren. In Kapitel 5 wird ein Block-Krylov-Unterraum-Verfahren zur Berechnung von randomisierten Niedrigrangapproximationen von allgemeinen Matrixfunktionen betrachtet. Wir werden probabilistische Fehlerabschätzungen für diese Methode bereitstellen.

Kapiteln 6 bis 8 befassen sich mit der der Schätzung von Spuren von Matrizen. In Kapitel 7 wird eine adaptive Version des Hutch++ Algorithmus vorgestellt. Dieser Algorithmus nimmt eine Fehlertoleranz als Eingabe und liefert eine Schätzung der Spur innerhalb der Fehlertoleranz mit einer kontrollierbaren Fehlerwahrscheinlichkeit, während er gleichzeitig die Anzahl der Matrix-Vektor-Produkte mit der Matrix minimiert. In Kapitel 8 stellen wir eine Single-Pass-Version des Hutch++-Algorithmus vor. Dieser Algorithmus verwendet die Nyström-Approximation anstelle der randomisierten SVD in der Phase der Niedrigrangapproximation von Hutch++ und wir beweisen, dass er eine ähnliche Komplexität wie Hutch++ garantiert.

Kapitel 9 befasst sich mit einer unendlich dimensionalen Verallgemeinerung der Nyström-Approximation, um randomisierte Low-Rank-Approximationen für selbstadjungierte nichtnegative Spurklassenoperatoren zu berechnen. Wir werden eine Fehlerabschätzung für die endlich dimensionale Nyström-Approximation angeben, für den Fall, dass diese mit nich standard normal-verteilen Gaußschen Zufallsvectoren implementiert ist. Anschließend verwenden wir diese Schranken, um den Fehler einer unendlich dimensionalen Verallgemeinerung der Nyström-Approximation abzuschätzen.

Stichwörter: Niedrigrangapproximation, randomisierte numerische lineare Algebra, Matrixfunktionen, Spurenschätzung, Hilbert-Schmidt-Operatoren, Spurklasseoperator.

Contents

Acknowledgements **i**

Abstract (English/Deutsch) **iii**

List of figures **xi**

List of tables **xv**

Notation **xvii**

1 Introduction **1**

1.1 Randomized low-rank approximation of matrix functions 2

1.2 Trace estimation 3

1.3 Randomized low-rank approximation of non-negative self-adjoint trace class operators 3

2 Preliminaries on randomized low-rank approximation **5**

2.1 Low-rank approximation of matrices 5

2.2 The randomized singular value decomposition 7

2.2.1 Truncating back 9

2.2.2 Symmetric matrices 9

2.2.3 Beyond the randomized SVD 10

2.3 The Nyström approximation 12

2.3.1 Truncating back 14

3 An introduction to low-rank approximation of matrix functions **15**

3.1 Computing matrix-vector products with matrix functions 16

3.1.1 Differential equations 16

3.1.2 Sampling from elliptical distributions 17

3.2 Trace and diagonal estimation of matrix functions 18

3.2.1 Nuclear norm 18

3.2.2 Statistical learning 18

3.2.3 Effective dimension 19

3.2.4 Triangle counting . . . 19
3.2.5 The matrix exponential . . . 19
3.2.6 The inverse . . . 19
3.3 Challenges with computing low-rank approximations of matrix functions . 19
3.4 Contributions . . . 21

4 funNyström: Low-rank approximation of operator monotone matrix functions 23
4.1 Operator monotone functions and the funNyström approximation . . . 23
4.2 Good Nyström approximations imply good funNyström approximations . 25
4.2.1 Frobenius and nuclear norm guarantees . . . 27
4.2.2 Operator norm guarantees . . . 29
4.2.3 Eigenvalue guarantees . . . 30
4.2.4 Wasserstein distance guarantees . . . 32
4.3 Good projections imply good Nyström approximations . . . 33
4.3.1 Frobenius norm guarantees . . . 34
4.3.2 Nuclear norm guarantees . . . 37
4.3.3 Operator norm guarantees . . . 37
4.3.4 Eigenvalue guarantee . . . 38
4.4 Explicit bounds . . . 39
4.5 General unitarily invariant norms . . . 43
4.6 Numerical experiments . . . 46
4.6.1 Test matrices . . . 46
4.6.2 Verifying theoretical results . . . 48
4.6.3 Comparing number of matrix vector products . . . 52
4.6.4 Comparing accuracy . . . 53
4.6.5 Fast computation of matrix-vector products . . . 53
4.6.6 Application to trace estimation . . . 56

5 Randomized block-Krylov subspace methods for low-rank approximation of matrix functions 59
5.1 Krylov aware low-rank approximation . . . 60
5.1.1 The randomized SVD for matrix functions . . . 60
5.1.2 Krylov aware low-rank approximation . . . 61
5.2 Error bounds . . . 62
5.2.1 Error bounds for inexact projections . . . 63
5.2.2 Structural bounds . . . 63
5.2.3 Probabilistic bounds . . . 66
5.2.4 Error bounds for Krylov aware low-rank approximation . . . 67
5.2.5 Simplified bounds for the matrix exponential . . . 69
5.2.6 Simplified bounds for the identity function . . . 71
5.3 Numerical experiments . . . 72
5.3.1 Test matrices . . . 72

5.3.2 Comparing relative errors . . . 74

6 An introduction to trace estimation 77

6.1 The stochastic trace estimator . . . 78

6.2 The Hutch++ algorithm . . . 79

6.3 Contributions . . . 80

6.3.1 Adaptive trace estimation . . . 80

6.3.2 A single pass algorithm . . . 81

7 A-Hutch++: An adaptive trace estimation algorithm 83

7.1 Derivation of adaptive Hutch++ . . . 83

7.1.1 Analysis of trace estimation . . . 84

7.1.2 Finding the minimum of $m(r)$ . . . 85

7.1.3 Estimating the Frobenius norm of the remainder . . . 86

7.1.4 A prototype algorithm . . . 87

7.2 A-Hutch++ . . . 89

7.3 Numerical experiments . . . 92

7.3.1 Synthetic matrices . . . 93

7.3.2 Triangle counting . . . 96

7.3.3 Estrada index . . . 96

7.3.4 Log-determinant . . . 98

7.3.5 Trace of inverses . . . 98

8 Nyström++: A single pass trace estimation algorithm 101

8.1 Analysis of Nyström++ . . . 102

8.1.1 Adaptive Nyström++ . . . 105

8.2 Numerical results . . . 105

9 Randomized Nyström approximation of non-negative self-adjoint trace class operators 109

9.1 Motivation . . . 109

9.2 The randomized Nyström approximation in finite dimensions with correlated Gaussian sketches . . . 110

9.2.1 Distribution of Gaussian sketches . . . 110

9.2.2 Nyström approximation with correlated Gaussian sketches . . . 111

9.3 The randomized Nyström approximation in infinite dimensions . . . 117

9.3.1 Quasimatrices . . . 117

9.3.2 Non-negative self-adjoint trace-class operators . . . 118

9.3.3 Gaussian processes . . . 118

9.3.4 Infinite-dimensional extension of the Nyström approximation . . . 121

9.3.5 Structural bound . . . 122

9.3.6 Probabilistic bounds . . . 123

9.4 Numerical experiments . . . 127

9.4.1 A pretty function 128
9.4.2 Matérn Kernels 128
9.4.3 Green's function for an elliptic differential operator 129

10 Conclusions and outlook **131**

A Properties of Gaussian matrices **135**

Bibliography **139**

List of Figures

4.1 Comparing $\varepsilon_{\text{projection}}, \varepsilon_{\text{Nyström}}$, and $\varepsilon_{\text{funNyström}}$ for column subset selection applied to $\boldsymbol{A}_{\text{disc}}$ defined in (4.25) and $f(x) = \frac{x}{x+1}$. Note that $\varepsilon_{\text{projection}}$ is significantly worse than the $\varepsilon_{\text{Nyström}}$ and $\varepsilon_{\text{funNyström}}$ since the orthogonal projection $\boldsymbol{Q}\boldsymbol{Q}^T$ zeros out all except ℓ rows of $\boldsymbol{A}$. In contrast, the Nyström approximation $\widehat{\boldsymbol{A}} = \boldsymbol{A}^{1/2}\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}^{1/2}$ effectively performs half a step of subspace iteration on $\boldsymbol{Q}$, giving a better approximation. 50

4.2 Comparing $\varepsilon_{\text{projection}}, \varepsilon_{\text{Nyström}}$, and $\varepsilon_{\text{funNyström}}$ for Krylov iteration applied $\boldsymbol{A}_{\text{alg}}$ defined in (4.24) with $n = 3000, C = c = 1$ and $f(x) = \log(1 + x)$. . . 51

4.3 Comparing $\varepsilon_{\text{projection}}, \varepsilon_{\text{Nyström}}$, and $\varepsilon_{\text{funNyström}}$ for subspace iteration applied to $\boldsymbol{A}_{\text{exp}}$ defined in (4.24) with $n = 3000, C = 1, \gamma = e^{-1}$ and $f(x) = x^{1/2}$. 52

4.4 Number of matrix-vector products with $\boldsymbol{A}$ vs. attained accuracy for low-rank approximations of $f(\boldsymbol{A})$ from funNyström approximation (blue) and $\widehat{\boldsymbol{B}}_{1,k}^{(d)}$ (red). 54

4.5 Error vs. prescribed rank of the approximation for the funNyström approximation applied to $\boldsymbol{A}$ (blue) and $\widehat{\boldsymbol{B}}_{q,k}$, the Nyström approximation applied to $f(\boldsymbol{A})$ (red), for $q = 1, 2$. 55

4.6 Number of matrix-vector products N vs. speed-up factor $T_{\text{Lanczos}}(N)/T_{\text{low-rank}}(N)$. 56

4.7 Approximation of $\text{tr}(\log(\boldsymbol{I}+\boldsymbol{A}))$ using the funNyström approximation (blue) and the method presented in [138] (red). The x-axis represents the number of matrix-vector products performed with $\boldsymbol{A}$ to obtain the approximation, and the y-axis represents the relative error of the approximation. 58

5.1 Comparing (5.23) for the the approximations returned by Algorithm 8 without truncation (untruncated), Algorithm 8 with truncation back to rank k (truncated), Algorithm 7, and Algorithm 6. The black line shows the optimal rank k approximation relative Frobenius norm error. The rank parameter k is visible as titles in the figures. In all experiments we set the oversampling parameter $p = 0$. All examples are outlined in Section 5.3.1. 75

5.2 Comparing (5.23) for the the approximations returned by Algorithm 8 without truncation (untruncated), Algorithm 8 with truncation back to rank k (truncated), Algorithm 7, and Algorithm 6. The black line shows the optimal rank k approximation relative Frobenius norm error. The rank parameter k is visible as titles in the figures. In all experiments we set the oversampling parameter $p = 5$. All examples are outlined in Section 5.3.1. 76

7.1 In this example we let $\boldsymbol{A} = \boldsymbol{U\Lambda U}^T \in \mathbb{R}^{1000\times1000}$ where $\boldsymbol{U}$ is a random orthogonal matrix and $\boldsymbol{\Lambda}$ is a diagonal matrix with $\boldsymbol{\Lambda}_{ii} = 1/i^2$. The x-axis shows the rank r, and the y-axis is the function $m(r)$ defined in (7.4) with $\delta = 0.01$, $\varepsilon = 0.05\,\mathrm{tr}(\boldsymbol{A})$ and $\upsilon = 0$. The function has its minimum at $r^* = 7$. 86

7.2 For different choices of δ, this plot demonstrates the relationship between k and the largest choice of α such that $\frac{\gamma(k/2,\alpha k/2)}{\Gamma(k/2)} \leq \delta$. 87

7.3 In this example we let $\boldsymbol{A} = \boldsymbol{U\Lambda U}^T \in \mathbb{R}^{1000\times1000}$ where $\boldsymbol{U}$ is a random orthogonal matrix and $\boldsymbol{\Lambda}$ is a diagonal matrix with $\boldsymbol{\Lambda}_{ii} = 1/i^{1.5}$. We run Algorithm 10 with $\varepsilon = 0.01\,\mathrm{tr}(\boldsymbol{A})$, $\delta = 0.05$ and $\upsilon = 0$ to obtain $\boldsymbol{A}_{\text{rest}}$ defined in line 16. The x-axis shows the number of matrix-vector products with $\boldsymbol{A}_{\text{rest}}$. The red line shows the evolution of the sequence M_k defined in (7.8), the blue line shows the linear line k against k and the black line is the number of matrix-vector products with $\boldsymbol{A}_{\text{rest}}$ to guarantee an error less than ε with probability at least $1 - \delta$. We stop the while loop in Algorithm 11 once the red and blue line cross. 91

7.4 Comparison of A-Hutch++ and Hutch++ for the estimation of the trace of the synthetic matrices with algebraic decay from Section 7.3.1. 93

7.5 Comparison of A-Hutch++ and Hutch++ for the triangle counting examples from Section 7.3.2. 97

7.6 Comparison of A-Hutch++ and Hutch++ for the estimation of the Estrada index of the matrix from Section 7.3.3. 97

7.7 Comparison of A-Hutch++ and Hutch++ for the log determinant estimation of the matrices from Section 7.3.4. 98

7.8 Comparison of A-Hutch++ and Hutch++ for the estimation of the trace of the inverse of the matrices described in Section 7.3.5. 99

8.1 Comparison of Hutch++, Single Pass Hutch++ and Nyström++ for the estimation of the trace of the synthetic matrices with algebraic decay described in Section 7.3.1. 107

8.2 Comparison of Hutch++, Single Pass Hutch++ and Nyström++ for the estimation of the Estrada index as described in Section 5.3.1. 107

8.3 Comparison of Hutch++, Single Pass Hutch++ and Nyström++ for the estimation of the trace of the inverse of the matrices described in Section 7.3.5. 108

8.4 Comparison of Hutch++, Single Pass Hutch++ and Nyström++ for the estimation of the trace of the synthetic matrices with exponential decay described in Section 8.2. 108

9.1 (a) Exact kernel defined by (9.27) along with convergence of the Nyström approximation for different values of ℓ (b). (c)-(e) Rank-40 Nyström approximations of the kernel for $\ell = 1, 0.1, 0.01$, respectively. 128

9.2 The contour plots show the exact kernel defined by (9.28)-(9.30) and the approximated kernels using rank 100 Nyström approximations. The error plots show the relative error in the Hilbert-Schmidt norm. *Optimal* denotes the best low-rank approximation error. 129

9.3 The contour plots show the exact kernel defined by (9.31) and the approximated kernels using rank 100 Nyström approximations. The error plots show the relative error in the Hilbert-Schmidt norm. *Optimal* denotes the best low-rank approximation error. 130

10.1 Comparing the subobtimality factor $\varepsilon = \frac{\|f(\boldsymbol{A})-\boldsymbol{B}\|_{\mathrm{F}}}{\|f(\boldsymbol{A})-f(\boldsymbol{A})_{(k)}\|_{\mathrm{F}}} - 1$ for a rank k approximations $\boldsymbol{B}$. We consider the rank k approximations given by (10.1) with $q = 2s$, the approximation returned by Algorithm 8 with $s = r$ with (with truncation), Algorithm 7, and Algorithm 1. The sketch matrix is chosen as a $n \times k$ standard Gaussian random matrix. The rank parameter k is visible as titles in the figures. 132

List of Tables

7.1 Empirical failure probabilities from 100000 repeats of applying A-Hutch++ on the synthetic matrices described in Section 7.3.1. 94

7.2 The average distribution of matrix-vector products between the low rank approximation phase and stochastic trace esimation phase of A-Hutch++ applied on the synthetic matrices with algebraic decay and input tolerance $\varepsilon = 2^{-t}\,\mathrm{tr}(\boldsymbol{A})$ for $t = 2, 3, \ldots, 10$. A-Hutch++ requires at least 6 matrix-vector products to detect a minimum of the function $\tilde{m}(r)$ in (7.6). 96

Notation

- SPSD is an abbreviation for *symmetric positive semi-definite*;
- $\boldsymbol{I}$ denotes the identity matrix;
- $\boldsymbol{A}_{(k)}$ denotes an optimal low-rank approximation of a matrix $\boldsymbol{A}$ in any unitarily invariant norm;
- $\boldsymbol{P}_{\boldsymbol{Y}}$ denotes the orthogonal projector onto range$(\boldsymbol{Y})$;
- $\boldsymbol{A}^T$ and $\boldsymbol{A}^*$ denote the transpose and Hermitian adjoint of a matrix $\boldsymbol{A}$;
- $\mathcal{A}^*$ denotes the adjoint of an operator $\mathcal{A}$;
- $\|\cdot\|_{(s)}$ denotes the Schatten s-norm;
- $\|\cdot\|_*$ denotes the nuclear norm;
- $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm;
- $\|\cdot\|_2$ denotes the operator norm;
- $\boldsymbol{A}^\dagger$ denotes the Moore-Penrose pseudoinverse of a matrix $\boldsymbol{A}$;
- $\mathcal{N}(0,1)$ denotes the standard Gaussian distribution;
- $\mathcal{N}(\mathbf{0},\boldsymbol{K})$ denotes the distribution of Gaussian random vectors with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{K}$;
- $\mathcal{GP}(0,K)$ denotes the distribution of Gaussian processes with mean 0 and covariance kernel K;
- i.i.d. is an abbreviation for *independent identically distributed.*

1 Introduction

This thesis explores the use of randomized low-rank approximation to compute approximations to matrix functions, as a variance reduction technique for Monte-Carlo estimators in trace estimation, and to compute approximations to non-negative self-adjoint trace class operators. In this chapter, we will briefly outline low-rank approximation and the problems considered in this thesis.

Nearly any algorithm that is designed to solve a real world problem will rely on matrix computations to be efficient. Matrices are versatile mathematical objects, that allow practitioners to represent data, operators, coordinates, functions, and many more mathematical objects. For a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, low-rank approximation is concerned with finding a low-rank matrix $\boldsymbol{B} \in \mathbb{R}^{m \times n}$ so that

$$\boldsymbol{B} \approx \boldsymbol{A}.$$

Nearly all matrices that appear in applications will be *analytically* full rank, but many will be *numerically* low rank [156], i.e., $\boldsymbol{A}$ can be well-approximated with a low-rank matrix $\boldsymbol{B}$. Such matrices appear in genomics [30, 62, 64, 164], discretizations of partial differential equations [33, 72], movie preferences [22], statistical machine learning [65], multiscale physics [75], and many more. The advantages of low-rank approximations are two-fold. Firstly, low-rank approximations yield advantages in terms of storage. Storing a dense matrix requires $O(mn)$ units of memory, wheras storing a rank k matrix requires $O((m+n)k)$ units of memory. Secondly, low-rank approximations yield advantages in terms of computational efficiency. For example, computing a matrix-vector product with a dense matrix requires $O(mn)$ operations, whereas computing a matrix-vector product with a rank k matrix requires $O((m+n)k)$ operations. When $k \ll \min\{m, n\}$, this yields a dramatic reduction in terms of storage and computational cost, especially in today's applications when m and n are very large.

Classical deterministic algorithms, employing for example Golub-Kahan bidiagonalization, aim at obtaining *optimal* low-rank approximations through the truncated singular value decomposition [149, Lecture 31]. These algorithms are accurate, but generally costs $O(mn^2)$ operations, assuming $m \geq n$. They are consequently prohibitively expensive for large scale matrices. In today's applications, high accuracy is not always required and the matrices that appear are generally too large for classical deterministic algorithms. On the other hand, randomized low-rank approximation algorithms are very fast and return near-optimal low-rank approximations with high probability. These algorithms are therefore suitable for the large-scale matrices that appear in today's applications.

The randomized SVD is the prototypical randomized low-rank approximation algorithm, and due to its simplicity and strong theoretical guarantees it has been extremely successful. It builds on the observation that one can exactly recover the SVD of a rank k matrix with only k matrix-vector products. Therefore, if $\boldsymbol{A}$ is very close to a rank k matrix, one should be able to nearly recover the SVD of $\boldsymbol{A}$ with only k matrix-vector products. If $\boldsymbol{A}$ is dense, this algorithm requires only $O(mnk)$ operations to obtain an approximation to $\boldsymbol{A}$. This compares favorably to the $O(mn^2)$ operations required by classical deterministic algorithms. The landmark paper by Halko, Martinsson, and Tropp [79] presented and analysed the randomized SVD and theoretically guaranteed that it will nearly recover matrices that admit accurate low-rank approximations.

In this thesis we will explore how randomized low-rank approximation can be used to approximate matrix functions, as a variance reduction technique for Monte-Carlo estimators in trace estimation, and how it can be used to approximate non-negative self-adjoint trace class operators. In the next subsections we will give a brief introduction to these three applications.

1.1 Randomized low-rank approximation of matrix functions

Matrix functions appear in numerous areas of applied mathematics, including differential equations [82, 85], statistics [127], network science [54, 55], machine learning [65, 168], quantum mechanics [141, 160] and many more. Matrix functions are generalizations of analytic scalar functions to matrices [84], and common examples include the inverse, the matrix square-root, and the matrix exponential.

Having access to a good low-rank approximation to a matrix function is beneficial as a variance reduction technique in trace estimation [111] and when one needs to compute repeated matrix-vector products with the matrix function. However, standard randomized low-rank approximation methods assume that we have access to the matrix of interest through (exact) matrix-vector products, which is usually not the case for matrix functions. In practice, we have access to the matrix $\boldsymbol{A}$ and not the matrix function $f(\boldsymbol{A})$. In this thesis we will present two algorithms that compute a low-rank approximation to a matrix function $f(\boldsymbol{A})$ using only matrix-vector products with $\boldsymbol{A}$. More details will be given in Chapters 3 to 5.

1.2 Trace estimation

Trace estimation is concerned with estimating the trace of a matrix implicitly given through matrix-vector products. This task arises in a wide variety of applications, such as triangle counting in graphs [8], Frobenius norm estimation [31, 74], quantum chromodynamics [146], computing the Estrada index of a graph [120, 54], computing the log-determinant [2, 41, 138, 159, 168] and many more [155]. Early methods rely on Monte-Carlo estimation, which, due to its slow convergence, required many matrix-vector products with the matrix to obtain a good estimate of the trace. Meyer, Musco, Musco, and Woodruff [111] showed that randomized low-rank approximation can be used to reduce the variance of the Monte-Carlo estimator, and consequently reduce the number of matrix-vector products with the matrix of interest. In this thesis we will explore two improved variants of the algorithm presented in [111]. More details will be given in Chapters 6 to 8.

1.3 Randomized low-rank approximation of non-negative self-adjoint trace class operators

Hilbert-Schmidt operators constitute a special class of compact operators between two Hilbert spaces [89]. Loosely speaking, they are infinite-dimensional analogs of matrices with a sufficiently fast singular value decay. They frequently appear in, for example, partial differential equations [29] and Gaussian process regression [53, 65, 119, 157, 162, 163]. Boullé and Townsend generalized the randomized SVD to Hilbert-Schmidt operators [28, 29], and in this thesis we will present and analyze an infinite-dimensional analog of the Nyström approximation [68] applied to self-adjoint non-negative trace class operators. More details will be given in Chapter 9.

Organization of thesis

We begin with providing preliminaries of randomized low-rank approximation in Chapter 2. We present the randomized SVD and the Nyström approximation, both of which will play central roles in this thesis.

Chapters 3 to 5 are concerned with matrix functions. In Chapter 3 we begin with giving a brief overview of matrix functions and their applications. In Chapter 4 we present funNyström, which is a method to compute low-rank approximations to a certain class of matrix functions: *operator monotone functions*. This chapter is based on the work in [124, 125]. In Chapter 5 we describe and analyze a Krylov subspace method to compute low-rank approximations to general matrix functions. This chapter is based on the work in [122].

Chapters 6 to 8 are concerned with trace estimation. In Chapter 6 we begin with giving a brief overview of trace estimation and its applications. In Chapter 7 we present A-

Hutch++, which is an adaptive method to approximate the trace of a matrix up to a prescribed accuracy. In Chapter 8 we describe and analyze Nyström++, which is a single pass algorithm to estimate the trace of a matrix. These chapters are based on the work in [123].

In Chapter 9 we present and analyse an infinite dimensional analog of the Nyström approximation to compute low-rank approximations of self-adjoint non-negative trace class operators. This chapter is based on the work in [121].

In Chapter 10 serves as the conclusion of this thesis and provides an outlook for future research directions.

2 Preliminaries on randomized low-rank approximation

In this chapter we present the notation and preliminary results regarding deterministic and randomized low-rank approximation.

In Section 2.1, we introduce the singular value decomposition (SVD) and how it can be used to obtain optimal low-rank approximations. We establish the notation for low-rank approximation that will serve as the framework for the subsequent chapters.

In Section 2.2, we explain how randomization can be used to compute *near-optimal* low-rank approximations. Here, we introduce the randomized SVD and present important theoretical guarantees, which will provide a foundation for many of the theoretical results in this thesis.

Subsequently, in Section 2.3 we will recall the Nyström approximation for computing low-rank approximations to symmetric positive semi-definite (SPSD) matrices. We will explain why the Nyström approximation is preferable over the randomized SVD for approximating SPSD matrices. We conclude with showing how the theoretical guarantees for the randomized SVD can be used to establish theoretical guarantees for the Nyström approximation.

2.1 Low-rank approximation of matrices

In this section we establish a few useful results and the notation for low-rank approximation of matrices. Consider $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, where we assume without loss of generality that $m \geq n$. Low-rank approximation is concerned with finding a low-rank matrix $\boldsymbol{B}$ that is a good approximation to the matrix $\boldsymbol{A}$. For our discussion, it will be useful to recall the singular value decomposition.

Theorem 2.1 (Singular Value Decomposition, [149, Theorem 4.1])**.** *Every matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ admits a decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, where $\boldsymbol{U} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{V} \in \mathbb{R}^{n \times n}$ are matrices with orthonormal columns and $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ is a diagonal matrix containing the* *singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$ on its diagonal.*

With the concept of the SVD established, we state the definition of a unitarily invariant norm and introduce an important example known as Schatten norms.

Definition 2.1 (Unitarily invariant norms and Schatten norms)**.** *A norm $\|\cdot\|$ is said to be unitarily invariant if for any matrix $\boldsymbol{A}$ with SVD $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ we have $\|\boldsymbol{A}\| = \|\boldsymbol{\Sigma}\|$. Furthermore, for $s \in [1, \infty]$ the Schatten-s norm of $\boldsymbol{A}$ is defined as*

$$\|\boldsymbol{A}\|_{(s)} = \left(\sum_{i=1}^{n} \sigma_i^s\right)^{1/s}.$$

We write $\|\boldsymbol{A}\|_* = \|\boldsymbol{A}\|_{(1)}, \|\boldsymbol{A}\|_\mathrm{F} = \|\boldsymbol{A}\|_{(2)}$, and $\|\boldsymbol{A}\|_2 = \|\boldsymbol{A}\|_{(\infty)} = \sigma_1$ to denote the nuclear, Frobenius, and operator norm, respectively.

The singular value decomposition is central to low-rank approximation in unitarily invariant norms, since it allows us to construct optimal low-rank approximations. In particular, the famous Eckart-Young-Mirsky Theorem asserts that the truncated SVD provides an optimal low-rank approximation in any unitarily invariant norm.

Theorem 2.2 (Eckart-Young-Mirsky Theorem, [50, 113, 140])**.** *Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ with SVD*

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = \begin{bmatrix} \boldsymbol{U}_1 & \boldsymbol{U}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_1 & \\ & \boldsymbol{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_1^T \\ \boldsymbol{V}_2^T \end{bmatrix}, \tag{2.1}$$

where $\boldsymbol{\Sigma}_1 = \operatorname{diag}(\sigma_1, \ldots, \sigma_k)$ and $\boldsymbol{U}_1 \in \mathbb{R}^{m \times k}$ and $\boldsymbol{V}_1 \in \mathbb{R}^{n \times k}$ contain the dominant k left and right singular vectors, respectively. Define

$$\boldsymbol{A}_{(k)} := \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^T = \boldsymbol{U}_1\boldsymbol{U}_1^T\boldsymbol{A}, \tag{2.2}$$

Then for any unitarily invariant norm $\|\cdot\|$ we have

$$\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\| = \|\boldsymbol{\Sigma}_2\| = \min_{\boldsymbol{B}:\operatorname{rank}(\boldsymbol{B}) \leq k} \|\boldsymbol{A} - \boldsymbol{B}\|. \tag{2.3}$$

Theorem 2.2 implies that that we can obtain an optimal rank k approximation to $\boldsymbol{A}$ by keeping only the dominant k singular vectors and singular values. Unfortunately, computing $\boldsymbol{A}_{(k)}$ in (2.2) generally costs $O(mn^2)$ operations [149, Lecture 31], which becomes prohibitively expensive for the large scale matrices that frequently appear in today's applications.

However, in cases when $\boldsymbol{A}$ admits an accurate low-rank approximation, which, in view of (2.3), implies that the singular values $\sigma_{k+1} \geq \sigma_{k+2} \geq \cdots \geq \sigma_n$ are small, it is usually preferable to find a *near-optimal* low-rank approximation that is significantly cheaper to compute. For this task, randomized low-rank approximation has proven to be highly successful, as we will discuss in the subsequent section.

2.2 The randomized singular value decomposition

The randomized SVD is a simple and extremely successful method to obtain cheap, yet accurate, low-rank approximations of matrices that have a rapid singular value decay [79, 100]. The basic idea is to find a matrix $\boldsymbol{Y}$ whose range contains a good approximation to the range of $\boldsymbol{A}$. In this case, if

$$\boldsymbol{P_Y} = \boldsymbol{Y}\boldsymbol{Y}^{\dagger},$$

denotes the orthogonal projection onto range$(\boldsymbol{Y})$, where † denotes the Moore-Penrose pseudoinverse, then $\boldsymbol{P_Y}\boldsymbol{A} \approx \boldsymbol{A}$. In its original form, the randomized SVD samples a random matrix $\boldsymbol{\Omega}$ and forms the product $\boldsymbol{A\Omega}$. When $\boldsymbol{A}$ admits an accurate low-rank approximation, then the range of $\boldsymbol{A\Omega}$ is a very good approximation to the range of $\boldsymbol{A}$, with high probability. Therefore, we expect $\boldsymbol{P_{A\Omega}}\boldsymbol{A} \approx \boldsymbol{A}$. The algorithm is outlined in Algorithm 1.

Algorithm 1 The randomized SVD

```
input: $\boldsymbol{A} \in \mathbb{R}^{m \times n}$. Target rank $k$. Oversampling parameter $p$.
output: Rank $k+p$ approximation to $\boldsymbol{A}$ in factored form $\widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^T$.
Sample a random $n \times (k+p)$ sketch matrix $\boldsymbol{\Omega}$.
$\boldsymbol{Y} = \boldsymbol{A\Omega}$.
Compute an orthonormal basis $\boldsymbol{Q}$ for range$(\boldsymbol{Y})$.
$\boldsymbol{X} = \boldsymbol{Q}^T\boldsymbol{A}$.
Compute the SVD of $\boldsymbol{X} = \boldsymbol{W}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^T$.
$\widehat{\boldsymbol{U}} = \boldsymbol{QW}$
return $\boldsymbol{P_{A\Omega}}\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A} = \boldsymbol{QX} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^T$.
```

Remark 2.1. *We note that the low-rank approximation returned by Algorithm 1 has rank $k+p$ instead of k. The oversampling parameter p improves the statistical performance of the algorithm, and can in practice be set to $p = 5$ or $p = 10$ [79]. When an exact rank k approximation is desired, one can simply return $(\boldsymbol{P_{A\Omega}}\boldsymbol{A})_{(k)}$ as a rank k approximation to $\boldsymbol{A}$. We provide a more detailed discussion in Section 2.2.1.*

Under some mild assumptions on the sketch matrix, one can derive *deterministic* error bounds for the error $\|\boldsymbol{A} - \boldsymbol{P_{A\Omega}}\boldsymbol{A}\|$ in the operator and Frobenius norm; similar bounds for general Schatten norms are proved in [137].

Theorem 2.3 ([79, Theorem 9.1])**.** *Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ have an SVD as partitioned in* (2.1) *and let $\boldsymbol{\Omega}$ be a sketch matrix. Define*

$$\boldsymbol{\Omega}_1 = \boldsymbol{V}_1^T\boldsymbol{\Omega}, \qquad \boldsymbol{\Omega}_2 = \boldsymbol{V}_2^T\boldsymbol{\Omega}, \tag{2.4}$$

and assume that $\operatorname{rank}(\boldsymbol{\Omega}_1) = k$. *Then, for* $\xi \in \{2, \mathrm{F}\}$ *we have*

$$\|\boldsymbol{A} - \boldsymbol{P}_{\boldsymbol{A}\boldsymbol{\Omega}}\boldsymbol{A}\|_\xi^2 \leq \|\boldsymbol{\Sigma}_2\|_\xi^2 + \|\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_\xi^2.$$

Theorem 2.3 allows us to derive statistical bounds on the error, since it is sufficient to derive bounds on $\|\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|$. In the special case when $\boldsymbol{\Omega}$ is a standard Gaussian random matrix, i.e. each entry of $\boldsymbol{\Omega}$ is drawn independently from $\mathcal{N}(0,1)$, deriving bounds for $\|\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|$ is particularly easy. From the unitary invariance of random Gaussian vectors, we know that $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ are *independent* standard Gaussian random matrices; see [79, Proof of Theorem 10.5]. With this fact in mind, we can derive probabilistic bounds by first conditioning on $\boldsymbol{\Omega}_1$, which has no effect on the distribution of $\boldsymbol{\Omega}_2$. For example, we have the following well-known results; see [79, Sections 10.2-10.3].

Lemma 2.4. *Let* $\boldsymbol{\Omega}_1 \in \mathbb{R}^{k\times(k+p)}$ *and* $\boldsymbol{\Omega}_2^{(n-k)\times(k+p)}$ *are independent standard Gaussian random matrices and* $\boldsymbol{D}$ *be any matrix with* $n-k$ *columns. Then, if* $p \geq 2$ *we have*

$$\mathbb{E}\|\boldsymbol{D}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_\mathrm{F}^2 = \frac{k}{p-1}\|\boldsymbol{D}\|_\mathrm{F}^2.$$

If $k \geq 2$ *and* $p \geq 4$, *then for all* $u, t \geq 1$ *we have*

$$\|\boldsymbol{D}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_\mathrm{F} \leq t\sqrt{\frac{3k}{p+1}}\|\boldsymbol{D}\|_\mathrm{F} + ut\frac{e\sqrt{k+p}}{p+1}\|\boldsymbol{D}\|_2,$$

with probability at least $1 - 2t^{-p} - e^{-u^2/2}$. *In particular, if* $\boldsymbol{\Omega} \in \mathbb{R}^{n\times(k+p)}$ *is a standard Gaussian random matrix, then these bounds hold for* $\boldsymbol{\Omega}_1$ *and* $\boldsymbol{\Omega}_2$ *defined in* (2.4).

This allows us to prove the following result for the Frobenius norm error; similar results are true for the operator norm [79, Theorem 10.6, Theorem 10.8].

Theorem 2.5 ([79, Theorem 10.5, Theorem 10.7])**.** *Let* $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ *and let* $\boldsymbol{\Omega}$ *be a random* $n \times (k+p)$ *standard Gaussian matrix. If* $p \geq 2$ *we have*

$$\mathbb{E}\|\boldsymbol{A} - \boldsymbol{P}_{\boldsymbol{A}\boldsymbol{\Omega}}\boldsymbol{A}\|_\mathrm{F}^2 \leq \left(1 + \frac{k}{p-1}\right)\|\boldsymbol{\Sigma}_2\|_\mathrm{F}^2.$$

Furthermore, if $k \geq 2$ *and* $p \geq 4$, *then for all* $u, t \geq 1$ *we have*

$$\|\boldsymbol{A} - \boldsymbol{P}_{\boldsymbol{A}\boldsymbol{\Omega}}\boldsymbol{A}\|_\mathrm{F} \leq \left(1 + t\sqrt{\frac{3k}{p+1}}\right)\|\boldsymbol{\Sigma}_2\|_\mathrm{F} + ut\frac{e\sqrt{k+p}}{p+1}\|\boldsymbol{\Sigma}_2\|_2,$$

with probability at least $1 - 2t^{-p} - e^{-u^2/2}$.

Other distributions, such as Rademacher and SRFT matrices have also been studied, but they generally obey weaker guarantees; see e.g. [79, 136].

2.2.1 Truncating back

As previously noted, the low-rank approximation returned by Algorithm 1 has a rank higher than the target rank k. An exact rank k approximation may in some cases be desirable. As mentioned in Remark 2.1, one simple remedy is to return the best rank k approximation $(\boldsymbol{P}_{\boldsymbol{A\Omega}}\boldsymbol{A})_{(k)}$ to $\boldsymbol{P}_{\boldsymbol{A\Omega}}\boldsymbol{A}$ and it comes at no additional cost. In fact, one can show that $\|\boldsymbol{A} - (\boldsymbol{P}_{\boldsymbol{A\Omega}})_{(k)}\|_F$ satisfies similar bounds as in Theorem 2.3.

Theorem 2.6. *Consider the setting of Theorem 2.3. Then,*

$$\|\boldsymbol{A} - (\boldsymbol{P}_{\boldsymbol{A\Omega}}\boldsymbol{A})_{(k)}\|_{\mathrm{F}}^2 \leq \|\boldsymbol{\Sigma}_2\|_{\mathrm{F}}^2 + \|\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}^2. \tag{2.5}$$

Proof. The proof follows from first applying [76, Equation (3.6)] and then [137, Equation (14)] with $q = 0$. □

Deriving statistical bounds for (2.5) can be done in an identical fashion as done for Theorem 2.5. We have the following immediate corollary of Theorem 2.6.

Theorem 2.7. *Consider the setting of Theorem 2.5. If $p \geq 2$ we have*

$$\mathbb{E}\|\boldsymbol{A} - (\boldsymbol{P}_{\boldsymbol{A\Omega}}\boldsymbol{A})_{(k)}\|_{\mathrm{F}}^2 \leq \left(1 + \frac{k}{p-1}\right)\|\boldsymbol{\Sigma}_2\|_{\mathrm{F}}^2.$$

Furthermore, if $k \geq 2$ and $p \geq 4$, then for all $u, t \geq 1$ we have

$$\|\boldsymbol{A} - (\boldsymbol{P}_{\boldsymbol{A\Omega}}\boldsymbol{A})_{(k)}\|_{\mathrm{F}} \leq \left(1 + t\sqrt{\frac{3k}{p+1}}\right)\|\boldsymbol{\Sigma}_2\|_{\mathrm{F}} + ut\frac{e\sqrt{k+p}}{p+1}\|\boldsymbol{\Sigma}_2\|_2,$$

with probability at least $1 - 2t^{-p} - e^{-u^2/2}$.

Unfortunately, the same argument to prove Theorem 2.6 does not carry over to show a similar bound in other norms. The proof relies on the [76, Theorem 3.5], which states that for any matrix $\boldsymbol{Q}$ we have

$$\min_{\boldsymbol{B}:\mathrm{rank}(\boldsymbol{B})\leq k} \|\boldsymbol{A} - \boldsymbol{QB}\|_{\mathrm{F}} = \|\boldsymbol{A} - (\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{A})_{(k)}\|_{\mathrm{F}},$$

which is not true in other norms; see [93] for a counter example in the operator norm. Bounds in other norms exists, see e.g. [79, Theorem 9.3], but they are generally weaker.

2.2.2 Symmetric matrices

When $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is symmetric it is usually preferable to obtain an eigenvalue decomposition instead of an SVD [79, Section 5.3]. For symmetric matrices, instead of an SVD we will consider an eigenvalue decomposition

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T = \begin{bmatrix}\boldsymbol{U}_1 & \boldsymbol{U}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\Lambda}_1 & \\ & \boldsymbol{\Lambda}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{U}_1^T \\ \boldsymbol{U}_2^T\end{bmatrix}, \tag{2.6}$$

where $\boldsymbol{\Lambda}_1 = \operatorname{diag}(\lambda_1, \ldots, \lambda_k)$ contain the largest magnitude eigenvalues and $\boldsymbol{U}_1 \in \mathbb{R}^{n\times k}$ the corresponding orthonormal eigenvectors. Hence, an optimal low-rank approximation to $\boldsymbol{A}$ in a unitarily invariant norm is $\boldsymbol{A}_{(k)} = \boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1^T$. When $\boldsymbol{A}$ is SPSD the eigenvalue decomposition is an SVD. To preserve the symmetry of the low-rank approximation of $\boldsymbol{A}$ we can project $\boldsymbol{A}$ from the left and right and obtain the approximation

$$\boldsymbol{A} \approx \boldsymbol{P}_{\boldsymbol{A\Omega}}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{A\Omega}} = \boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}\boldsymbol{Q}^T, \tag{2.7}$$

where $\boldsymbol{Q}$ is an orthonormal basis for range($\boldsymbol{A\Omega}$). Obtaining an eigenvalue decomposition of $\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}\boldsymbol{Q}^T$ can be done first by computing an eigenvalue decomposition of the smaller matrix $\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{W}\widehat{\boldsymbol{\Lambda}}\boldsymbol{W}^T$. The eigenvalue decomposition of $\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}\boldsymbol{Q}^T$ is therefore $(\boldsymbol{Q}\boldsymbol{W})\widehat{\boldsymbol{\Lambda}}(\boldsymbol{Q}\boldsymbol{W})^T$. The approximation (2.7) satisfies similar bounds as in Theorem 2.3 and Theorem 2.6. We will provide such bounds in a more general setting in Chapter 5. The pseudocode for the this version of the randomized SVD will be presented in Section 5.1.

2.2.3 Beyond the randomized SVD

Before proceeding, we emphasize that many variants of Algorithm 1 have been studied in the literature. For example, when $\boldsymbol{A}$ is symmetric, to obtain a better low-rank approximation one can replace $\boldsymbol{Q}$ in Algorithm 1 with an orthonormal basis for range($\boldsymbol{A}^q\boldsymbol{\Omega}$) for some $q \geq 0$, which comes at the cost of more matrix-vector products with $\boldsymbol{A}$. Performing q subspace iterations on $\boldsymbol{A}$ will make $\boldsymbol{Q}$ more closely aligned with the dominant singular vectors, which, in view of (2.2), makes $\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A}$ closer an optimal low-rank approximation; this has been discussed in [79, Section 4.5] and [76, 150]. One has to be careful with how one obtains the orthonormal basis $\boldsymbol{Q}$ for range($\boldsymbol{A}^q\boldsymbol{\Omega}$), as a naive implementation can be numerically unstable. Algorithm 2 provides a numerically stable way of computing an orthonormal basis for range($\boldsymbol{A}^q\boldsymbol{\Omega}$).

Algorithm 2 Subspace iteration

```
input: Symmetric $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Number of subspace iterations $q \geq 0$. Sketch matrix $\boldsymbol{\Omega}$.
output: Orthonormal basis for range($\boldsymbol{A}^q\boldsymbol{\Omega}$)
Compute a thin QR decomposition of $\boldsymbol{\Omega} = \boldsymbol{Q}\boldsymbol{R}$.
for $q_{\text{count}} = 1, \ldots, q$ do
    $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Q}$
    Compute a thin QR decomposition of $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$.
end for
return $\boldsymbol{Q}$.
```

Once an orthonormal basis $\boldsymbol{Q}$ for range$(\boldsymbol{A}^q\boldsymbol{\Omega})$ is obtained, one can show that the approximation $\boldsymbol{P}_{\boldsymbol{A}^q\boldsymbol{\Omega}}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{A}^q\boldsymbol{\Omega}} = \boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}\boldsymbol{Q}^T$ satisfies a similar bound as Theorem 2.3 and Theorem 2.6.

Theorem 2.8. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be symmetric and have an eigenvalue decomposition as partitioned in* (2.6) *and let $\boldsymbol{\Omega}$ be a sketch matrix. Define*

$$\boldsymbol{\Omega}_1 = \boldsymbol{U}_1^T\boldsymbol{\Omega}, \qquad \boldsymbol{\Omega}_2 = \boldsymbol{U}_2^T\boldsymbol{\Omega}, \tag{2.8}$$

and assume that $\operatorname{rank}(\boldsymbol{\Omega}_1) = k$. *Then,*

$$\begin{aligned}\|\boldsymbol{A} - \boldsymbol{P}_{\boldsymbol{A}^q\boldsymbol{\Omega}}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{A}^q\boldsymbol{\Omega}}\|_{\mathrm{F}}^2 \leq & \|\boldsymbol{A} - (\boldsymbol{P}_{\boldsymbol{A}^q\boldsymbol{\Omega}}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{A}^q\boldsymbol{\Omega}})_{(k)}\|_{\mathrm{F}}^2 \leq \\ & \|\boldsymbol{\Lambda}_2\|_{\mathrm{F}}^2 + 5\left|\frac{\lambda_{k+1}}{\lambda_k}\right|^{2(q-1)} \|\boldsymbol{\Lambda}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\|_{\mathrm{F}}^2.\end{aligned}$$

Theorem 2.8 is an immediate corollary of a more general result that will be proven in Chapter 5; see Remark 5.1. Furthermore, to obtain statistical bounds one only require bounds for $\|\boldsymbol{\Lambda}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\|_{\mathrm{F}}^2$, which are given in Lemma 2.4 when $\boldsymbol{\Omega}$ is a standard Gaussian matrix.

Intuitively, the reason why subspace iteration improves the low-rank approximation is because the polynomial x^q is small on the small eigenvalues of $\boldsymbol{A}$ and large on the large eigenvalues of $\boldsymbol{A}$. Hence, the polynomial x^q effectively denoises the contribution from the small eigenvalues. However, there are potentially much better polynomials that achieve this task. For example, scaled and shifted Chebyshev polynomials are usually much better at denoising the contribution from the small eigenvalues of $\boldsymbol{A}$, since Chebyshev polynomials are very small on $[-1, 1]$ and grow very quickly outside this interval; a fact that has been frequently used to analyse Krylov subspace methods in the context of eigenvalue problems [71, 96, 101]. Therefore, in order to obtain even better low-rank approximations, one can allow $\boldsymbol{Q}$ to be an orthonormal basis for a block-Krylov subspace range$\left(\begin{bmatrix}\boldsymbol{\Omega} & \boldsymbol{A}\boldsymbol{\Omega} & \cdots & \boldsymbol{A}^{q-1}\boldsymbol{\Omega}\end{bmatrix}\right)$, which contains range$(g(\boldsymbol{A})\boldsymbol{\Omega})$ for any polynomial g of degree at most $q-1$. $\boldsymbol{Q}$ can be obtained using the block Lanczos algorithm, which will be outlined in Chapter 3. The use of Krylov subspace methods in the context of low-rank approximation has been studied in [109, 115, 150], all of which make heavy use of properties of scaled and shifted Chebyshev polynomials. For example, one can show the following result, which is very similar to [150, Theorem 9.2], but it allows for truncation and a two-sided projection.

Theorem 2.9. *Consider the setting of Theorem 2.8 and let* $\boldsymbol{K} = \begin{bmatrix}\boldsymbol{\Omega} & \boldsymbol{A}\boldsymbol{\Omega} & \cdots & \boldsymbol{A}^q\boldsymbol{\Omega}\end{bmatrix}$.

Then, if $\gamma = \frac{|\lambda_k|-|\lambda_{k+1}|}{|\lambda_k|+|\lambda_{k+1}|}$

$$\|\boldsymbol{A} - \boldsymbol{P_K}\boldsymbol{A}\boldsymbol{P_K}\|_{\mathrm{F}}^2 \leq \|\boldsymbol{A} - (\boldsymbol{P_K}\boldsymbol{A}\boldsymbol{P_K})_{(k)}\|_{\mathrm{F}}^2 \leq \\ \|\boldsymbol{\Lambda}_2\|_{\mathrm{F}}^2 + 20e^{-4(q-1)\sqrt{\gamma}}\|\boldsymbol{\Lambda}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\|_{\mathrm{F}}^2.$$

Remark 2.2. *The matrix* $\boldsymbol{K}$ *in Theorem 2.9 is never explicitly formed. In practice, one uses the block-Lanczos algorithm to construct an orthonormal basis* $\boldsymbol{Q}$ *for* $\mathrm{range}(\boldsymbol{K})$ *and the matrix* $\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}$*; see Section 3.3.*

Once again, Theorem 2.9 is an immediate corollary of a more general result that will be proven in Chapter 5; see Theorem 5.9 for the proof. Furthermore, as discussed before, to obtain statistical bounds one only require bounds on $\|\boldsymbol{\Lambda}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\|_{\mathrm{F}}^2$, which are given in Lemma 2.4 when $\boldsymbol{\Omega}$ is a standard Gaussian matrix.

2.3 The Nyström approximation

For SPSD matrices, it is usually preferable to use the Nyström approximation instead of the randomized SVD. One can show that with the same computational cost, the Nyström approximation can always return a more accurate approximation to $\boldsymbol{A}$ compared to the randomized SVD; more details will be given in Chapter 4. The Nyström approximation will play a central role in this thesis, and will be discussed in Chapters 4, 8 and 9. For a SPSD matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, the Nyström approximation with respect to a matrix $\boldsymbol{Q}$ is defined as [104, Section 14]

$$\boldsymbol{A} \approx \widehat{\boldsymbol{A}} := \boldsymbol{A}\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q})^{\dagger}\boldsymbol{Q}^T\boldsymbol{A} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{U}}^T, \tag{2.9}$$

where $\widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{U}}^T$ is the eigenvalue decomposition of $\widehat{\boldsymbol{A}}$ and can be obtained using Algorithm 3.

Algorithm 3 Nyström approximation

```
input: SPSD $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Matrix $\boldsymbol{Q} \in \mathbb{R}^{n\times \ell}$.
output: Nyström approximation to $\boldsymbol{A}$ in factored form $\widehat{\boldsymbol{A}} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{U}}^T$.
1: Compute a matrix $\boldsymbol{C}$ so that $\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{C}^T\boldsymbol{C}$.
2: Set $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{Q}\boldsymbol{C}^{\dagger}$
3: Compute the SVD of $\boldsymbol{B} = \widehat{\boldsymbol{U}}\boldsymbol{\Sigma}\boldsymbol{W}^T$. Set $\widehat{\boldsymbol{\Lambda}} = \boldsymbol{\Sigma}^2$
4: return $\widehat{\boldsymbol{U}}, \widehat{\boldsymbol{\Lambda}}$.
```

Remark 2.3. *In our numerical experiments, we let* $\boldsymbol{C}$ *be a square-root of* $\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}$ *obtained by an eigenvalue decomposition, or we let* $\boldsymbol{C}$ *be the Cholesky factor of* $\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}$*. However, in exact arithmetic,* any $\boldsymbol{C}$ *satisfying* $\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{C}^T\boldsymbol{C}$ *suffices, even a rectangular* $\boldsymbol{C}$*. Furthermore, the Nyström approximation can be prone to numerical issues due to the appearance of the pseudo-inverse in line 2. To mitigate numerical issues when* $\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}$

in line 1 is highly ill-conditioned, in our implementation we use different regularization techniques. The first regularization technique is to compute the ϵ-pseudoinverse of $\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}$ where $\epsilon = 5 \cdot 10^{-16} \cdot \|\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}\|_2$. The second alternative is to compute the Nyström approximation of the shifted matrix $\boldsymbol{A} + \epsilon\|\boldsymbol{A}\boldsymbol{Q}\|_2\boldsymbol{I}$, where ϵ is the machine precision. Then in the final step we shift back $\boldsymbol{\Lambda} \to \max\{\boldsymbol{\Lambda} - \epsilon\|\boldsymbol{A}\boldsymbol{Q}\|_2\boldsymbol{I}, 0\}$. The latter technique has been described in detail in [104, Algorithm 16]. Both techniques work well in practice.

The Nyström approximation of $\boldsymbol{A}$ defined in (2.9) depends only on range($\boldsymbol{Q}$) [151, Proposition A.2], and we may therefore assume that $\boldsymbol{Q}$ has orthonormal columns. In the ideal case when $\boldsymbol{Q}$ spans the dominant eigenvectors to $\boldsymbol{A}$ the Nyström approximation (2.9) returns an optimal low-rank approximation. So, the goal is to efficiently find $\boldsymbol{Q}$ that approximately spans these vectors. For example, $\boldsymbol{Q}$ might be chosen to be an orthonormal basis for range($\boldsymbol{A}^q\boldsymbol{\Omega}$) or a Krylov subspace range $\left(\begin{bmatrix}\boldsymbol{\Omega} & \boldsymbol{A}\boldsymbol{\Omega} & \dots & \boldsymbol{A}^q\boldsymbol{\Omega}\end{bmatrix}\right)$ for some $q \geq 0$ and random sketching matrix $\boldsymbol{\Omega}$. However, in its original form, $\boldsymbol{Q}$ consisted of a carefully chosen subset of columns of the identity matrix. This form of Nyström approximation frequently appears in applications involving kernel matrices; see [12, 32, 38, 40, 46, 116, 163, 167]. Analyses for various choices of $\boldsymbol{Q}$ can be found in [68, 124, 150, 151].

There is a close relationship between the randomized SVD and the Nyström approximation. First note that since $\boldsymbol{A}$ is SPSD it has a unique square root. Hence, by the definition of orthogonal projectors we can write

$$\widehat{\boldsymbol{A}} = \boldsymbol{A}^{1/2}\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}^{1/2} = (\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}^{1/2})^T(\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}^{1/2}); \tag{2.10}$$

see e.g. [68, Equation (4)]. Thus,

$$\boldsymbol{A} - \widehat{\boldsymbol{A}} = \boldsymbol{A}^{1/2}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}})\boldsymbol{A}^{1/2}. \tag{2.11}$$

Hence, for any Schatten norm $\|\cdot\|_{(s)}$ we have

$$\|\boldsymbol{A} - \widehat{\boldsymbol{A}}\|_{(s)} = \|\boldsymbol{A}^{1/2}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}})\boldsymbol{A}^{1/2}\|_{(s)} = \|\boldsymbol{A}^{1/2} - \boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}^{1/2}\|_{(2s)}^2.$$

Consequently, bounds for the randomized SVD in $\|\cdot\|_{(2s)}$ can be turned into bounds for the Nyström approximation in $\|\cdot\|_{(s)}$. In particular, we have the following theorem, which was proven in [68, 150, 151].

Theorem 2.10. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be SPSD and let $\boldsymbol{Q}$ be an orthonormal basis for the range of a random $n \times (k+p)$ standard Gaussian matrix. If $p \geq 2$ we have*

$$\mathbb{E}\|\boldsymbol{A} - \widehat{\boldsymbol{A}}\|_* \leq \left(1 + \frac{k}{p-1}\right)\|\boldsymbol{\Lambda}_2\|_*.$$

Bounds for other norms exist, see e.g. [68], and will also be covered in Chapter 4 and Chapter 9.

2.3.1 Truncating back

As discussed for the randomized SVD, if k is the target rank $\boldsymbol{Q}$ often has more than k columns. This means that $\widehat{\boldsymbol{A}}$ has a higher rank than k. In order to recover an exactly rank k approximation, we would return the best rank k approximation to $\widehat{\boldsymbol{A}}$, denoted as $\widehat{\boldsymbol{A}}_{(k)}$, instead of $\widehat{\boldsymbol{A}}$ itself. The matrix $\widehat{\boldsymbol{A}}_{(k)}$ can be computed efficiently using Algorithm 3 by truncating the eigenvalue decomposition of $\widehat{\boldsymbol{A}}$. As with the randomized SVD, the truncated version of the Nyström approximation satisfies similar guarantees as the untruncated Nyström approximation. In particular, we have the following theorem, which is proven in a very similar fashion to Theorem 2.6.

Theorem 2.11 ([151, Theorem 4.1])**.** *Consider the setting of Theorem 2.10. If* $p \geq 2$ *we have*

$$\mathbb{E}\|\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}\|_* \leq \left(1 + \frac{k}{p-1}\right)\|\boldsymbol{\Lambda}_2\|_*.$$

3 An introduction to low-rank approximation of matrix functions

This chapters serves as an introduction to low-rank approximation of matrix functions. We begin with recalling the definition of matrix functions.

Definition 3.1 (Matrix function of a symmetric matrix $\boldsymbol{A}$)**.** *Given a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ with eigenvalue decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, and a scalar function f defined on the eigenvalues of $\boldsymbol{A}$. Then, the matrix function $f(\boldsymbol{A})$ is defined as*

$$f(\boldsymbol{A}) := \boldsymbol{U} f(\boldsymbol{\Lambda}) \boldsymbol{U}^T,$$

where $f(\boldsymbol{\Lambda}) = \operatorname{diag}(f(\lambda_1), \ldots, f(\lambda_n))$.

This definition extends to non-symmetric matrices. However, in this thesis we will be concerned with matrix functions of symmetric matrices, and we therefore omit a discussion on the non-symmetric case. We refer to the famous book by Nicholas Higham for a more detailed discussion [84].

Matrix functions are ubiquitous in applied mathematics. Consequently, a lot of research has been devoted to computing matrix functions, approximating matrix-vector products with matrix functions, and estimating quantities associated with matrix functions. In the next two sections we outline a few different examples when low-rank approximations to matrix functions can bring computational advantages. In Section 3.1, we present two applications when low-rank approximation can be beneficial in approximating matrix-vector products with $f(\boldsymbol{A})$. In Section 3.2, we briefly present six examples of when low-rank approximation can be beneficial in trace and diagonal estimation of $f(\boldsymbol{A})$. In Section 3.3, we will outline the challenges with computing low-rank approximations to matrix functions.

3.1 Computing matrix-vector products with matrix functions

Many applications involve computing matrix-vector products with a matrix-function $f(\boldsymbol{A})$. Explicitly computing $f(\boldsymbol{A})$ generally requires $O(n^3)$ operations using standard algorithms [70] and becomes prohibitively expensive for large n. However, approximating matrix-vector products using Krylov subspace methods [69, 78, 84] generally costs $O(n^2 d)$ operations, where d is a parameter controlling the accuracy; a more detailed outline of Krylov subspace methods will be given Section 3.3. Unfortunately, in some cases the Krylov subspace method may converge slowly, resulting in a large d. When $f(\boldsymbol{A})$ admits an accurate rank k approximation $\boldsymbol{B} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^T$, where $\widehat{\boldsymbol{\Sigma}} \in \mathbb{R}^{k\times k}$, one can cheaply and accurately approximate matrix-vector products with $f(\boldsymbol{A})$ using the approximation

$$f(\boldsymbol{A})\boldsymbol{x} \approx \boldsymbol{B}\boldsymbol{x} = \widehat{\boldsymbol{U}}\left[\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{V}}^T\boldsymbol{x}\right)\right],$$

which, costs only $O(nk)$ operations. This compares favorably to the $O(n^2 d)$ operations required by Krylov subspace methods, even for moderate d. Below we will give two examples of when computing low-rank approximations to matrix-functions can dramatically reduce the computational cost of computing matrix-vector products with $f(\boldsymbol{A})$.

3.1.1 Differential equations

One notable example of an application to matrix functions is differential equations; see e.g [48, 82]. Consider the following ordinary differential equation (ODE)

$$\begin{aligned} \dot{\boldsymbol{u}}(t) &= \boldsymbol{A}\boldsymbol{u}(t) \\ \boldsymbol{u}(0) &= \boldsymbol{u}_0, \end{aligned} \tag{3.1}$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{u}_0 \in \mathbb{R}^n$. The ODE (3.1) may arise from, for example, a discretization of a partial differential equation $u_t = \mathcal{L}u$, where $\boldsymbol{A}$ represents the discretization of the (space-)differential operator $\mathcal{L}$ and $\boldsymbol{u}$ denotes the discretization in space of the solution $u = u(x,t)$. It is well-known that the solution to (3.1) is

$$\boldsymbol{u}(t) = \exp(t\boldsymbol{A})\boldsymbol{u}_0; \tag{3.2}$$

see e.g. [87]. Because the relative eigenvalue gaps of $\exp(t\boldsymbol{A})$ can be much larger than the relative eigenvalue gaps of $\boldsymbol{A}$, in some cases, $\exp(t\boldsymbol{A})$ admits an accurate relative low-rank approximation, even if $\boldsymbol{A}$ does not admit an accurate low-rank approximation. For example, consider the following parabolic differential equation, which is inspired by the numerical experiments in [5, 49, 138]

$$\begin{aligned} &u_t = \kappa\Delta u + \lambda u \text{ in } [0,1]^2 \times [0,2] \\ &u(\cdot,0) = \theta \text{ in } [0,1]^2 \\ &u = 0 \text{ on } \Gamma_1 \\ &\frac{\partial u}{\partial \boldsymbol{n}} = 0 \text{ on } \Gamma_2, \end{aligned} \tag{3.3}$$

for $\kappa, \lambda > 0$ and $\Gamma_2 = \{(x,1) \in \mathbb{R}^2 : x \in [0,1]\}$ and $\Gamma_1 = \partial[0,1]^2 \setminus \Gamma_2$. Discretizing (3.3) in space using finite differences on 40×40 equispaced grid yields an ODE of the form (3.1) and can be solved using (3.2). In fact, in this case the first 100 singular values of $\exp(t\boldsymbol{A})$ decay exponentially; for $\kappa = 0.01$ and $\lambda = 1$ the i^{th} singular value of $\exp(t\boldsymbol{A})$ is approximately $0.876^{t(i-1)}\|\exp(t\boldsymbol{A})\|_2$ for $i \leq 100$. Therefore, $\exp(t\boldsymbol{A})$ admits an accurate relative low-rank approximation for sufficiently large t. Hence, unless t is too small, we can compute an accurate low-rank approximation to $\exp(t\boldsymbol{A})$ and efficiently solve (3.3) by approximating (3.2) using our low-rank approximation.

3.1.2 Sampling from elliptical distributions

Computing matrix-vector products with matrix functions also arises in statistics. For example, sampling from non-standard Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{A})$ [114] is a task that arises when sampling from (discretized) Gaussian processes. It is well-known that if $\boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{C}$ is a matrix so that $\boldsymbol{C}\boldsymbol{C}^T = \boldsymbol{A}$, then $\boldsymbol{\mu} + \boldsymbol{C}\boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{A})$. Hence, a possible choice is $\boldsymbol{C} = \boldsymbol{A}^{1/2}$. Once again, one can approximate matrix-vector products with $\boldsymbol{A}^{1/2}$ using Krylov subspace methods, such as the Lanczos method [84, Chapter 13]. As the error of Lanczos is linked to polynomial approximations of f [135, Proposition 6.3], one may observe slow convergence when $f(x) = \sqrt{x}$ and $\boldsymbol{A}$ has eigenvalues close to 0. To avoid this, one could resort to rational approximations such as rational Krylov subspace methods [78] or quadrature methods [127], but these methods require the solution of a (shifted) linear system with $\boldsymbol{A}$ in every iteration, which comes with challenges on its own. On the other hand, as previously mentioned, if we have access to an accurate SPSD low-rank approximation $\boldsymbol{B}^{1/2}$ to $\boldsymbol{A}^{1/2}$, we can cheaply approximate matrix-vector products with $\boldsymbol{A}^{1/2}$ by computing matrix-vector products with $\boldsymbol{B}^{1/2}$. In fact, for accurate low-rank approximations one can show that one has a small mean-squared error: if $\boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, $\boldsymbol{\psi} = \boldsymbol{\mu} + \boldsymbol{A}^{1/2}\boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{A})$, and $\widehat{\boldsymbol{\psi}} = \boldsymbol{\mu} + \boldsymbol{B}^{1/2}\boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{B})$ then the mean-squared error equals

$$\mathbb{E}\|\boldsymbol{\psi} - \widehat{\boldsymbol{\psi}}\|_2^2 = \|\boldsymbol{A}^{1/2} - \boldsymbol{B}^{1/2}\|_{\text{F}}^2, \tag{3.4}$$

which is small for accurate low-rank approximation $\boldsymbol{B}^{1/2}$. The same technique can be used to sample from a general elliptical distribution [106]; for a general elliptical distribution the mean-squared error equals $c\|\boldsymbol{A}^{1/2} - \boldsymbol{B}^{1/2}\|_{\text{F}}^2$ where c is a constant depending on the elliptical distribution.

Furthermore, for accurate approximations $\boldsymbol{B}^{1/2}$ one also has a small Wasserstein 2-distance [158]. To see this, first recall for two probability distributions ν_1 and ν_2 the Wasserstein distance is defined as

$$\mathcal{W}_2(\nu_1,\nu_2)^2 = \inf_{\boldsymbol{X}_1\sim\nu_1,\boldsymbol{X}_2\sim\nu_2} \mathbb{E}\|\boldsymbol{X}_1-\boldsymbol{X}_2\|_2^2.$$

Therefore, if $\nu_1 = \mathcal{N}(\boldsymbol{\mu},\boldsymbol{A})$ and $\nu_2 = \mathcal{N}(\boldsymbol{\mu},\boldsymbol{B})$ we have

$$\mathcal{W}_2(\nu_1,\nu_2)^2 \leq \mathbb{E}\|\boldsymbol{\psi}-\widehat{\boldsymbol{\psi}}\|_2^2 = \|\boldsymbol{A}^{1/2}-\boldsymbol{B}^{1/2}\|_{\mathrm{F}}^2,$$

where $\boldsymbol{\psi}$ and $\widehat{\boldsymbol{\psi}}$ are as in (3.4). Hence, $\mathcal{W}_2(\nu_1,\nu_2)$ is small if $\boldsymbol{B}^{1/2}$ is an accurate low-rank approximations to $\boldsymbol{A}^{1/2}$. In fact, one can show stronger results for general elliptical distributions; see Theorem 4.11. The matrix square-root will be discussed further in Chapter 4.

3.2 Trace and diagonal estimation of matrix functions

Quite a few applications of matrix functions only require quantities associated with $f(\boldsymbol{A})$ instead of the full matrix function. Notable examples include the trace $\operatorname{tr}(f(\boldsymbol{A}))$ and the diagonal elements of $f(\boldsymbol{A})$, which can be estimated with Monte Carlo methods [16, 44, 20, 41, 80, 154, 155]. In recent years, there has been increased attention to the use of randomized low-rank approximation techniques in this context, for estimating these quantities [99, 138] or as a variance reduction technique for Monte Carlo methods [36, 92, 111, 123]. We will provide a more detailed discussion on trace estimation in Chapter 6. Below we outline a few examples when approximating $\operatorname{tr}(f(\boldsymbol{A}))$ or the diagonal entries of $f(\boldsymbol{A})$ are important. As will be discussed in Chapter 6, this is a task that greatly benefits from low-rank approximation.

3.2.1 Nuclear norm

The matrix square root also appears when estimating the nuclear norm $\|\boldsymbol{X}\|_*$ of a matrix [154, 155]. Because of the relation

$$\|\boldsymbol{X}\|_* = \operatorname{tr}(\boldsymbol{A}^{1/2}), \quad \text{where } \boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{X},$$

nuclear norm estimation is equivalent to performing trace estimation on $\boldsymbol{A}^{1/2}$.

3.2.2 Statistical learning

The matrix function $\log(\boldsymbol{I}+\boldsymbol{A})$ frequently appears in statistical learning [65, 168] and in Bayesian inverse problems [138]. In these applications one typically aims at estimating $\log\det(\boldsymbol{I}+\boldsymbol{A}) = \operatorname{tr}(\log(\boldsymbol{I}+\boldsymbol{A}))$.

3.2.3 Effective dimension

The effective dimension $d_{\text{eff}}(\mu)$, also called statistical dimension, is defined as

$$d_{\text{eff}}(\mu) = \operatorname{tr}(f_\mu(\boldsymbol{A})), \quad f_\mu(x) = \frac{x}{x+\mu}, \quad \mu > 0.$$

This quantity appears in kernel learning [4, 9, 10] and inverse problems [107]. The effective dimension can once again be estimated using trace estimation. Another important quantity is the diagonal of $f_\mu(\boldsymbol{A})$; its entries are called the Ridge leverage scores.

3.2.4 Triangle counting

Given the adjacency matrix $\boldsymbol{A}$ for a graph, then $\frac{1}{6}\operatorname{tr}(\boldsymbol{A}^3)$ is equal to the number of triangles in the graph. Counting the number of triangles in a graph is an important task in data mining applications [8], and can be done using trace estimation.

3.2.5 The matrix exponential

Estimating $\operatorname{tr}(\exp(\boldsymbol{A}))$ is a task that appears in several applications. In mathematical biology, for an undirected graph with adjacency matrix $\boldsymbol{A}$, the Estrada index is defined as $\operatorname{tr}(\exp(\boldsymbol{A}))$. The Estrada index is a measure of the degree of protein folding [54] and it frequently appears in network analysis [55]. Furthermore, estimating the partition function $Z(\beta) = \operatorname{tr}(\exp(-\beta\boldsymbol{A}))$ is an important task in quantum mechanics [141]. In addition, the diagonal entries of $\exp(\boldsymbol{A})$ is important in measuring the centrality of the nodes in a graph and it is called the *exponential subgraph centrality* [7, 23, 24, 43].

3.2.6 The inverse

Estimating the diagonal entries of the matrix inverse $\boldsymbol{A}^{-1}$ is important in uncertainty quantification [21, 130]. Furthermore, the diagonal entries of the inverse are also a measure of centrality of the nodes in a graph and is called the *resolvent subgraph centrality* [7, 23, 24].

3.3 Challenges with computing low-rank approximations of matrix functions

As established, low-rank approximations of $f(\boldsymbol{A})$ are useful. However, most existing algorithms for computing them require at least some access to the matrix $f(\boldsymbol{A})$, for example in the form of matrix-vector products. This is true of the randomized SVD and the Nyström approximation introduced in Chapter 2. Since in general we cannot directly compute matrix-vector products with $f(\boldsymbol{A})$ we need to resort to approximating them. This can be done, for example, with the (block-)Lanczos method [84, 69] or rational Krylov subspace methods [78]. Given a $n \times b$ matrix $\boldsymbol{\Omega}$, the block-Lanczos algorithm

(Algorithm 4) can be used to iteratively obtain an orthonormal basis for the Krylov subspace

$$\mathcal{K}_d(\boldsymbol{A}, \boldsymbol{\Omega}) = \operatorname{range}\left(\begin{bmatrix} \boldsymbol{\Omega} & \boldsymbol{A\Omega} & \cdots & \boldsymbol{A}^{d-1}\boldsymbol{\Omega} \end{bmatrix}\right).$$

In particular, using at most db matrix vector products with $\boldsymbol{A}$, the algorithm produces a basis $\boldsymbol{Q}_d$ and a block-tridiagonal matrix $\boldsymbol{T}_d$

$$\boldsymbol{Q}_d = \begin{bmatrix} \boldsymbol{V}_0 & \boldsymbol{V}_1 & \cdots & \boldsymbol{V}_{d-1} \end{bmatrix}, \quad \boldsymbol{T}_d = \boldsymbol{Q}_d^T \boldsymbol{A} \boldsymbol{Q}_d = \operatorname{tridiag}\begin{pmatrix} & \boldsymbol{R}_1^T & \cdots & \boldsymbol{R}_{d-1}^T & \\ \boldsymbol{M}_1 & & \cdots & \cdots & \boldsymbol{M}_d \\ & \boldsymbol{R}_1 & \cdots & \boldsymbol{R}_{d-1} & \end{pmatrix}, \tag{3.5}$$

where $\boldsymbol{R}_0$ is a quantity also output by the algorithm and is given by the decomposition $\boldsymbol{\Omega} = \boldsymbol{V}_0 \boldsymbol{R}_0$, where $\boldsymbol{V}_0$ is an orthonormal basis for range($\boldsymbol{\Omega}$).

Algorithm 4 Block-Lanczos Algorithm

```
input: Symmetric $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. Matrix $\boldsymbol{\Omega} \in \mathbb{R}^{n \times b}$. Number of iterations $d$.
output: Orthonormal basis $\boldsymbol{Q}_d$ for $\mathcal{K}_d(\boldsymbol{A}, \boldsymbol{\Omega})$, and block tridiagonal $\boldsymbol{T}_d$.
Compute an orthonormal basis $\boldsymbol{V}_0$ for range($\boldsymbol{\Omega}$) and $\boldsymbol{R}_0 = \boldsymbol{V}_0^T \boldsymbol{\Omega}$.
for $i = 1, \dots, d$ do
    $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{V}_{i-1} - \boldsymbol{V}_{i-2}\boldsymbol{R}_{i-1}^T$            ▷ $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{V}_{i-1}$ if $i = 1$
    $\boldsymbol{M}_i = \boldsymbol{V}_{i-1}^T \boldsymbol{Y}$
    $\boldsymbol{Y} = \boldsymbol{Y} - \boldsymbol{V}_{i-1}\boldsymbol{M}_i$
    $\boldsymbol{Y} = \boldsymbol{Y} - \sum_{j=0}^{i-1} \boldsymbol{V}_j \boldsymbol{V}_j^T \boldsymbol{Y}$            ▷ reorthogonalize (repeat as needed)
    Compute an orthonormal basis $\boldsymbol{V}_i$ for range($\boldsymbol{Y}$) and $\boldsymbol{R}_i = \boldsymbol{V}_i^T \boldsymbol{Y}$.
end for
return $\boldsymbol{Q}_d = \begin{bmatrix} \boldsymbol{V}_0 & \boldsymbol{V}_1 & \cdots & \boldsymbol{V}_{d-1} \end{bmatrix}$ and the block-tridiagonal matrix $\boldsymbol{T}_d = \boldsymbol{Q}_d^T \boldsymbol{A} \boldsymbol{Q}_d$
as in (3.5).
```

The block-Lanczos algorithm can be used to approximate matrix-vector products and quadratic forms with $f(\boldsymbol{A})$ using the approximations

$$\boldsymbol{Q}_d f(\boldsymbol{T}_d)_{:,1:b} \boldsymbol{R}_0 \approx f(\boldsymbol{A})\boldsymbol{\Omega}, \tag{3.6}$$

$$\boldsymbol{R}_0^T f(\boldsymbol{T}_d)_{1:b,1:b} \boldsymbol{R}_0 \approx \boldsymbol{\Omega}^T f(\boldsymbol{A})\boldsymbol{\Omega}, \tag{3.7}$$

where we $f(\boldsymbol{T}_d)_{:,1:b}$ is the submatrix consisting of the first b columns and $f(\boldsymbol{T}_d)_{1:b,1:b}$ is the submatrix consisting of the first b rows and columns.[1] If f is a low-degree polynomial, then the approximations (3.6) and (3.7) are exact.[2]

Lemma 3.1 ([37, Lemma 2.1]). *The approximation* (3.6) *is exact if f is a polynomial of degree at most $d-1$, and the approximation* (3.7) *is exact if f is a polynomial of degree at most $2d-1$.*

[1] (3.6) and (3.7) are written out under the assumption that $\boldsymbol{\Omega}$ has rank b. If rank($\boldsymbol{\Omega}$) $= r < b$, then the index set $1 : b$ should be replaced with $1 : r$ in both (3.6) and (3.7).

[2] In fact, a stronger result for multipolynomials holds, see [58, Theorem 2.7].

Consequently, (3.6) and (3.7) are good approximations if f is well approximated by polynomials. In particular, one can obtain bounds in terms of the best polynomial approximation to f on $[\lambda_{\min}, \lambda_{\max}]$; see e.g. [134, Lemma 4.1]. Such a bound will be given in Lemma 5.5.

However, as previously mentioned, such methods may converge slowly for "difficult" functions and therefore require many matrix-vector products with $\boldsymbol{A}$ to accurately approximate matrix-vectors products with $f(\boldsymbol{A})$. Therefore, we will investigate alternative methods to more efficiently compute low-rank approximations of matrix functions.

3.4 Contributions

In Chapter 4 we will study a method called *funNyström*, which computes a low-rank approximation of a special class of matrix-functions called operator monotone. This method does not require any matrix-vector products with $f(\boldsymbol{A})$. Instead, it constructs a low-rank approximation to $f(\boldsymbol{A})$ immediately from a low-rank approximation to $\boldsymbol{A}$.

In Chapter 5 we will investigate an alternative method for general matrix-functions. This method is effectively a more efficient version of the randomized SVD applied to a matrix-function with matrix-vector products approximated using the block-Lanczos method. This algorithm, which was initially presented in the context of trace estimation by Chen and Hallman [37], is derived by taking into account that matrix-vector products with $f(\boldsymbol{A})$ are computed using a Krylov subspace method. From such introspection, one can derive a significantly more efficient version of the randomized SVD. One of our contribution is to provide an analysis of the method.

4 funNyström: Low-rank approximation of operator monotone matrix functions

In this chapter we present and analyze funNyström, a method to compute low-rank approximations of operator monotone matrix functions. Throughout this chapter we consider a SPSD matrix $\boldsymbol{A}$ with eigenvalue decomposition, and equivalently an SVD, as partitioned in (2.6) where the eigenvalues are ordered as $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n \geq 0$. In this chapter, $\widehat{\boldsymbol{A}}$ usually denotes an approximation to $\boldsymbol{A}$ so that $\boldsymbol{A} \succeq \widehat{\boldsymbol{A}} \succeq \boldsymbol{0}$. In particular, $\widehat{\boldsymbol{A}}$ could be a Nyström approximation of the form (2.9) since it satisfies $\boldsymbol{A} \succeq \widehat{\boldsymbol{A}} \succeq \boldsymbol{0}$ as a consequence of (2.11). By $\widehat{\lambda}_i$ we denote the i^{th} eigenvalue of $\widehat{\boldsymbol{A}}$.

This chapter is outlined as follows: In Section 4.1 we introduce operator monotone matrix functions and the funNyström approximation. Sections 4.2 to 4.5 are concerned with the theoretical results. We conclude with the numerical experiments in Section 4.6.

This chapter is based on the work in [124, 125]. The Sections 4.2.4, 4.4 and 4.5 contain a few new results that are not presented in [124, 125].

4.1 Operator monotone functions and the funNyström approximation

As discussed in Chapter 3, Krylov subspace methods to compute matrix-vector products with $f(\boldsymbol{A})$ may converge slowly and therefore require many matrix-vector products with $\boldsymbol{A}$ to accurately approximate matrix-vectors products with $f(\boldsymbol{A})$. Consequently, the cost of obtaining a low-rank approximation to $f(\boldsymbol{A})$ can be significantly higher than obtaining a low-rank approximation to $\boldsymbol{A}$.

However, for non-negative monotonically increasing functions of SPSD matrices it is possible to obtain a low-rank approximation of $f(\boldsymbol{A})$ directly from a low-rank approximation of $\boldsymbol{A}$. This is a key observation that potentially allows us to completely bypass the need for performing matrix-vector products with $f(\boldsymbol{A})$. The following basic lemma provides a first result in this direction, for the special case of *best* low-rank approximations (with respect to a unitarily invariant norm).

Lemma 4.1. *Consider a SPSD matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ with eigenvalue decomposition partitioned as* (2.6) *and a best rank k approximation $\boldsymbol{A}_{(k)} = \boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1^T$. Then, for monotonically increasing $f : [0,\infty) \mapsto [0,\infty)$ it holds that $f(\boldsymbol{A}_{(k)})_{(k)} = \boldsymbol{U}_1 f(\boldsymbol{\Lambda}_1)\boldsymbol{U}_1^T$ is a best rank-k approximation of $f(\boldsymbol{A})$.*

Proof. By the spectral decomposition of $\boldsymbol{A}$, we can write $\boldsymbol{A} = \boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1^T + \boldsymbol{U}_2\boldsymbol{\Lambda}_2\boldsymbol{U}_2^T$, with diagonal $\boldsymbol{\Lambda}_2$ and orthonormal $\boldsymbol{U}_2$. Because the first term is a best low-rank approximation $\boldsymbol{A}_{(k)}$, none of the eigenvalues of $\boldsymbol{\Lambda}_2$ is larger than any of the eigenvalues of $\boldsymbol{\Lambda}_1$. Because of monotonicity, the same statement holds for the relation between the eigenvalues of $f(\boldsymbol{\Lambda}_2)$ and $f(\boldsymbol{\Lambda}_1)$. Using the spectral decomposition $f(\boldsymbol{A}) = \boldsymbol{U}_1 f(\boldsymbol{\Lambda}_1)\boldsymbol{U}_1^T + \boldsymbol{U}_2 f(\boldsymbol{\Lambda}_2)\boldsymbol{U}_2^T$, this implies that $\boldsymbol{U}_1 f(\boldsymbol{\Lambda}_1)\boldsymbol{U}_1^T$ is an optimal low-rank approximation to $f(\boldsymbol{A})$. Furthermore, note that $f(\boldsymbol{A}_{(k)}) = \boldsymbol{U}_1 f(\boldsymbol{\Lambda}_1)\boldsymbol{U}_1^T + f(0)(\boldsymbol{I} - \boldsymbol{U}_1\boldsymbol{U}_1^T)$ and since for any $i = 1, \dots, k$ we have $f(\lambda_i) \geq f(0) \geq 0$ we know that $f(\boldsymbol{A}_{(k)})_{(k)} = \boldsymbol{U}_1 f(\boldsymbol{\Lambda}_1)\boldsymbol{U}_1^T$. □

The result of Lemma 4.1 is constrained to *best* rank-k approximations and does not extend to the quasi-optimal rank-k approximations $\widehat{\boldsymbol{A}}_{(k)}$ of $\boldsymbol{A}$ usually returned by (randomized) numerical algorithms. One still hopes that if $\widehat{\boldsymbol{A}}_{(k)}$ is a near-optimal rank k approximation to $\boldsymbol{A}$, then $f(\widehat{\boldsymbol{A}})_{(k)}$ remains a near-optimal rank k approximation to $f(\boldsymbol{A})$. A similar idea was used in [138], which analyzes approximations of the form $\operatorname{tr}(\log(\boldsymbol{I} + \widehat{\boldsymbol{A}})) \approx \operatorname{tr}(\log(\boldsymbol{I} + \boldsymbol{A}))$; see also [99]. In this chapter, we will present and analyze funNyström, a simple and effective method to compute low-rank approximations of a special class of functions known as operator monotone functions. Formally, we define an operator monotone function as follows.

Definition 4.1 (Operator monotone matrix function [25, p.112])**.** *A function f is called operator monotone if $\boldsymbol{B} \succeq \boldsymbol{C}$ for symmetric $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{n\times n}$ implies $f(\boldsymbol{B}) \succeq f(\boldsymbol{C})$, where $\boldsymbol{B} \succeq \boldsymbol{C}$ means that $\boldsymbol{B} - \boldsymbol{C}$ is SPSD; see [88, Definition 7.7.1].*

Trivially, any operator monotone function is monotonically increasing, but the converse is not true. For example, the functions $\exp(x)$ and x^2 are monotonically increasing on $[0,\infty)$, but not operator monotone. Examples of operator monotone functions include $\sqrt{x}, \log(1+x)$ and $\frac{x}{x+\lambda}$ for $\lambda > 0$ [25, Section V.1] and were discussed in Chapter 3.

In Section 4.2 we will show that if $\widehat{\boldsymbol{A}}$ is a matrix satisfying $\boldsymbol{A} \succeq \widehat{\boldsymbol{A}} \succeq \boldsymbol{0}$ and its rank k truncation $\widehat{\boldsymbol{A}}_{(k)} = \widehat{\boldsymbol{U}}_1\widehat{\boldsymbol{\Lambda}}_1\widehat{\boldsymbol{U}}_1^T$ is a near-optimal low-rank approximation to $\boldsymbol{A}$, then for any positive, continuous operator monotone function, $f(\widehat{\boldsymbol{A}})_{(k)} = \widehat{\boldsymbol{U}}_1 f(\widehat{\boldsymbol{\Lambda}}_1)\widehat{\boldsymbol{U}}_1^T$ is a near-optimal low-rank approximation to $f(\boldsymbol{A})$, *independently of how $\widehat{\boldsymbol{A}}$ was computed.* In particular, we show that if for some $\varepsilon \geq 0$

$$\|\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}\| \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|,$$

then for any operator monotone function $f : [0, \infty) \to [0, \infty)$ we have

$$\|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}})_{(k)}\| \leq (1 + \varepsilon)\|f(\boldsymbol{A}) - f(\boldsymbol{A})_{(k)}\|, \tag{4.1}$$

where $\|\cdot\|$ is the nuclear, Frobenius, or operator norm. Importantly, as discussed in the introduction to this chapter, any Nyström approximation of the form (2.9) satisfies $\boldsymbol{A} \succeq \widehat{\boldsymbol{A}} \succeq \boldsymbol{0}$, so our guarantees extend to all possible Nyström approximations to $\boldsymbol{A}$, no matter how it is obtained. Therefore, the approximation $f(\widehat{\boldsymbol{A}})_{(k)}$ is called the *funNyström* approximation. If $\widehat{\boldsymbol{A}}$ is the Nyström approximation as defined in (2.9), then the funNyström approximation can be computed using Algorithm 5. A strength of our results is that any guarantee for the Nyström approximation can immediately be turned into a guarantee for the funNyström approximation; more details will be given in Section 4.4. Furthermore, a major advantage of funNyström is that it only requires access to $\boldsymbol{A}$ and not with $f(\boldsymbol{A})$, and can therefore be significantly cheaper than naively implementing the Nyström approximation directly to $f(\boldsymbol{A})$ with matrix-vector products approximated using, for example, the (block-)Lanczos method.

In addition, to facilitate the application of these theorems, we present sufficient conditions for an orthonormal basis $\boldsymbol{Q}$ to produce a near-optimal Nyström approximation of the form (2.9). In particular, we show that if $\boldsymbol{Q} \in \mathbb{R}^{n \times \ell}$, where $\ell \geq k$, is an orthonormal basis satisfying

$$\|\boldsymbol{A} - (\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A})_{(k)}\| \leq (1 + \varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|, \tag{4.2}$$

then a Nyström approximation $\widehat{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q})^\dagger\boldsymbol{Q}^T\boldsymbol{A}$ satisfies

$$\|\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}\| \leq (1 + \varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|,$$

where $\|\cdot\|$ is the nuclear or Frobenius norm; for the operator norm we show that such result is impossible. Similar guarantees are proven in [150], but they are constrained to the case when $\boldsymbol{Q}$ has exactly k columns. In contrast, our results allow $\boldsymbol{Q}$ to have more than k columns. Guarantees of the form (4.2) are common in the literature; see e.g. [14, 15, 91, 109, 115]. These results allow us to translate these existing results into guarantees for the Nyström approximation. Then, by using guarantees of the form (4.1), these existing results for $\boldsymbol{Q}$ translate all the way into results for the funNyström approximation; more details are given in Section 4.4. In particular, this means that if $\boldsymbol{Q}$ satisfies (4.2), then Algorithm 5 returns a low-rank approximation of $f(\boldsymbol{A})$ satisfying (4.1) if $\|\cdot\| = \|\cdot\|_*$ or $\|\cdot\| = \|\cdot\|_{\mathrm{F}}$.

4.2 Good Nyström approximations imply good funNyström approximations

In this section we prove that near optimal Nyström approximations imply near optimal funNyström approximations for the nuclear, Frobenius, and operator norms. We also

Algorithm 5 funNyström approximation

```
input: SPSD $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. Matrix $\boldsymbol{Q} \in \mathbb{R}^{n \times \ell}$. Target rank $k$.
output: Rank $k$ funNyström approximation to $f(\boldsymbol{A})$ in factored form $f(\widehat{\boldsymbol{A}})_{(k)} = \widehat{\boldsymbol{U}}_1 f(\widehat{\boldsymbol{\Lambda}}_1) \widehat{\boldsymbol{U}}_1^T$.
1: Use Algorithm 3 to compute the Nyström approximation $\widehat{\boldsymbol{A}} = \widehat{\boldsymbol{U}} \widehat{\boldsymbol{\Lambda}} \widehat{\boldsymbol{U}}^T$
2: Perform a rank $k$ truncation $\widehat{\boldsymbol{A}}_{(k)} = \widehat{\boldsymbol{U}}_1 \widehat{\boldsymbol{\Lambda}}_1 \widehat{\boldsymbol{U}}_1^T$, where $\widehat{\boldsymbol{\Lambda}}_1 \in \mathbb{R}^{k \times k}$.
3: return $f(\widehat{\boldsymbol{A}})_{(k)} = \widehat{\boldsymbol{U}}_1 f(\widehat{\boldsymbol{\Lambda}}_1) \widehat{\boldsymbol{U}}_1^T$.
```

prove analogous guarantees for eigenvalue estimation and for approximations of elliptical distributions in the Wasserstein distance. We begin with establishing a few useful lemmas. We start by recalling some basic properties of operator monotone and concave functions.

Lemma 4.2. *Let $f : [0, \infty) \to [0, \infty)$ be a continuous operator monotone function. Then,*

(i) f is concave;

(ii) $f \in C^{\infty}(0, \infty)$.

Proof. (i) This is due to [25, Theorem V.2.5]. (ii) This is due to [25, p.134-135]. □

By the above, the following fact about concave functions also extends to continuous operator monotone functions.

Lemma 4.3. *Let $f : [0, \infty) \to [0, \infty)$ be a concave function. Then,*

(i) $\frac{f(x)}{x}$ is decreasing for $x > 0$;

(ii) $f(tx) \leq t f(x)$ for $t \geq 1$;

(iii) $f(tx) \geq t f(x)$ for $0 \leq t \leq 1$.

(iv) For $t \geq 0$, the function $f(x) - f(x - t)$ is decreasing.

Proof. (i): f is concave if and only if the function

$$R(x_1, x_2) = \frac{f(x_2) - f(x_1)}{x_2 - x_1} \tag{4.3}$$

is decreasing in x_2 for any fixed x_1 (or vice versa). Hence, since $f(0) \geq 0$ we have that $\frac{f(x)}{x} = R(0, x) + \frac{f(0)}{x}$ is decreasing, as required.

(ii): For any fixed x we know that the function $h_x(t) = f(tx)$ is concave. Hence, by (i) we know that $\frac{h_x(t)}{t}$ is decreasing. Therefore, if $t \geq 1$ we have $\frac{f(tx)}{t} = \frac{h_x(t)}{t} \leq \frac{h_x(1)}{1} = f(x)$, which yields the desired result.

(*iii*): Proven in an analogous way to (*ii*) but using $\frac{h_x(t)}{t} \geq \frac{h_x(1)}{1}$ if $t \leq 1$.

(*iv*): Since $R(x_1, x_2)$ in (4.3) is decreasing in x_1 for any fixed x_2 and vice versa we have for any $x \geq y \geq t$

$$\frac{f(y) - f(y-t)}{t} = R(y-t, y) \geq R(x-t, y) \geq R(x-t, x) = \frac{f(x) - f(x-t)}{t},$$

which yields the desired result. □

The following lemma provides an upper bound for the Schatten norm difference of two ordered SPSD matrices. In particular, it bounds $\|\boldsymbol{B} - \boldsymbol{C}\|_{(s)}^s$ by the difference of the Schatten norms of $\boldsymbol{B}$ and $\boldsymbol{C}$. The latter is easier to analyze because it allows us to separate the eigenvalues of $\boldsymbol{B}$ and $\boldsymbol{C}$. This will be especially useful for proving results for the nuclear and Frobenius norm.

Lemma 4.4. *Let* $\boldsymbol{B} \succeq \boldsymbol{C} \succeq \boldsymbol{0}$*, then for* $s \geq 1$

$$\|\boldsymbol{B} - \boldsymbol{C}\|_{(s)} \leq \left(\|\boldsymbol{B}\|_{(s)}^s - \|\boldsymbol{C}\|_{(s)}^s\right)^{1/s}.$$

When $s = 1$ *the inequality becomes an equality.*

Proof. By a result by McCarthy [105, Lemma 2.6] we know that if $\boldsymbol{X}, \boldsymbol{Y} \succeq \boldsymbol{0}$,

$$\operatorname{tr}((\boldsymbol{X} + \boldsymbol{Y})^s) \geq \operatorname{tr}(\boldsymbol{X}^s) + \operatorname{tr}(\boldsymbol{Y}^s).$$

Setting $\boldsymbol{X} = \boldsymbol{B} - \boldsymbol{C}$ and $\boldsymbol{Y} = \boldsymbol{C}$ yields the desired inequality. When $s = 1$, using the fact that $\boldsymbol{B} \succeq \boldsymbol{C} \succeq \boldsymbol{0}$, we find that, as required,

$$\|\boldsymbol{B} - \boldsymbol{C}\|_* = \operatorname{tr}(\boldsymbol{B} - \boldsymbol{C}) = \operatorname{tr}(\boldsymbol{B}) - \operatorname{tr}(\boldsymbol{C}) = \|\boldsymbol{B}\|_* - \|\boldsymbol{C}\|_*.$$ □

4.2.1 Frobenius and nuclear norm guarantees

In this section we prove two results. First, we prove the following nuclear norm result.

Theorem 4.5. *Suppose that, for* $\varepsilon \geq 0$, $\boldsymbol{A} \succeq \widehat{\boldsymbol{A}} \succeq \boldsymbol{0}$ *satisfy*

$$\|\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}\|_* \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_*.$$

Then for any continuous operator monotone function $f : [0, \infty) \to [0, \infty)$*,*

$$\|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}})_{(k)}\|_* \leq (1+\varepsilon)\|f(\boldsymbol{A}) - f(\boldsymbol{A})_{(k)}\|_*.$$

Additionally, we prove an analogous guarantee for the Frobenius norm.

Theorem 4.6. *Suppose that, for $\varepsilon \geq 0$, $\boldsymbol{A} \succeq \widehat{\boldsymbol{A}} \succeq \boldsymbol{0}$ satisfy*

$$\|\boldsymbol{A}\|_{\mathrm{F}}^2 - \|\widehat{\boldsymbol{A}}_{(k)}\|_{\mathrm{F}}^2 \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_{\mathrm{F}}^2.$$

Then for any continuous operator monotone function $f : [0,\infty) \to [0,\infty)$,

$$\|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}})_{(k)}\|_{\mathrm{F}}^2 \leq (1+\varepsilon)\|f(\boldsymbol{A}) - f(\boldsymbol{A})_{(k)}\|_{\mathrm{F}}^2.$$

We note that, by Lemma 4.4, $\|\boldsymbol{A}\|_{\mathrm{F}}^2 - \|\widehat{\boldsymbol{A}}_{(k)}\|_{\mathrm{F}}^2 \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_{\mathrm{F}}^2$ is a *stronger* assumption than $\|\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}\|_{\mathrm{F}}^2 \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_{\mathrm{F}}^2$. I.e., Theorem 4.6 requires assuming more than that $\widehat{\boldsymbol{A}}_{(k)}$ is a near optimal low-rank approximation to $\boldsymbol{A}$ in the Frobenius norm. However, the assumption is still reasonable because, as we show in Section 4.3, many standard low-rank algorithms return results that satisfy this stronger guarantee.

We prove Theorems 4.5 and 4.6 as special cases of a single general theorem about Schatten norms. In particular, by Lemma 4.4 we know that $\|\boldsymbol{A}\|_* - \|\widehat{\boldsymbol{A}}\|_* = \|\boldsymbol{A} - \widehat{\boldsymbol{A}}\|_*$, so the following theorem about Schatten norms immediately implies both Theorems 4.5 and 4.6 by taking $s = 1$ and $s = 2$, respectively.

Theorem 4.7. *Fix $s \in [1,\infty)$. Suppose that, for $\varepsilon \geq 0$, $\boldsymbol{A} \succeq \widehat{\boldsymbol{A}} \succeq \boldsymbol{0}$ satisfy*

$$\|\boldsymbol{A}\|_{(s)}^s - \|\widehat{\boldsymbol{A}}_{(k)}\|_{(s)}^s \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_{(s)}^s$$

Then for any continuous operator monotone function $f : [0,\infty) \to [0,\infty)$,

$$\|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}})_{(k)}\|_{(s)}^s \leq (1+\varepsilon)\|f(\boldsymbol{A}) - f(\boldsymbol{A})_{(k)}\|_{(s)}^s.$$

Remark 4.1. *Before we give the proof, we will rule out some pathological cases. Note that if* $\operatorname{rank}(\boldsymbol{A}) \leq k$ *then by the assumption in Theorem 4.7 we have $\boldsymbol{A} = \widehat{\boldsymbol{A}}_{(k)}$ by Lemma 4.4 and the result trivially holds. Now if* $\operatorname{rank}(\boldsymbol{A}) > k$ *and* $\operatorname{rank}(f(\boldsymbol{A})) \leq k$ *we have $\lambda_{k+1} > 0$, but $f(\lambda_{k+1}) = 0$. By monotonicity and non-negativity of f we have $f(x) = 0$ for any $t \in [0, \lambda_{k+1}]$. Now for any $x \geq \lambda_{k+1} > 0$ we can write $x = t\lambda_{k+1}$ for some $t \geq 1$. Thus, by Lemma 4.3 (ii) we have $f(x) = f(t\lambda_{k+1}) \leq tf(\lambda_{k+1}) = 0$. We can therefore conclude $f \equiv 0$. Therefore, the only interesting cases are when* $\operatorname{rank}(\boldsymbol{A}), \operatorname{rank}(f(\boldsymbol{A})) > k$*, which we will assume throughout this section.*

Proof of Theorem 4.7. Let λ_i and $\widehat{\lambda}_i$ denote the i^{th} largest eigenvalues of $\boldsymbol{A}$ and $\widehat{\boldsymbol{A}}$, respectively. Our assumption on $\widehat{\boldsymbol{A}}$ implies that

$$\sum_{i=1}^{k}(\lambda_i^s - \widehat{\lambda}_i^s) \leq \varepsilon \sum_{i=k+1}^{n} \lambda_i^s. \tag{4.4}$$

By Weyl's monotonicity principle, $\boldsymbol{A} \succeq \widehat{\boldsymbol{A}}$ implies that $\lambda_i \geq \widehat{\lambda}_i$ [88, Corollary 7.7.4 (c)]. Let us define $(1-\delta_i) = \left(\frac{\widehat{\lambda}_i}{\lambda_i}\right)^s$ for $\delta_i \in [0,1]$ for $i = 1, \dots, k$. Hence, (4.4) implies that

$$\sum_{i=1}^{k} \delta_i \lambda_i^s \leq \varepsilon \sum_{i=k+1}^{n} \lambda_i^s. \tag{4.5}$$

By Lemma 4.3 (*iii*), we know that $f(\widehat{\lambda}_i)^s \geq (1-\delta_i) f(\lambda_i)^s$, and so

$$\sum_{i=1}^{k} (f(\lambda_i)^s - f(\widehat{\lambda}_i)^s) \leq \sum_{i=1}^{k} \delta_i f(\lambda_i)^s. \tag{4.6}$$

Lemma 4.3 (*i*) also shows that, for all $i = 1, \dots, k$ and $j = k+1, \dots, n$, $\frac{\lambda_j}{\lambda_i} \leq \frac{f(\lambda_j)}{f(\lambda_i)}$. So, we have,

$$\begin{aligned} \frac{\lambda_j^s}{\delta_i \lambda_i^s} &\leq \frac{f(\lambda_j)^s}{\delta_i f(\lambda_i)^s} \\ \Rightarrow \frac{\sum_{j=k+1}^{n} \lambda_j^s}{\delta_i \lambda_i^s} &\leq \frac{\sum_{j=k+1}^{n} f(\lambda_j)^s}{\delta_i f(\lambda_i)^s} \\ \Rightarrow \frac{\delta_i \lambda_i^s}{\sum_{j=k+1}^{n} \lambda_j^s} &\geq \frac{\delta_i f(\lambda_i)^s}{\sum_{j=k+1}^{n} f(\lambda_j)^s} \end{aligned} \tag{4.7}$$

Combining (4.7) with (4.6) and (4.5) gives

$$\sum_{i=1}^{k} (f(\lambda_i)^s - f(\widehat{\lambda}_i)^s) \leq \sum_{i=1}^{k} \delta_i f(\lambda_i)^s \leq \frac{\sum_{i=1}^{k} \delta_i \lambda_i^s}{\sum_{j=k+1}^{n} \lambda_j^s} \sum_{j=k+1}^{n} f(\lambda_j)^s \leq \varepsilon \sum_{i=k+1}^{n} f(\lambda_i)^s. \tag{4.8}$$

Hence, (4.8) implies that

$$\|f(\boldsymbol{A})\|_{(s)}^s - \|f(\widehat{\boldsymbol{A}})_{(k)}\|_{(s)}^s \leq (1+\varepsilon) \|f(\boldsymbol{A}) - f(\boldsymbol{A})_{(k)}\|_{(s)}^s.$$

We have $f(\boldsymbol{A}) \succeq f(\widehat{\boldsymbol{A}})_{(k)} \succeq \boldsymbol{0}$ since f is operator monotone. The desired inequality follows from applying Lemma 4.4. □

4.2.2 Operator norm guarantees

We next present an analogue to Theorems 4.5 and 4.6 for the operator norm. Our operator norm result is actually more general since we do not require access to an approximation $\widehat{\boldsymbol{A}}$ satisfying $\widehat{\boldsymbol{A}} \preceq \boldsymbol{A}$. The result applies to *any* $\boldsymbol{B}$ such that $\|\boldsymbol{A} - \boldsymbol{B}\|_2 \leq (1+\varepsilon) \|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_2$.

We do not even require that $\boldsymbol{B}$ is rank k.

Theorem 4.8. *Suppose that, for* $\varepsilon \geq 0$, $\boldsymbol{A}, \boldsymbol{B} \succeq \mathbf{0}$ *satisfy*

$$\|\boldsymbol{A} - \boldsymbol{B}\|_2 \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_2.$$

Let $r = \operatorname{rank}(\boldsymbol{B})$. *For any continuous operator monotone function* $f : [0, \infty) \to [0, \infty)$,

$$\|f(\boldsymbol{A}) - f(\boldsymbol{B})_{(r)}\|_2 \leq (1+\varepsilon)\|f(\boldsymbol{A}) - f(\boldsymbol{A})_{(k)}\|_2.$$

Proof. First assume that $f(0) = 0$. As a consequence of this assumption, we have $f(\boldsymbol{B})_{(r)} = f(\boldsymbol{B})$. We leverage [25, Theorem X.1.1] about operator monotone functions, which says that $\|f(\boldsymbol{A}) - f(\boldsymbol{B})\|_2 \leq f(\|\boldsymbol{A} - \boldsymbol{B}\|_2)$. Then, recalling from Lemma 4.2 that f is increasing and concave, we have that

$$\begin{aligned} \|f(\boldsymbol{A}) - f(\boldsymbol{B})\|_2 &\leq f(\|\boldsymbol{A} - \boldsymbol{B}\|_2) && \text{([25, Theorem X.1.1])} \\ &\leq f((1+\varepsilon)\lambda_{k+1}) \\ &\leq (1+\varepsilon) f(\lambda_{k+1}) && \text{(Lemma 4.3 (\textit{ii}))} \\ &= (1+\varepsilon)\|f(\boldsymbol{A}) - f(\boldsymbol{A})_{(k)}\|_2, \end{aligned}$$

which yields the desired inequality for the case when $f(0) = 0$. We now consider the general case when $f(0) \geq 0$. Write $f(x) = g(x) + f(0)$, where $g(x)$ is a continuous operator monotone function satisfying $g(0) = 0$. Let $\boldsymbol{P}_{\boldsymbol{B}}$ be the orthogonal projector onto range($\boldsymbol{B}$). We have that $f(\boldsymbol{B})_{(r)} = g(\boldsymbol{B}) + f(0)\boldsymbol{P}_{\boldsymbol{B}}$, so

$$f(\boldsymbol{A}) - f(\boldsymbol{B})_{(r)} = g(\boldsymbol{A}) - g(\boldsymbol{B}) + f(0)(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{B}}).$$

So by the triangle inequality we have and by the result for operator monotone functions satisfying $g(0) = 0$ we have

$$\begin{aligned} &\|f(\boldsymbol{A}) - f(\boldsymbol{B})_{(r)}\|_2 = \|g(\boldsymbol{A}) - g(\boldsymbol{B})\|_2 + f(0)\|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{B}})\|_2 \leq \\ &(1+\varepsilon)g(\lambda_{k+1}) + f(0) \leq (1+\varepsilon)(g(\lambda_{k+1}) + f(0)) = (1+\varepsilon)f(\lambda_k) = \\ &(1+\varepsilon)\|f(\boldsymbol{A}) - f(\boldsymbol{A})_{(k)}\|_2, \end{aligned}$$

as required. □

4.2.3 Eigenvalue guarantees

We next establish guarantees for eigenvalue estimation. In particular, we show that if the eigenvalues of a SPSD matrix $\widehat{\boldsymbol{A}}$ are good approximations to the eigenvalues of $\boldsymbol{A}$, then the eigenvalues of $f(\widehat{\boldsymbol{A}})$ are even better approximations to the eigenvalues of $f(\boldsymbol{A})$. This result could be combined with results that prove eigenvalue approximation guarantees for algorithms including subspace iteration and block Krylov subspace methods [115].

Theorem 4.9. *Suppose that, for $\varepsilon \in [0,1]$, we have estimates $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \ldots \geq \widehat{\lambda}_k \geq 0$ of the k largest eigenvalues of $\boldsymbol{A}$ satisfying*

$$0 \leq \lambda_i - \widehat{\lambda}_i \leq \varepsilon \lambda_{k+1} \quad \text{for } i = 1, \ldots, k.$$

Then for any non-decreasing concave function $f : [0,\infty) \to [0,\infty)$,

$$0 \leq f(\lambda_i) - f(\widehat{\lambda}_i) \leq \varepsilon f(\lambda_{k+1}) \quad \text{for } i = 1, \ldots, k.$$

Proof. Note that by Lemma 4.3 (*iv*) the function

$$g(t) = f(t) - f(t - \varepsilon \lambda_{k+1})$$

is decreasing. Hence, for $i = 1, \ldots, k$ we have

$$\begin{aligned} 0 &\leq f(\lambda_i) - f(\widehat{\lambda}_i) && (f \text{ is non-decreasing}) \\ &\leq f(\lambda_i) - f(\lambda_i - \varepsilon\lambda_{k+1}) && (f \text{ is non-decreasing}) \\ &\leq f(\lambda_{k+1}) - f((1-\varepsilon)\lambda_{k+1}) && (g(\lambda_i) \leq g(\lambda_{k+1}) \text{ since } \lambda_i \geq \lambda_{k+1}) \\ &\leq \varepsilon f(\lambda_{k+1}), && (\text{Lemma 4.3 } (iii)) \end{aligned}$$

as required. □

The assumption in Theorem 4.9 can be weakened to the case when we have small relative errors

$$0 \leq \lambda_i - \widehat{\lambda}_i \leq \varepsilon \lambda_i. \tag{4.9}$$

By the monotonicity of f and Lemma 4.3 (*iii*), we have that (4.9) implies

$$f(\lambda_i) - f(\widehat{\lambda}_i) \leq f(\lambda_i) - f((1-\varepsilon)\lambda_i) \leq \varepsilon f(\lambda_i).$$

Hence, we also have the following result.

Theorem 4.10. *Suppose that, for $\varepsilon \in [0,1]$, we have estimates $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \ldots \geq \widehat{\lambda}_k \geq 0$ of the k largest eigenvalues of $\boldsymbol{A}$ satisfying*

$$0 \leq \lambda_i - \widehat{\lambda}_i \leq \varepsilon \lambda_i \quad \text{for } i = 1, \ldots, k.$$

Then for any non-decreasing concave function $f : [0,\infty) \to [0,\infty)$,

$$0 \leq f(\lambda_i) - f(\widehat{\lambda}_i) \leq \varepsilon f(\lambda_i) \quad \text{for } i = 1, \ldots, k.$$

4.2.4 Wasserstein distance guarantees

As mentioned in Chapter 3, an application of low-rank approximations to the square-root is that we can cheaply sample from multivariate Gaussian distributions with covariance matrix $\boldsymbol{A}$. For this reason, we would like a guarantee on how close our approximation is to the original distribution. For example, if we want to approximate samples from $\mathcal{N}(\mathbf{0},\boldsymbol{A})$ by sampling from $\mathcal{N}(\mathbf{0},\widehat{\boldsymbol{A}}_{(k)})$, where $\widehat{\boldsymbol{A}}_{(k)}$ is a good rank k approximation to $\boldsymbol{A}$, we would like to guarantee that the distribution $\mathcal{N}(\mathbf{0},\widehat{\boldsymbol{A}}_{(k)})$ is "close" to the original distribution $\mathcal{N}(\mathbf{0},\boldsymbol{A})$. An important measure of the difference between probability distributions is the Wasserstein distance. In this section we present a guarantee for using the Nyström approximation to approximate elliptical distributions in the Wasserstein distance. We proceed with the definition of the Wasserstein 2-distance for probability distrubutions of random vectors.

Definition 4.2 (Wasserstein 2-distance [158]). *Consider two probability distributions ν and γ on $(\mathbb{R}^n,\mathcal{B})$, where $\mathcal{B}$ denotes the Borel σ-algebra on $\mathbb{R}^n$. Then, the Wasserstein 2-distance between ν and γ is defined as*

$$\mathcal{W}_2(\nu,\gamma)=\inf_{\boldsymbol{X}\sim\nu,\boldsymbol{Y}\sim\gamma}\left(\mathbb{E}\|\boldsymbol{X}-\boldsymbol{Y}\|_2^2\right)^{1/2}.$$

We will consider the Wasserstein distance between a family of distributions known as *elliptical distributions* [34], formally defined as follows.

Definition 4.3 (Elliptical distributions). *A random vector $\boldsymbol{\omega}$ of length n is said to elliptical if it has the same distribution as $\boldsymbol{\mu}+R\boldsymbol{A}^{1/2}\boldsymbol{u}$, where R is a positive univariate random variable, $\boldsymbol{A}$ is a SPSD matrix, and $\boldsymbol{u}$ is a uniform random vector on the sphere that is independent of R. Furthermore, by scaling $\boldsymbol{A}$ if necessary we may assume without loss of generality that $\mathbb{E}R^2=n$ so that $\mathbb{E}\left[(\boldsymbol{\omega}-\boldsymbol{\mu})(\boldsymbol{\omega}-\boldsymbol{\mu})^T\right]=\boldsymbol{A}$. If $R\sim\mathcal{D}$ we write $\boldsymbol{\omega}\sim E_{\mathcal{D}}(\boldsymbol{\mu},\boldsymbol{A})$.*

With the concept of elliptical distributions established, we are ready to state and prove a theorem about optimal low-rank approximations in the Wasserstein 2-distance.

Theorem 4.11. *Suppose that, for $\varepsilon\geq 0$, $\boldsymbol{A}\succeq\widehat{\boldsymbol{A}}\succeq\mathbf{0}$ satisfy*

$$\|\boldsymbol{A}-\widehat{\boldsymbol{A}}_{(k)}\|_*\leq(1+\varepsilon)\|\boldsymbol{A}-\boldsymbol{A}_{(k)}\|_*.$$

Then we have

$$\mathcal{W}_2(E_{\mathcal{D}}(\boldsymbol{\mu},\boldsymbol{A}),E_{\mathcal{D}}(\boldsymbol{\mu},\widehat{\boldsymbol{A}}_{(k)}))^2\leq(1+\varepsilon)\min_{\boldsymbol{B}\succeq\mathbf{0}:\operatorname{rank}(\boldsymbol{B})\leq k}\mathcal{W}_2(E_{\mathcal{D}}(\boldsymbol{\mu},\boldsymbol{A}),E_{\mathcal{D}}(\boldsymbol{\mu},\boldsymbol{B}))^2,$$

where

$$\min_{\boldsymbol{B}\succeq\mathbf{0}:\operatorname{rank}(\boldsymbol{B})\leq k}\mathcal{W}_2(E_{\mathcal{D}}(\boldsymbol{\mu},\boldsymbol{A}),E_{\mathcal{D}}(\boldsymbol{\mu},\boldsymbol{B}))^2=\|\boldsymbol{A}-\boldsymbol{A}_{(k)}\|_*.$$

Before we proceed with the proof of Theorem 4.11, we acknowledge [95, Lemma 2.2], which also proved that good approximations to the covariance kernel $\boldsymbol{A}$ yields good approximations to $\mathcal{N}(\boldsymbol{0}, \boldsymbol{A})$ in the Wasserstein distance. However, our result is more general since it applies to *any* elliptical distribution and it shows that near-optimal low-rank approximations imply near-optimal approximations in the Wasserstein distance. We will now proceed with the proof.

Proof. Throughout the proof we denote by $\lambda_i(\boldsymbol{B})$ the i^{th} eigenvalue of a general SPSD matrix $\boldsymbol{B}$. First note for a SPSD matrix $\boldsymbol{B}$ we have by a result by Gelbrich [66, Theorem 2.1 and p.193]

$$\mathcal{W}_2(E_{\mathcal{D}}(\boldsymbol{\mu}, \boldsymbol{A}), E_{\mathcal{D}}(\boldsymbol{\mu}, \boldsymbol{B}))^2 = \operatorname{tr}(\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{B}) - 2\|\boldsymbol{A}^{1/2}\boldsymbol{B}^{1/2}\|_*.$$

By a singular value inequality we have [103, p.342]

$$\|\boldsymbol{A}^{1/2}\boldsymbol{B}^{1/2}\|_* \leq \sum_{i=1}^{n} \lambda_i^{1/2}\lambda_i(\boldsymbol{B})^{1/2}.$$

Recall that λ_i denotes the i^{th} eigenvalue of $\boldsymbol{A}$. Now, if $\operatorname{rank}(\boldsymbol{B}) \leq k$ we know that $\lambda_i(\boldsymbol{B}) = 0$ for $i \geq k$. Hence, we get

$$\begin{aligned}&\operatorname{tr}(\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{B}) - 2\|\boldsymbol{A}^{1/2}\boldsymbol{B}^{1/2}\|_* \geq \\ &\sum_{i=1}^{k}(\lambda_i^{1/2} - \lambda_i(\boldsymbol{B})^{1/2})^2 + \sum_{i=k+1}^{n} \lambda_i \geq \sum_{i=k+1}^{n} \lambda_i.\end{aligned}$$

Thus, $\mathcal{W}_2(E_{\mathcal{D}}(\boldsymbol{\mu}, \boldsymbol{A}), E_{\mathcal{D}}(\boldsymbol{\mu}, \boldsymbol{B}))^2 \geq \sum_{i=k+1}^{n} \lambda_i$ for any SPSD $\boldsymbol{B}$ with $\operatorname{rank}(\boldsymbol{B}) \leq k$ and one can verify that the lower bound can be achieved by putting $\boldsymbol{B} = \boldsymbol{A}_{(k)}$. Hence,

$$\min_{\boldsymbol{B} \succeq \boldsymbol{0}:\operatorname{rank}(\boldsymbol{B}) \leq k} \mathcal{W}_2(E_{\mathcal{D}}(\boldsymbol{\mu}, \boldsymbol{A}), E_{\mathcal{D}}(\boldsymbol{\mu}, \boldsymbol{B}))^2 = \|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_*.$$

Now note that by [95, Lemma 2.2] we have

$$\begin{aligned}&\mathcal{W}_2(E_{\mathcal{D}}(\boldsymbol{\mu}, \boldsymbol{A}), E_{\mathcal{D}}(\boldsymbol{\mu}, \boldsymbol{B}))^2 \leq \operatorname{tr}(\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}) = \|\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}\|_* \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_* = \\ &(1+\varepsilon) \min_{\boldsymbol{B} \succeq \boldsymbol{0}:\operatorname{rank}(\boldsymbol{B}) \leq k} \mathcal{W}_2(E_{\mathcal{D}}(\boldsymbol{\mu}, \boldsymbol{A}), E_{\mathcal{D}}(\boldsymbol{\mu}, \boldsymbol{B}))^2,\end{aligned}$$

as required. □

4.3 Good projections imply good Nyström approximations

In this section we show that if $\boldsymbol{Q}$ is an orthonormal basis so that $(\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A})_{(k)}$ is a good rank k approximation to $\boldsymbol{A}$, then $\widehat{\boldsymbol{A}}_{(k)}$ is a *better* rank k approximation to $\boldsymbol{A}$, where $\widehat{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q})^\dagger\boldsymbol{Q}^T\boldsymbol{A}$ is the Nyström approximation to $\boldsymbol{A}$. Existing low-rank approximation literature commonly provides guarantees for the error $\|\boldsymbol{A} - (\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A})_{(k)}\|$, where $\boldsymbol{Q}$ is the output of some algorithm, see e.g. [14, 15, 76, 91, 109, 115].[1] Hence, this result allows us to transform many known low-rank approximation guarantees into low-rank approximation guarantees for the rank k truncated Nyström approximation $\widehat{\boldsymbol{A}}_{(k)}$. Further, by the results in Section 4.2 we therefore extend these guarantees to the funNyström approximation.

We point out that whenever $\boldsymbol{Q}$ has exactly k columns, many of the results in this section would follow from [150, Lemma 5.2], which shows that $\|\boldsymbol{A} - \widehat{\boldsymbol{A}}\| \leq \|\boldsymbol{A} - \boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A}\|$ for any unitarily invariant norm $\|\cdot\|$. However, often $\boldsymbol{Q}$ has more than k columns, e.g. when $\boldsymbol{Q}$ is an orthonormal basis for a Krylov subspace, and we want to establish guarantees when we truncate $\widehat{\boldsymbol{A}}$ back to rank k. Truncation is desirable when the low-rank approximation is needed for downstream applications like data visualization or k-means clustering [128].

We show that $\|\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}\| \leq \|\boldsymbol{A} - (\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A})_{(k)}\|$ for the nuclear and Frobenius norms. Perhaps surprisingly, the inequality is false in the operator norm and we provide a counterexample. Lastly, we also provide an analogous guarantee for estimating the eigenvalues of $\boldsymbol{A}$. Before doing so, we repeat (2.10), which is a standard fact about Nyström approximation used throughout this section.

Lemma 4.12 ([68, Equation (4)]). *For any $\boldsymbol{Q} \in \mathbb{R}^{n\times\ell}$, the Nyström approximation satisfies $\widehat{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q})^\dagger\boldsymbol{Q}^T\boldsymbol{A} = \boldsymbol{A}^{1/2}\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}^{1/2}$.*

4.3.1 Frobenius norm guarantees

We first prove the following result on Frobenius norm low-rank approximation.

Theorem 4.13. *Let $\boldsymbol{A} \succeq \boldsymbol{0}$ and let $\boldsymbol{Q}$ be an orthonormal basis so that, for $\varepsilon \geq 0$,*

$$\|\boldsymbol{A} - (\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A})_{(k)}\|_{\mathrm{F}}^2 \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_{\mathrm{F}}^2. \tag{4.10}$$

Then if $\widehat{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q})^\dagger\boldsymbol{Q}^T\boldsymbol{A}$ we have

$$\|\boldsymbol{A}\|_{\mathrm{F}}^2 - \|\widehat{\boldsymbol{A}}_{(k)}\|_{\mathrm{F}}^2 \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_{\mathrm{F}}^2.$$

Theorem 4.13 establishes that the condition needed for Theorem 4.6 can be achieved with many low-rank approximation algorithms, including e.g. block Krylov subspace methods, sketching methods, and sampling methods [14, 15, 39, 109, 115, 139, 165]. Further, by

[1] We recall that $(\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A})_{(k)} = \boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{A})_{(k)}$ and $(\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}\boldsymbol{Q}^T)_{(k)} = \boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q})_{(k)}\boldsymbol{Q}^T$. Both $\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{A})_{(k)}$ and $\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q})_{(k)}\boldsymbol{Q}^T$ are preferable for computational purposes since we only have to compute the best rank k approximation of a smaller matrix. However, in the following sections we use $(\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A})_{(k)}$ and $(\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}\boldsymbol{Q}^T)_{(k)}$, since it simplifies our notation.

Lemma 4.4 we know that $\|\boldsymbol{A}-\widehat{\boldsymbol{A}}_{(k)}\|_\mathrm{F}^2 \leq \|\boldsymbol{A}\|_\mathrm{F}^2 - \|\widehat{\boldsymbol{A}}_{(k)}\|_\mathrm{F}^2$, which shows that if (4.10) is satisfied then $\|\boldsymbol{A}-\widehat{\boldsymbol{A}}_{(k)}\|_\mathrm{F}^2 \leq (1+\varepsilon)\|\boldsymbol{A}-\boldsymbol{A}_{(k)}\|_\mathrm{F}^2$. That is, $\widehat{\boldsymbol{A}}_{(k)}$ is a near-optimal rank k approximation.

We begin by recalling a fact about the best rank k approximation to $\boldsymbol{A}$ constrained to range($\boldsymbol{Q}$).

Lemma 4.14. *Let $\boldsymbol{Q}$ be an orthonormal basis. Then,*

$$\|\boldsymbol{B}-\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{B})_{(k)}\|_\mathrm{F}^2 = \min_{\boldsymbol{C}:\mathrm{rank}(\boldsymbol{C})\leq k} \|\boldsymbol{B}-\boldsymbol{Q}\boldsymbol{C}\|_\mathrm{F}^2, \tag{4.11}$$

and

$$\|\boldsymbol{B}-\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{B}\boldsymbol{Q})_{(k)}\boldsymbol{Q}^T\|_\mathrm{F}^2 = \min_{\boldsymbol{C}:\mathrm{rank}(\boldsymbol{C})\leq k} \|\boldsymbol{B}-\boldsymbol{Q}\boldsymbol{C}\boldsymbol{Q}^T\|_\mathrm{F}^2. \tag{4.12}$$

Proof. (4.11) was proven in [76, Theorem 3.5] and (4.12) is proven in a similar fashion: Let $\boldsymbol{C}$ be any matrix so that rank($\boldsymbol{C}$) $\leq k$. Notice that

$$\langle \boldsymbol{B}-\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{B}\boldsymbol{Q}\boldsymbol{Q}^T, \boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{B}\boldsymbol{Q}\boldsymbol{Q}^T - \boldsymbol{Q}\boldsymbol{C}\boldsymbol{Q}^T\rangle = 0$$

Therefore, using the Pythagorean theorem we obtain

$$\begin{aligned}&\|\boldsymbol{B}-\boldsymbol{Q}\boldsymbol{C}\boldsymbol{Q}^T\|_\mathrm{F}^2 = \|\boldsymbol{B}-\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{B}\boldsymbol{Q}\boldsymbol{Q}^T + \boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{B}\boldsymbol{Q}\boldsymbol{Q}^T - \boldsymbol{Q}\boldsymbol{C}\boldsymbol{Q}^T\|_\mathrm{F}^2 = \\ &\|\boldsymbol{B}-\boldsymbol{Q}\boldsymbol{Q}\boldsymbol{B}\boldsymbol{Q}\boldsymbol{Q}^T\|_\mathrm{F}^2 + \|\boldsymbol{Q}^T\boldsymbol{B}\boldsymbol{Q}-\boldsymbol{C}\|_\mathrm{F}^2.\end{aligned}$$

Thus, to minimize $\|\boldsymbol{B}-\boldsymbol{Q}\boldsymbol{C}\boldsymbol{Q}^T\|_\mathrm{F}^2$ we should choose $\boldsymbol{C} = (\boldsymbol{Q}^T\boldsymbol{B}\boldsymbol{Q})_{(k)}$. □

With Lemma 4.14 at hand, we can show that error of the rank k truncated Nyström approximation is sandwiched between the error of two projection based rank k approximations.

Lemma 4.15. *Let $\boldsymbol{Q}$ be an orthonormal basis and let $\widehat{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q})^\dagger\boldsymbol{Q}^T\boldsymbol{A}$ and suppose $\boldsymbol{A}\succeq \boldsymbol{0}$. Then the following holds*

$$\begin{aligned}&\|\boldsymbol{A}-(\boldsymbol{P}_{\boldsymbol{A}\boldsymbol{Q}}\boldsymbol{A})_{(k)}\|_\mathrm{F}^2 \leq \|\boldsymbol{A}-(\boldsymbol{P}_{\boldsymbol{A}\boldsymbol{Q}}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{A}\boldsymbol{Q}})_{(k)}\|_\mathrm{F}^2 \leq \|\boldsymbol{A}-\widehat{\boldsymbol{A}}_{(k)}\|_\mathrm{F}^2 \leq \\ &\|\boldsymbol{A}\|_\mathrm{F}^2 - \|\widehat{\boldsymbol{A}}_{(k)}\|_\mathrm{F}^2 = \|\boldsymbol{A}-(\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}})_{(k)}\|_\mathrm{F}^2 \leq \|\boldsymbol{A}-(\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{A})_{(k)}\|_\mathrm{F}^2.\end{aligned}$$

Proof. The first inequality is immediate from (4.11) since $(\boldsymbol{P}_{\boldsymbol{A}\boldsymbol{Q}}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{A}\boldsymbol{Q}})_{(k)}$ is a rank k approximation whose range is contained in range($\boldsymbol{A}\boldsymbol{Q}$). By a similar argument, the second inequality is immediate from (4.12) since $\widehat{\boldsymbol{A}}_{(k)}$ is a rank k approximation whose range and co-range is contained in range($\boldsymbol{A}\boldsymbol{Q}$). The third inequality is a consequence of Lemma 4.4 for $s=2$ since $\boldsymbol{A}\succeq\widehat{\boldsymbol{A}}_{(k)}\succeq\boldsymbol{0}$.

To prove the equality, note that by Lemma 4.12 $\widehat{\boldsymbol{A}} = \boldsymbol{A}^{1/2}\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}^{1/2} = (\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}^{1/2})^T(\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}^{1/2})$.

Hence, $\widehat{\boldsymbol{A}}_{(k)} = (\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}^{1/2})_{(k)}^T(\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}^{1/2})_{(k)}$. Then observe that there always exists an orthogonal projection $\boldsymbol{P}$ so that range$(\boldsymbol{P}) \subseteq$ range$(\boldsymbol{A}^{1/2}\boldsymbol{Q})$ and

$$\widehat{\boldsymbol{A}}_{(k)} = (\boldsymbol{P}\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}^{1/2})^T(\boldsymbol{P}\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}^{1/2}) = (\boldsymbol{P}\boldsymbol{A}^{1/2})^T(\boldsymbol{P}\boldsymbol{A}^{1/2}).$$

Hence,

$$\begin{aligned}\|\boldsymbol{A}\|_{\mathrm{F}}^2 - \|\widehat{\boldsymbol{A}}_{(k)}\|_{\mathrm{F}}^2 = \|\boldsymbol{A}\|_{\mathrm{F}}^2 - \|(\boldsymbol{P}\boldsymbol{A}^{1/2})^T(\boldsymbol{P}\boldsymbol{A}^{1/2})\|_{\mathrm{F}}^2 &= \|\boldsymbol{A}\|_{\mathrm{F}}^2 - \|\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}\|_{\mathrm{F}}^2 \\ &= \|\boldsymbol{A} - \boldsymbol{P}\boldsymbol{A}\boldsymbol{P}\|_{\mathrm{F}}^2.\end{aligned}$$

Finally, noting that $\boldsymbol{P}\boldsymbol{A}\boldsymbol{P} = (\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}})_{(k)}$ yields the desired equality.

For the last inequality in Lemma 4.15, we let $\bar{\boldsymbol{P}}$ be an orthogonal projection so that range$(\bar{\boldsymbol{P}}) \subseteq$ range$(\boldsymbol{Q})$ and $\bar{\boldsymbol{P}}\boldsymbol{A} = (\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{A})_{(k)}$. Note that $\boldsymbol{A}^{1/2}\bar{\boldsymbol{P}}\boldsymbol{A}^{1/2}$ is a rank k approximation to $\boldsymbol{A}$ whose range and co-range are both contained in range$(\boldsymbol{A}^{1/2}\boldsymbol{Q})$. By (4.12) we have that, as required,

$$\begin{aligned}&\|\boldsymbol{A} - (\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}})_{(k)}\|_{\mathrm{F}}^2 \leq \|\boldsymbol{A} - \boldsymbol{A}^{1/2}\bar{\boldsymbol{P}}\boldsymbol{A}^{1/2}\|_{\mathrm{F}}^2 = \\ &\|(\boldsymbol{I} - \bar{\boldsymbol{P}})\boldsymbol{A}(\boldsymbol{I} - \bar{\boldsymbol{P}})\|_{\mathrm{F}}^2 \leq \|(\boldsymbol{I} - \bar{\boldsymbol{P}})\boldsymbol{A}\|_{\mathrm{F}}^2 = \|\boldsymbol{A} - (\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{A})_{(k)}\|_{\mathrm{F}}^2.\end{aligned}$$ □

Proof of Theorem 4.13. The proof of our main result in this section follows immediately from Lemma 4.15. In particular, we have since

$$\|\boldsymbol{A}\|_{\mathrm{F}}^2 - \|\widehat{\boldsymbol{A}}_{(k)}\|_{\mathrm{F}}^2 \leq \|\boldsymbol{A} - (\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{A})_{(k)}\|_{\mathrm{F}}^2 = \|\boldsymbol{A} - (\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A})_{(k)}\|_{\mathrm{F}}^2.$$ □

Remark 4.2. *We remark on a few additional consequence of Lemma 4.15 that may be of independent interest.*

1. *The lemma implies that $\|\boldsymbol{A} - (\boldsymbol{P}_{\boldsymbol{AQ}}\boldsymbol{A})_{(k)}\|_{\mathrm{F}}^2 \leq \|\boldsymbol{A} - (\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{A})_{(k)}\|_{\mathrm{F}}^2$ and that $\|\boldsymbol{A} - (\boldsymbol{P}_{\boldsymbol{AQ}}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{AQ}})_{(k)}\|_{\mathrm{F}}^2 \leq \|\boldsymbol{A} - (\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{Q}})_{(k)}\|_{\mathrm{F}}^2$. Hence, if we approximate $\boldsymbol{A}$ via either a one-sided or two-sided projection onto $\boldsymbol{AQ}$, the error is always better than if we simply project onto $\boldsymbol{Q}$. In other words, we establish an intuitive fact: that subspace iteration monotonically improves Frobenius norm low-rank approximation error. Via a change of basis, a similar result is true for rectangular matrices. One can show that $\|\boldsymbol{A} - (\boldsymbol{P}_{(\boldsymbol{A}\boldsymbol{A}^T)^{q/2}\boldsymbol{Q}}\boldsymbol{A})_{(k)}\|_{\mathrm{F}}^2 \leq \|\boldsymbol{A} - (\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{A})_{(k)}\|_{\mathrm{F}}^2$ for any positive integer q.*

2. *If one has obtained an orthonormal basis $\boldsymbol{Q}$ with ℓ columns so that $\|\boldsymbol{A} - (\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{A})_{(k)}\|_{\mathrm{F}} \leq \epsilon$ then $\|\boldsymbol{A} - (\boldsymbol{P}_{\boldsymbol{AQ}}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{AQ}})_{(k)}\|_{\mathrm{F}} \leq \epsilon$. This implies that a relative error low-rank approximation guarantee for one-sided projection translates to a guarantee for two-sided projection, at the cost of at most ℓ extra matrix-vector products with $\boldsymbol{A}$ to form $\boldsymbol{AQ}$.*

3. *Given a basis $\boldsymbol{Q}$, we require ℓ matrix-vector multiplications with $\boldsymbol{A}$ to either form*

the one-sided projection $(\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A})_{(k)}$ or to form the rank k truncated Nyström approximation $\widehat{\boldsymbol{A}}_{(k)}$ where $\widehat{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q})^\dagger\boldsymbol{Q}^T\boldsymbol{A}$. However, truncated Nyström approximation always provides better error in the Frobenius norm, so should be preferred.

4.3.2 Nuclear norm guarantees

In this section we establish similar guarantees as in the previous section, but for the nuclear norm. Specifically, we prove the following theorem.

Theorem 4.16. *Let $\boldsymbol{A} \succeq \boldsymbol{0}$ and let $\boldsymbol{Q}$ be an orthonormal basis so that, for $\varepsilon \geq 0$,*

$$\|\boldsymbol{A} - (\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A})_{(k)}\|_* \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_*.$$

Then if $\widehat{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q})^\dagger\boldsymbol{Q}^T\boldsymbol{A}$ we have

$$\|\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}\|_* \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_*.$$

Proof. Since $\boldsymbol{A} \succeq \widehat{\boldsymbol{A}}_{(k)} \succeq \boldsymbol{0}$ we know that

$$\|\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}\|_* = \operatorname{tr}(\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}).$$

Then, since $\widehat{\boldsymbol{A}}_{(k)} = (\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}^{1/2})_{(k)}^T(\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}^{1/2})_{(k)}$ by Lemma 4.12, we have

$$\operatorname{tr}(\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}) = \|\boldsymbol{A}^{1/2}\|_{\mathrm{F}}^2 - \|(\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}^{1/2})_{(k)}\|_{\mathrm{F}}^2 = \|\boldsymbol{A}^{1/2} - (\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}^{1/2})_{(k)}\|_{\mathrm{F}}^2.$$

Choose an orthogonal projector $\boldsymbol{P}$ so that $\operatorname{range}(\boldsymbol{P}) \subseteq \operatorname{range}(\boldsymbol{Q})$ and $\boldsymbol{P}\boldsymbol{A} = (\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A})_{(k)}$. Finally, by Lemma 4.15 and Lemma 4.14 we have

$$\begin{aligned}
&\|\boldsymbol{A}^{1/2} - (\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}^{1/2})_{(k)}\|_{\mathrm{F}}^2 \leq \|\boldsymbol{A}^{1/2} - (\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{A}^{1/2})_{(k)}\|_{\mathrm{F}}^2 \leq \\
&\|\boldsymbol{A}^{1/2} - \boldsymbol{P}\boldsymbol{A}^{1/2}\|_{\mathrm{F}} = \operatorname{tr}(\boldsymbol{A}) - \operatorname{tr}(\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}) = \operatorname{tr}(\boldsymbol{A}) - \operatorname{tr}(\boldsymbol{P}\boldsymbol{A}) \leq \\
&\|\boldsymbol{A} - (\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A})_{(k)}\|_* \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_*,
\end{aligned}$$

which yields the desired inequality. □

4.3.3 Operator norm guarantees

In this section we consider the operator norm. When $\boldsymbol{Q}$ has exactly k columns, [150] establishes the following guarantee.

Theorem 4.17 ([150, Lemma 5.2])**.** *Let $\boldsymbol{A} \succeq \boldsymbol{0}$ and let $\boldsymbol{Q} \in \mathbb{R}^{n \times k}$ be an orthonormal basis so that, for some $\varepsilon \geq 0$,*

$$\|\boldsymbol{A} - \boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A}\|_2 \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_2.$$

Then if $\widehat{\boldsymbol{A}} = \boldsymbol{AQ}(\boldsymbol{Q}^T\boldsymbol{AQ})^\dagger\boldsymbol{Q}^T\boldsymbol{A}$ *we have*

$$\|\boldsymbol{A} - \widehat{\boldsymbol{A}}\|_2 \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_2.$$

Ideally, we would extend this guarantee to the case when $\boldsymbol{Q}$ has $\ell > k$ columns, as in Theorems 4.13 and 4.16 for the Frobenius and nuclear norms. I.e., we might hope to prove that $\|\boldsymbol{A} - (\boldsymbol{QQ}^T\boldsymbol{A})_{(k)}\|_2 \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_2$ implies $\|\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}\|_2 \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_2$. Interestingly, however, we show that doing so is impossible. In particular, consider the following counterexample.

$$\boldsymbol{A} = \begin{bmatrix} 9.627 & 1.538 & -0.717 & 1.418 & -0.309 \\ 1.538 & 8.084 & 1.904 & -1.868 & 0.573 \\ -0.717 & 1.904 & 1.353 & -1.538 & -1.300 \\ 1.418 & -1.868 & -1.538 & 2.534 & 0.169 \\ -0.309 & 0.573 & -1.300 & 0.169 & 6.055 \end{bmatrix}, \quad \boldsymbol{Q} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{4.13}$$

For these matrices, $\|\boldsymbol{A} - (\boldsymbol{QQ}^T\boldsymbol{A})_{(2)}\|_2 \approx (1 + 2.59 \times 10^{-8})\|\boldsymbol{A} - \boldsymbol{A}_{(2)}\|_2$ whereas $\|\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(2)}\|_2 \approx (1 + 5.75 \times 10^{-3})\|\boldsymbol{A} - \boldsymbol{A}_{(2)}\|_2$, so $(\boldsymbol{QQ}^T\boldsymbol{A})_{(k)}$ is a better rank k approximation to $\boldsymbol{A}$ compared to $\widehat{\boldsymbol{A}}_{(k)}$. As guaranteed by Theorem 4.17, we do at least have that $3.75 \approx \|\boldsymbol{A} - \widehat{\boldsymbol{A}}\|_2 < \|\boldsymbol{A} - \boldsymbol{QQ}^T\boldsymbol{A}\|_2 \approx 6.24$. Via the same counterexample, we also have the following remark.

Remark 4.3. *In Section 4.3.1 we showed that Frobenius norm low-rank approximation error decreases monotonically in the number of subspace iterations (see Remark 4.2). The same is not true in the operator norm. To see this, let* $\boldsymbol{A}$ *and* $\boldsymbol{Q}$ *be as in* (4.13). *We can check that*

$$\|\boldsymbol{A} - (\boldsymbol{P_Q A})_{(2)}\|_2 \approx 6.449 < 6.455 \approx \|\boldsymbol{A} - (\boldsymbol{P_{AQ} A})_{(2)}\|_2.$$

4.3.4 Eigenvalue guarantee

Now we provide a guarantee for eigenvalue estimation. Specifically, if we have a basis $\boldsymbol{Q}$ so that top k singular values of $\boldsymbol{QQ}^T\boldsymbol{A}$ are estimates of the eigenvalues of $\boldsymbol{A}$, then the eigenvalues of the Nyström approximation $\widehat{\boldsymbol{A}} = \boldsymbol{AQ}(\boldsymbol{Q}^T\boldsymbol{AQ})^\dagger\boldsymbol{Q}^T\boldsymbol{A}$ can only be better estimates.

Theorem 4.18. *Let* $\boldsymbol{A} \succeq \boldsymbol{0}$ *and let* $\boldsymbol{Q}$ *be an orthonormal basis so that* $\lambda_i - \varepsilon_i \leq \sigma_i(\boldsymbol{Q}^T\boldsymbol{A}) \leq \lambda_i$ *for* $i = 1, ..., k$ *and for* $\varepsilon_1, \ldots, \varepsilon_k \geq 0$. *Then if* $\widehat{\lambda}_i$ *is the* i^{th} *eigenvalue of* $\widehat{\boldsymbol{A}} = \boldsymbol{AQ}(\boldsymbol{Q}^T\boldsymbol{AQ})^\dagger\boldsymbol{Q}^T\boldsymbol{A}$, *we have* $\lambda_i - \varepsilon_i \leq \widehat{\lambda}_i \leq \lambda_i$ *for* $i = 1, ..., k$.

Proof. Notice that, by Lemma 4.12, we have

$$\boldsymbol{Q}^T\widehat{\boldsymbol{A}} = (\boldsymbol{Q}^T\boldsymbol{A}^{1/2}\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}})\boldsymbol{A}^{1/2} = \boldsymbol{Q}^T\boldsymbol{A}.$$

Therefore, by applying a standard singular value inequality [88, p.452] we have

$$\widehat{\lambda}_i = \sigma_i(\widehat{\boldsymbol{A}}) \geq \sigma_i(\boldsymbol{Q}^T\widehat{\boldsymbol{A}}) = \sigma_i(\boldsymbol{Q}^T\boldsymbol{A}) \geq \lambda_i - \varepsilon_i.$$

We complete the proof by noting that $\widehat{\lambda}_i \leq \lambda_i$ because $\widehat{\boldsymbol{A}} \preceq \boldsymbol{A}$ [88, Corollary 7.7.4 (c)]. □

4.4 Explicit bounds

We will now demonstrate corollaries of our results from Sections 4.2 and 4.3. In particular, we will show how one can obtain explicit error bounds for the funNyström approximation immediately from existing results for the Nyström approximation. Furthermore, we will also demonstrate how one can obtain guarantees for the funNyström approximation immediately from guarantees for the error $||\boldsymbol{A} - (\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A})_{(k)}\|$, where $\boldsymbol{Q}$ is an orthonormal basis.

We begin with stating and proving the following result, which is very similar to [151, Theorem 4.1].

Theorem 4.19. *Let $\boldsymbol{A} \succeq \boldsymbol{0}$ and let $\gamma = \lambda_{k+1}/\lambda_k$ denote the k^{th} spectral gap of $\boldsymbol{A}$. Let $\boldsymbol{Q}$ be an orthonormal basis for* range$(\boldsymbol{A}^q\boldsymbol{\Omega})$*, where $q \geq 0$ and $\boldsymbol{\Omega}$ is a $n \times (k+p)$ random matrix whose entries are independent identically distributed (i.i.d.) $\mathcal{N}(0,1)$ random variables. If $\widehat{\boldsymbol{A}}$ is defined as in* (2.9) *and $p \geq 2$, then we have*

$$\mathbb{E}\|\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}\|_* \leq \left(1 + \gamma^{2q}\frac{k}{p-1}\right)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_*.$$

Proof. Let $\boldsymbol{\Psi} = \boldsymbol{A}^q\boldsymbol{\Omega}$. Since $\widehat{\boldsymbol{A}}$ depends only on range$(\boldsymbol{Q})$, we know that $\widehat{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{\Psi}\left(\boldsymbol{\Psi}^T\boldsymbol{A}\boldsymbol{\Psi}\right)^\dagger\boldsymbol{\Psi}^T\boldsymbol{A}$. First assume that rank$(\boldsymbol{A}) \leq k$. Then range$(\boldsymbol{\Psi})$ = range$(\boldsymbol{A})$ almost surely, so $\boldsymbol{A} = \widehat{\boldsymbol{A}}_{(k)}$ and the bound trivially holds. Now assume rank$(\boldsymbol{A}) > k$. By [151, Proof of Theorem 4.1] we have

$$\|\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}\|_* \leq \|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_* + \|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{\Psi}})\boldsymbol{A}_{(k)}^{1/2}\|_{\mathrm{F}}^2.$$

Let $\boldsymbol{\Psi}_1 = \boldsymbol{U}_1^T\boldsymbol{\Psi}$ and $\boldsymbol{\Psi}_2 = \boldsymbol{U}_2^T\boldsymbol{\Psi}$. Since $\boldsymbol{\Omega}$ is a standard Gaussian and rank$(\boldsymbol{A}) > k$ we know $\boldsymbol{\Psi}_1$ has rank k almost surely. Hence, by [137, Theorem 7] we get $\|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{\Psi}})\boldsymbol{A}_{(k)}^{1/2}\|_{\mathrm{F}}^2 \leq \|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Psi}_2\boldsymbol{\Psi}_1^\dagger\|_{\mathrm{F}}^2$, where $\boldsymbol{\Lambda}_2$ contains the smallest $n-k$ eigenvalues of $\boldsymbol{A}$ as in (2.6). Note that $\boldsymbol{\Psi}_2 = \boldsymbol{\Lambda}_2^q\boldsymbol{\Omega}_2$ and $\boldsymbol{\Psi}_1^\dagger = \boldsymbol{\Omega}_1^\dagger\boldsymbol{\Lambda}_1^{-q}$, where we define $\boldsymbol{\Omega}_1 = \boldsymbol{U}_1^T\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_2 = \boldsymbol{U}_2^T\boldsymbol{\Omega}$ as in (2.8). Hence, by strong submultiplicativity we have $\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Psi}_2\boldsymbol{\Psi}_1^\dagger\|_{\mathrm{F}}^2 \leq \gamma^{2q}\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}^2$. Note that by unitary invariance $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ are independent standard Gaussian random matrices. Taking expectation and applying Lemma 2.4 yields the desired result. □

By applying Theorem 4.5 and Theorem 4.19 we get the following immediate corollary.

Corollary 4.20. *Consider the setting of Theorem 4.19. Then for any continuous operator monotone function* $f : [0,\infty) \to [0,\infty)$*, we have*

$$\mathbb{E}\|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}})_{(k)}\|_* \leq \left(1 + \gamma^{2q}\frac{k}{p-1}\right)\|f(\boldsymbol{A}) - f(\boldsymbol{A})_{(k)}\|_*.$$

Proof. By Remark 4.1 we can assume $\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_*, \|f(\boldsymbol{A}) - f(\boldsymbol{A})_{(k)}\|_* > 0$. By Theorem 4.5 we have

$$\frac{\|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}})_{(k)}\|_*}{\|f(\boldsymbol{A}) - f(\boldsymbol{A})_{(k)}\|_*} \leq \frac{\|\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}\|_*}{\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_*}.$$

Taking expectation and applying Theorem 4.19 yields the desired result. □

Note also that in an identical fashion one can use Theorem 4.19 and Theorem 4.11 to derive a bound on $\mathbb{E}\mathcal{W}_2(E_{\mathcal{D}}(\boldsymbol{\mu}, \boldsymbol{A}), E_{\mathcal{D}}(\boldsymbol{\mu}, \widehat{\boldsymbol{A}}_{(k)}))^2$. We omit a detailed discussion.

One can derive similar bounds for the Frobenius norm. Consider the following bound.

Theorem 4.21. *Let* $\boldsymbol{A} \succeq \boldsymbol{0}$ *and let* $\gamma = \frac{\lambda_{k+1}}{\lambda_k}$ *denote the* k^{th} *spectral gap of* $\boldsymbol{A}$*. Let* $\boldsymbol{Q}$ *be an orthonormal basis for* range $(\boldsymbol{A}^q\boldsymbol{\Omega})$*, where* $q \geq 1$ *and* $\boldsymbol{\Omega}$ *is a* $n \times (k+p)$ *random matrix whose entries are i.i.d.* $\mathcal{N}(0,1)$ *random variables. If* $\widehat{\boldsymbol{A}}$ *is defined as in* (2.9) *and* $p \geq 2$*, then we have*

$$\mathbb{E}\|\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}\|_{\mathrm{F}}^2 \leq \mathbb{E}\left[\|\boldsymbol{A}\|_{\mathrm{F}}^2 - \|\widehat{\boldsymbol{A}}_{(k)}\|_{\mathrm{F}}^2\right] \leq \left(1 + \gamma^{2q-1}\frac{5k}{p-1}\right)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_{\mathrm{F}}^2.$$

Proof. Letting $\boldsymbol{Y} = \boldsymbol{A}^{q+1/2}\boldsymbol{\Omega}$, Lemma 4.15 states that $\|\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}\|_{\mathrm{F}}^2 \leq \|\boldsymbol{A}\|_{\mathrm{F}}^2 - \|\widehat{\boldsymbol{A}}_{(k)}\|_{\mathrm{F}}^2 \leq \|\boldsymbol{A} - (\boldsymbol{P}_{\boldsymbol{Y}}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{Y}})_{(k)}\|_{\mathrm{F}}^2$.

Partition $\boldsymbol{A}$ as in (2.6) and $\boldsymbol{\Omega}$ as in (2.8). Setting

$$\boldsymbol{Z} = \boldsymbol{U}^T\boldsymbol{Y}\boldsymbol{\Omega}_1^\dagger\boldsymbol{\Lambda}_1^{-(q+1/2)} = \begin{bmatrix}\boldsymbol{I}\\ \boldsymbol{F}\end{bmatrix}, \quad \boldsymbol{F} = \boldsymbol{\Lambda}_2^{q+1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\boldsymbol{\Lambda}_1^{-(q+1/2)}.$$

we note that range$(\boldsymbol{U}\boldsymbol{Z}) \subseteq$ range$(\boldsymbol{Y})$ and by (4.12) we know

$$\begin{aligned}\|\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}\|_{\mathrm{F}}^2 &\leq \|\boldsymbol{A} - (\boldsymbol{P}_{\boldsymbol{Y}}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{Y}})_{(k)}\|_{\mathrm{F}}^2 \leq \|\boldsymbol{A} - \boldsymbol{P}_{\boldsymbol{U}\boldsymbol{Z}}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{U}\boldsymbol{Z}}\|_{\mathrm{F}}^2\\ &= \|\boldsymbol{\Lambda} - \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{\Lambda}\boldsymbol{P}_{\boldsymbol{Z}}\|_{\mathrm{F}}^2\\ &= \|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{\Lambda}\|_{\mathrm{F}}^2 + \|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{\Lambda}\boldsymbol{P}_{\boldsymbol{Z}}\|_{\mathrm{F}}^2, \qquad (4.14)\end{aligned}$$

where we used the unitary invariance of the Frobenius norm and that $\boldsymbol{U}^T\boldsymbol{P}_{\boldsymbol{U}\boldsymbol{Z}}\boldsymbol{U} = \boldsymbol{P}_{\boldsymbol{Z}}$ [79, Proposition 8.4].

For treating the first term in the sum (4.14), we recall from [79, Proposition 8.2] that

$$\begin{aligned}
&\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{Z}}=\begin{bmatrix}\boldsymbol{I}-(\boldsymbol{I}+\boldsymbol{F}^T\boldsymbol{F})^{-1} & -(\boldsymbol{I}+\boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T\\ -\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^T\boldsymbol{F})^{-1} & \boldsymbol{I}-\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T\end{bmatrix},\\
&\boldsymbol{I}-(\boldsymbol{I}+\boldsymbol{F}^T\boldsymbol{F})^{-1}\preceq\boldsymbol{F}^T\boldsymbol{F}, \qquad (4.15)\\
&\boldsymbol{I}-\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T\preceq\boldsymbol{I}.
\end{aligned}$$

Hence,

$$\|(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{\Lambda}\|_{\mathrm{F}}^2=\mathrm{tr}(\boldsymbol{\Lambda}(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{\Lambda})\leq\|\boldsymbol{F}\boldsymbol{\Lambda}_1\|_F^2+\|\boldsymbol{\Lambda}_2\|_{\mathrm{F}}^2. \qquad (4.16)$$

Utilizing $q\geq 1$ we obtain

$$\begin{aligned}
&\|\boldsymbol{F}\boldsymbol{\Lambda}_1\|_{\mathrm{F}}\leq\|\boldsymbol{\Lambda}_2^{q-1/2}\|_2\|\boldsymbol{\Lambda}_1^{-(q-1/2)}\|_2\|\boldsymbol{\Lambda}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}\\
\leq&\gamma^{q-1/2}\|\boldsymbol{\Lambda}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}},
\end{aligned}$$

where the second inequality relies on $q\geq 1$. Plugging this inequality into (4.16) yields

$$\|(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{\Lambda}\|_{\mathrm{F}}^2\leq\|\boldsymbol{\Lambda}_2\|_{\mathrm{F}}^2+\gamma^{2q-1}\|\boldsymbol{\Lambda}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}^2. \qquad (4.17)$$

For treating the second term in the sum (4.14), we decompose $\boldsymbol{\Lambda}$ into

$$\widetilde{\boldsymbol{\Lambda}}_1=\begin{bmatrix}\boldsymbol{\Lambda}_1 & \\ & \boldsymbol{0}\end{bmatrix},\quad \widetilde{\boldsymbol{\Lambda}}_2=\begin{bmatrix}\boldsymbol{0} & \\ & \boldsymbol{\Lambda}_2\end{bmatrix},$$

which gives

$$\begin{aligned}
\|(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{\Lambda}\boldsymbol{P}_{\boldsymbol{Z}}\|_{\mathrm{F}}&\leq\|(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{Z}})\widetilde{\boldsymbol{\Lambda}}_1\boldsymbol{P}_{\boldsymbol{Z}}\|_{\mathrm{F}}+\|(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{Z}})\widetilde{\boldsymbol{\Lambda}}_2\boldsymbol{P}_{\boldsymbol{Z}}\|_{\mathrm{F}}\\
&\leq\|(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{Z}})\widetilde{\boldsymbol{\Lambda}}_1\|_{\mathrm{F}}+\|\widetilde{\boldsymbol{\Lambda}}_2\boldsymbol{P}_{\boldsymbol{Z}}\|_{\mathrm{F}}.
\end{aligned}$$

Replacing $\boldsymbol{\Lambda}$ by $\widetilde{\boldsymbol{\Lambda}}_1$ in (4.17) shows $\|(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{Z}})\widetilde{\boldsymbol{\Lambda}}_1\|_{\mathrm{F}}\leq\gamma^{q-1/2}\|\boldsymbol{\Lambda}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}$. It turns out that the second term $\|\widetilde{\boldsymbol{\Lambda}}_2\boldsymbol{P}_{\boldsymbol{Z}}\|_{\mathrm{F}}=\|\boldsymbol{P}_{\boldsymbol{Z}}\widetilde{\boldsymbol{\Lambda}}_2\|_{\mathrm{F}}$ obeys the same bound:

$$\begin{aligned}
&\|\boldsymbol{P}_{\boldsymbol{Z}}\widetilde{\boldsymbol{\Lambda}}_2\|_{\mathrm{F}}^2=\mathrm{tr}(\boldsymbol{\Lambda}_2\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T\boldsymbol{\Lambda}_2)\\
\leq&\,\mathrm{tr}(\boldsymbol{\Lambda}_2\boldsymbol{F}\boldsymbol{F}^T\boldsymbol{\Lambda}_2)=\|\boldsymbol{\Lambda}_2\boldsymbol{F}\|_{\mathrm{F}}^2\leq\gamma^{2(q-1/2)}\gamma^2\|\boldsymbol{\Lambda}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}^2\\
\leq&\gamma^{2(q-1/2)}\|\boldsymbol{\Lambda}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}^2,
\end{aligned}$$

where we used $(\boldsymbol{I}+\boldsymbol{F}^T\boldsymbol{F})^{-1}\preceq\boldsymbol{I}$ and the monotonicity of f. Overall one obtains

$$\|(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{\Lambda}\boldsymbol{P}_{\boldsymbol{Z}}\|_{\mathrm{F}}\leq 2\gamma^{q-1/2}\|\boldsymbol{\Lambda}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}.$$

Plugging this inequality and inequality (4.17) into (4.14) yields the following structural bound

$$\|\boldsymbol{A}-\widehat{\boldsymbol{A}}_{(k)}\|_{\mathrm{F}}^2 \leq \|\boldsymbol{A}\|_{\mathrm{F}}^2 - \|\widehat{\boldsymbol{A}}_{(k)}\|_{\mathrm{F}}^2 \leq \|\boldsymbol{A}-\boldsymbol{A}_{(k)}\|_{\mathrm{F}}^2 + 5\gamma^{2q-1}\|\boldsymbol{\Lambda}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}^2.$$

Applying expectation and using Lemma 2.4 gives $\mathbb{E}\|\boldsymbol{\Lambda}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}^2 = \frac{k}{p-1}\|\boldsymbol{A}-\boldsymbol{A}_{(k)}\|_{\mathrm{F}}^2$, which yields the desired result.

□

Once again, using an entirely identical argument as in the proof of Corollary 4.20 we can combine Theorem 4.21 and Theorem 4.6 to obtain the following bound on the funNyström approximation.

Corollary 4.22. *Consider the setting of Theorem 4.21. Then for any continuous operator monotone function $f:[0,\infty)\to[0,\infty)$, we have*

$$\mathbb{E}\|f(\boldsymbol{A})-f(\widehat{\boldsymbol{A}})_{(k)}\|_{\mathrm{F}}^2 \leq \left(1+\gamma^{2q-1}\frac{5k}{p-1}\right)\|f(\boldsymbol{A})-f(\boldsymbol{A})_{(k)}\|_{\mathrm{F}}^2.$$

Alternatively, one can combine the results from Section 4.2 and Section 4.3 to translate bounds for $\|\boldsymbol{A}-(\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A})_{(k)}\|$ into a bound for the funNyström approximation where $\widehat{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q})^\dagger\boldsymbol{Q}^T\boldsymbol{A}$. Consider the following two known result.

Theorem 4.23 ([14, Theorem 4.2]). *Let $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ and $\varepsilon\in(0,1)$. There is an algorithm which performs $O(\frac{k\log(n/\varepsilon)}{\varepsilon^{1/3}})$ matrix vector products with $\boldsymbol{A}$ and returns an orthonormal matrix $\boldsymbol{Q}\in\mathbb{R}^{n\times k}$ such that with probability at least* 0.9

$$\|\boldsymbol{A}-\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A}\|_* \leq (1+\varepsilon)\|\boldsymbol{A}-\boldsymbol{A}_{(k)}\|_*.$$

Theorem 4.24 ([115, Theorems 11-12]). *Let $\boldsymbol{A}\succeq\boldsymbol{0}$ be SPSD and $\varepsilon\in(0,1)$. Then [115, Algorithm 2] performs $\ell = O\left(\frac{k\log(n)}{\varepsilon^{1/2}}\right)$ matrix vector products with $\boldsymbol{A}$ and returns an orthonormal matrix $\boldsymbol{Q}\in\mathbb{R}^{n\times\ell}$ such that with probability at least* 0.99 *we have*

$$\|\boldsymbol{A}-(\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A})_{(k)}\|_{\mathrm{F}} \leq (1+\varepsilon)\|\boldsymbol{A}-\boldsymbol{A}_{(k)}\|_{\mathrm{F}}, \tag{4.18}$$

and for $i=1,\ldots,k$ we have

$$|\lambda_i^2 - \lambda_i(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{Q})| \leq \varepsilon\lambda_{k+1}^2 \tag{4.19}$$

where $\lambda_i(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{Q})$ is the i^{th} eigenvalue of $\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{Q}$.

By using Theorems 4.5, 4.6, 4.9, 4.13, 4.16 and 4.18, we get the following corollaries.

Corollary 4.25. *Let $\boldsymbol{A}\succeq\boldsymbol{0}$ and $\varepsilon\in(0,1)$. Let $f:[0,\infty)\to[0,\infty)$ be a continuous* *operator monotone function. There is an algorithm which performs $O(\frac{k\log(n/\varepsilon)}{\varepsilon^{1/3}})$ matrix-vector products with $\boldsymbol{A}$ and returns an orthonormal basis $\boldsymbol{Q} \in \mathbb{R}^{n\times k}$ so that the rank k Nyström approximation $\widehat{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q})^\dagger\boldsymbol{Q}^T\boldsymbol{A}$ satisfies with probability at least* 0.9

$$\|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}})_{(k)}\|_* \leq (1+\varepsilon)\|f(\boldsymbol{A}) - f(\boldsymbol{A})_{(k)}\|_*.$$

Corollary 4.26. *Let $\boldsymbol{A} \succeq \boldsymbol{0}$ and $\varepsilon \in (0,1)$. Let $f : [0,\infty) \to [0,\infty)$ be a continuous operator monotone function. Then [115, Algorithm 2] performs $\ell = O\left(\frac{k\log(n)}{\varepsilon^{1/2}}\right)$ matrix vector products with $\boldsymbol{A}$ and returns an orthonormal matrix $\boldsymbol{Q} \in \mathbb{R}^{n\times \ell}$ such that $\widehat{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q})^\dagger\boldsymbol{Q}^T\boldsymbol{A}$ satisfies with probability at least* 0.99

$$\|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}})_{(k)}\|_{\mathrm{F}} \leq (1+\varepsilon)\|f(\boldsymbol{A}) - f(\boldsymbol{A})_{(k)}\|_{\mathrm{F}}, \tag{4.20}$$

and for $i = 1,\dots,k$ we have

$$|f(\lambda_i) - f(\widehat{\lambda}_i)| \leq \varepsilon f(\lambda_{k+1}) \tag{4.21}$$

Proof. We focus on the proof of Corollary 4.26, since the proof of Corollary 4.25 is analogous. We begin with showing (4.20). Conditioned on the inequality (4.18) we know by Theorem 4.13 that we have $(\|\boldsymbol{A}\|_{\mathrm{F}}^2 - \|\widehat{\boldsymbol{A}}_{(k)}\|_{\mathrm{F}}^2)^{1/2} \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_{\mathrm{F}}$. Consequently, by Theorem 4.6 we also have (4.20). Since (4.18) happens with probability at least 0.99, we know that (4.20) happens with probability at least 0.99.

We proceed with showing (4.21). Conditioned on the inequality (4.19) we know by Theorem 4.9 that we have $|\lambda_i - \sigma_i(\boldsymbol{Q}^T\boldsymbol{A})| \leq \varepsilon\lambda_{k+1}$. By Theorem 4.18 we know that this implies $|\lambda_i - \widehat{\lambda}_i| \leq \varepsilon\lambda_{k+1}$. Finally, by Theorem 4.9 we have (4.21). Since (4.19) happens with probability at least 0.99, we know that (4.21) happens with probability at least 0.99. □

4.5 General unitarily invariant norms

In this section we will present an expectation bound for $\|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}})\|$, where $\widehat{\boldsymbol{A}}$ denotes a Nyström approximation as defined in (2.9) where $\boldsymbol{Q}$ is an orthonormal basis for range($\boldsymbol{A}^q\boldsymbol{\Omega}$) and $\|\cdot\|$ denotes *any* unitarily invariant norm; see Definition 2.1. We begin with two useful results that hold in any unitarily invariant norms.

Lemma 4.27. *Consider $n\times n$ SPSD matrices $\boldsymbol{B}, \boldsymbol{C}$ satisfying $\boldsymbol{B} \succeq \boldsymbol{C}$. Then*

(i) $\|\boldsymbol{B}\| \geq \|\boldsymbol{C}\|$;

(ii) $\|f(\boldsymbol{B})\| \geq \|f(\boldsymbol{C})\|$ *for any increasing function $f : [0,\infty) \to [0,\infty)$;*

(iii) $\|f(\boldsymbol{B}) - f(\boldsymbol{C})\| \leq \|f(\boldsymbol{B} - \boldsymbol{C})\|$, *for any operator monotone function $f : [0,\infty) \to$*

$[0, \infty)$.

Proof. (i) Let $\lambda_i(\boldsymbol{B})$ and $\lambda_i(\boldsymbol{C})$ denote the ith largest eigenvalues of $\boldsymbol{B}$ and $\boldsymbol{C}$, respectively. By [88, Corollary 7.7.4 (c)], $\lambda_i(\boldsymbol{B}) \geq \lambda_i(\boldsymbol{C}) \geq 0$ for $i = 1, \ldots, n$. By Fan's dominance theorem [25, Theorem IV.2.2], this implies $\|\boldsymbol{B}\| \geq \|\boldsymbol{C}\|$.

(ii) Because of the monotonicity and non-negativity of f, $\lambda_i(f(\boldsymbol{B})) \geq \lambda_i(f(\boldsymbol{C})) \geq 0$ and hence the arguments from (i) apply.

(iii) This is a consequence of a result by Ando [6, Theorem 1]. □

Lemma 4.28 ([98, Theorem 2.1])**.** *Let $f : [0, \infty) \to [0, \infty)$ be concave. Then given a partitioned SPSD matrix $\begin{bmatrix} \boldsymbol{B} & \boldsymbol{X} \\ \boldsymbol{X}^T & \boldsymbol{C} \end{bmatrix}$ with square $\boldsymbol{B}$ and $\boldsymbol{C}$, one has*

$$\left\| f\left(\begin{bmatrix} \boldsymbol{B} & \boldsymbol{X} \\ \boldsymbol{X}^T & \boldsymbol{C} \end{bmatrix} \right) \right\| \leq \|f(\boldsymbol{B})\| + \|f(\boldsymbol{C})\|.$$

With these two lemmas available, we proceed with proving a structural bound that hold in any unitarily invariant norm.

Lemma 4.29. *Let $\boldsymbol{A} \succeq \boldsymbol{0}$ have eigenvalue partitioned as* (2.6)*. Let $\boldsymbol{Q}$ be an orthonormal basis for* $\operatorname{range}(\boldsymbol{A}^q\boldsymbol{\Omega})$*, for $q \geq 0$, and define $\widehat{\boldsymbol{A}}$ as in* (2.9)*. Partition $\boldsymbol{\Omega}$ as in* (2.8) *and assume that* $\operatorname{rank}(\boldsymbol{\Omega}_1) = k$*. Let $\boldsymbol{F} = \boldsymbol{\Lambda}_2^{q+1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\boldsymbol{\Lambda}_1^{-(q+1/2)}$. Then, for any continuous operator monotone function $f : [0, \infty) \to [0, \infty)$ we have*

$$\|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}})\| \leq \|f(\boldsymbol{\Lambda}_2)\| + \|f(\boldsymbol{\Lambda}_1^{1/2}\boldsymbol{F}^T\boldsymbol{F}\boldsymbol{\Lambda}_1^{1/2})\|.$$

Proof. From (2.10) it follows that

$$\boldsymbol{A} - \widehat{\boldsymbol{A}} = \boldsymbol{A}^{1/2}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Y}})\boldsymbol{A}^{1/2} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{\boldsymbol{Y}}})\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^T,$$

where we set $\boldsymbol{Y} = \boldsymbol{A}^{q+1/2}\boldsymbol{\Omega}$ and $\widetilde{\boldsymbol{Y}} = \boldsymbol{U}^T\boldsymbol{Y}$. Combined with Lemma 4.27, this gives

$$\|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}})\| \leq \|f(\boldsymbol{A} - \widehat{\boldsymbol{A}})\| = \|f(\boldsymbol{\Lambda}^{1/2}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{\boldsymbol{Y}}})\boldsymbol{\Lambda}^{1/2})\|.$$

As in the proof of Theorem 4.21, we set $\boldsymbol{Z} = \widetilde{\boldsymbol{Y}}\boldsymbol{\Omega}_1^\dagger\boldsymbol{\Lambda}_1^{-(q+1/2)} = \begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{F} \end{bmatrix}$. Using $\operatorname{range}(\boldsymbol{Z}) \subseteq \operatorname{range}(\widetilde{\boldsymbol{Y}})$, we obtain $\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}} \succeq \boldsymbol{I} - \boldsymbol{P}_{\widetilde{\boldsymbol{Y}}} \succeq \boldsymbol{0}$ and, in turn, $\boldsymbol{\Lambda}^{1/2}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{\Lambda}^{1/2} \succeq \boldsymbol{\Lambda}^{1/2}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{\boldsymbol{Y}}})\boldsymbol{\Lambda}^{1/2} \succeq \boldsymbol{0}$. Using Lemma 4.27 (ii), this gives

$$\|f(\boldsymbol{\Lambda}^{1/2}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{\boldsymbol{Y}}})\boldsymbol{\Lambda}^{1/2})\| \leq \|f(\boldsymbol{\Lambda}^{1/2}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{\Lambda}^{1/2})\|. \tag{4.22}$$

Exploiting the 2×2 block structure (4.15) of the SPSD matrix $\boldsymbol{I} - \boldsymbol{P_Z}$ and applying Lemma 4.28 yields

$$\begin{aligned}&\|f(\boldsymbol{\Lambda}^{1/2}(\boldsymbol{I} - \boldsymbol{P_Z})\boldsymbol{\Lambda}^{1/2})\| \\ \leq &\|f(\boldsymbol{\Lambda}_1^{1/2}(\boldsymbol{I} - (\boldsymbol{I} + \boldsymbol{F}^T\boldsymbol{F})^{-1})\boldsymbol{\Lambda}_1^{1/2})\| + \|f(\boldsymbol{\Lambda}_2^{1/2}(\boldsymbol{I} - \boldsymbol{F}(\boldsymbol{I} + \boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T)\boldsymbol{\Lambda}_2^{1/2})\|.\end{aligned}$$

The proof is completed using the inequalities

$$\begin{aligned}&\|f(\boldsymbol{\Lambda}_1^{1/2}(\boldsymbol{I} - (\boldsymbol{I} + \boldsymbol{F}^T\boldsymbol{F})^{-1})\boldsymbol{\Lambda}_1^{1/2})\| \leq \|f(\boldsymbol{\Lambda}_1^{1/2}\boldsymbol{F}^T\boldsymbol{F}\boldsymbol{\Lambda}_1^{1/2})\| \\ &\|f(\boldsymbol{\Lambda}_2^{1/2}(\boldsymbol{I} - \boldsymbol{F}(\boldsymbol{I} + \boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T)\boldsymbol{\Lambda}_2^{1/2})\| \leq \|f(\boldsymbol{\Lambda}_2)\|,\end{aligned}$$

which are derived from the inequalities in (4.15) with the same arguments used for (4.22). □

In order to obtain the expectation bound we make use of the following result established in [56, Proof of Proposition 2.2].

Lemma 4.30. *Let $\boldsymbol{\Omega}_1 \in \mathbb{R}^{k\times(k+p)}$ and $\boldsymbol{\Omega}_2 \in \mathbb{R}^{(n-k)\times(k+p)}$ be independent random matrices whose entries are i.i.d. $\mathcal{N}(0,1)$ random variables. If $\boldsymbol{D}$ is a matrix, then if $p \geq 2$*

$$\mathbb{E}\|\boldsymbol{D}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_2^2 \leq \frac{2k}{p-1}\|\boldsymbol{D}\|_2^2 + \frac{2e^2(k+p)}{p^2-1}\|\boldsymbol{D}\|_\mathrm{F}^2 \tag{4.23}$$

With the structural bound at hand, we are ready to present an expectation bound in any unitarily invariant norm. For conciseness we only state an expectation bound. From the proof of Theorem 4.31, it follows that a deviation bound can be obtained from a deviation bound on the quantity $\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_2$; see, e.g., [79, Theorem 10.8] for such a bound.

Theorem 4.31. *Let $\boldsymbol{A} \succeq \boldsymbol{0}$ have eigenvalue partitioned as (2.6) and let $\gamma = \frac{\lambda_{k+1}}{\lambda_k}$ denote the k^{th} spectral gap of $\boldsymbol{A}$. Let $\boldsymbol{Q}$ be an orthonormal basis for $\mathrm{range}\,(\boldsymbol{A}^q\boldsymbol{\Omega})$, where $q \geq 0$ and $\boldsymbol{\Omega}$ is a $n \times (k+p)$ random matrix whose entries are i.i.d. $\mathcal{N}(0,1)$ random variables. If $\widehat{\boldsymbol{A}}$ is the Nyström approximation as defined in (2.9) and $p \geq 2$, then for any operator monotone function $f : [0,\infty) \to [0,\infty)$ we have*

$$\begin{aligned}\mathbb{E}\|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}})\| \leq &\|f(\boldsymbol{\Lambda}_2)\| + \\ &C_k\left(\left\|f\left(\gamma^{2q}\frac{2k}{p-1}\boldsymbol{\Lambda}_2\right)\right\|_2 + \left\|f\left(\gamma^{2q}\frac{2e^2(k+p)}{p^2-1}\boldsymbol{\Lambda}_2\right)\right\|_*\right),\end{aligned}$$

where C_k is the equivalence constant so that $\|\boldsymbol{B}\| \leq C_k\|\boldsymbol{B}\|_2$ for any $\boldsymbol{B} \in \mathbb{R}^{k\times k}$.

Proof. Note that

$$\begin{aligned}&\|f(\mathbf{\Lambda}_1^{1/2}\boldsymbol{F}^T\boldsymbol{F}\mathbf{\Lambda}_1^{1/2})\| \leq C_k\|f(\mathbf{\Lambda}_1^{1/2}\boldsymbol{F}^T\boldsymbol{F}\mathbf{\Lambda}_1^{1/2})\|_2 = \\ &C_k f(\|\mathbf{\Lambda}_1^{1/2}\boldsymbol{F}^T\boldsymbol{F}\mathbf{\Lambda}_1^{1/2}\|_2) \leq C_k f\left(\gamma^{2q}\|\mathbf{\Lambda}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_2^2\right).\end{aligned}$$

Using Jensen's inequality we obtain

$$\begin{aligned}\mathbb{E}\|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}})\| \leq &\|f(\mathbf{\Lambda}_2)\|_2 + C_k\mathbb{E}f\left(\gamma^{2q}\|\mathbf{\Lambda}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_2^2\right) \\ \leq &\|f(\mathbf{\Lambda}_2)\|_2 + C_k f\left(\gamma^{2q}\mathbb{E}\|\mathbf{\Lambda}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_2^2\right).\end{aligned}$$

Bounding $\mathbb{E}\|\mathbf{\Lambda}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_2^2$ via (4.23) and using the subadditivity of f on $[0,\infty)$ as well as the relations $\|\mathbf{\Lambda}_2^{1/2}\|_2^2 = \|\mathbf{\Lambda}_2\|_2$ and $\|\mathbf{\Lambda}_2^{1/2}\|_{\mathrm{F}}^2 = \|\mathbf{\Lambda}_2\|_*$ we obtain

$$\begin{aligned}&f\left(\gamma^{2q}\mathbb{E}\|\mathbf{\Lambda}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_2^2\right) \leq f\left(\gamma^{2q}\left(\frac{2k}{p-1}\|\mathbf{\Lambda}_2\|_2 + \frac{2e^2(k+p)}{p^2-1}\|\mathbf{\Lambda}_2\|_*\right)\right) \\ \leq &f\left(\gamma^{2q}\frac{2k}{p-1}\|\mathbf{\Lambda}_2\|_2\right) + f\left(\gamma^{2q}\frac{2e^2(k+p)}{p^2-1}\|\mathbf{\Lambda}_2\|_*\right).\end{aligned}$$

Noting that $f\left(\gamma^{2q}\frac{2k}{p-1}\|\mathbf{\Lambda}_2\|_2\right) = \left\|f\left(\gamma^{2q}\frac{2k}{p-1}\mathbf{\Lambda}_2\right)\right\|_2$ and using once again the subadditivity of f we have

$$f\left(\gamma^{2q}\frac{2e^2(k+p)}{p^2-1}\|\mathbf{\Lambda}_2\|_*\right) \leq \left\|f\left(\gamma^{2q}\frac{2e^2(k+p)}{p^2-1}\mathbf{\Lambda}_2\right)\right\|_*,$$

which completes the proof. □

4.6 Numerical experiments

In this section we numerically verify the theoretical results in Sections 4.2 and 4.3. We also demonstrate the strong performance of funNyström. All experiments have been performed in MATLAB (version 2020a) on a MacBook Pro with a 2.3 GHz Intel Core i7 processor with 4 cores. Scripts to reproduce all figures in Section 4.6.2 are available at https://github.com/davpersson/funNystrom-v2 and scripts to reproduce the figures in Sections 4.6.3 to 4.6.5 are available at https://github.com/davpersson/funNystrom. In our implementation we let $\boldsymbol{C}$ in Algorithm 3 be the square root of $\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}$ obtained by diagonalization. Furthermore, to deal with potential numerical issues due to the appearance of the pseudo-inverse in the Nyström approximation, we compute the ϵ-pseudoinverse of $\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}$, where $\epsilon = 5\cdot 10^{-16}\cdot\|\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}\|_2$.

4.6.1 Test matrices

In the following, we describe the test matrices used in our experiments.

Synthetic matrices

We consider synthetic matrices with prescribed algebraic and exponential eigenvalue decays. Let $\boldsymbol{\Lambda}_{\text{alg}}$ and $\boldsymbol{\Lambda}_{\text{exp}}$ be diagonal matrices with diagonal entries

$$(\boldsymbol{\Lambda}_{\text{alg}})_{ii} = Ci^{-c}, \quad (\boldsymbol{\Lambda}_{\text{exp}})_{ii} = C\gamma^{i}, \quad i = 1, \ldots, n,$$

for parameters $s, c > 0$ and $\gamma \in (0,1)$. Letting $\boldsymbol{U}$ denote the orthogonal matrix generated by the MATLAB-command `gallery('orthog',n)`, we set

$$\boldsymbol{A}_{\text{alg}} = \boldsymbol{U}\boldsymbol{\Lambda}_{\text{alg}}\boldsymbol{U}^{T}, \quad \boldsymbol{A}_{\text{exp}} = \boldsymbol{U}\boldsymbol{\Lambda}_{\text{exp}}\boldsymbol{U}^{T}. \tag{4.24}$$

We also consider the following matrix $\boldsymbol{A}_{\text{disc}} \in \mathbb{R}^{1000\times 1000}$, which is a discretization of a function and is defined by

$$(\boldsymbol{A}_{\text{disc}})_{ij} = \left(\left(\frac{i}{1000}\right)^{10} + \left(\frac{j}{1000}\right)^{10}\right)^{\frac{1}{10}}, \tag{4.25}$$

where $(\boldsymbol{A}_{\text{disc}})_{ij}$ denotes the (i,j)-entry of $\boldsymbol{A}_{\text{disc}}$. This example is inspired by the numerical experiments on column subset selection in [40].

Gaussian process covariance kernels

We consider two classes of matrices that arise from the discretization of the squared exponential and Matérn Gaussian process covariance kernels [142]. For this purpose, we generate $n = 5000$ i.i.d. data points $x_1, \ldots, x_{5000} \sim \mathcal{N}(0,1)$ and set

$$\boldsymbol{A}_{\text{SE}} \in \mathbb{R}^{n\times n}, \quad (\boldsymbol{A}_{\text{SE}})_{ij} = \exp\left(-|x_i - x_j|^2/(2\ell^2)\right), \tag{4.26}$$

$$\mathbf{A}_{\text{Mat}} \in \mathbb{R}^{n\times n}, \quad (\mathbf{A}_{\text{Mat}})_{ij} = \frac{\pi^{1/2}\left(\alpha|x_i - x_j|\right)^{\nu} K_{\nu}(\alpha|x_i - x_j|)}{2^{\nu-1}\Gamma(\nu + 1/2))\alpha^{2\nu}}, \tag{4.27}$$

for $i, j = 1, \ldots, n$, and parameters $\ell, \alpha, \nu > 0$. Note that K_{ν} is the modified Bessel function of the second kind. Computing $\operatorname{tr}(\log(\boldsymbol{I} + \boldsymbol{A}))$, where $\boldsymbol{A}$ is a matrix arising from discretizing a covariance kernel, is an important task in Bayesian optimization and maximum likelihood estimation for Gaussian processes [65, 161].

Bayesian inverse problem

Motivated by the numerical experiments in [5, 49, 138], this test matrix arises from a Bayesian inverse problem. Recall the parabolic differential equation presented in

(3.3)

$$
\begin{aligned}
&u_t = \kappa \Delta u + \lambda u \text{ in } [0,1]^2 \times [0,2] \\
&u(\cdot, 0) = \theta \text{ in } [0,1]^2 \\
&u = 0 \text{ on } \Gamma_1 \\
&\frac{\partial u}{\partial \boldsymbol{n}} = 0 \text{ on } \Gamma_2
\end{aligned}
\tag{4.28}
$$

for $\kappa, \lambda > 0$ and $\Gamma_2 = \{(x, 1) \in \mathbb{R}^2 : x \in [0,1]\}$ and $\Gamma_1 = \partial \mathcal{D} \setminus \Gamma_2$. We place 49 sensors at $(i/8, j/8) \in [0,1]^2$ for $i, j = 1, \ldots, 7$ to take measurements of u at these sensor locations at times $t = 1, 1.5, 2$. We gather all $3 \times 49 = 147$ measurements in a vector $\boldsymbol{d} \in \mathbb{R}^{147}$.

Recall that discretizing (4.28) in space using finite differences on 40×40 equispaced grid yields an ordinary differential equation of the form

$$
\begin{aligned}
&\dot{\boldsymbol{u}}(t) = \boldsymbol{A}\boldsymbol{u}(t) \text{ for } t \in [0,2], \\
&\boldsymbol{u}(0) = \boldsymbol{\theta}.
\end{aligned}
\tag{4.29}
$$

The solution to (4.29) is $\boldsymbol{u}(t) = \exp(t\boldsymbol{A})\boldsymbol{\theta}$. Let $\boldsymbol{u}_{\text{measure}} \in \mathbb{R}^{147}$ contain the values of $\boldsymbol{u}$ corresponding to sensor locations at times $t = 1, 1.5, 2$. Then, by linearity, we can write $\boldsymbol{u}_{\text{measure}} = \boldsymbol{C}\boldsymbol{\theta}$ for a matrix $\boldsymbol{C}$.

Assume that $\boldsymbol{\theta} \sim N(\boldsymbol{\theta}_{\text{original}}, \boldsymbol{\Sigma}_{\text{original}})$, the discretization error is negligible, and that the measurements $\boldsymbol{d}$ are distorted by some noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\text{noise}})$ so that

$$
\boldsymbol{d} = \boldsymbol{u}_{\text{measure}} + \boldsymbol{\varepsilon}.
$$

It is well known that the posterior distribution of $\boldsymbol{\theta}$ is given by $\boldsymbol{\theta}|\boldsymbol{d} \sim N(\boldsymbol{\theta}_{\text{post}}, \boldsymbol{\Sigma}_{\text{post}})$ with

$$
\boldsymbol{\theta}_{\text{post}} = \boldsymbol{\Sigma}_{\text{post}}(\boldsymbol{C}^T \boldsymbol{\Sigma}_{\text{noise}}^{-1}\boldsymbol{d} + \boldsymbol{\Sigma}_{\text{original}}^{-1}\boldsymbol{\theta}_{\text{original}}), \quad \boldsymbol{\Sigma}_{\text{post}} = (\boldsymbol{C}^T \boldsymbol{\Sigma}_{\text{noise}}^{-1}\boldsymbol{C} + \boldsymbol{\Sigma}_{\text{original}}^{-1})^{-1};
$$

see [145]. Now let

$$
\boldsymbol{A}_{\text{pde}} = \boldsymbol{\Sigma}_{\text{original}}^{1/2}\boldsymbol{C}^T \boldsymbol{\Sigma}_{\text{noise}}^{-1}\boldsymbol{C}\boldsymbol{\Sigma}_{\text{original}}^{1/2}. \tag{4.30}
$$

Then, $\operatorname{tr}(\log(\boldsymbol{I} + \boldsymbol{A}_{\text{pde}}))$ is related to the expected information gain from the posterior distribution relative to the prior distribution [5]. For fine discretization grids, the matrix $\boldsymbol{C}$, and thus $\boldsymbol{A}_{\text{pde}}$, cannot be formed explicitly. Instead, one only implicitly performs matrix-vector products with $\boldsymbol{A}_{\text{pde}}$ via solving (4.29).

4.6.2 Verifying theoretical results

We will now verify the theoretical results proven in Section 4.2 and Section 4.3. In all our experiments, we begin with computing an orthonormal basis $\boldsymbol{Q}$ with $\ell \geq k$ columns. We will outline three different algorithms for doing so. Next, using the orthonormal basis $\boldsymbol{Q}$, we construct a Nyström approximation $\widehat{\boldsymbol{A}}$ as defined in (2.9) and the projection based approximation $\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A}$. Note that once $\boldsymbol{Q}$ is computed, constructing the two approximations comes at the same computational cost. Then, we truncate $\widehat{\boldsymbol{A}}$ and $\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A}$ to rank k to obtain $\widehat{\boldsymbol{A}}_{(k)}$ and $(\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A})_{(k)}$. Finally, we compare the following quantities

$$\begin{aligned}
\varepsilon_{\text{projection}} &= \frac{\|\boldsymbol{A} - (\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A})_{(k)}\|}{\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|} - 1;\\
\varepsilon_{\text{Nyström}} &= \frac{\|\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}\|}{\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|} - 1;\\
\varepsilon_{\text{funNyström}} &= \frac{\|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}}_{(k)})\|}{\|f(\boldsymbol{A}) - f(\boldsymbol{A}_{(k)})\|} - 1,
\end{aligned}$$

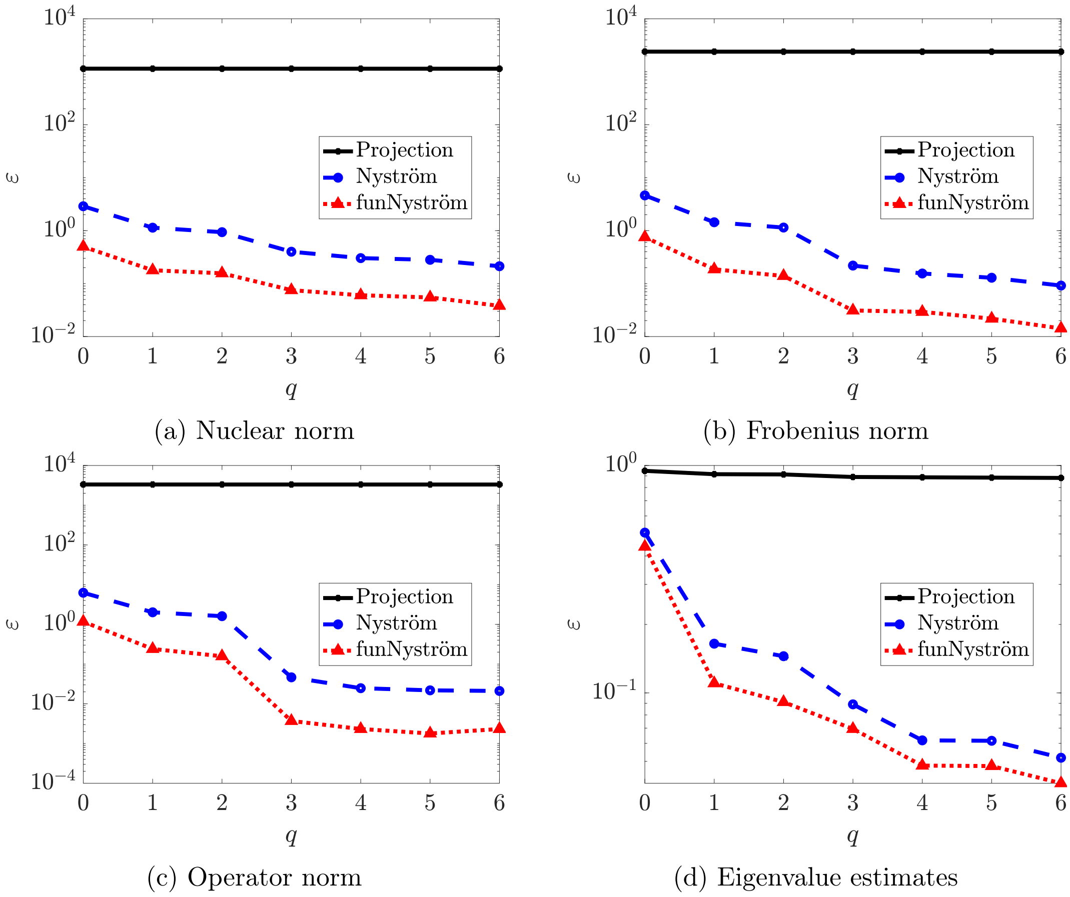
where $\|\cdot\| = \|\cdot\|_*, \|\cdot\|_{\text{F}}$ or $\|\cdot\|_2$. For comparing accuracy in estimating eigenvalues we use the alternative metrics:

$$\begin{aligned}
\varepsilon_{\text{projection}} &= \max_{i=1,\dots,k}\left\{\frac{\lambda_i - \sigma_i(\boldsymbol{Q}^T\boldsymbol{A})}{\lambda_i}\right\};\\
\varepsilon_{\text{Nyström}} &= \max_{i=1,\dots,k}\left\{\frac{\lambda_i - \widehat{\lambda}_i}{\lambda_i}\right\};\\
\varepsilon_{\text{funNyström}} &= \max_{i=1,\dots,k}\left\{\frac{f(\lambda_i) - f(\widehat{\lambda}_i)}{f(\lambda_i)}\right\}.
\end{aligned}$$

Our theory suggests that, for the Frobenius norm, nuclear norm, or for eigenvalue estimation, $\varepsilon_{\text{projection}} \geq \varepsilon_{\text{Nyström}} \geq \varepsilon_{\text{funNyström}}$. For the operator norm, we expect the second inequality to hold, but we have shown a counterexample to the first in Section 4.3.3. However, in our experiments we generally observe that $\varepsilon_{\text{projection}} \geq \varepsilon_{\text{Nyström}}$ even when $\|\cdot\| = \|\cdot\|_2$.

Column subset selection

In this experiment we compute the orthonormal basis $\boldsymbol{Q}$ using the randomly pivoted Cholesky algorithm [38, Algorithm 2.1]. In this setting, $\boldsymbol{Q} = \begin{bmatrix}\boldsymbol{e}_{i_1} & \dots & \boldsymbol{e}_{i_\ell}\end{bmatrix}$ where $\{i_1, \dots, i_\ell\} \subseteq \{1, \dots, n\}$ is an index set returned by the algorithm, and $\boldsymbol{e}_i$ is the i^{th} standard basis vector. We set $k = 10$ and $\ell = k + q$ for $q = 0, \dots, 6$. We let $\boldsymbol{A} = \boldsymbol{A}_{\text{disc}}$ be defined as in Section 4.6.1. We set the matrix function to be $f(x) = \frac{x}{x+1}$, which is operator monotone. The results are presented in Figure 4.1, which shows that $\varepsilon_{\text{projection}} \geq \varepsilon_{\text{Nyström}} \geq \varepsilon_{\text{funNyström}}$ for all norms and q we consider, which confirms our theoretical results.



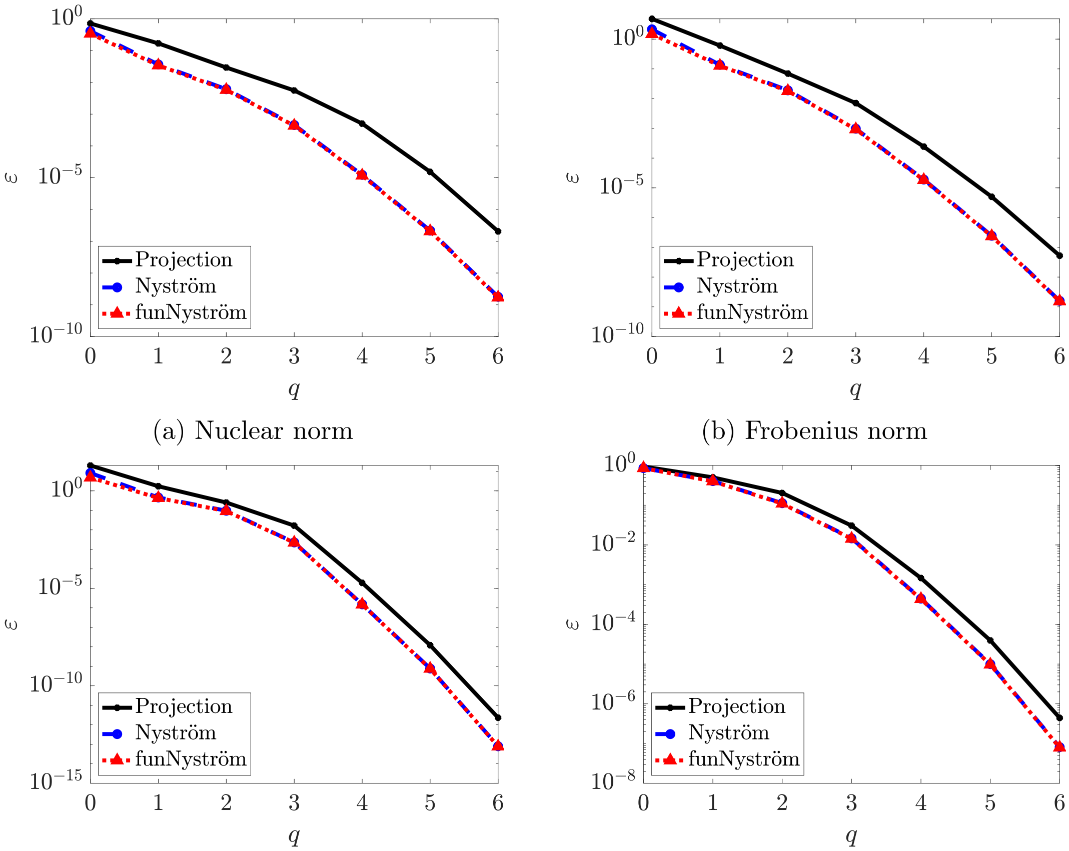
(a) Nuclear norm

(b) Frobenius norm

(c) Operator norm

(d) Eigenvalue estimates

Figure 4.1: Comparing $\varepsilon_{\text{projection}}, \varepsilon_{\text{Nyström}}$, and $\varepsilon_{\text{funNyström}}$ for column subset selection applied to $\boldsymbol{A}_{\text{disc}}$ defined in (4.25) and $f(x) = \frac{x}{x+1}$. Note that $\varepsilon_{\text{projection}}$ is significantly worse than the $\varepsilon_{\text{Nyström}}$ and $\varepsilon_{\text{funNyström}}$ since the orthogonal projection $\boldsymbol{Q}\boldsymbol{Q}^T$ zeros out all except ℓ rows of $\boldsymbol{A}$. In contrast, the Nyström approximation $\widehat{\boldsymbol{A}} = \boldsymbol{A}^{1/2}\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{Q}}\boldsymbol{A}^{1/2}$ effectively performs half a step of subspace iteration on $\boldsymbol{Q}$, giving a better approximation.



(a) Nuclear norm

(b) Frobenius norm

(c) Operator norm

(d) Eigenvalue estimates

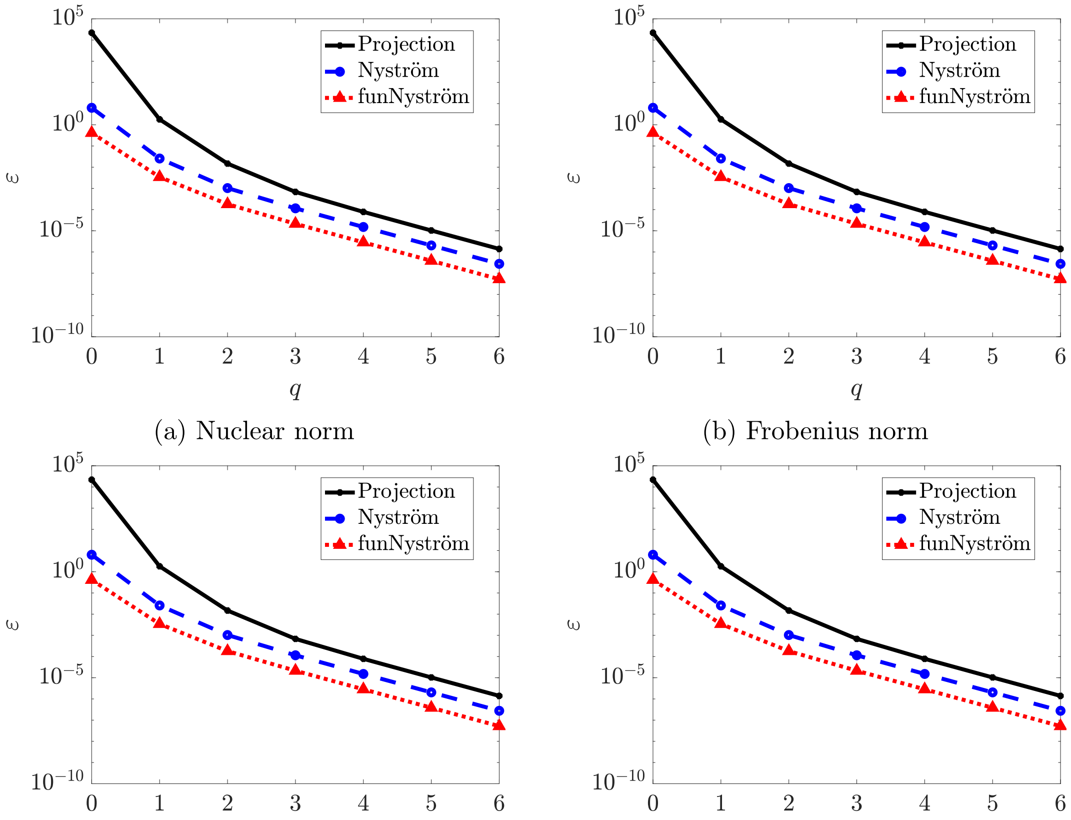
Figure 4.2: Comparing $\varepsilon_{\text{projection}}, \varepsilon_{\text{Nyström}}$, and $\varepsilon_{\text{funNyström}}$ for Krylov iteration applied $\boldsymbol{A}_{\text{alg}}$ defined in (4.24) with $n = 3000, C = c = 1$ and $f(x) = \log(1 + x)$.

Krylov subspace iteration

In this experiment we set $\boldsymbol{Q}$ to be an orthonormal basis for the Krylov subspace range($\begin{bmatrix}\boldsymbol{\Omega} & \boldsymbol{A\Omega} & \ldots & \boldsymbol{A}^q\boldsymbol{\Omega}\end{bmatrix}$) where $\boldsymbol{\Omega}$ is a random $3000 \times k$ matrix whose entries are i.i.d. $\mathcal{N}(0, 1)$ random variables. Recall that $\boldsymbol{Q}$ can be constructed using Algorithm 4. We set $k = 10$ and vary $q = 0, 1, \ldots, 6$. We set $\boldsymbol{A} = \boldsymbol{A}_{\text{alg}}$ with $n = 3000$, $C = c = 1$. We set the matrix function to be $f(x) = \log(1 + x)$, which is operator monotone. The results are presented in Figure 4.2. Again, for all norms and all choices of q, we see that $\varepsilon_{\text{projection}} \geq \varepsilon_{\text{Nyström}} \geq \varepsilon_{\text{funNyström}}$.

Subspace iteration

Finally, we set $\boldsymbol{Q}$ to be an orthonormal basis for range($\boldsymbol{A}^q\boldsymbol{\Omega}$) where $\boldsymbol{\Omega}$ is a random $3000 \times k$ matrix whose entries are i.i.d. $\mathcal{N}(0, 1)$ random variables. We set $k = 10$ and vary $q = 0, 1, \ldots, 6$. We set $\boldsymbol{A} = \boldsymbol{A}_{\text{exp}}$ with $n = 3000, C = 1, \gamma = e^{-1}$. We set the matrix function to be $f(x) = x^{1/2}$, which is operator monotone. The results are



(a) Nuclear norm

(b) Frobenius norm

(c) Operator norm

(d) Eigenvalue estimates

Figure 4.3: Comparing $\varepsilon_{\text{projection}}, \varepsilon_{\text{Nyström}}$, and $\varepsilon_{\text{funNyström}}$ for subspace iteration applied to $\boldsymbol{A}_{\text{exp}}$ defined in (4.24) with $n = 3000, C = 1, \gamma = e^{-1}$ and $f(x) = x^{1/2}$.

presented in Figure 4.3. Once again, for all norms and all choices of q, we see that $\varepsilon_{\text{projection}} \geq \varepsilon_{\text{Nyström}} \geq \varepsilon_{\text{funNyström}}$.

4.6.3 Comparing number of matrix vector products

In this section we compare the funNyström approximation $f(\widehat{\boldsymbol{A}})_{(k)}$ where $\widehat{\boldsymbol{A}}$ is defined as in (2.9) with $\boldsymbol{Q} \in \mathbb{R}^{n \times k}$ is an orthonormal basis for range($\boldsymbol{A}^{q-1}\boldsymbol{\Omega}$), for some $q \geq 1$.[2] We compare the approximation $f(\widehat{\boldsymbol{A}})_{(k)}$ returned by funNyström with the following references. Applying Nyström directly to $f(\boldsymbol{A})$ with an $n \times k$ Gaussian random matrix $\boldsymbol{\Omega}$ yields the

[2] Usually we let $\boldsymbol{Q}$ be an orthonormal basis for range($\boldsymbol{A}^{q}\boldsymbol{\Omega}$) for some $q \geq 0$ and $\boldsymbol{\Omega}$ with, say, k columns. Constructing the Nyström approximation as defined in (2.9) with this $\boldsymbol{Q}$ would require $(q+1)k$ matrix-vector products with $\boldsymbol{A}$. In the following sections we instead say that $\boldsymbol{Q}$ is an orthonormal basis for range($\boldsymbol{A}^{q-1}\boldsymbol{\Omega}$), where $q \geq 1$. Now, constructing the Nyström approximation as defined in (2.9) with this $\boldsymbol{Q}$ would require qk matrix-vector products with $\boldsymbol{A}$. We make this change since it simplifies counting the number of matrix-vector products with $\boldsymbol{A}$.

rank-k approximation

$$\widehat{\boldsymbol{B}}_{q,k} = f(\boldsymbol{A})^q \boldsymbol{\Omega}(\boldsymbol{\Omega}^T f(\boldsymbol{A})^{2q-1}\boldsymbol{\Omega})^\dagger (f(\boldsymbol{A})^q \boldsymbol{\Omega})^T. \tag{4.31}$$

This assumes that matrix-vector products with $f(\boldsymbol{A})$ are carried out *exactly*. If each matrix-vector product with $f(\boldsymbol{A})$ needed in (4.31) is approximated using d iterations of the Lanczos method, one obtains a different approximation, which will be denoted by $\widehat{\boldsymbol{B}}_{q,k}^{(d)}$. In our implementation, $\boldsymbol{B}_{q,k}^{(d)}$ is constructed by running k separate Lanczos iterations (Algorithm 4) for each column in $\boldsymbol{\Omega}$.

Recall that computing $f(\widehat{\boldsymbol{A}})_{(k)}$ requires qk matrix-vector products with $\boldsymbol{A}$. In contrast, the approximation $\widehat{\boldsymbol{B}}_{q,k}^{(d)}$ – obtained via applying Nyström to $f(\boldsymbol{A})$ – requires dqk matrix-vector products with $\boldsymbol{A}$. The choice of d, the number of Lanczos iterations, needs to be chosen in dependence of q, k such that the impact on the overall accuracy remains negligible. For the purpose of our numerical comparison, we have precomputed the matrix $\widehat{\boldsymbol{B}}_{q,k}$ obtained without the additional Lanczos approximation and choose d such that

$$\|f(\boldsymbol{A}) - \widehat{\boldsymbol{B}}_{q,k}^{(d)}\| \leq 1.1 \cdot \|f(\boldsymbol{A}) - \widehat{\boldsymbol{B}}_{q,k}\|. \tag{4.32}$$

In practice, $\widehat{\boldsymbol{B}}_{q,k}$ is not available and one needs to employ heuristic and potentially less reliable criteria. In our implementation we increase d by 5 until (4.32) is satisfied.

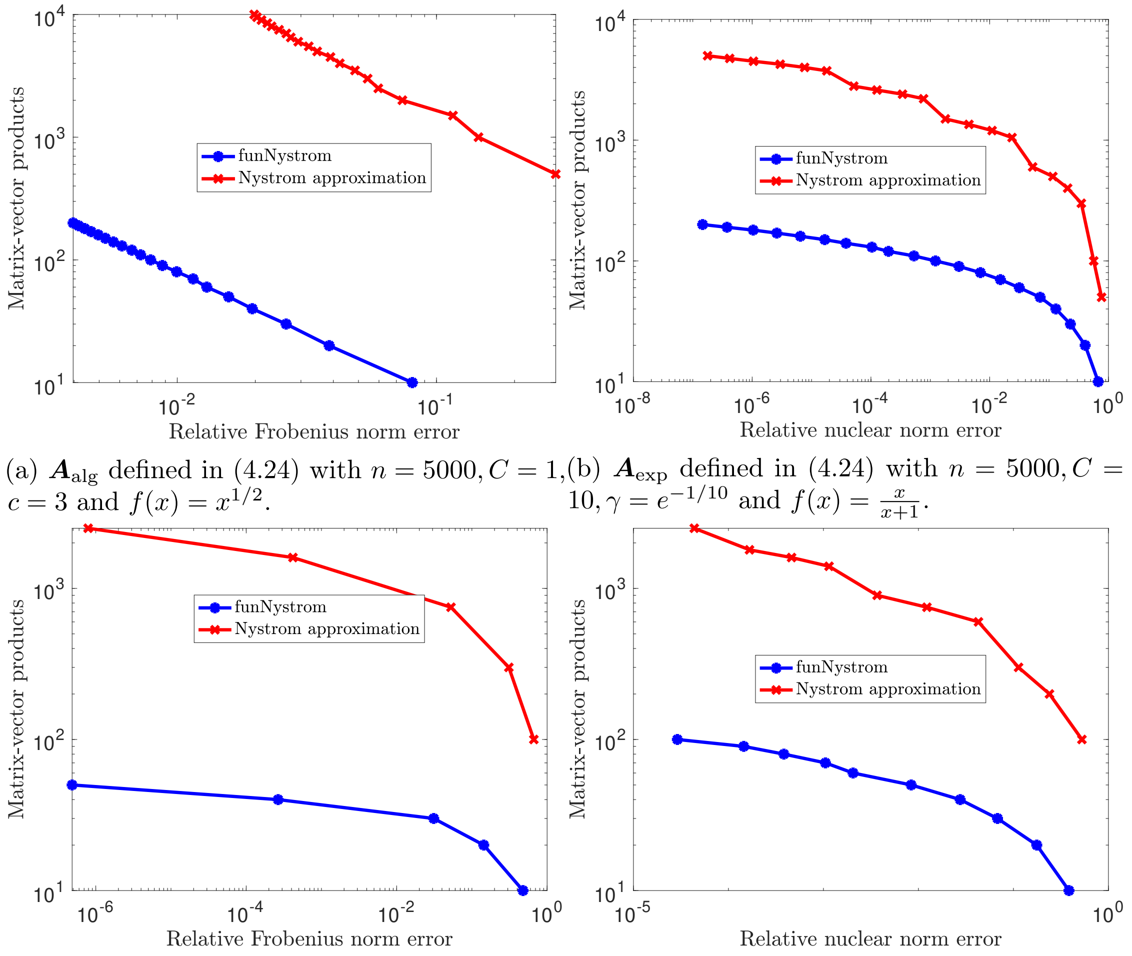
Given an approximation $\boldsymbol{B}$ of $f(\boldsymbol{A})$, we will measure the relative error $\|f(\boldsymbol{A})-\boldsymbol{B}\|/\|f(\boldsymbol{A})\|$ for some norm $\|\cdot\|$. The results obtained for $q = 1$ are reported in Figure 4.4. Clearly, funNyström needs fewer matrix-vector products; the difference can be up to three orders of magnitude.

4.6.4 Comparing accuracy

In Figure 4.5, we compare the approximation error of the funNyström approximation with the (significantly more expensive) approximation $\widehat{\boldsymbol{B}}_{q,k}$. It can be observed that funNyström is never worse than $\widehat{\boldsymbol{B}}_{q,k}$, and sometimes even better. This suggests that even when matrix-vector products with $f(\boldsymbol{A})$ can be performed very efficiently, funNyström may still be the preferred choice.

4.6.5 Fast computation of matrix-vector products

In this section we show that funNyström can be used to compute fast matrix-vector products with $f(\boldsymbol{A})$. We let $f(x) = x^{1/2}$ and $\boldsymbol{A} = \boldsymbol{A}_{\exp}$ defined in (4.24) with $C = 1, \gamma = e^{-1}$ and $n = 10000$. We let $\boldsymbol{\Phi} \in \mathbb{R}^{n \times N}$ be the matrix containing the first N columns of the identity matrix. Hence, computing $\boldsymbol{A}^{1/2}\boldsymbol{\Phi}$ requires N matrix-vector products with $\boldsymbol{A}^{1/2}$. We compare the computation times of the following two methods for approximating $\boldsymbol{A}^{1/2}\boldsymbol{\Phi}$:



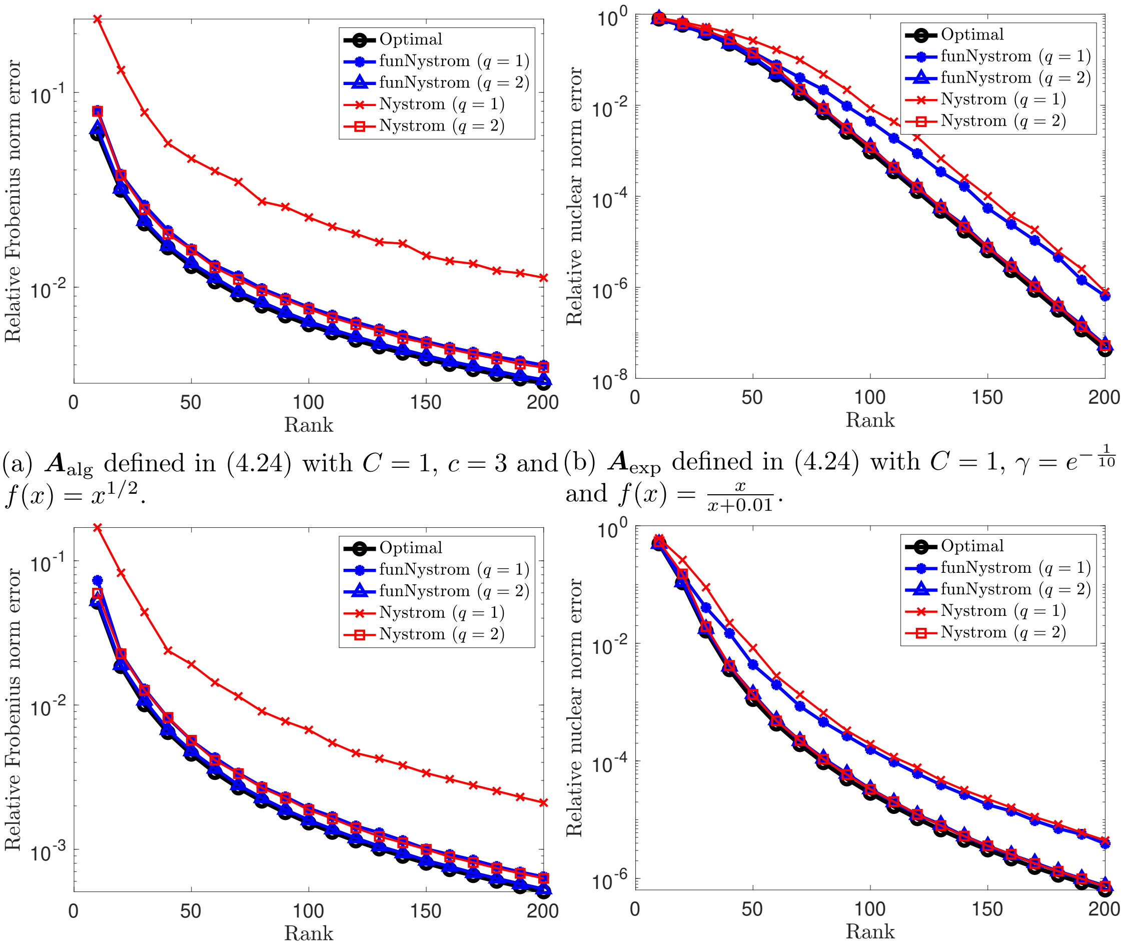
(a) $\boldsymbol{A}_{\text{alg}}$ defined in (4.24) with $n = 5000, C = 1, c = 3$ and $f(x) = x^{1/2}$.

(b) $\boldsymbol{A}_{\text{exp}}$ defined in (4.24) with $n = 5000, C = 10, \gamma = e^{-1/10}$ and $f(x) = \frac{x}{x+1}$.

(c) $\boldsymbol{A}_{\text{SE}}$ defined in (4.26) with $\ell^2 = 0.1$ and $f(x) = \log(1+x)$.

(d) $\boldsymbol{A}_{\text{pde}}$ defined in (4.30) with $\kappa = 0.01, \lambda = 1$, $\boldsymbol{\Sigma}_{\text{noise}} = \boldsymbol{I}$ and $f(x) = \log(1+x)$.

Figure 4.4: Number of matrix-vector products with $\boldsymbol{A}$ vs. attained accuracy for low-rank approximations of $f(\boldsymbol{A})$ from funNyström approximation (blue) and $\widehat{\boldsymbol{B}}_{1,k}^{(d)}$ (red).



(a) $\boldsymbol{A}_{\text{alg}}$ defined in (4.24) with $C=1$, $c=3$ and $f(x)=x^{1/2}$.

(b) $\boldsymbol{A}_{\text{exp}}$ defined in (4.24) with $C=1$, $\gamma=e^{-\frac{1}{10}}$ and $f(x)=\frac{x}{x+0.01}$.

(c) $\boldsymbol{A}_{\text{Mat}}$ defined in (4.27) with $\alpha=1$, $\nu=3/2$ and $f(x)=x^{1/2}$.

(d) $\boldsymbol{A}_{\text{Mat}}$ defined in (4.27) with $\alpha=1$, $\nu=5/2$ and $f(x)=\frac{x}{x+0.01}$.

Figure 4.5: Error vs. prescribed rank of the approximation for the funNyström approximation applied to $\boldsymbol{A}$ (blue) and $\widehat{\boldsymbol{B}}_{q,k}$, the Nyström approximation applied to $f(\boldsymbol{A})$ (red), for $q=1,2$.

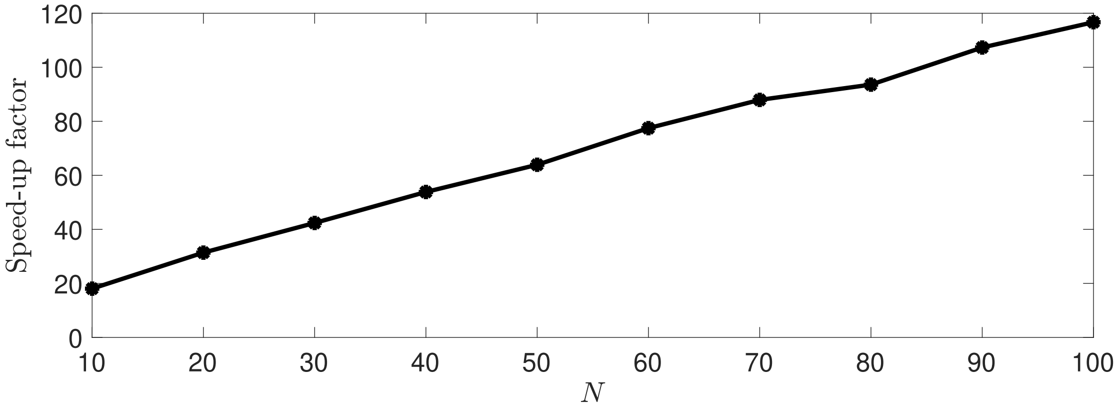


Figure 4.6: Number of matrix-vector products N vs. speed-up factor $T_{\text{Lanczos}}(N)/T_{\text{low-rank}}(N)$.

1. Approximating $\boldsymbol{A}^{1/2}\boldsymbol{\Phi}$ using the Lanczos method with d iterations. This comes at a computational cost of $O(dn^2N)$. The implementation we use for the Lanczos method is the same implementation used for the numerical expriments in [111], which approximates the N matrix-vector products with $\boldsymbol{A}^{1/2}$ simultaneously by vectorizing all computations, rather than approximating the N matrix-vector products subsequently. This significantly speeds up the computation.

2. Computing $\widehat{\boldsymbol{A}}^{1/2}$ using funNyström and approximate $\widehat{\boldsymbol{A}}^{1/2}\boldsymbol{\Phi} \approx \boldsymbol{A}^{1/2}\boldsymbol{\Phi}$. If $\widehat{\boldsymbol{A}}$ is the Nyström approximation defined in (2.9) with $\boldsymbol{Q}$ being an orthonormal basis for range$(\boldsymbol{A}^{q-1}\boldsymbol{\Omega})$ for $q \geq 1$ and a random standard Gaussian $n \times k$ matrix $\boldsymbol{\Omega}$, this comes at a computational cost of $O(qkn^2 + nkN)$.

If we let $T_{\text{Lanczos}}(N)$ and $T_{\text{low-rank}}(N)$ be wall-clock time to approximate N matrix-vector products with $\boldsymbol{A}^{1/2}$ using the Lanczos method and low-rank approximation respectively, the speed-up factor will be

$$\frac{T_{\text{Lanczos}}(N)}{T_{\text{low-rank}}(N)} = O\left(\frac{dn^2N}{qkn^2 + nkN}\right) = O(N),$$

if we keep d, q and k constant and assume $N \ll n$.

In our numerical experiments we set $k = 14, q = 1, d = 21$. This choice of parameters yields a similar relative error $\|\boldsymbol{A}^{1/2}\boldsymbol{\Phi} - \boldsymbol{Y}\|_{\text{F}}/\|\boldsymbol{A}^{1/2}\boldsymbol{\Phi}\|_{\text{F}} \approx 0.01$ for both methods and for all N, where $\boldsymbol{Y}$ is the approximation to $\boldsymbol{A}^{1/2}\boldsymbol{\Phi}$. We set $N = 10, 20, \ldots, 100$. The results are presented in Figure 4.6, which confirm the $O(N)$ speed-up factor.

4.6.6 Application to trace estimation

When an $n \times n$ matrix $\boldsymbol{B}$ admits an excellent rank-k approximation $\widehat{\boldsymbol{B}}_{(k)}$ for $k \ll n$, it is sensible to approximate $\text{tr}(\boldsymbol{B})$ by $\text{tr}(\widehat{\boldsymbol{B}}_{(k)})$. Setting $\boldsymbol{B} = f(\boldsymbol{A})$, this motivates the approximation

$$\operatorname{tr}(f(\boldsymbol{A})) \approx \operatorname{tr}(f(\widehat{\boldsymbol{A}}) = f(\hat{\lambda}_1) + \cdots + f(\hat{\lambda}_{k+p}), \tag{4.33}$$

where $\widehat{\boldsymbol{A}} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{U}}^T = \boldsymbol{AQ}(\boldsymbol{Q}^T\boldsymbol{AQ})^\dagger\boldsymbol{Q}^T\boldsymbol{A}$ denotes the funNyström approximation for some orthonormal basis $\boldsymbol{Q} \in \mathbb{R}^{n\times(k+p)}$. Using that $f(\boldsymbol{A}) \succeq f(\widehat{\boldsymbol{A}})$ we get

$$\operatorname{tr}(f(\boldsymbol{A})) - \operatorname{tr}(f(\widehat{\boldsymbol{A}}) = \|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}})\|_*.$$

Hence, Corollary 4.20 yields probabilistic bounds for the error of this trace approximation.

It is instructive to compare our results with the bounds from [138] for the special case $f(x) = \log(1+x)$. In particular, Theorem 1 from [138] states that

$$\begin{aligned}&\mathbb{E}\left[\operatorname{tr}\left(\log(\boldsymbol{I}+\boldsymbol{A})\right) - \operatorname{tr}\left(\log(\boldsymbol{I}+\boldsymbol{Q}^T\boldsymbol{AQ})\right)\right]\\ &\leq \operatorname{tr}\left(\log(\boldsymbol{I}+\boldsymbol{\Lambda}_2)\right) + \operatorname{tr}\left(\log(\boldsymbol{I}+\gamma^{2q-1}K\boldsymbol{\Lambda}_2)\right),\end{aligned} \tag{4.34}$$

where $\boldsymbol{Q}$ is an orthonormal basis for range($\boldsymbol{A}^q\boldsymbol{\Omega}$), and

$$K = \frac{e^2(k+p)}{(p+1)(p-1)}\left(\frac{1}{2\pi(p+1)}\right)^{\frac{2}{p+1}}(\sqrt{n-k}+\sqrt{k+p}+\sqrt{2})^2.$$

Constructing $\boldsymbol{Q}^T\boldsymbol{AQ}$ requires a total of $(q+1)(k+p)$ matrix-vector products with $\boldsymbol{A}$. On the other hand, within the same budget one obtains the more accurate low-rank approximation $\widehat{\boldsymbol{A}} = \boldsymbol{AQ}(\boldsymbol{Q}^T\boldsymbol{AQ})^\dagger\boldsymbol{Q}^T\boldsymbol{A}$. This also translates into tighter probabilistic bounds for trace estimation. To see this, note that Theorem 4.20 gives

$$\mathbb{E}\left[\operatorname{tr}\Big(\log(\boldsymbol{I}+\boldsymbol{A})\Big) - \operatorname{tr}\Big(\log(\boldsymbol{I}+\widehat{\boldsymbol{A}})\Big)\right] \leq \left(1 + \frac{\gamma^{2q}k}{p-1}\right)\operatorname{tr}\left(\log(\boldsymbol{I}+\boldsymbol{\Lambda}_2)\right).$$

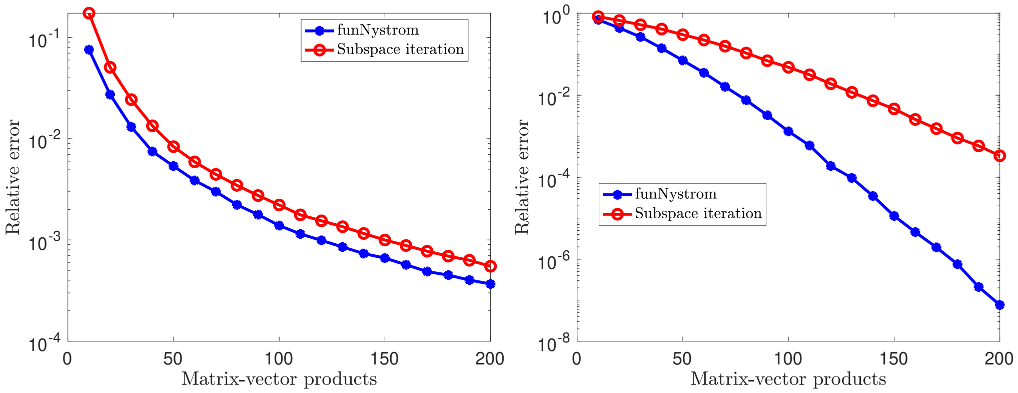
The difference between (4.34) and this bound satisfies

$$\sum_{i=k+1}^{n}\left[\log(1+\gamma^{2q-1}C\lambda_i) - \frac{\gamma^{2q}k\log(1+\lambda_i)}{p-1}\right] \approx \gamma^{2q-1}\sum_{i=k+1}^{n}\left(K - \frac{\gamma k}{p-1}\right)\lambda_i$$

for $\lambda_{k+1} \approx 0$.[3] Because $K \geq \frac{0.55kn}{(p+1)(p-1)}$ and usually $n \gg p$, this shows that we obtain a much tighter bound for our method compared to [138]. Similarly, it can be shown that our deviation bounds are tighter than those in [138]. Similar bounds for $f(x) = \frac{x}{x+1}$ exist in [83, Theorem A.1]. By an identical argument we can show that our bounds are tighter than those in [83].

In Figure 4.7 we compare our approach, funNystrom combined with (4.33), with the method presented in [138] to approximate $\operatorname{tr}(\log(\boldsymbol{I}+\boldsymbol{A}))$. For the method in [138] we let $\boldsymbol{Q}$ be an orthonormal basis for range($\boldsymbol{A\Omega}$) where $\boldsymbol{\Omega}$ is a random standard Gaussian matrix

[3] We use $\log(1+x) \approx x$ for small x.



(a) $\boldsymbol{A}_{\text{alg}}$ defined in (4.24) with $n = 5000, C = 100$ and $c = 2$.

(b) $\boldsymbol{A}_{\text{exp}}$ defined in (4.24) with $n = 5000, C = 100$ and $\gamma = 0.9$.

Figure 4.7: Approximation of $\operatorname{tr}(\log(\boldsymbol{I} + \boldsymbol{A}))$ using the funNyström approximation (blue) and the method presented in [138] (red). The x-axis represents the number of matrix-vector products performed with $\boldsymbol{A}$ to obtain the approximation, and the y-axis represents the relative error of the approximation.

with $m/2$ columns. For the funNyström approximation we let $\boldsymbol{Q}$ be an orthonormal basis for the range($\boldsymbol{\Phi}$), where $\boldsymbol{\Phi}$ is random standard Gaussian matrix with m columns. We make this choice since we have observed that increasing the rank of the low-rank approximation often yields a more accurate low-rank approximation than increasing the number of subspace iterations q. A budget of m matrix-vector products allows one to choose $k+p = m$ in the funNyström approximation while one can only choose $k+p = m/2$ in the method from [138]. This explains the better performance of funNystrom observed in Figure 4.7; a similar observation has been made in [123, Section 3].

5 Randomized block-Krylov subspace methods for low-rank approximation of matrix functions

Recall that the funNyström approximation presented in Chapter 4 can only be used for non-negative monotonically increasing functions, and the theory is only valid non-negative operator monotone functions. However, many matrix functions are not monotonically increasing and still admit accurate low-rank approximations. One such example is $\exp(-\beta \boldsymbol{A})$, which is numerically low-rank for large enough β. In this section we will consider an alternative algorithm that can compute low-rank approximations of general matrix functions of a symmetric matrix $\boldsymbol{A}$.

We proceed with highlighting an important difference in notation in this chapter compared to the other chapters in this thesis. We will consider a function f and a symmetric matrix $\boldsymbol{A}$ with eigenvalue decomposition partitioned as

$$\begin{aligned} \boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T &= \begin{bmatrix} \boldsymbol{U}_{f,1} & \boldsymbol{U}_{f,2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda}_{f,1} & \\ & \boldsymbol{\Lambda}_{f,2} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_{f,1}^T \\ \boldsymbol{U}_{f,2}^T \end{bmatrix}, \\ \boldsymbol{\Lambda}_{f,1} &= \operatorname{diag}(\lambda_{f,1}, \lambda_{f,2}, \cdots, \lambda_{f,k}), \\ \boldsymbol{\Lambda}_{f,2} &= \operatorname{diag}(\lambda_{f,k+1}, \lambda_{f,k+2}, \cdots, \lambda_{f,n}), \end{aligned} \tag{5.1}$$

where the eigenvalues are ordered so that

$$|f(\lambda_{f,1})| \geq |f(\lambda_{f,2})| \geq \ldots \geq |f(\lambda_{f,n})|.$$

This ordering of the eigenvalues and eigenvectors will be particularly useful for stating the theoretical results in this chapter. In particular, with this ordering we have the following equality for any unitarily invariant norm $\|\cdot\|$

$$\min_{\boldsymbol{B}:\operatorname{rank}(\boldsymbol{B})\leq k} \|f(\boldsymbol{A}) - \boldsymbol{B}\| = \|f(\boldsymbol{A}) - \boldsymbol{U}_{f,1} f(\boldsymbol{\Lambda}_{f,1})\boldsymbol{U}_{f,1}^T\| = \|f(\boldsymbol{\Lambda}_{f,2})\|.$$

In addition, we would also like to define the following variations of (2.4): for a sketching matrix $\boldsymbol{\Omega} \in \mathbb{R}^{n\times(k+p)}$, we define

$$\boldsymbol{\Omega}_{f,1} = \boldsymbol{U}_{f,1}^T\boldsymbol{\Omega}, \quad \boldsymbol{\Omega}_{f,2} = \boldsymbol{U}_{f,2}^T\boldsymbol{\Omega}. \tag{5.2}$$

In Section 5.1 we will present the randomized SVD applied to matrix functions and present an alternative and more efficient algorithm. This algorithm was initially presented by Chen and Hallman in the context of trace estimation [37]. One of our contributions is to provide an analysis of the method. The work [37] used the term *Krylov-aware* to describe the algorithm. Inspired by this, we call this low-rank approximation method *Krylov-aware low-rank approximation.* In Section 5.2 we present an error analysis of the algorithm. Finally, in Section 5.3 we present the numerical experiments, which confirms the excellent performance of the algorithm.

This chapter is based on the work in [122].

5.1 Krylov aware low-rank approximation

We now describe and motivate the algorithm that we will analyse. In Section 5.1.1 we outline how one would naively implement the randomized SVD for a matrix function $f(\boldsymbol{A})$. Next, in Section 5.1.2 we present the Krylov-aware low-rank approximation algorithm and why this method allows us to gain efficiencies. As previously mentioned, this algorithm was initially proposed by Chen and Hallman [37].

5.1.1 The randomized SVD for matrix functions

Recall from Section 2.2.2 that it is sometimes preferable for symmetric matrices to return a symmetric low-rank approximation as presented in (2.7). Algorithm 6 is a modification of Algorithm 1 applied to a symmetric matrix and returns a symmetric low-rank approximation. The algorithm returns either the rank $k+p \geq k$ approximation $\boldsymbol{QXQ}^T$ or the rank k approximation $\boldsymbol{QX}_{(k)}\boldsymbol{Q}^T$, depending on the needs of the user.

Algorithm 6 Symmetric Randomized SVD

```
input: Symmetric $\boldsymbol{B} \in \mathbb{R}^{n\times n}$. Target rank $k$. Oversampling parameter $p$.
output: Low-rank approximation to $\boldsymbol{B}$ in factored form $\boldsymbol{QXQ}^T = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{U}}^T$ or $\boldsymbol{QX}_{(k)}\boldsymbol{Q}^T = \widehat{\boldsymbol{U}}_1\widehat{\boldsymbol{\Lambda}}_1\widehat{\boldsymbol{U}}_1^T$.
Sample a random $n \times (k+p)$ sketch matrix $\boldsymbol{\Omega}$.
$\boldsymbol{Y} = \boldsymbol{B\Omega}$.
Compute an orthonormal basis $\boldsymbol{Q}$ for range$(\boldsymbol{Y})$.
Compute $\boldsymbol{X} = \boldsymbol{Q}^T\boldsymbol{BQ}$.
Compute the eigenvalue decomposition of $\boldsymbol{X} = \boldsymbol{W}\widehat{\boldsymbol{\Lambda}}\boldsymbol{W}^T$.
$\widehat{\boldsymbol{U}} = \boldsymbol{QW}$
return $\boldsymbol{P}_{\boldsymbol{B\Omega}}\boldsymbol{BP}_{\boldsymbol{B\Omega}} = \boldsymbol{QXQ}^T = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{U}}^T$ or $(\boldsymbol{P}_{\boldsymbol{B\Omega}}\boldsymbol{BP}_{\boldsymbol{B\Omega}})_{(k)} = \boldsymbol{QX}_{(k)}\boldsymbol{Q}^T = \widehat{\boldsymbol{U}}_1\widehat{\boldsymbol{\Lambda}}_1\widehat{\boldsymbol{U}}_1^T$.
```

The dominant cost of Algorithm 6 is the number of matrix-vector products with $\boldsymbol{B}$. Note that Algorithm 6 requires $2(k+p)$ matrix-vector products with $\boldsymbol{B}$: $(k+p)$ matrix-vector products in line 2 and $(k+p)$ matrix-vector products in line 4. When Algorithm 6 is applied to a matrix function $\boldsymbol{B} = f(\boldsymbol{A})$ these matrix-vector products cannot be performed exactly, but need to be approximated using, for example, the block Lanczos method as done in (3.6). Algorithm 7 implements the randomized SVD applied to $f(\boldsymbol{A})$ with approximate matrix-vector products using d and r iterations of the block Lanczos method. The cost is now $(d+r)(k+p)$ matrix-vector products with $\boldsymbol{A}$, where d and r should be set sufficiently large so that the approximations (3.6) and (3.7) are accurate.

Algorithm 7 Randomized SVD on a matrix function $f(\boldsymbol{A})$

```
input: Symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. Target rank $k$. Oversampling parameter $p$. Matrix function $f : \mathbb{R} \to \mathbb{R}$. Accuracy parameters $d$ and $r$.
output: Low-rank approximation to $f(\boldsymbol{A})$ in factored form $\boldsymbol{Q}\boldsymbol{X}\boldsymbol{Q}^T = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{U}}^T$ or $\boldsymbol{Q}\boldsymbol{X}_{(k)}\boldsymbol{Q}^T = \widehat{\boldsymbol{U}}_1\widehat{\boldsymbol{\Lambda}}_1\widehat{\boldsymbol{U}}_1^T$.
Sample a random $n \times (k+p)$ sketch matrix $\boldsymbol{\Omega}$.
Run Algorithm 4 for $d$ iterations to obtain an orthonormal basis $\boldsymbol{Q}_d$ for $\mathcal{K}_d(\boldsymbol{A}, \boldsymbol{\Omega})$, a block tridiagonal matrix $\boldsymbol{T}_d$ and an upper triangular matrix $\boldsymbol{R}_0$.
Compute the approximation $\boldsymbol{Y} = \boldsymbol{Q}_d f(\boldsymbol{T}_d)_{:,1:(k+p)}\boldsymbol{R}_0 \approx f(\boldsymbol{A})\boldsymbol{\Omega}$.
Compute an orthonormal basis $\boldsymbol{Q}$ for range($\boldsymbol{Y}$).
Run Algorithm 4 for $r$ iterations with starting block $\boldsymbol{Q}$ to obtain a block tridiagonal matrix $\widetilde{\boldsymbol{T}}_r$.
Compute the approximation $\boldsymbol{X} = f(\widetilde{\boldsymbol{T}}_r)_{1:(k+p),1:(k+p)} \approx \boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q}$.
Compute the eigenvalue decomposition of $\boldsymbol{X} = \boldsymbol{W}\widehat{\boldsymbol{\Lambda}}\boldsymbol{W}^T$.
$\widehat{\boldsymbol{U}} = \boldsymbol{Q}\boldsymbol{W}$
return $\boldsymbol{Q}\boldsymbol{X}\boldsymbol{Q}^T = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{U}}^T$ or $\boldsymbol{Q}\boldsymbol{X}_{(k)}\boldsymbol{Q}^T = \widehat{\boldsymbol{U}}_1\widehat{\boldsymbol{\Lambda}}_1\widehat{\boldsymbol{U}}_1^T$.
```

5.1.2 Krylov aware low-rank approximation

A key observation in [37] was that range($\boldsymbol{Q}$) $\subseteq$ range($\boldsymbol{Q}_d$), where $\boldsymbol{Q}$ and $\boldsymbol{Q}_d$ are as in Algorithm 7. Therefore, by Lemma 4.14 one has

$$\|f(\boldsymbol{A}) - \boldsymbol{Q}_d\boldsymbol{Q}_d^T f(\boldsymbol{A})\boldsymbol{Q}_d\boldsymbol{Q}_d^T\|_{\mathrm{F}} \leq \|f(\boldsymbol{A}) - \boldsymbol{Q}\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q}\boldsymbol{Q}^T\|_{\mathrm{F}},$$
$$\|f(\boldsymbol{A}) - \boldsymbol{Q}_d(\boldsymbol{Q}_d^T f(\boldsymbol{A})\boldsymbol{Q}_d)_{(k)}\boldsymbol{Q}_d^T\|_{\mathrm{F}} \leq \|f(\boldsymbol{A}) - \boldsymbol{Q}(\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q})_{(k)}\boldsymbol{Q}^T\|_{\mathrm{F}}.$$

Hence, assuming that the quadratic form $\boldsymbol{Q}_d^T f(\boldsymbol{A})\boldsymbol{Q}_d$ can be computed accurately, the naive implementation of the randomized SVD outlined in Algorithm 7 will yield a worse error than using $\boldsymbol{Q}_d\boldsymbol{Q}_d^T f(\boldsymbol{A})\boldsymbol{Q}_d\boldsymbol{Q}_d^T$ as an approximation to $f(\boldsymbol{A})$.

Since $\boldsymbol{Q}_d$ could have as many as $d(k+p)$ columns, an apparent downside to using $\boldsymbol{Q}_d$ to construct a low-rank approximation to $f(\boldsymbol{A})$ is that computing $f(\boldsymbol{A})\boldsymbol{Q}_d$ might require $rd(k+p)$ matrix-vector products with $\boldsymbol{A}$ if we run r iterations of the block Lanczos method. The key observation which allows Krylov-aware algorithms to be implemented efficiently is the following result.

Lemma 5.1 ([37, Section 3]). *Suppose that $\boldsymbol{Q}_d$ is the output of Algorithm 4 with starting block $\boldsymbol{\Omega}$ and d iterations. Then, running $r+1$ iterations of Algorithm 4 with starting block $\boldsymbol{Q}_d$ yields the same output as running $d+r$ iterations of Algorithm 4 with starting block $\boldsymbol{\Omega}$.*

This observation enables us to approximate $\boldsymbol{Q}_d^T f(\boldsymbol{A})\boldsymbol{Q}_d$ with just $r(k+p)$ additional matrix-vector products with $\boldsymbol{A}$, even though $\boldsymbol{Q}_d$ has many more than $k+p$ columns. Hence, approximating $\boldsymbol{Q}_d^T f(\boldsymbol{A})\boldsymbol{Q}_d$ is essentially as costly, in terms of the number of matrix-vector products with $\boldsymbol{A}$, as approximating $\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q}$ as done in line 6 of Algorithm 7.

We can now present the Krylov-aware low-rank approximation algorithm; see Algorithm 8. The total number of matrix-vector products with $\boldsymbol{A}$ is $(d+r)(k+p)$, the same as Algorithm 7. However, as noted above, Algorithm 8 uses a better projection space. We further note that the function f in Algorithm 8 does not need to be fixed; one can compute a low-rank approximation for many different functions f at minimal additional cost.

Algorithm 8 Krylov aware low-rank approximation

```
input: Symmetric $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Target rank $k$. Oversampling parameter $p$. Matrix function $f : \mathbb{R} \to \mathbb{R}$. Number of iterations $q = d + r$.
output: Low-rank approximation to $f(\boldsymbol{A})$ in factor form $\boldsymbol{Q}_d\boldsymbol{X}\boldsymbol{Q}_d^T = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{U}}^T$ or $\boldsymbol{Q}_d\boldsymbol{X}_{(k)}\boldsymbol{Q}_d^T = \widehat{\boldsymbol{U}}_1\widehat{\boldsymbol{\Lambda}}_1\widehat{\boldsymbol{U}}_1^T$.
1: Sample a random $n \times (k+p)$ sketch matrix $\boldsymbol{\Omega}$.
2: Run Algorithm 4 for $q = d + r$ iterations to obtain an orthonormal basis $\boldsymbol{Q}_d$ for $\mathcal{K}_d(\boldsymbol{A}, \boldsymbol{\Omega})$ and a block tridiagonal matrix $\boldsymbol{T}_q$.
3: Compute $\boldsymbol{X} = f(\boldsymbol{T}_q)_{1:n_d,1:n_d}$ where $n_d = \dim(\mathcal{K}_d(\boldsymbol{A}, \boldsymbol{\Omega}))$.   ▷ $\approx \boldsymbol{Q}_d^T f(\boldsymbol{A})\boldsymbol{Q}_d$
4: Compute the eigenvalue decomposition of $\boldsymbol{X} = \boldsymbol{W}\widehat{\boldsymbol{\Lambda}}\boldsymbol{W}^T$.
5: $\widehat{\boldsymbol{U}} = \boldsymbol{Q}_d\boldsymbol{W}$.
6: return $\boldsymbol{Q}_d\boldsymbol{X}\boldsymbol{Q}_d^T = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{U}}^T$ or $\boldsymbol{Q}_d\boldsymbol{X}_{(k)}\boldsymbol{Q}_d^T = \widehat{\boldsymbol{U}}_1\widehat{\boldsymbol{\Lambda}}_1\widehat{\boldsymbol{U}}_1^T$.
```

5.2 Error bounds

In this section we will establish Frobenius norm error bounds for Algorithm 8. In Section 5.2.1 we derive error bounds for approximations of $f(\boldsymbol{A})$ when projections $\boldsymbol{Q}\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q}\boldsymbol{Q}^T$ cannot be computed exactly. In Section 5.2.2 we provide structural bounds for the errors $\|f(\boldsymbol{A}) - \boldsymbol{Q}_d\boldsymbol{Q}_d^T f(\boldsymbol{A})\boldsymbol{Q}_d\boldsymbol{Q}_d^T\|_{\mathrm{F}}$ and $\|f(\boldsymbol{A}) - \boldsymbol{Q}_d(\boldsymbol{Q}_d^T f(\boldsymbol{A})\boldsymbol{Q}_d)_{(k)}\boldsymbol{Q}_d^T\|_{\mathrm{F}}$ that hold with probability 1, and in Section 5.2.3 we derive the corresponding probabilistic bounds. Next, we combine the results from Sections 5.2.1 to 5.2.3 to derive error bounds for Algorithm 8. Finally, in Sections 5.2.5 and 5.2.6 we apply our results to the matrix exponential and the identity function.

5.2.1 Error bounds for inexact projections

In this section we will derive error bounds for $\|f(\boldsymbol{A})-\boldsymbol{Q}\boldsymbol{X}\boldsymbol{Q}^T\|_{\mathrm{F}}$ and $\|f(\boldsymbol{A})-\boldsymbol{Q}\boldsymbol{X}_{(k)}\boldsymbol{Q}^T\|_{\mathrm{F}}$ where $\boldsymbol{Q}$ is *any* orthonormal basis and $\boldsymbol{X}$ is *any* matrix. By Lemma 4.14 we know that the optimal choice of $\boldsymbol{X}$ is $\boldsymbol{X} = \boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q}$. However, since $\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q}$ can only be approximated we need to show that the errors $\|f(\boldsymbol{A}) - \boldsymbol{Q}\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q}\boldsymbol{Q}^T\|_{\mathrm{F}}$ and $\|f(\boldsymbol{A}) - \boldsymbol{Q}(\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q})_{(k)}\boldsymbol{Q}^T\|_{\mathrm{F}}$ are robust against perturbations in $\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q}$. Theorem 5.2 provides such a result.

Theorem 5.2. *Given an orthonormal basis $\boldsymbol{Q}$ and a matrix $\boldsymbol{X}$ of the same size as the matrix $\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q}$. Then,*

$$\|f(\boldsymbol{A}) - \boldsymbol{Q}\boldsymbol{X}\boldsymbol{Q}^T\|_{\mathrm{F}}^2 = \|f(\boldsymbol{A}) - \boldsymbol{Q}\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q}\boldsymbol{Q}^T\|_{\mathrm{F}}^2 + \|\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q} - \boldsymbol{X}\|_{\mathrm{F}}^2, \tag{5.3}$$

and

$$\|f(\boldsymbol{A}) - \boldsymbol{Q}\boldsymbol{X}_{(k)}\boldsymbol{Q}^T\|_{\mathrm{F}} \leq \|f(\boldsymbol{A}) - \boldsymbol{Q}(\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q})_{(k)}\boldsymbol{Q}^T\|_{\mathrm{F}} + 2\|\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q} - \boldsymbol{X}\|_{\mathrm{F}}. \tag{5.4}$$

Proof. (5.3) was proven in the proof of Lemma 4.14, but we repeat the proof for completeness. Note that for any matrix $\boldsymbol{B}$ we have $\langle f(\boldsymbol{A}) - \boldsymbol{Q}\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q}\boldsymbol{Q}^T, \boldsymbol{Q}\boldsymbol{B}\boldsymbol{Q}\rangle = 0$. Hence, by the Pythagorean theorem, we have

$$\begin{aligned}\|f(\boldsymbol{A}) - \boldsymbol{Q}\boldsymbol{X}\boldsymbol{Q}^T\|_{\mathrm{F}}^2 &= \|f(\boldsymbol{A}) - \boldsymbol{Q}\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q}\boldsymbol{Q}^T + \boldsymbol{Q}(\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q} - \boldsymbol{X})\boldsymbol{Q}^T\|_{\mathrm{F}}^2 \\ &= \|f(\boldsymbol{A}) - \boldsymbol{Q}\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q}\boldsymbol{Q}^T\|_{\mathrm{F}}^2 + \|\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q} - \boldsymbol{X}\|_{\mathrm{F}}^2,\end{aligned}$$

as required.

We now proceed with proving (5.4) using a similar argument to [153, Proof of Theorem 5.1]. Define $\boldsymbol{C} = f(\boldsymbol{A}) - \boldsymbol{Q}\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q}\boldsymbol{Q}^T + \boldsymbol{Q}\boldsymbol{X}\boldsymbol{Q}^T$. Note that $\|\boldsymbol{C} - f(\boldsymbol{A})\|_{\mathrm{F}} = \|\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q} - \boldsymbol{X}\|_{\mathrm{F}}$ and $\boldsymbol{Q}^T\boldsymbol{C}\boldsymbol{Q} = \boldsymbol{X}$. Hence,

$$\begin{aligned}&\|f(\boldsymbol{A}) - \boldsymbol{Q}\boldsymbol{X}_{(k)}\boldsymbol{Q}^T\|_{\mathrm{F}} = \|f(\boldsymbol{A}) - \boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{C}\boldsymbol{Q})_{(k)}\boldsymbol{Q}^T\|_{\mathrm{F}} \\ \leq&\|f(\boldsymbol{A}) - \boldsymbol{C}\|_{\mathrm{F}} + \|\boldsymbol{C} - \boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{C}\boldsymbol{Q})_{(k)}\boldsymbol{Q}^T\|_{\mathrm{F}} \\ =&\|\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q} - \boldsymbol{X}\|_{\mathrm{F}} + \|\boldsymbol{C} - \boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{C}\boldsymbol{Q})_{(k)}\boldsymbol{Q}^T\|_{\mathrm{F}} \\ \leq&\|\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q} - \boldsymbol{X}\|_{\mathrm{F}} + \|\boldsymbol{C} - \boldsymbol{Q}(\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q})_{(k)}\boldsymbol{Q}^T\|_{\mathrm{F}} \qquad \text{(Lemma 4.14)} \\ \leq&2\|\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q} - \boldsymbol{X}\|_{\mathrm{F}} + \|f(\boldsymbol{A}) - \boldsymbol{Q}(\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q})_{(k)}\boldsymbol{Q}^T\|_{\mathrm{F}},\end{aligned}$$

as required. □

5.2.2 Structural bounds

In this section we derive a structural bound for $\|f(\boldsymbol{A})-\boldsymbol{Q}_d\boldsymbol{Q}_d^T f(\boldsymbol{A})\boldsymbol{Q}_d\boldsymbol{Q}_d^T\|_{\mathrm{F}}$ and $\|f(\boldsymbol{A})-\boldsymbol{Q}_d(\boldsymbol{Q}_d^T f(\boldsymbol{A})\boldsymbol{Q}_d)_{(k)}\boldsymbol{Q}_d^T\|_{\mathrm{F}}$ that is true for *any* sketch matrix $\boldsymbol{\Omega}$ as long as $\boldsymbol{\Omega}_{f,1}$ defined in

(5.2) has rank k. This bound will allow us to obtain probabilistic bounds.

Introduce the quantity

$$\mathcal{E}_{\mathbf{\Omega}}(d;f) = \min_{g\in\mathbb{P}_{d-1}} \left[\|g(\mathbf{\Lambda}_{f,2})\mathbf{\Omega}_{f,2}\mathbf{\Omega}_{f,1}^{\dagger}\|_{\mathrm{F}}^2 \max_{i=1,\ldots,k} \left| \frac{f(\lambda_{f,i})}{g(\lambda_{f,i})} \right|^2 \right], \tag{5.5}$$

which relates to how well a polynomial can be large (relative to f) on the eigenvalues $\lambda_{f,1},\ldots,\lambda_{f,k}$ and small on the remaining eigenvalues.

Lemma 5.3. *Consider $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ as defined in* (5.1). *Assuming $\mathbf{\Omega}_{f,1}$ in* (5.2) *has rank k, for all functions $f : \mathbb{R} \mapsto \mathbb{R}$, we have*

$$\begin{aligned} &\|f(\boldsymbol{A}) - \boldsymbol{Q}_d\boldsymbol{Q}_d^T f(\boldsymbol{A})\boldsymbol{Q}_d\boldsymbol{Q}_d^T\|_{\mathrm{F}}^2 \\ &\qquad \le \|f(\boldsymbol{A}) - \boldsymbol{Q}_d(\boldsymbol{Q}_d^T f(\boldsymbol{A})\boldsymbol{Q}_d)_{(k)}\boldsymbol{Q}_d^T\|_{\mathrm{F}}^2 \\ &\qquad\qquad \le \|f(\mathbf{\Lambda}_{f,2})\|_{\mathrm{F}}^2 + 5\mathcal{E}_{\mathbf{\Omega}}(d;f). \end{aligned} \tag{5.6}$$

Proof. The first inequality is immediate due to the fact that $\boldsymbol{Q}_d\boldsymbol{Q}_d^T f(\boldsymbol{A})\boldsymbol{Q}_d\boldsymbol{Q}_d^T$ is the nearest matrix to $f(\boldsymbol{A})$ in the Frobenius norm whose range and co-range is contained in range$(\boldsymbol{Q}_d)$ (Lemma 4.14).

We proceed with proving the second inequality. Choose any $g \in \mathbb{P}_{d-1}$. Note that if we choose g so that $g(\lambda_{f,i}) = 0$ for some $i = 1,\ldots,k$ then the right hand side of (5.6) is infinite and the bound trivially holds. Hence, we may assume that $g(\lambda_{f,i}) \neq 0$ for $i = 1,\ldots,k$. Consequently, $g(\mathbf{\Lambda}_{f,1})$ is non-singular. Define $\boldsymbol{Z} = g(\boldsymbol{A})\mathbf{\Omega}\mathbf{\Omega}_{f,1}^{\dagger}g(\mathbf{\Lambda}_{f,1})^{-1}$ and let $\widetilde{\boldsymbol{P}}$ be the orthogonal projector onto range$(\boldsymbol{Z}) \subseteq$ range$(\boldsymbol{Q}_s)$. Note that rank$(\boldsymbol{Z}) \le k$ and $\boldsymbol{Q}_d(\boldsymbol{Q}_d^T f(\boldsymbol{A})\boldsymbol{Q}_d)_{(k)}\boldsymbol{Q}_d^T$ is the best rank k approximation to $f(\boldsymbol{A})$ whose range and co-range is contained in range$(\boldsymbol{Q}_d)$ (Lemma 4.14). Hence,

$$\|f(\boldsymbol{A}) - \boldsymbol{Q}_d(\boldsymbol{Q}_d^T f(\boldsymbol{A})\boldsymbol{Q}_d)_{(k)}\boldsymbol{Q}_d^T\|_{\mathrm{F}}^2 \le \|f(\boldsymbol{A}) - \widetilde{\boldsymbol{P}} f(\boldsymbol{A})\widetilde{\boldsymbol{P}}\|_{\mathrm{F}}^2.$$

Now define $\widehat{\boldsymbol{P}} = \boldsymbol{U}^T\widetilde{\boldsymbol{P}}\boldsymbol{U}$, which is the orthogonal projector onto range$(\boldsymbol{U}^T\boldsymbol{Z})$. By the unitary invariance of the Frobenius norm we have

$$\|f(\boldsymbol{A}) - \widetilde{\boldsymbol{P}} f(\boldsymbol{A})\widetilde{\boldsymbol{P}}\|_{\mathrm{F}}^2 = \|f(\mathbf{\Lambda}) - \widehat{\boldsymbol{P}} f(\mathbf{\Lambda})\widehat{\boldsymbol{P}}\|_{\mathrm{F}}^2.$$

Furthermore,

$$\|f(\mathbf{\Lambda}) - \widehat{\boldsymbol{P}} f(\mathbf{\Lambda})\widehat{\boldsymbol{P}}\|_{\mathrm{F}}^2 = \|(\boldsymbol{I} - \widehat{\boldsymbol{P}})f(\mathbf{\Lambda})\|_{\mathrm{F}}^2 + \|\widehat{\boldsymbol{P}} f(\mathbf{\Lambda})(\boldsymbol{I} - \widehat{\boldsymbol{P}})\|_{\mathrm{F}}^2. \tag{5.7}$$

We are going to bound the two terms on the right hand side of (5.7) separately, as done in Theorem 4.21. We begin with bounding the first term in (5.7). Our analysis is similar to the proof of [79, Theorem 9.1].

Note that since $\operatorname{rank}(\boldsymbol{\Omega}_{f,1}) = k$ we have $\boldsymbol{\Omega}_{f,1}\boldsymbol{\Omega}_{f,1}^\dagger = \boldsymbol{I}$. Hence,

$$\boldsymbol{U}^T\boldsymbol{Z} = \boldsymbol{U}^T g(\boldsymbol{A})\boldsymbol{\Omega}\boldsymbol{\Omega}_{f,1}^\dagger g(\boldsymbol{\Lambda}_{f,1})^{-1} = \begin{bmatrix} \boldsymbol{I} \\ g(\boldsymbol{\Lambda}_{f,2})\boldsymbol{\Omega}_{f,2}\boldsymbol{\Omega}_{f,1}^\dagger g(\boldsymbol{\Lambda}_{f,1})^{-1} \end{bmatrix} := \begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{F} \end{bmatrix}.$$

Hence,

$$\begin{aligned} \boldsymbol{I} - \widehat{\boldsymbol{P}} &= \begin{bmatrix} \boldsymbol{I} - (\boldsymbol{I} + \boldsymbol{F}^T\boldsymbol{F})^{-1} & -(\boldsymbol{I} + \boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T \\ -\boldsymbol{F}(\boldsymbol{I} + \boldsymbol{F}^T\boldsymbol{F})^{-1} & \boldsymbol{I} - \boldsymbol{F}(\boldsymbol{I} + \boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T \end{bmatrix} \\ &\preceq \begin{bmatrix} \boldsymbol{F}^T\boldsymbol{F} & -(\boldsymbol{I} + \boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T \\ -\boldsymbol{F}(\boldsymbol{I} + \boldsymbol{F}^T\boldsymbol{F})^{-1} & \boldsymbol{I} \end{bmatrix}, \end{aligned} \tag{5.8}$$

where the inequality is due to [79, Proposition 8.2]. Consequently, using (5.8) we have

$$\begin{aligned} \|(\boldsymbol{I} - \widehat{\boldsymbol{P}})f(\boldsymbol{\Lambda})\|_\mathrm{F}^2 &= \operatorname{tr}(f(\boldsymbol{\Lambda})(\boldsymbol{I} - \widehat{\boldsymbol{P}})f(\boldsymbol{\Lambda})) \\ &\leq \|\boldsymbol{F}f(\boldsymbol{\Lambda}_{f,1})\|_\mathrm{F}^2 + \|f(\boldsymbol{\Lambda}_{f,2})\|_\mathrm{F}^2 \\ &\leq \|f(\boldsymbol{\Lambda}_{f,2})\|_\mathrm{F}^2 + \|g(\boldsymbol{\Lambda}_{f,2})\boldsymbol{\Omega}_{f,2}\boldsymbol{\Omega}_{f,1}^\dagger\|_\mathrm{F}^2 \|g(\boldsymbol{\Lambda}_{f,1})^{-1}f(\boldsymbol{\Lambda}_{f,1})\|_2^2 \\ &= \|f(\boldsymbol{\Lambda}_{f,2})\|_\mathrm{F}^2 + \|g(\boldsymbol{\Lambda}_{f,2})\boldsymbol{\Omega}_{f,2}\boldsymbol{\Omega}_{f,1}^\dagger\|_\mathrm{F}^2 \max_{i=1,\ldots,k} \left|\frac{f(\lambda_{f,i})}{g(\lambda_{f,i})}\right|^2. \end{aligned} \tag{5.9}$$

We proceed with bounding the second term in (5.7). By the triangle inequality we have

$$\|\widehat{\boldsymbol{P}}f(\boldsymbol{\Lambda})(\boldsymbol{I} - \widehat{\boldsymbol{P}})\|_\mathrm{F} \leq \left\| \begin{bmatrix} \boldsymbol{0} & \\ & f(\boldsymbol{\Lambda}_{f,2}) \end{bmatrix} \widehat{\boldsymbol{P}} \right\|_\mathrm{F} + \left\| (\boldsymbol{I} - \widehat{\boldsymbol{P}}) \begin{bmatrix} f(\boldsymbol{\Lambda}_{f,1}) & \\ & \boldsymbol{0} \end{bmatrix} \right\|_\mathrm{F}.$$

Using a similar argument as in (5.9) we have

$$\left\| (\boldsymbol{I} - \widehat{\boldsymbol{P}}) \begin{bmatrix} f(\boldsymbol{\Lambda}_{f,1}) & \\ & \boldsymbol{0} \end{bmatrix} \right\|_\mathrm{F} \leq \|g(\boldsymbol{\Lambda}_{f,2})\boldsymbol{\Omega}_{f,2}\boldsymbol{\Omega}_{f,1}^\dagger\|_\mathrm{F} \max_{i=1,\ldots,k} \left|\frac{f(\lambda_{f,i})}{g(\lambda_{f,i})}\right|, \tag{5.10}$$

and since $\boldsymbol{F}(\boldsymbol{I} + \boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T \preceq \boldsymbol{F}\boldsymbol{F}^T$ we have

$$\begin{aligned} \left\| \begin{bmatrix} \boldsymbol{0} & \\ & f(\boldsymbol{\Lambda}_{f,2}) \end{bmatrix} \widehat{\boldsymbol{P}} \right\|_\mathrm{F}^2 &= \operatorname{tr}\left(\begin{bmatrix} \boldsymbol{0} & \\ & f(\boldsymbol{\Lambda}_{f,2}) \end{bmatrix} \widehat{\boldsymbol{P}} \begin{bmatrix} \boldsymbol{0} & \\ & f(\boldsymbol{\Lambda}_{f,2}) \end{bmatrix} \right) \\ &= \operatorname{tr}(f(\boldsymbol{\Lambda}_{f,2})\boldsymbol{F}(\boldsymbol{I} + \boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T f(\boldsymbol{\Lambda}_{f,2})) \\ &\leq \operatorname{tr}(f(\boldsymbol{\Lambda}_{f,2})\boldsymbol{F}\boldsymbol{F}^T f(\boldsymbol{\Lambda}_{f,2})) = \|f(\boldsymbol{\Lambda}_{f,2})\boldsymbol{F}\|_\mathrm{F} \\ &\leq \|f(\boldsymbol{\Lambda}_{f,2})\|_2^2 \|g(\boldsymbol{\Lambda}_{f,1})^{-1}\|_2^2 \|g(\boldsymbol{\Lambda}_{f,2})\boldsymbol{\Omega}_{f,2}\boldsymbol{\Omega}_{f,1}^\dagger\|_\mathrm{F}^2 \\ &\leq \|g(\boldsymbol{\Lambda}_{f,2})\boldsymbol{\Omega}_{f,2}\boldsymbol{\Omega}_{f,1}^\dagger\|_\mathrm{F}^2 \max_{i=1,\ldots,k} \left|\frac{f(\lambda_{f,i})}{g(\lambda_{f,i})}\right|^2. \end{aligned} \tag{5.11}$$

Inserting the bounds (5.9), (5.10), and (5.11) into (5.7) and optimizing over $\mathbb{P}_{d-1}$ yields the desired inequality. □

Remark 5.1. *Note that the proof of Lemma 5.3 allows us to prove Theorem 2.8 by choosing $g(x) = x^q$ and using that* $\|\mathbf{\Lambda}_2^q \mathbf{\Omega}_2 \mathbf{\Omega}_1^\dagger\|_{\mathrm{F}}^2 \max\limits_{i=1,\dots,k} \left|\frac{\lambda_i}{\lambda_i^q}\right|^2 \leq \left|\frac{\lambda_{k+1}}{\lambda_k}\right|^{2(q-1)} \|\mathbf{\Lambda}_2 \mathbf{\Omega}_2 \mathbf{\Omega}_1^\dagger\|_{\mathrm{F}}^2$.

5.2.3 Probabilistic bounds

With the structural bound available, we are ready to derive the probabilistic bounds for $\|f(\boldsymbol{A}) - \boldsymbol{Q}_d \boldsymbol{Q}_d^T f(\boldsymbol{A}) \boldsymbol{Q}_d \boldsymbol{Q}_d^T\|_{\mathrm{F}}$ and $\|f(\boldsymbol{A}) - \boldsymbol{Q}_d (\boldsymbol{Q}_d^T f(\boldsymbol{A}) \boldsymbol{Q}_d)_{(k)} \boldsymbol{Q}_d^T\|_{\mathrm{F}}$. Note that by Lemma 5.3 it is sufficient to derive a probabilistic bound for $\mathcal{E}_{\mathbf{\Omega}}(d; f)$ defined in (5.5).

We will bound $\mathcal{E}_{\mathbf{\Omega}}(d; f)$ in terms of a deterministic quantity

$$\mathcal{E}(d; f) = \min_{g \in \mathbb{P}_{d-1}} \left[\|g(\mathbf{\Lambda}_{f,2})\|_{\mathrm{F}}^2 \max_{i=1,\dots,k} \left| \frac{f(\lambda_{f,i})}{g(\lambda_{f,i})} \right|^2 \right], \tag{5.12}$$

which also relates to how well a polynomial can be large (relative to f) on the eigenvalues $\lambda_{f,1}, \dots, \lambda_{f,k}$ where f has the largest magnitude and small on the remaining eigenvalues, but does not depend on the randomness used by the algorithm.

Lemma 5.4. *If $\mathbf{\Omega}$ is a random $n \times (k+p)$ matrix whose entries are i.i.d. $\mathcal{N}(0,1)$ random variables, and $\mathbf{\Omega}_{f,1}$ and $\mathbf{\Omega}_{f,2}$ are defined as in* (5.2)*, then with $\mathcal{E}_{\mathbf{\Omega}}(d; f)$ and $\mathcal{E}(d; f)$ as defined in* (5.5) *and* (5.12)*,*

(i) for any $u, t \geq 0$, with probability at least $1 - e^{-(u-2)/4} - \sqrt{\pi k} \left(\frac{t}{e}\right)^{-(p+1)/2}$ we have

$$\mathcal{E}_{\mathbf{\Omega}}(d; f) \leq \frac{utk}{p+1} \mathcal{E}(d; f);$$

(ii) if $p \geq 2$ we have

$$\mathbb{E}[\mathcal{E}_{\mathbf{\Omega}}(d; f)] \leq \frac{k}{p-1} \mathcal{E}(d; f).$$

Proof. (i): For any polynomial $g \in \mathbb{P}_{d-1}$ by [150, Proposition 8.6] we have with probability at least $1 - e^{-(u-2)/4} - \sqrt{\pi k} \left(\frac{t}{e}\right)^{-(p+1)/2}$

$$\left[\|g(\mathbf{\Lambda}_{f,2}) \mathbf{\Omega}_{f,2} \mathbf{\Omega}_{f,1}^\dagger\|_{\mathrm{F}}^2 \max_{i=1,\dots,k} \left| \frac{f(\lambda_{f,i})}{g(\lambda_{f,i})} \right|^2 \right] \leq \frac{utk}{p+1} \left[\|g(\mathbf{\Lambda}_{f,2})\|_{\mathrm{F}}^2 \max_{i=1,\dots,k} \left| \frac{f(\lambda_{f,i})}{g(\lambda_{f,i})} \right|^2 \right].$$

The inequality is respected if we minimize both sides over all polynomials.

(ii): This is proven in an identical fashion utilizing the expectation bound in [150, Proposition 8.6]. □

Before proceeding, we note that if E and F are disjoint sets containing $\lambda_{f,k+1},\ldots,\lambda_{f,n}$ and $\lambda_{f,1},\ldots,\lambda_{f,k}$ respectively, we can upper bound (5.12) by

$$\mathcal{E}(d;f) \leq n\|f(\boldsymbol{A})\|_2^2 \min_{g\in\mathbb{P}_{d-1}} \left[\frac{\sup\limits_{x\in E}|g(x)|^2}{\inf\limits_{x\in F}|g(x)|^2}\right] := n\|f(\boldsymbol{A})\|_2^2 \widetilde{Z}_{d-1}^2(E,F), \tag{5.13}$$

where $\widetilde{Z}_{d-1}(E,F)$ is similar to the Zolotarev number of E and F [19, 77], but the ratio is minimized over polynomials instead of rational functions. We could proceed with bounding $\widetilde{Z}(E,F)$. However, due to the appearance of n in (5.13) we expect that such a bound would be loose and we therefore omit a more detailed discussion.

5.2.4 Error bounds for Krylov aware low-rank approximation

With Theorem 5.2, Lemma 5.3, and Lemma 5.4 we can now derive a probabilistic error bound for the approximation returned by Algorithm 8. We focus on the truncated approximation, as deriving a bound for the untruncated approximation can be done in an entirely identical fashion.

We begin with a result that is an immediate consequence of Lemma 5.1. The proof is very similar to the proof of [134, Lemma 4.1].

Lemma 5.5. *Let $\lambda_{\max}$ and $\lambda_{\min}$ denote the largest and smallest eigenvalue of $\boldsymbol{A}$. Let $q=d+r$ and let $\boldsymbol{T}_q$ and $\boldsymbol{Q}_d$ be computed using Algorithm 4. Then,*

$$\|\boldsymbol{Q}_d^T f(\boldsymbol{A})\boldsymbol{Q}_d - f(\boldsymbol{T}_q)_{1:n_d,1:n_d}\|_{\mathrm{F}} \leq 2\sqrt{(k+p)d} \inf_{g\in\mathbb{P}_{2r+1}} \|f(x)-g(x)\|_{L^\infty([\lambda_{\min},\lambda_{\max}])}.$$

Proof. By Lemmas 3.1 and 5.1 we know that for any polynomial $g\in\mathbb{P}_{2r+1}$ we have $\boldsymbol{Q}_d^T g(\boldsymbol{A})\boldsymbol{Q}_d = g(\boldsymbol{T}_q)_{1:n_d,1:n_d}$, where n_d is the number of columns in $\boldsymbol{Q}_d$. Therefore, since $\|\boldsymbol{Q}_d\|_{\mathrm{F}} \leq \sqrt{n_d} \leq \sqrt{(k+p)d}$ and $\|\boldsymbol{Q}_d\|_2 \leq 1$ we have

$$\begin{aligned}
&\|\boldsymbol{Q}_d^T f(\boldsymbol{A})\boldsymbol{Q}_d - f(\boldsymbol{T}_q)_{1:n_d,1:n_d}\|_{\mathrm{F}} \\
&\qquad = \|\boldsymbol{Q}_d^T f(\boldsymbol{A})\boldsymbol{Q}_d - g(\boldsymbol{T}_q)_{1:n_d,1:n_d} + g(\boldsymbol{T}_q)_{1:n_d,1:n_d} - f(\boldsymbol{T}_q)_{1:n_d,1:n_d}\|_{\mathrm{F}} \\
&\qquad \leq \|\boldsymbol{Q}_d^T f(\boldsymbol{A})\boldsymbol{Q}_d - \boldsymbol{Q}_d^T g(\boldsymbol{A})\boldsymbol{Q}_d\|_{\mathrm{F}} + \|(g(\boldsymbol{T}_q) - f(\boldsymbol{T}_q))_{1:n_d,1:n_d}\|_{\mathrm{F}} \\
&\qquad \leq \sqrt{(k+p)d}\,(\|f(\boldsymbol{A}) - g(\boldsymbol{A})\|_2 + \|g(\boldsymbol{T}_q) - f(\boldsymbol{T}_q)\|_2) \\
&\qquad \leq 2\sqrt{(k+p)d}\|f(x)-g(x)\|_{L^\infty([\lambda_{\min},\lambda_{\max}])},
\end{aligned}$$

where the last inequality is due to the fact that the spectrum of $\boldsymbol{T}_q$ is contained in $[\lambda_{\min},\lambda_{\max}]$. Optimizing over $g\in\mathbb{P}_{2r+1}$ gives the result. □

We note that for block-size $k+p>1$, the Krylov subspace is not equivalent to $\cup\{\mathrm{range}(g(\boldsymbol{A})\boldsymbol{\Omega}) : g \in \mathbb{P}_{d-1}\}$, and bounds based on best approximation may be pessimistic due to this fact. In fact, deriving stronger bounds is an active area of research; see e.g. [35, 57, 58, 60, 61, 86]. However, in this thesis we will stick with this simple and well known bound.

We proceed with proving the following error bound for Algorithm 8.

Theorem 5.6. *Consider $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ as defined in* (5.1) *and let $\lambda_{\max}$ and $\lambda_{\min}$ be the largest and smallest eigenvalue of $\boldsymbol{A}$ respectively. Let $\boldsymbol{Q}_d\boldsymbol{X}_{(k)}\boldsymbol{Q}_d^T$ be the rank k approximation returned by Algorithm 8 where $\boldsymbol{\Omega}$ is a random matrix with i.i.d. $\mathcal{N}(0,1)$ entries. Then, with $\mathcal{E}(d;f)$ as defined in* (5.12)

(i) with probability at least $1-e^{-(u-2)/4}-\sqrt{\pi k}\left(\frac{t}{e}\right)^{-(p+1)/2}$,

$$\begin{aligned}\|f(\boldsymbol{A})-\boldsymbol{Q}_d\boldsymbol{X}_{(k)}\boldsymbol{Q}_d^T\|_{\mathrm{F}} \leq & 4\sqrt{(k+p)d}\inf_{g\in\mathbb{P}_{2r+1}}\|f(x)-g(x)\|_{L^\infty([\lambda_{\min},\lambda_{\max}])} \\ & +\sqrt{\|f(\boldsymbol{\Lambda}_{f,2})\|_{\mathrm{F}}^2+\frac{5utk}{p+1}\mathcal{E}(d;f)};\end{aligned}$$

(ii) if $p\geq 2$ that

$$\begin{aligned}\mathbb{E}\|f(\boldsymbol{A})-\boldsymbol{Q}_d\boldsymbol{X}_{(k)}\boldsymbol{Q}_d^T\|_{\mathrm{F}} \leq & 4\sqrt{(k+p)d}\inf_{g\in\mathbb{P}_{2r+1}}\|f(x)-g(x)\|_{L^\infty([\lambda_{\min},\lambda_{\max}])}+ \\ & \sqrt{\|f(\boldsymbol{\Lambda}_{f,2})\|_{\mathrm{F}}^2+\frac{5k}{p-1}\mathcal{E}(d;f)}.\end{aligned}$$

Proof. (i): By applying Theorem 5.2, Lemma 5.3, and Lemma 5.5 we obtain the following structural bound

$$\begin{aligned}\|f(\boldsymbol{A})-\boldsymbol{Q}_d\boldsymbol{X}_{(k)}\boldsymbol{Q}_d^T\|_{\mathrm{F}} \leq & 4\sqrt{(k+p)d}\inf_{g\in\mathbb{P}_{2r+1}}\|f(x)-g(x)\|_{L^\infty([\lambda_{\min},\lambda_{\max}])}+ \\ & \sqrt{\|f(\boldsymbol{\Lambda}_{f,2})\|_{\mathrm{F}}^2+5\mathcal{E}_{\boldsymbol{\Omega}}(d;f)}.\end{aligned} \tag{5.14}$$

Applying Lemma 5.4 *(i)* yields the desired bound.

(ii): By applying the Cauchy-Schwarz inequality to (5.14) and Lemma 5.4 *(ii)* yields the desired inequality. □

We proceed with commenting on the three terms appearing in the bounds in Theorem 5.6. As noted above, $4\sqrt{(k+p)d}\inf\limits_{g\in\mathbb{P}_{2r+1}}\|f(x)-g(x)\|_{L^\infty([\lambda_{\min},\lambda_{\max}])}$ has to do with how well quadratic forms involving matrix functions are approximated by the Lanczos method.

Note that if we know that $\|f(\boldsymbol{T}_q)_{1:n_d,1:n_d} - \boldsymbol{Q}_d^T f(\boldsymbol{A})\boldsymbol{Q}_d\|_{\mathrm{F}} \leq \epsilon$ almost surely then this term can be replaced with 2ϵ. The $\|f(\boldsymbol{\Lambda}_{f,2})\|_{\mathrm{F}}$ term tells us that the error can never be below the optimal rank k approximation error. Finally, $\mathcal{E}(d; f)$ tells us that $\boldsymbol{Q}_d$ is a good orthonormal basis for low-rank approximation if there is a polynomial of degree at most $d-1$ that is very large on the eigenvalues $\lambda_{f,1}, \ldots, \lambda_{f,k}$ and is very small on the eigenvalues $\lambda_{f,k+1}, \ldots, \lambda_{f,n}$, which effectively denoises the contribution from the small eigenvalues of $f(\boldsymbol{A})$. A similar intuition was used in [115, 150] when $f(x) = x$.

5.2.5 Simplified bounds for the matrix exponential

By constructing particular polynomials of degree $< d$, we can obtain more explicit bounds that depend only on how accurately $f(x)$ can be approximated by polynomials. These bounds are reminiscent of standard bounds that might be obtained if we could do exact products with $f(\boldsymbol{A})$, except that they have small error terms accounting for the fact that $f(x)$ might not be a polynomial. We will provide such bounds for $\exp(\boldsymbol{A})$. For simplicity, we focus on expectation bounds. However, using an almost identical argument, one can obtain the corresponding tailbounds.

We begin with the following result, which shows that by Theorem 5.6 can recover the bounds of the randomized SVD [79, Theorem 10.5].

Corollary 5.7. *Consider the setting of Theorem 5.6 with $f(x) = \exp(x)$. Let $\boldsymbol{A}$ have eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Then, if $p \geq 2$, $\beta = \lambda_{\max} - \lambda_{\min} = \lambda_1 - \lambda_n$, and $d > e\beta$ we have*

$$\mathbb{E}\|\exp(\boldsymbol{A}) - \boldsymbol{Q}_d \boldsymbol{X}_{(k)} \boldsymbol{Q}_d^T\|_{\mathrm{F}} \leq \\ \frac{\sqrt{(k+p)d}\beta^{2r+2}}{2^{4r+1}(2r+2)!}\|\exp(\boldsymbol{A})\|_2 + \sqrt{1 + \frac{1}{(1-\frac{\beta^d}{d!})^2}\frac{5k}{p-1}}\|\exp(\boldsymbol{\Lambda}_{f,2})\|_{\mathrm{F}}.$$

Proof. First note that by a standard Chebyshev interpolation bound [144, Lecture 20]

$$\inf_{g \in \mathbb{P}_{2r+1}} \|\exp(x) - p(x)\|_{L^\infty([\lambda_{\min},\lambda_{\max}])} \leq \frac{\beta^{2r+r}}{2^{4r+3}(2r+2)!}\|\exp(\boldsymbol{A})\|_2.$$

Hence,

$$4\sqrt{(k+p)d} \inf_{g \in \mathbb{P}_{2r+1}} \|f(x) - p(x)\|_{L^\infty([\lambda_{\min},\lambda_{\max}])} \leq 4\sqrt{(k+p)d}\frac{\beta^{2r+2}}{(2r+2)!}\|\exp(\boldsymbol{A})\|_2.$$

We proceed with bounding $\mathcal{E}(d; \exp(x))$. Let $g(x) = \sum_{i=0}^{d-1} \frac{(x-\lambda_{\min})^i}{i!}$. Then for $x \in [\lambda_{\min}, \lambda_{\max}]$ we have $0 \leq g(x) \leq \exp(x - \lambda_{\min})$. Consequently, $\|g(\boldsymbol{\Lambda}_{f,2})\|_{\mathrm{F}} \leq \|\exp(\boldsymbol{\Lambda}_{f,2})\|_{\mathrm{F}} \exp(-\lambda_{\min})$.

Furthermore, for $x \in [\lambda_{\min}, \lambda_{\max}]$ we have

$$0 \leq \exp(x - \lambda_{\min}) - g(x) = \sum_{i=d}^{\infty} \frac{(x-\lambda_{\min})^i}{i!} \leq \frac{(x-\lambda_{\min})^d}{d!} \exp(x - \lambda_{\min}) \leq \frac{\beta^d}{d!} \exp(x - \lambda_{\min}).$$

Note that by the assumption on d and a Stirling approximation $d! \geq \sqrt{2\pi d} \left(\frac{d}{e}\right)^d$ [131] we have $\frac{\beta^d}{d!} < 1$. Hence,

$$\max_{i=1,\ldots,k} \left| \frac{\exp(\lambda_i)}{g(\lambda_i)} \right| = \exp(\lambda_{\min}) \max_{i=1,\ldots,k} \left| \frac{\exp(\lambda_i - \lambda_{\min})}{g(\lambda_i)} \right| = \frac{\exp(\lambda_{\min})}{1 - \frac{\beta^d}{d!}}. \tag{5.15}$$

Therefore, $\mathcal{E}(d; \exp(x))$ is bounded above by

$$\mathcal{E}(d; \exp(x)) \leq \frac{1}{(1 - \frac{\beta^d}{d!})^2} \| \exp(\mathbf{\Lambda}_{f,2}) \|_{\mathrm{F}}^2. \tag{5.16}$$

Plugging the inequalities (5.15) and (5.16) into Theorem 5.6 yields the desired inequality. □

In the proof of Corollary 5.7 we upper bounded $\mathcal{E}(d; \exp(x))$ by choosing the truncated Taylor series of the exponential. However, there are other polynomials that can achieve a significantly tighter upper bound of $\mathcal{E}(d; \exp(x))$. For example, we can choose a scaled and shifted Chebyshev polynomial to obtain the following bound, which shows that as d and r grow larger the low-rank approximation returned by Algorithm 8 converges to an optimal low-rank approximation.

Corollary 5.8. *Consider the setting of Theorem 5.6 with $f(x) = \exp(x)$. Let $\boldsymbol{A}$ have eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Then, if $p \geq 2$, $\beta = \lambda_{\max} - \lambda_{\min} = \lambda_1 - \lambda_n$, and $\gamma = \frac{\lambda_k - \lambda_{k+1}}{\lambda_k + \lambda_{k+1} - 2\lambda_{\min}}$ we have*

$$\mathbb{E}\| \exp(\boldsymbol{A}) - \boldsymbol{Q}_d \boldsymbol{X}_{(k)} \boldsymbol{Q}_d^T \|_{\mathrm{F}} \leq$$
$$\frac{\sqrt{(k+p)d}\,\beta^{2r+2}}{2^{4r+1}(2r+2)!} \| \exp(\boldsymbol{A}) \|_2 + \sqrt{1 + e^{2\beta - 4(d-2)\sqrt{\gamma}} \frac{20k}{p-1}} \| \exp(\mathbf{\Lambda}_{f,2}) \|_{\mathrm{F}}.$$

Proof. Bounding $4\sqrt{(k+p)d} \inf_{g \in \mathbb{P}_{2r+1}} \| \exp(x) - g(x) \|_{L^\infty([\lambda_{\min}, \lambda_{\max}])}$ is done identical to as done in Corollary 5.7. We proceed with bounding $\mathcal{E}(d; \exp(x))$ by choosing the polynomial $g(x) = (1 + x - \lambda_{\min}) T_{d-2}\left(\frac{x - \lambda_{\min}}{\lambda_{k+1} - \lambda_{\min}}\right)$, where T_{d-2} is the Chebyshev polynomial of degree $d-2$. Hence, recalling the definition (5.12) of $\mathcal{E}(d; \exp(x))$, since $0 \leq 1 + x \leq e^x$ for $x \geq 0$,

$$\|g(\mathbf{\Lambda}_{f,2})\|_{\mathrm{F}} \leq \| \exp(\mathbf{\Lambda}_{f,2} - \lambda_{\min} \boldsymbol{I}) \|_{\mathrm{F}} \left\| T_{d-2}\left(\frac{\mathbf{\Lambda}_{f,2} - \lambda_{\min}\boldsymbol{I}}{\lambda_{k+1} - \lambda_{\min}} \right) \right\|_2.$$

Hence, using that $|T_{d-2}(x)| \leq 1$ for $x \in [-1, 1]$,

$$\begin{aligned}
\mathcal{E}(d; \exp(x)) &\leq \|g(\mathbf{\Lambda}_{f,2})\|_{\mathrm{F}}^2 \max_{i=1,\ldots,k} \left| \frac{e^{\lambda_i}}{g(\lambda_i)} \right|^2 \\
&\leq \| \exp(\mathbf{\Lambda}_{f,2} - \lambda_{\min} \boldsymbol{I})\|_{\mathrm{F}}^2 \max_{i=1,\ldots,k} e^{2\lambda_{\min}} \left| \frac{e^{\lambda_i - \lambda_{\min}}}{g(\lambda_i)} \right|^2 \\
&= \| \exp(\mathbf{\Lambda}_{f,2})\|_{\mathrm{F}}^2 \max_{i=1,\ldots,k} \left| \frac{e^{\lambda_i - \lambda_{\min}}}{g(\lambda_i)} \right|^2. \qquad (5.17)
\end{aligned}$$

Finally, using that $\frac{e^x}{1+x} \leq e^x$ for $x \geq 0$,that $T_{d-2}(x)$ is increasing for $x \geq 1$, and [150, Lemma 9.3] we have

$$\begin{aligned}
\max_{i=1,\ldots,k} \left| \frac{e^{\lambda_i - \lambda_{\min}}}{g(\lambda_i)} \right|^2 &\leq \max_{i=1,\ldots,k} \left| \frac{e^{\lambda_i - \lambda_{\min}}}{T_{d-2}\left(\frac{x - \lambda_{\min}}{\lambda_{k+1} - \lambda_{\min}} \right)} \right|^2 \\
&\leq \frac{e^{2\beta}}{T_{d-2}\left(\frac{\lambda_k - \lambda_{\min}}{\lambda_{k+1} - \lambda_{\min}} \right)^2} \leq 4e^{2\beta - 4(d-2)\sqrt{\gamma}},
\end{aligned}$$

Plugging this inequality into (5.17) and then (5.17) into Theorem 5.6 yields the desired inequality. □

5.2.6 Simplified bounds for the identity function

By applying our results to the function $f(x) = x$, one can derive bounds for low-rank approximation of a symmetric matrix $\boldsymbol{A}$. These bounds are reminiscent of the bounds in [150, Theorem 9.2], but they allow for a symmetric and truncated low-rank approximation, and the bounds can therefore be of independent interest. In particular, we have the following result when $f(x) = x$.

Theorem 5.9 (Theorem 2.9 restated)**.** *Consider the setting of Theorem 5.6 with $f(x) = x$ and $r = 0$. Let the eigenvalues of $\boldsymbol{A}$ be ordered so that $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$. Let $\gamma = \frac{|\lambda_k| - |\lambda_{k+1}|}{|\lambda_k| + |\lambda_{k+1}|}$. Then,*

(i) with probability at least $1 - e^{-(u-2)/4} - \sqrt{\pi k} \left(\frac{t}{e}\right)^{-(p+1)/2}$,

$$\|\boldsymbol{A} - \boldsymbol{Q}_d(\boldsymbol{Q}_d^T \boldsymbol{A} \boldsymbol{Q}_d)_{(k)} \boldsymbol{Q}_d^T\|_{\mathrm{F}} \leq \sqrt{1 + e^{-4(d-2)\sqrt{\gamma}} \frac{20utk}{p+1}} \|\mathbf{\Lambda}_2\|_{\mathrm{F}};$$

(ii) if $p \geq 2$ that

$$\mathbb{E}\|\boldsymbol{A} - \boldsymbol{Q}_d(\boldsymbol{Q}_d^T\boldsymbol{A}\boldsymbol{Q}_d)_{(k)}\boldsymbol{Q}_d^T\|_{\mathrm{F}} \leq \sqrt{1 + e^{-4(d-2)\sqrt{\gamma}}\frac{20k}{p-1}}\|\boldsymbol{\Lambda}_2\|_{\mathrm{F}}.$$

Proof. First note that the polynomial approximation term in Theorem 5.6 is 0. Moreover, in this case we have $\boldsymbol{\Lambda}_{f,2} = \boldsymbol{\Lambda}_2$, where $\boldsymbol{\Lambda}_2$ is as in (2.6). Hence, by Theorem 5.6 it is sufficient to show that

$$\min_{g\in\mathbb{P}_{d-1}}\left[\|g(\boldsymbol{\Lambda}_2)\|_{\mathrm{F}}^2 \max_{i=1,\dots,k}\left|\frac{\lambda_i}{g(\lambda_i)}\right|^2\right] \leq 4e^{-4(d-2)\sqrt{\gamma}}\|\boldsymbol{\Lambda}_2\|_{\mathrm{F}}^2. \tag{5.18}$$

Let $g(x) = xT_{d-2}(x/|\lambda_{k+1}|)$, where T_{d-2} is the Chebyshev polynomial of degree $d-2$. We have

$$\min_{g\in\mathbb{P}_{d-1}}\left[\|g(\boldsymbol{\Lambda}_2)\|_{\mathrm{F}}^2 \max_{i=1,\dots,k}\left|\frac{\lambda_i}{g(\lambda_i)}\right|^2\right] \leq \|g(\boldsymbol{\Lambda}_2)\|_{\mathrm{F}}^2 \max_{i=1,\dots,k}\left|\frac{\lambda_i}{g(\lambda_i)}\right|^2. \tag{5.19}$$

Note that since $g(x)$ is even or odd we have $|g(x)| = |g(|x|)|$. Hence,

$$\|g(\boldsymbol{\Lambda}_2)\|_{\mathrm{F}}^2 = \|g(|\boldsymbol{\Lambda}_2|)\|_{\mathrm{F}}^2 \leq \|\boldsymbol{\Lambda}_2\|_{\mathrm{F}}^2\|T_{d-2}(|\boldsymbol{\Lambda}_2|/|\lambda_{k+1}|)\|_2^2 \leq \|\boldsymbol{\Lambda}_2\|_{\mathrm{F}}^2, \tag{5.20}$$

where $|\boldsymbol{\Lambda}_2|$ denotes the matrix where we take the absolute values of the diagonal of $\boldsymbol{\Lambda}_2$ and where we used that $|T_{d-2}(x)| \leq 1$ for $x \in [-1,1]$ [150, Lemma 9.3]. Furthermore,

$$\max_{i=1,\dots,k}\left|\frac{\lambda_i}{g(\lambda_i)}\right|^2 = \left|\frac{1}{T_{d-2}(|\lambda_k|/|\lambda_{k+1}|)}\right|^2 \leq 4e^{-4(d-2)\sqrt{\gamma}}, \tag{5.21}$$

where we used that $T_{d-2}(x)$ is increasing for $x \geq 1$ and [150, p.46]. Combining (5.19), (5.20), and (5.21) yields (5.18), as required. □

5.3 Numerical experiments

In this section we compare the Krylov aware low-rank approximation (Algorithm 8) and Algorithm 1 (assuming exact matrix-vector products with $f(\boldsymbol{A})$) and Algorithm 7 (inexact matrix-vector products with $f(\boldsymbol{A})$). All experiments have been performed in MATLAB (version 2020a) and scripts to reproduce the figures are available at https://github.com/davpersson/Krylov_aware_LRA.git.

5.3.1 Test matrices

We begin with outlining the test matrices and matrix functions used in our examples.

Exponential integrator

Consider the parabolic differential equation outlined in (3.3)

$$
\begin{aligned}
&u_t = \kappa \Delta u + \lambda u \text{ in } [0,1]^2 \times [0,2] \\
&u(\cdot, 0) = \theta \text{ in } [0,1]^2 \\
&u = 0 \text{ on } \Gamma_1 \\
&\frac{\partial u}{\partial \boldsymbol{n}} = 0 \text{ on } \Gamma_2
\end{aligned}
$$

for $\kappa, \lambda > 0$ and $\Gamma_2 = \{(x,1) \in \mathbb{R}^2 : x \in [0,1]\}$ and $\Gamma_1 = \partial \mathcal{D} \setminus \Gamma_2$. As mentioned in Section 5.3.1, discretizing in space using finite differences on a 100 × 100 grid we obtain an ordinary differential equation of the form (3.1)

$$
\begin{aligned}
\dot{\boldsymbol{u}}(t) &= \boldsymbol{A}\boldsymbol{u}(t) \text{ for } t \geq 0, \\
\boldsymbol{u}(0) &= \boldsymbol{\theta},
\end{aligned}
$$

for symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{9900 \times 9900}$. It is well known that the solution to (4.29) is given by $\boldsymbol{u}(t) = \exp(t\boldsymbol{A})\boldsymbol{\theta}$. Suppose that we want to compute the solution for $t \geq 1$. One can verify that $\exp(\boldsymbol{A})$ admits a good rank 60 approximation

$$
\frac{\|\exp(\boldsymbol{A}) - (\exp(\boldsymbol{A}))_{(60)}\|_{\mathrm{F}}}{\|\exp(\boldsymbol{A})\|_{\mathrm{F}}} \approx 4 \times 10^{-4}.
$$

Hence, we can use Algorithm 8 to compute $\boldsymbol{Q}_s$ and $\boldsymbol{T}_q$ and use them to efficiently construct a rank 60 approximation to $\exp(t\boldsymbol{L})$ for any t *with almost no additional cost.*

In the experiments we set $\kappa = 0.01$ and $\lambda = 1$.

Estrada index

Recall as outlined in Section 3.2.5 or an (undirected) graph with adjacency matrix $\boldsymbol{A}$ the Estrada index is defined as $\mathrm{tr}(\exp(\boldsymbol{A}))$. It is used to measure the degree of protein folding [54]. One can estimate the Estrada index of a network by the Hutch++ algorithm or its variations [37, 52, 111, 123], which requires computing a low-rank approximation of $\exp(\boldsymbol{A})$. Motivated by the numerical experiments in [111] we let $\boldsymbol{A}$ be the adjacency matrix of Roget's Thesaurus semantic graph [118].

Quantum spin system

We use an example from [52, Section 4.3]; a similar example is found in [37]. We want to approximate $\exp(-\beta\boldsymbol{A})$ where

$$
\boldsymbol{A} = -\sum_{i=1}^{N-1} \boldsymbol{Z}_i \boldsymbol{Z}_{i+1} - h \sum_{i=1}^{N} \boldsymbol{X}_i \in \mathbb{R}^{n \times n}, \tag{5.22}
$$

where

$$\boldsymbol{X}_i = \boldsymbol{I}_{2^{i-1}} \otimes \boldsymbol{X} \otimes \boldsymbol{I}_{2^{N-i}}, \quad \boldsymbol{Z}_i = \boldsymbol{I}_{2^{i-1}} \otimes \boldsymbol{Z} \otimes \boldsymbol{I}_{2^{N-i}}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the Pauli operators

$$\boldsymbol{X} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{Z} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Estimating the partition function $Z(\beta) = \mathrm{tr}(\exp(-\beta \boldsymbol{A}))$ is an important task in quantum mechanics [126], which once again can benefit from computing a low-rank approximation of $\exp(-\beta \boldsymbol{A})$.

In the experiments we set $N = 14$ so that $n = 2^{14}$, $\beta = 0.3$, and $h = 10$.

Synthetic example for the matrix logarithm

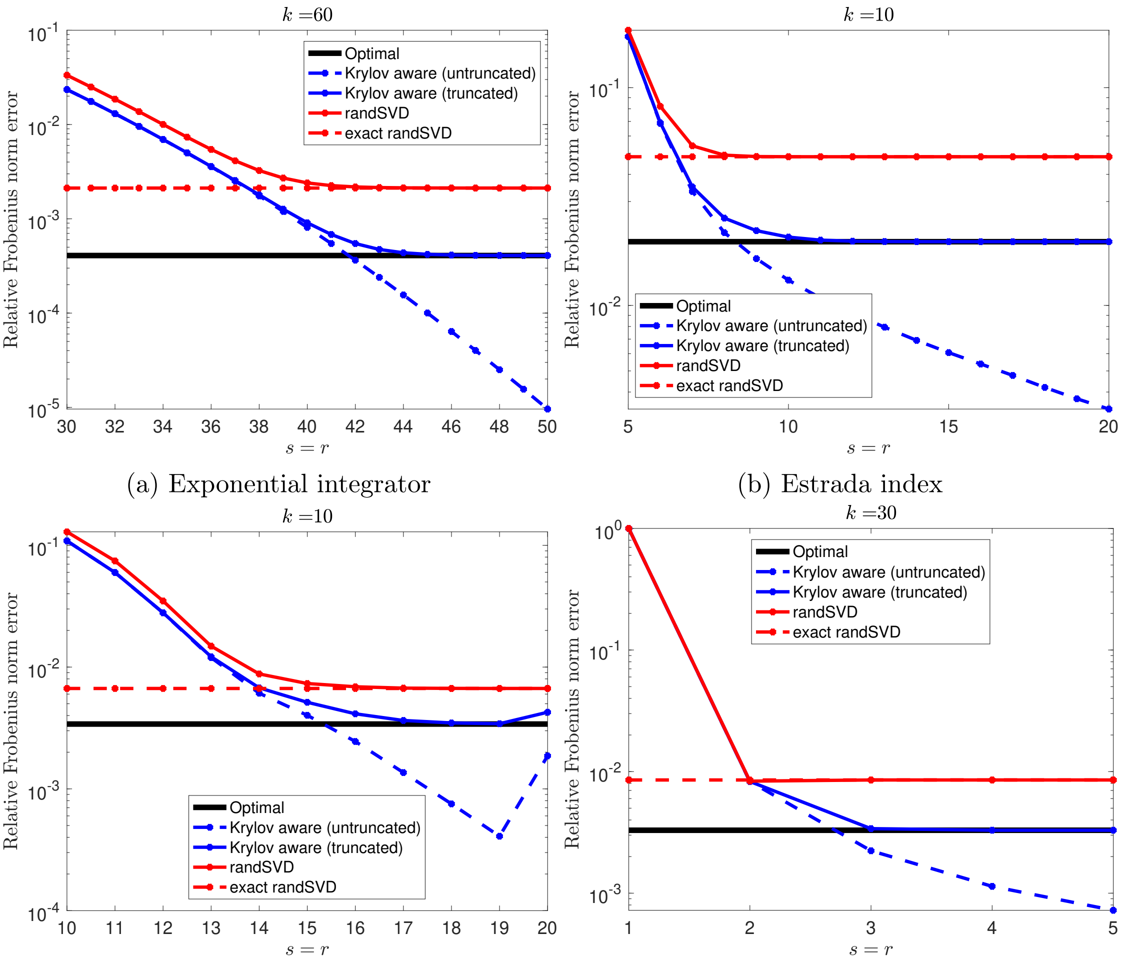
We generate a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{5000 \times 5000}$ with eigenvalues $\lambda_{f,i} = \exp(\frac{1}{i^2})$ for $i = 1, \ldots, n$. We let $f(x) = \log(x)$ so that the eigenvalues of $f(\boldsymbol{A})$ are $f(\lambda_{f,i}) = \frac{1}{i^2}$ for $i = 1, \ldots, n$.

5.3.2 Comparing relative errors

In this section we compare the error returned by Algorithm 6 (with truncation), Algorithm 7, and Algorithm 8. If $\boldsymbol{C}$ is a low-rank approximation returned by one of the algorithms then we compare the relative error

$$\frac{\|f(\boldsymbol{A}) - \boldsymbol{C}\|_{\mathrm{F}}}{\|f(\boldsymbol{A})\|_{\mathrm{F}}}. \tag{5.23}$$

In all experiments we set the parameters in Algorithm 8 and Algorithm 7 to be $p = 0$ or $p = 5$ and $d = r$ so that the total number of matrix vector products with $\boldsymbol{A}$ is $2(k+p)d$. We run *Algorithm* 6 on $f(\boldsymbol{A})$ with exact matrix-vector products, which cannot be done in practice. Hence, the results from this algorithm are only used as a reference for Algorithm 8 and Algorithm 7. The results are presented in Figure 5.1 for $p = 0$ and Figure 5.2 for $p = 5$. All results confirm that Algorithm 8 returns a more accurate approximation than Algorithm 7, and can even be more accurate than Algorithm 1. Note that for (5.22) the error for the untruncated version of the approximation returned by Algorithm 8 stagnates. This is because the error from the approximation of the quadratic form dominates the error. In this case, r should be chosen larger than d.



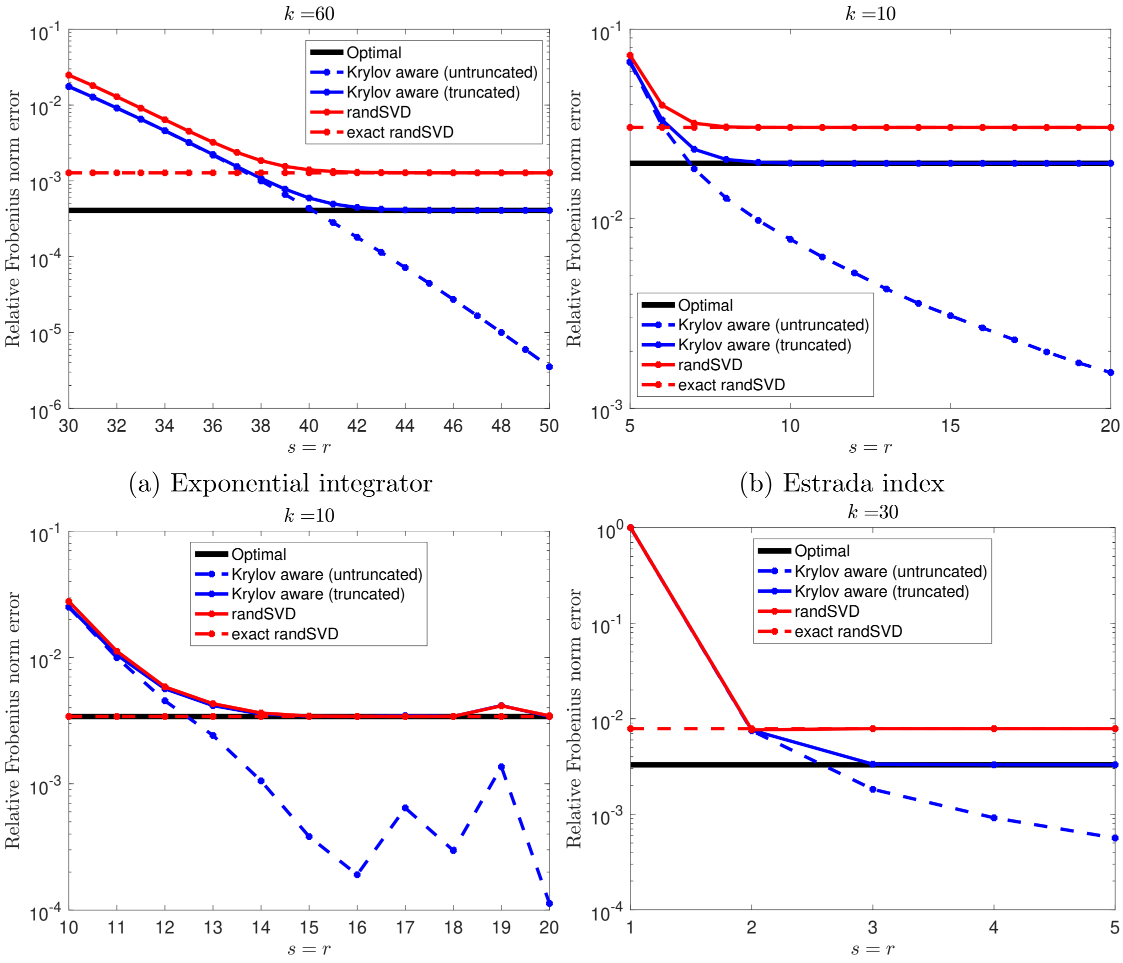
(a) Exponential integrator

(b) Estrada index

(c) Quantum spin system

(d) Synthetic example for the matrix logarithm

Figure 5.1: Comparing (5.23) for the the approximations returned by Algorithm 8 without truncation (untruncated), Algorithm 8 with truncation back to rank k (truncated), Algorithm 7, and Algorithm 6. The black line shows the optimal rank k approximation relative Frobenius norm error. The rank parameter k is visible as titles in the figures. In all experiments we set the oversampling parameter $p = 0$. All examples are outlined in Section 5.3.1.



(a) Exponential integrator

(b) Estrada index

(c) Quantum spin system

(d) Synthetic example for the matrix logarithm

Figure 5.2: Comparing (5.23) for the the approximations returned by Algorithm 8 without truncation (untruncated), Algorithm 8 with truncation back to rank k (truncated), Algorithm 7, and Algorithm 6. The black line shows the optimal rank k approximation relative Frobenius norm error. The rank parameter k is visible as titles in the figures. In all experiments we set the oversampling parameter $p = 5$. All examples are outlined in Section 5.3.1.

6 An introduction to trace estimation

In this chapter we will give an introduction to trace estimation. We will outline its applications, present the stochastic trace estimator, and present the Hutch++ algorithm.

Computing or estimating the trace of a large symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$,

$$\operatorname{tr}(\boldsymbol{A}) := \sum_{i=1}^{n} \boldsymbol{A}_{ii},$$

is an important problem that arises in a wide variety of applications, such as triangle counting in graphs [8], Frobenius norm estimation [31, 74], quantum chromodynamics [146], computing the Estrada index of a graph [120, 54], computing the log-determinant [2, 41, 138, 159, 168] and many more [155].

Computing $\operatorname{tr}(\boldsymbol{A})$ is of course a trivial task if the matrix $\boldsymbol{A}$ is explicitly available to us. However, the difficulty arises when we do not have explicit access to the entries of $\boldsymbol{A}$, but can only access $\boldsymbol{A}$ through matrix-vector products $\boldsymbol{x} \mapsto \boldsymbol{A}\boldsymbol{x}$. This appears when, for example, $\boldsymbol{A}$ is a function of another matrix $\boldsymbol{C}$, such as $\boldsymbol{A} = \exp(\boldsymbol{C})$, $\boldsymbol{A} = \log(\lambda \boldsymbol{I} + \boldsymbol{C})$, $\boldsymbol{A} = \boldsymbol{C}^{-1}$ or $\boldsymbol{A} = \boldsymbol{C}^3$. Computing $\boldsymbol{A}$ (or even only its diagonal entries) explicitly in these situations is typically too expensive and may require up to $O(n^3)$ operations, as discussed in Chapter 3. On the other hand, computing (approximate) matrix-vector products $\boldsymbol{A}\boldsymbol{x}$ is tractable using, for example, Lanczos methods [84, 86], as discussed in Section 3.3. One can of course exactly recover $\operatorname{tr}(\boldsymbol{A})$ by computing n matrix-vector products with $\boldsymbol{A}$. However, computing $O(n)$ matrix-vector products is usually too expensive, and we want to obtain an estimate of the trace using significantly fewer matrix-vector products.

In Section 6.1 we will introduce the stochastic trace estimator, which is an unbiased Monte-Carlo estimator of $\operatorname{tr}(\boldsymbol{A})$. Then, in Section 6.2 we will present the Hutch++ algorithm, which combines the stochastic trace estimator with randomized low-rank approximation to achieve a faster convergence of the estimator.

6.1 The stochastic trace estimator

The stochastic trace estimator, sometimes called the Girard-Hutchinson estimator [67, 90], builds on the following observation: if $\boldsymbol{\omega}$ is a random vector of length n satisfying $\mathbb{E}\boldsymbol{\omega}\boldsymbol{\omega}^T = \boldsymbol{I}$ then

$$\mathbb{E}\boldsymbol{\omega}^T\boldsymbol{A}\boldsymbol{\omega} = \operatorname{tr}(\boldsymbol{A}).$$

Therefore, sampling m such quadratic forms and computing the sample mean yields the following unbiased estimator of the trace:

$$\operatorname{tr}_m(\boldsymbol{A}) := \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\omega}_i^T\boldsymbol{A}\boldsymbol{\omega}_i = \frac{1}{m}\operatorname{tr}\left(\boldsymbol{\Omega}^T\boldsymbol{A}\boldsymbol{\Omega}\right) \approx \operatorname{tr}(\boldsymbol{A}), \tag{6.1}$$

where $\boldsymbol{\Omega} = \begin{bmatrix}\boldsymbol{\omega}_1 & \cdots & \boldsymbol{\omega}_m\end{bmatrix}$ contains m independent copies of $\boldsymbol{\omega}$. Common choices for the random vector $\boldsymbol{\omega}$ are standard Gaussians; the entries in $\boldsymbol{\omega}$ are independent identically distributed (i.i.d.) samples from $\mathcal{N}(0,1)$, Rademacher vectors; the entries in $\boldsymbol{\omega}$ are independently chosen to be -1 or $+1$ with equal probability, and spherical random vectors: the vector $\boldsymbol{\omega}$ is drawn uniformly from a sphere of radius $\sqrt{n}$. Among random vectors whose entries are i.i.d., Rademacher vectors achieves the smallest variance. Furthermore, uniform random vectors achieve the smallest variance among spherical distributions. See [51] for a more detailed discussion. However, for simplicity, in this thesis, we choose $\boldsymbol{\omega}$ to be standard Gaussian. In this case, the variance of $\operatorname{tr}_m(\boldsymbol{A})$ is given by

$$\operatorname{Var}(\operatorname{tr}_m(\boldsymbol{A})) = \frac{2}{m}\|\boldsymbol{A}\|_{\mathrm{F}}^2. \tag{6.2}$$

Under the assumption that $\boldsymbol{A}$ is symmetric positive semi-definite, one can derive bounds on m that guarantee a small relative error with high probability:

$$\mathbb{P}\left(|\operatorname{tr}_m(\boldsymbol{A}) - \operatorname{tr}(\boldsymbol{A})| \leq \varepsilon\operatorname{tr}(\boldsymbol{A})\right) \geq 1-\delta; \tag{6.3}$$

see, e.g., [11, 74, 132, 133]. When $\boldsymbol{A}$ is indefinite, aiming for such a relative bound is unrealistic, as can be easily seen for a non-zero matrix $\boldsymbol{A}$ with $\operatorname{tr}(\boldsymbol{A}) = 0$. Instead, one aims at deriving bounds on m that guarantee a small *absolute* error:

$$\mathbb{P}\left(|\operatorname{tr}_m(\boldsymbol{A}) - \operatorname{tr}(\boldsymbol{A})| \leq \varepsilon\right) \geq 1-\delta. \tag{6.4}$$

It is well known that the number of samples needed to attain (6.3) or (6.4) grows at a rate proportional to ε^{-2} as $\varepsilon \to 0$, which is confirmed by the following tailbound.

Theorem 6.1 ([41, Theorem 1]). *Let $\boldsymbol{A}$ be a symmetric matrix. Then for any $\varepsilon \geq 0$ we have*

$$\mathbb{P}\left(|\operatorname{tr}_m(\boldsymbol{A}) - \operatorname{tr}(\boldsymbol{A})| \geq \varepsilon\right) \leq 2\exp\left(-\frac{m\varepsilon^2}{4\|\boldsymbol{A}\|_{\mathrm{F}}^2 + 4\varepsilon\|\boldsymbol{A}\|_2}\right).$$

Algorithm 9 Hutch++

```
input: Symmetric $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Number of matrix-vector products $m \in \mathbb{N}$ (multiple of 3).
output: An approximation to $\mathrm{tr}(\boldsymbol{A}) : \mathrm{tr}_m^{\mathsf{h}++}(\boldsymbol{A})$.
1: Sample $\boldsymbol{\Omega} \in \mathbb{R}^{n\times \frac{m}{3}}$ with i.i.d. $\mathcal{N}(0,1)$ or Rademacher entries.
2: Compute $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\Omega}$.
3: Get an orthonormal basis $\boldsymbol{Q} \in \mathbb{R}^{n\times \frac{m}{3}}$ for range($\boldsymbol{Y}$).
4: Sample $\boldsymbol{\Psi} \in \mathbb{R}^{n\times \frac{m}{3}}$ with i.i.d. $\mathcal{N}(0,1)$ or Rademacher entries.
5: return $\mathrm{tr}_m^{\mathsf{h}++}(\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}) + \frac{3}{m}\,\mathrm{tr}(\boldsymbol{\Psi}^T(\boldsymbol{I}-\boldsymbol{Q}\boldsymbol{Q}^T)\boldsymbol{A}(\boldsymbol{I}-\boldsymbol{Q}\boldsymbol{Q}^T)\boldsymbol{\Psi})$
```

Hence, if $\boldsymbol{A}$ is SPSD, Theorem 6.1 implies that if

$$m = 4\varepsilon^{-2}\left(\frac{\|\boldsymbol{A}\|_{\mathrm{F}}^2}{\mathrm{tr}(\boldsymbol{A})^2} + \varepsilon\frac{\|\boldsymbol{A}\|_2}{\mathrm{tr}(\boldsymbol{A})}\right)\log(2\delta^{-1}) \le 4\varepsilon^{-2}(1+\varepsilon)\log(2\delta^{-1}),$$

then (6.3) is satisfied. For small ε, the ε^{-2} becomes very large. Therefore, the number of required samples grows quickly as $\varepsilon \to 0$. To reduce the number of samples (and, in turn, the number of matrix-vector products), different variance reduction techniques were studied [63, 102, 111, 166], which we will outline in the subsequent section.

6.2 The Hutch++ algorithm

Variance reduction techniques for trace estimation usually aim at finding a decomposition

$$\mathrm{tr}(\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{A}_1) + \mathrm{tr}(\boldsymbol{A}_2), \tag{6.5}$$

such that $\mathrm{tr}(\boldsymbol{A}_1)$ can be computed explicitly and the stochastic estimator for $\mathrm{tr}(\boldsymbol{A}_2)$ has reduced variance, which – in view of (6.2) – means that $\boldsymbol{A}_2$ has reduced Frobenius norm. Among these techniques is the Hutch++ algorithm presented in [111]. In Hutch++, the matrix $\boldsymbol{A}_1$ in (6.5) is chosen to be a low-rank approximation of $\boldsymbol{A}$ obtained with the randomized SVD (Algorithm 1), and $\boldsymbol{A}_2 = \boldsymbol{A} - \boldsymbol{A}_1$. The resulting method is presented in Algorithm 9. Hutch++ consists of two phases. The first phase is concerned with obtaining a low-rank approximation $\boldsymbol{A} \approx \boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A}$ and exploits the cyclic property of the trace: $\mathrm{tr}(\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q})$. It uses $\frac{2m}{3}$ matrix-vector products with $\boldsymbol{A}$: $\boldsymbol{A}\boldsymbol{\Omega}$ in line 2 of Algorithm 9 and $\boldsymbol{A}\boldsymbol{Q}$ to compute $\mathrm{tr}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q})$ in line 5. The second phase is concerned with estimating $\mathrm{tr}(\boldsymbol{A} - \boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A}) = \mathrm{tr}((\boldsymbol{I}-\boldsymbol{Q}\boldsymbol{Q}^T)\boldsymbol{A}(\boldsymbol{I}-\boldsymbol{Q}\boldsymbol{Q}^T))$ via the stochastic trace estimator (6.1). It uses the remaining $\frac{m}{3}$ matrix-vector products with $\boldsymbol{A}$ to compute $\boldsymbol{A}((\boldsymbol{I}-\boldsymbol{Q}\boldsymbol{Q}^T)\boldsymbol{\Psi})$ in line 5 of Algorithm 9. A major breakthrough in [111] was that Hutch++ guarantees an ε-relative error, as in (6.3), with only $O(\varepsilon^{-1})$ matrix-vector products, provided that $\boldsymbol{A}$ is SPSD. This compares favorably with the $O\left(\varepsilon^{-2}\right)$ matrix-vector products that are required when using stochastic trace estimation alone.

Theorem 6.2 ([111, Theorem 1.1])**.** *Suppose that $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is SPSD. If Algorithm 9 is* *implemented with* $m = O\left(\varepsilon^{-1}\sqrt{\log(\delta^{-1})} + \log(\delta^{-1})\right)$ *matrix-vector products then*

$$\left|\mathrm{tr}_m^{\mathsf{h}++}(\boldsymbol{A}) - \mathrm{tr}(\boldsymbol{A})\right| \leq \varepsilon\,\mathrm{tr}(\boldsymbol{A}).$$

holds with probability at least $1-\delta$.

Before proceeding, we point out that alternatives to Algorithm 9 have been studied before [111] was published. For example, Lin Lin [102] proposed to use the Nyström approximation instead of the randomized SVD in Algorithm 9; we will consider this algorithm in Chapter 8. However, the authors in [111] were the first to prove a $O(\varepsilon^{-1})$ upper bound on the required matrix-vector products with $\boldsymbol{A}$. Other trace estimation techniques have also been studied. For example, the authors in [138] proposed to use a low-rank approximation $\boldsymbol{B}$ to $\boldsymbol{A}$ and to use the approximation $\mathrm{tr}(\boldsymbol{B}) \approx \mathrm{tr}(\boldsymbol{A})$. However, this method only works well whenever $\boldsymbol{A}$ has a rapid eigenvalue decay. Furthermore, the authors in [52] presented the XTrace algorithm. This algorithm is a version of the Hutch++ algorithm that uses all random vectors for both trace estimation and low-rank approximation. This makes the resulting estimator a symmetric function of the random test vectors, which is a requirement for a minimum variance unbiased estimator if the random test vectors are exchangeable.[1] Chapters 7 and 8 are based on [123], and because [52] was published after [123] we omit a more detailed discussion on the XTrace algorithm.

6.3 Contributions

6.3.1 Adaptive trace estimation

The effectiveness of the two phases of Hutch++ depends on the singular values of $\boldsymbol{A}$. When $\boldsymbol{A}$ admits an accurate low-rank approximation (e.g., when its singular values decay quickly), it would be sufficient to perform the approximation $\mathrm{tr}(\boldsymbol{A}) \approx \mathrm{tr}(\boldsymbol{A}_1)$, as suggested by [138] and skip the second phase of Hutch++. On the other hand, when all singular values of $\boldsymbol{A}$ are nearly equal, the variance reduction achieved during the first phase of Hutch++ is insignificant and all effort should be spent on the second phase, the stochastic trace estimator (6.1). One can easily perceive a situation where it is preferable to spend maybe not all but most of the matrix-vector products on the stochastic trace estimator. Algorithm 9 does not recognize such situations; the number of matrix-vector products is fixed a priori and distributed in a prescribed fashion among the two phases.

Furthermore, the results in [111] are of significant theoretical importance, but since the $O(\varepsilon^{-1})$ bound comes without explicit constants it gives practitioners little indication of how many matrix-vector products to use when estimating the trace of a given matrix $\boldsymbol{A}$. One can work out the constants, for example by using results in [79] if Gaussian random

[1] A set of random vectors $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_m$ is said to be exchangeable if the family $(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_m)$ has the same distribution as the family $(\boldsymbol{\omega}_{\pi(1)}, \ldots, \boldsymbol{\omega}_{\pi(m)})$, where π is any permutation.

vectors are used, and conclude that, for fixed failure probability δ, $m = C/\varepsilon$ matrix-vector products are sufficient to get an estimate of the trace with a relative error at most ε with high probability, where C is a constant depending only on δ. However, this bound is in some cases a significant overestimation of the number of required matrix-vector products. To see this, consider the case when $\boldsymbol{A}$ has rapidly decaying singular values. In this case it would be sufficient to perform the approximation $\operatorname{tr}(\boldsymbol{A}) \approx \operatorname{tr}(\boldsymbol{A}_1)$, with potentially much fewer matrix-vector products than suggested by the C/ε bound. On the other hand, when all singular values of $\boldsymbol{A}$ are nearly equal, the standard deviation of the stochastic trace estimator, which is proportional to $\|\boldsymbol{A}\|_\mathrm{F}$, is much smaller than $\operatorname{tr}(\boldsymbol{A})$. Therefore, the relative error of the estimate produced by the stochastic trace estimator with only a few matrix-vector products, potentially much fewer than suggested by the C/ε bound, will give a sufficiently accurate estimate of the trace with high probability.[2]

In Chapter 7, we develop an adaptive version of Hutch++ to address the above mentioned issues. This algorithm takes an input tolerance ε and a tolerated failure probability δ, and outputs an estimate of the trace with an error bounded by ε with probability at least $1 - \delta$, while splitting the matrix-vector products in a near-optimal way among the two phases.

6.3.2 A single pass algorithm

Another aspect we address in this work is that the Hutch++ algorithm requires several passes over the matrix $\boldsymbol{A}$; in Algorithm 9 the matrix-vector products carried out in line 5 depend on earlier ones. In the streaming model it is desirable to design an algorithm that requires only one pass over $\boldsymbol{A}$ and if the matrix of interest is modified by a linear update $\boldsymbol{A} + \boldsymbol{E}$ one does not have to revisit $\boldsymbol{A}$ to update the output of the algorithm. Such a single pass property also increases parallelism. A single pass trace estimation algorithm was presented in [111] and we will call it *Single Pass Hutch++* in this thesis. For a symmetric positive semidefinite matrix this algorithm comes with nearly the same theoretical guarantees as Hutch++, but performs worse in practice. In the case of a symmetric positive semidefinite matrix we develop a variation of Hutch++, Nyström++, utilizing the Nyström approximation defined in Chapter 2. Nyström++ requires only one pass over $\boldsymbol{A}$ and satisfies, up to constants, the theoretical guarantees of Hutch++. This new variation of Hutch++ significantly outperforms Single Pass Hutch++ and often outperforms Hutch++.

Remark 6.1. *Note that the word adaptive is used differently in [111], where Hutch++ itself is already called adaptive because the matrix-vector products $\boldsymbol{AQ}$ depend on (and thus adapt to) the previously computed $\boldsymbol{A\Omega}$. In this work, we follow the convention where the term adaptive refers to an algorithm that adapts to a desired error bound. The Single Pass*

[2] To see this, recall that the standard deviation of the stochastic trace estimator with m samples equals $\sqrt{2/m}\|\boldsymbol{A}\|_F$. This can be much smaller than $\varepsilon \operatorname{tr}(\boldsymbol{A})$ with m potentially much smaller than C/ε, provided ε is not too small.

Hutch++ mentioned above is called NA-Hutch++ (non-adaptive variant of Hutch++) in [111].

7 A-Hutch++: An adaptive trace estimation algorithm

In this chapter, we develop an adaptive version of Hutch++ to address the issues mentioned in Section 6.3.1. In Section 7.1, we start with developing a prototype algorithm which given a prescribed tolerance ε and failure probability δ outputs an estimate of the trace of $\boldsymbol{A}$, denoted $\operatorname{tr}_{\mathsf{adap}}(\boldsymbol{A})$, such that

$$|\operatorname{tr}_{\mathsf{adap}}(\boldsymbol{A}) - \operatorname{tr}(\boldsymbol{A})| \leq \varepsilon$$

holds, provably, with probability at least $1-\delta$. At the same time, our algorithm attempts to minimize the overall number of matrix-vector products by distributing them between the two phases in a near-optimal fashion. Then, in Section 7.2, we modify the prototype algorithm to develop a more efficient adaptive trace estimation algorithm, which will be A-Hutch++. Note, however, that the potential for improving Hutch++ is limited, in [111] the $O(\varepsilon^{-1})$ bound mentioned above is proven to be optimal up to a $\log(\varepsilon^{-1})$ factor. In Section 7.3 we present the numerical experiments. In practice, we observe that our adaptive version of Hutch++ is never worse than the original Hutch++ and often outperforms it. Possibly more importantly, the output of our prototype algorithm comes with a probabilistic guarantee on the error of the estimate of $\operatorname{tr}(\boldsymbol{A})$ without requiring the user to know a priori how many matrix-vector products are needed. Our algorithm does not assume that $\boldsymbol{A}$ is positive definite, which is why we focus on estimating $\operatorname{tr}(\boldsymbol{A})$ up to a given absolute error as in (6.4).

This chapter is based on the work in [123].

7.1 Derivation of adaptive Hutch++

The aim of this section is to develop adaptive variants of Hutch++ (Algorithm 9). In a first step, we derive a prototype algorithm that aims at minimizing the number of matrix-vector products and comes with a guaranteed bound on the failure probability. The latter requires to estimate the variance or, equivalently (see (6.2)), the Frobenius norm, and this estimate needs additional matrix-vector products. Our final algorithm

A-Hutch++ reuses these matrix-vector products for trace estimation and chooses the number of them in an adaptive fashion. In turn, this creates dependencies that complicate the analysis but do not lead to observed failure probabilities that are above the prescribed failure probability.

The first phase of Algorithm 9 requires $2r$ matrix-vector products with $\boldsymbol{A}$ to obtain a rank-r approximation $\boldsymbol{Q}^{(r)}\boldsymbol{Q}^{(r)T}\boldsymbol{A}$, where we have added a superscript to emphasize the dependence on r. Let $M(r)$ be the number of matrix-vector products with $\boldsymbol{A}$ in the second phase such that the stochastic trace estimator of

$$\boldsymbol{A}_{\text{rest}}^{(r)} := (\boldsymbol{I} - \boldsymbol{Q}^{(r)}\boldsymbol{Q}^{(r)T})\boldsymbol{A}(\boldsymbol{I} - \boldsymbol{Q}^{(r)}\boldsymbol{Q}^{(r)T})$$

attains a prescribed accuracy and success probability. Then the total number of matrix-vector products with $\boldsymbol{A}$ is

$$m(r) = 2r + M(r). \tag{7.1}$$

We aim at minimizing $m(r)$ in order to obtain a near-optimal distribution of matrix-vector products between the two phases.[1] For this purpose, we first derive a suitable expression for $M(r)$.

7.1.1 Analysis of trace estimation

The tightest tail bound available in the literature for the stochastic trace estimator $\text{tr}_m(\boldsymbol{B})$ for a symmetric matrix $\boldsymbol{B}$ is Theorem 6.1. In most situations of interest, the term involving $\|\boldsymbol{B}\|_2$ will be insignificant. The following lemma is a variation of Theorem 6.1 that suppresses this term for sufficiently large m, similar to [111, Lemma 2.1].

Lemma 7.1. *Let $\rho(\boldsymbol{B}) = \frac{\|\boldsymbol{B}\|_F^2}{\|\boldsymbol{B}\|_2^2}$ denote the stable rank of $\boldsymbol{B}$. Given $\upsilon > 0$ assume that $m \geq \frac{4(1+\upsilon)\log(2/\delta)}{\upsilon^2\rho(\boldsymbol{B})}$. Then the inequality*

$$|\text{tr}_m(\boldsymbol{B}) - \text{tr}(\boldsymbol{B})| \leq 2\sqrt{1+\upsilon}\sqrt{\frac{\log(2/\delta)}{m}}\|\boldsymbol{B}\|_{\text{F}} \tag{7.2}$$

holds with probability at least $1-\delta$.

Proof. Inserting the right-hand side of (7.2), $\varepsilon := 2\sqrt{1+\upsilon}\sqrt{\frac{\log(2/\delta)}{m}}\|\boldsymbol{B}\|_{\text{F}}$, into Theo-

[1] In practice we perform randomized low-rank approximations. Consequently, $\boldsymbol{A}_{\text{rest}}^{(r)}$ is random and therefore the function m is a random variable. Hence, it can be ambiguous what it means to minimize m. To clarify this, first note that we always assume $r \leq n$, where $\boldsymbol{A}$ is $n \times n$, since when $r = n$ we are able to exactly compute $\text{tr}(\boldsymbol{A})$. Therefore, we will never sample more than n random vectors to obtain a low-rank approximation. Thus, let $\boldsymbol{\Omega} \in \mathbb{R}^{n\times n}$ be the random matrix from which we can construct $\boldsymbol{Q}^{(1)}, \boldsymbol{Q}^{(2)}, \ldots, \boldsymbol{Q}^{(n)}$. Conditioned on $\boldsymbol{\Omega}$ the function m becomes deterministic and has a minimum, which is what we aim to find. We will describe a heuristic strategy to find the minimum in Section 7.1.2.

rem 6.1 one obtains the desired result:

$$\begin{aligned}\mathbb{P}\left(|\operatorname{tr}_m(\boldsymbol{B})-\operatorname{tr}(\boldsymbol{B})|\geq\varepsilon\right) &\leq 2\exp\left(-\frac{(1+\upsilon)\log(2/\delta)\|\boldsymbol{B}\|_{\mathrm{F}}}{\|\boldsymbol{B}\|_{\mathrm{F}}+2\sqrt{1+\upsilon}\sqrt{\frac{\log(2/\delta)}{m}}\|\boldsymbol{B}\|_2}\right)\\ &\leq 2\exp\left(-\frac{(1+\upsilon)\log(2/\delta)\|\boldsymbol{B}\|_{\mathrm{F}}}{(1+\upsilon)\|\boldsymbol{B}\|_{\mathrm{F}}}\right)=\delta,\end{aligned}$$

where the second inequality utilizes

$$\upsilon\|\boldsymbol{B}\|_{\mathrm{F}}\geq 2\sqrt{1+\upsilon}\sqrt{\frac{\log(2/\delta)}{m}}\|\boldsymbol{B}\|_2,$$

a consequence of the assumption on m. □

Let

$$C(\varepsilon,\delta):=4(1+\upsilon)\varepsilon^{-2}\log(2/\delta). \tag{7.3}$$

By Lemma 7.1, for sufficiently small ε, $C(\varepsilon,\delta)\|\boldsymbol{B}\|_{\mathrm{F}}^2$ samples are sufficient to achieve $|\operatorname{tr}_m(\boldsymbol{B})-\operatorname{tr}(\boldsymbol{B})|\leq\varepsilon$ with probability at least $1-\delta$. In practice one cannot assume to know, or be able to compute, the stable rank appearing in the condition $m\geq\frac{4(1+\upsilon)\log(2/\delta)}{\upsilon^2\rho(\boldsymbol{B})}$. Since the stable rank is always larger than 1, requiring $m\geq\frac{4(1+\upsilon)\log(2/\delta)}{\upsilon^2}$ would be sufficient to ensure that $m\geq\frac{4(1+\upsilon)\log(2/\delta)}{\upsilon^2\rho(\boldsymbol{B})}$. However, in practice we set $\upsilon=0$ and completely omit the side condition $m\geq\frac{4(1+\upsilon)\log(2/\delta)}{\upsilon^2\rho(\boldsymbol{B})}$. While not justified by Lemma 7.1, we observe no significant loss in the success probabilities of our algorithm, see Section 7.3.1.

7.1.2 Finding the minimum of $m(r)$

Applying the results above to $\boldsymbol{B}=\boldsymbol{A}_{\text{rest}}^{(r)}$ implies that a suitable choice for the function $m(r)$ in (7.1) is given by

$$m(r)=2r+C(\varepsilon,\delta)\|\boldsymbol{A}_{\text{rest}}^{(r)}\|_{\mathrm{F}}^2. \tag{7.4}$$

In the idealistic scenario that $\boldsymbol{Q}^{(r)}$ contains the dominant r singular vectors, we have $\|\boldsymbol{A}_{\text{rest}}^{(r)}\|_F^2=\sigma_{r+1}^2+\cdots+\sigma_n^2$. This implies that the differences $m(r)-m(r-1)=2-C(\varepsilon,\delta)\sigma_{r+1}^2$ are monotonically increasing and switch sign at most once. In turn, r^* is a global minimum whenever it is a local minimum, that is, $m(r^*\pm 1)\geq m(r^*)$. Since $\boldsymbol{Q}^{(r)}$ only approximates the space spanned by the dominant r singular vectors of $\boldsymbol{A}$, these relations are not guaranteed to hold. In practice, we have observed $m(r^*\pm 1)\geq m(r^*)$ to remain a reliable criterion; see Figure 7.1 for an example.

Evaluating $m(r)$ involves the quantity $\|\boldsymbol{A}_{\text{rest}}^{(r)}\|_{\mathrm{F}}^2$, which is too expensive to evaluate. Using

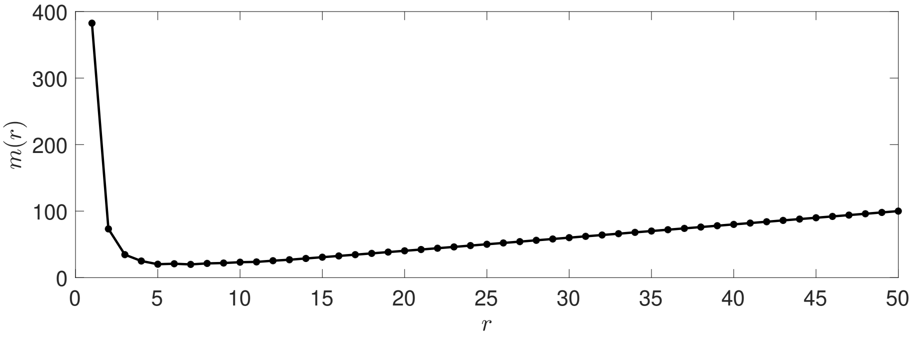


Figure 7.1: In this example we let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T \in \mathbb{R}^{1000\times 1000}$ where $\boldsymbol{U}$ is a random orthogonal matrix and $\boldsymbol{\Lambda}$ is a diagonal matrix with $\boldsymbol{\Lambda}_{ii} = 1/i^2$. The x-axis shows the rank r, and the y-axis is the function $m(r)$ defined in (7.4) with $\delta = 0.01$, $\varepsilon = 0.05\,\mathrm{tr}(\boldsymbol{A})$ and $v = 0$. The function has its minimum at $r^* = 7$.

the symmetry of $\boldsymbol{A}$ and the unitary invariance of the Frobenius norm we get

$$\|\boldsymbol{A}_{\text{rest}}^{(r)}\|_{\text{F}}^2 = \|\boldsymbol{A}\|_{\text{F}}^2 + \|\boldsymbol{Q}^{(r)T}\boldsymbol{A}\boldsymbol{Q}^{(r)}\|_{\text{F}}^2 - 2\|\boldsymbol{A}\boldsymbol{Q}^{(r)}\|_{\text{F}}^2. \tag{7.5}$$

In turn, $m(r)$ and the function

$$\tilde{m}(r) := 2r + C(\varepsilon,\delta)\Big(\|\boldsymbol{Q}^{(r)T}\boldsymbol{A}\boldsymbol{Q}^{(r)}\|_{\text{F}}^2 - 2\|\boldsymbol{A}\boldsymbol{Q}^{(r)}\|_{\text{F}}^2\Big) \tag{7.6}$$

have the same minimum. The latter can be cheaply computed by recursive updating, without any additional matrix-vector products with $\boldsymbol{A}$.

To summarize, we adapt the randomized SVD to build $\boldsymbol{Q}^{(r)}$ column-by-column, similar to as described in [79, Section 4.4], and stop the loop whenever a minimum of $\tilde{m}(r)$ is detected. By the heuristics discussed above, it is safe to stop at $r = r^*$ when $\tilde{m}(r^*) > \tilde{m}(r^*-1) > \tilde{m}(r^*-2)$.

7.1.3 Estimating the Frobenius norm of the remainder

Having found an approximate minimum r^* of $\tilde{m}(r)$ and computed $\boldsymbol{Q} \equiv \boldsymbol{Q}^{(r^*)}$, it remains to apply stochastic trace estimation to $\boldsymbol{A}_{\text{rest}} \equiv \boldsymbol{A}_{\text{rest}}^{(r^*)}$. By Lemma 7.1 it suffices to use $M \geq C(\varepsilon,\delta)\|\boldsymbol{A}_{\text{rest}}\|_{\text{F}}^2$ samples. Because computing $\|\boldsymbol{A}_{\text{rest}}\|_{\text{F}}$ is too expensive, we need to resort (once more) to a stochastic estimator utilizing only matrix-vector products. The following result is essential for that purpose.

Lemma 7.2. *Let $\boldsymbol{\Omega}$ be a random $n \times k$ matrix whose entries are i.i.d. $\mathcal{N}(0,1)$ random variables and let $\boldsymbol{B} \in \mathbb{R}^{n\times n}$. For any $\alpha \in (0,1)$ it holds that*

$$\mathbb{P}\left(\frac{1}{k}\|\boldsymbol{B}\boldsymbol{\Omega}\|_{\text{F}}^2 < \alpha\|\boldsymbol{B}\|_{\text{F}}^2\right) \leq \mathbb{P}(X < \alpha) = \frac{\gamma(k/2, \alpha k/2)}{\Gamma(k/2)},$$

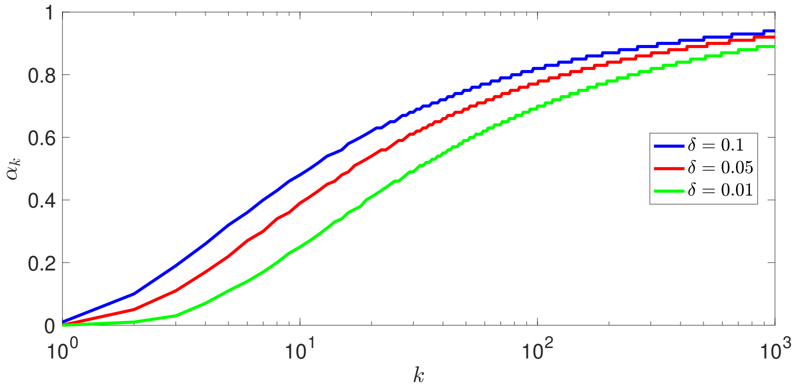


Figure 7.2: For different choices of δ, this plot demonstrates the relationship between k and the largest choice of α such that $\frac{\gamma(k/2,\alpha k/2)}{\Gamma(k/2)} \leq \delta$.

where $X \sim \Gamma(k/2, k/2)$ *(gamma distribution with shape and rate parameter* $k/2$*),* $\gamma(s,x) := \int_0^x t^{s-1}e^{-t}dt$ *is the lower incomplete gamma function and* $\Gamma(s)$ *is the standard gamma function.*

Proof. It is well known that

$$\frac{1}{k}\|\boldsymbol{B\Omega}\|_{\mathrm{F}}^2 = \frac{1}{k}\sum_{j=1}^{n}\sigma_j^2 Z_j,$$

where Z_j, $j = 1,\ldots,n$, denote i.i.d. χ_k^2 random variables; see, e.g., [74, Section 2]. Setting $X_j := \frac{1}{k}Z_j \sim \Gamma(k/2, k/2)$ and $\lambda_j = \frac{\sigma_j^2}{\|\boldsymbol{B}\|_F^2}$ for $j = 1,\ldots,n$ we rewrite

$$\mathbb{P}\left(\frac{1}{k}\|\boldsymbol{B\Omega}\|_{\mathrm{F}}^2 < \alpha\|\boldsymbol{B}\|_{\mathrm{F}}^2\right) = \mathbb{P}\left(\sum_{j=1}^{n}\lambda_j X_j < \alpha\right).$$

By [133, Theorem 2.2] the right-hand side is bounded for every $\alpha \in (0,1)$ by $\mathbb{P}\left(X_1 < \alpha\right)$, which completes the proof. □

Lemma 7.2 states that if $\frac{\gamma(k/2,\alpha k/2)}{\Gamma(k/2)} \leq \delta$ then $\frac{1}{k\alpha}\|\boldsymbol{B\Omega}\|_{\mathrm{F}}^2 > \|\boldsymbol{B}\|_{\mathrm{F}}^2$ with probability at least $1-\delta$. Hence, using $M := \lceil C(\varepsilon,\delta)\cdot\frac{1}{k\alpha}\|\boldsymbol{A}_{\mathrm{rest}}\boldsymbol{\Omega}\|_{\mathrm{F}}^2\rceil$ samples ensures an error of at most ε with low failure probability. See Figure 7.2 for the relationship between k, α and δ.

7.1.4 A prototype algorithm

Combining the results presented above we obtain the prototype algorithm presented in Algorithm 10. To reduce the number of passes over the matrix $\boldsymbol{A}$ the algorithm can be implemented in a block-wise fashion, which can in turn lead to a reduction of wall-clock time. For block-size $b = 1$ we use the heuristic stopping criteria for the low-rank approximation described above. For larger block-sizes it is sufficient to use $m(r^* - b) < m(r^*)$ as a stopping criteria. A simple probabilistic analysis yields the following result on the success probability of Algorithm 10:

Algorithm 10 Prototype algorithm

```
input: Symmetric $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Error tolerance $\varepsilon > 0$. Failure probability $\delta \in (0,1)$.
Parameter $\upsilon > 0$. Block-size $b$.
output: An approximation to $\operatorname{tr}(\boldsymbol{A})$ : $\operatorname{tr}_{\mathsf{adap}}(\boldsymbol{A})$.
 1: $\boldsymbol{Y}^{(b)} = \boldsymbol{A}\boldsymbol{\Omega}^{(b)}$ where $\boldsymbol{\Omega}^{(b)} \in \mathbb{R}^{n\times b}$ has i.i.d. $\mathcal{N}(0,1)$ entries.
 2: Obtain orthonormal basis $\widehat{\boldsymbol{Q}}^{(b)}$ for range$\left(\boldsymbol{Y}^{(b)}\right)$.
 3: $\boldsymbol{Q}^{(1)} = \widehat{\boldsymbol{Q}}^{(1)}$
 4: $\text{trest}_1 = \operatorname{tr}\left(\widehat{\boldsymbol{Q}}^{(1)T}\left(\boldsymbol{A}\widehat{\boldsymbol{Q}}^{(1)}\right)\right)$
 5: Compute $\tilde{m}(b)$.
 6: $r = b$
 7: while A minimum of $\tilde{m}(r)$ not detected do
 8:     $\boldsymbol{Y}^{(r+b)} = \boldsymbol{A}\boldsymbol{\Omega}^{(r+b)}$ where $\boldsymbol{\Omega}^{(r+b)} \in \mathbb{R}^{n\times b}$ has i.i.d. $\mathcal{N}(0,1)$ entries.
 9:     $\widetilde{\boldsymbol{Q}}^{(r+b)} = (\boldsymbol{I} - \boldsymbol{Q}^{(r)}\boldsymbol{Q}^{(r)T})\boldsymbol{Y}^{(r+b)}$
10:     Obtain orthonormal basis $\widehat{\boldsymbol{Q}}^{(r+b)}$ for range$\left(\widetilde{\boldsymbol{Q}}^{(r+b)}\right)$.
11:     $\boldsymbol{Q}^{(r+b)} = \begin{bmatrix}\boldsymbol{Q}^{(r)} & \widehat{\boldsymbol{Q}}^{(r+b)}\end{bmatrix}$
12:     $\text{trest}_1 = \text{trest}_1 + \operatorname{tr}\left(\widehat{\boldsymbol{Q}}^{(r+b)T}\left(\boldsymbol{A}\widehat{\boldsymbol{Q}}^{(r+b)}\right)\right)$
13:     Update $\tilde{m}(r+b)$ recursively.
14:     $r = r + b$
15: end while
16: Let $\boldsymbol{Q} = \boldsymbol{Q}^{(r)}$ and $\boldsymbol{A}_{\text{rest}} = (\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^T)\boldsymbol{A}(\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^T)$.    ▷ $\boldsymbol{A}_{\text{rest}}$ is never formed
    explicitly.
17: Choose $(k,\alpha) \in \mathbb{N} \times (0,1)$ such that $\frac{\gamma(k/2,\alpha k/2)}{\Gamma(k/2)} \leq \delta$.
18: $M = \max\left\{\frac{4(1+\upsilon)\log(2/\delta)}{\upsilon^2}, \left\lceil C(\varepsilon,\delta)\cdot\frac{1}{k\alpha}\|\boldsymbol{A}_{\text{rest}}\boldsymbol{\Psi}\|_{\mathrm{F}}^2\right\rceil\right\}$ where $\boldsymbol{\Psi} \in \mathbb{R}^{n\times k}$ has i.i.d.
    $\mathcal{N}(0,1)$ entries .
19: $\text{trest}_2 = \operatorname{tr}_M(\boldsymbol{A}_{\text{rest}})$
20: return $\operatorname{tr}_{\mathsf{adap}}(\boldsymbol{A}) = \text{trest}_1 + \text{trest}_2$
```

Lemma 7.3. *The output of Algorithm 10 satisfies* $|\operatorname{tr}_{\mathsf{adap}}(\boldsymbol{A}) - \operatorname{tr}(\boldsymbol{A})| \leq \varepsilon$ *with probability at least* $1 - 2\delta$.

Proof. For the moment, let us consider $\boldsymbol{Q}$ fixed and, hence, $\boldsymbol{A}_{\text{rest}}$ deterministic. For a fixed arbitrary integer N let us consider the event

$$S_N := \left\{|\operatorname{tr}_N(\boldsymbol{A}_{\text{rest}}) - \operatorname{tr}(\boldsymbol{A}_{\text{rest}})| \leq \varepsilon\right\}.$$

Let M be the random variable defined in line 18 of Algorithm 10. Therefore, S_M is the event that the estimate of $\operatorname{tr}(\boldsymbol{A}_{\text{rest}})$ from Algorithm 10 has an error at most ε. That is,

$$S_M = \{|\operatorname{tr}_M(\boldsymbol{A}_{\text{rest}}) - \operatorname{tr}(\boldsymbol{A}_{\text{rest}})| \leq \varepsilon\} = \bigcup_{N\geq 1} [S_N \cap \{M = N\}].$$

The analysis of $\mathbb{P}(S_M)$ is complicated by the fact that the integer M defined in line 18 of Algorithm 10 is also random. Letting

$$M_1 := \max\left\{\frac{4(1+\upsilon)\log(2/\delta)}{\upsilon^2\rho(\boldsymbol{A}_{\text{rest}})}, C(\varepsilon,\delta)\|\boldsymbol{A}_{\text{rest}}\|_{\mathrm{F}}^2\right\},$$

we know from Lemma 7.2 that $\mathbb{P}(M \geq M_1) \geq 1-\delta$ and from (7.3) that $\mathbb{P}(S_N) \geq 1-\delta$ for $N \geq M_1$. Moreover, it is important to remark that the events S_N and $M = N$ are independent. In particular, this implies $\mathbb{P}(S_M|M = N) = \mathbb{P}(S_N)$. Combining these observations yields

$$\begin{aligned}
\mathbb{P}(S_M) &\geq \mathbb{P}(S_M \cap \{M \geq M_1\}) \\
&= \sum_{N\geq M_1} \mathbb{P}(S_M \cap \{M = N\}) = \sum_{N\geq M_1} \mathbb{P}(S_M|M = N)\mathbb{P}(M = N) \\
&= \sum_{N\geq M_1} \mathbb{P}(S_N)\mathbb{P}(M = N) \geq (1-\delta)\sum_{N\geq M_1} \mathbb{P}(M = N) \\
&= (1-\delta)\mathbb{P}(M \geq M_1) \geq (1-\delta)^2 \geq 1-2\delta,
\end{aligned}$$

which holds independently of $\boldsymbol{Q}$ and thus completes the proof. □

7.2 A-Hutch++

To turn Algorithm 10 into a practical method, we need to address the choice of the pair (k,α) in line 17 and apply further modification to increase its efficiency by reusing the matrix vector products in the Frobenius norm estimation in line 18 in the trace estimation in line 19 of Algorithm 10.

For fixed k, it makes sense to choose α as large as possible because M decreases with increasing α; see line 18. Thus, we set

$$\alpha_k := \sup\left\{\alpha \in (0,1) : \frac{\gamma\left(k/2, \alpha k/2\right)}{\Gamma\left(k/2\right)} \leq \delta\right\}. \tag{7.7}$$

Lemma 7.4. *The sequence* $\{\alpha_k\}_{k\in\mathbb{N}}$ *defined by* (7.7) *increases monotonically and converges to* 1.

Proof. Letting $X := \frac{1}{k}\sum_{i=1}^{k} X_i \sim \Gamma(k/2, k/2)$ for i.i.d. χ_1^2 random variables X_i, we set

$$p_k(\alpha) := \mathbb{P}\left(X \leq \alpha\right) = \frac{\gamma\left(k/2, \alpha k/2\right)}{\Gamma\left(k/2\right)}.$$

By [133, Theorem 2.1] $p_{k+1}(\alpha) \leq p_k(\alpha)$ for every $\alpha \in (0,1]$. Furthermore, by continuity of p_k in α and monotonicity of $p_k(\alpha)$ in k we have

$$\delta = p_k(\alpha_k) = p_{k+1}(\alpha_{k+1}) \leq p_k(\alpha_{k+1}).$$

Thus, by monotonicity of p_k in α we have $\alpha_k \leq \alpha_{k+1}$, which proves the monotonicity of the sequence $\{\alpha_k\}_{k\in\mathbb{N}}$.

To show $\alpha_k \to 1$ as $k \to +\infty$, let $\alpha_\varepsilon := 1 - \varepsilon > 0$ for fixed arbitrary $0 < \varepsilon < 1$. By the law of large numbers, $p_k(\alpha_\varepsilon) \to 0$ and by the argument above this convergence is monotonic. Let $k_{\varepsilon,\delta} = \min\{k \in \mathbb{N} : p_k(\alpha_\varepsilon) \leq \delta\}$. Let $k \geq k_{\varepsilon,\delta}$. Then, $\delta \geq p_{k_{\varepsilon,\delta}}(\alpha_\varepsilon) \geq p_k(\alpha_\varepsilon)$. Thus, for all $k \geq k_{\varepsilon,\delta}$ we have $1 \geq \alpha_k \geq \alpha_\varepsilon \geq 1 - \varepsilon$, as required. □

Furthermore, define the following random sequence M_k:

$$M_k := C(\varepsilon,\delta) \cdot \frac{1}{k\alpha_k}\|\boldsymbol{A}_{\text{rest}}\boldsymbol{\Psi}^{(k)}\|_{\text{F}}^2, \quad \boldsymbol{\Psi}^{(k)} = \begin{bmatrix}\boldsymbol{\Psi}^{(k-1)} & \boldsymbol{\psi}^{(k)}\end{bmatrix}, \quad \boldsymbol{\psi}^{(k)} \sim N(\boldsymbol{0}, \boldsymbol{I}). \tag{7.8}$$

By the law of large numbers we have $M_k \to C(\varepsilon,\delta)\|\boldsymbol{A}_{\text{rest}}\|_{\text{F}}^2$ almost surely as $k \to +\infty$. If we reuse the matrix-vector products from line 18 in line 19 the total number of performed matrix vector products in the second phase of Algorithm 10 is

$$\max\left\{k, \lceil M_k \rceil\right\}. \tag{7.9}$$

Because of the monotonicity of α_k, and as seen in Figure 7.3, M_k is expected to decrease in k. Hence, in order to minimize (7.9) we choose k such that $k = \lceil M_k \rceil$. Thus, we evaluate M_k inside a while loop and stop the while loop once we detect $k > M_k$ for the first time. At this point we reuse the computation $\boldsymbol{A}_{\text{rest}}\boldsymbol{\Psi}^{(k)}$ to estimate $\text{tr}(\boldsymbol{A}_{\text{rest}})$. The resulting algorithm is presented in Algorithm 11. As with the prototype algorithm, Algorithm 11 can also be implemented to perform matrix-vector products in a block-wise fashion.

Due to the lack of independence between the Frobenius norm estimation and the stochastic trace estimation, the proof of Lemma 7.3 does not extend to Algorithm 11. In turn, this algorithm does not come with the same type of success guarantee. However, as presented in Section 7.3.1 the empirical failure probabilities remain well below the prescribed failure probability.

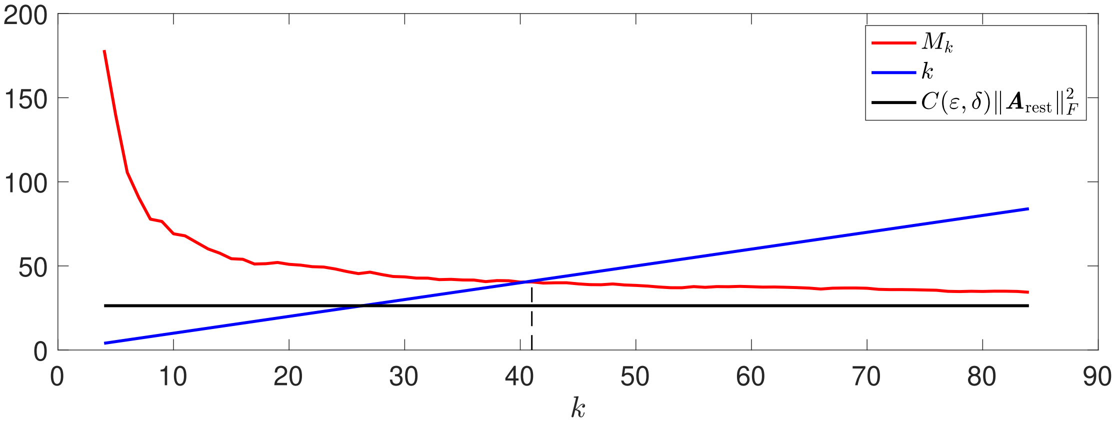


Figure 7.3: In this example we let $\boldsymbol{A} = \boldsymbol{U\Lambda U}^T \in \mathbb{R}^{1000\times1000}$ where $\boldsymbol{U}$ is a random orthogonal matrix and $\boldsymbol{\Lambda}$ is a diagonal matrix with $\boldsymbol{\Lambda}_{ii} = 1/i^{1.5}$. We run Algorithm 10 with $\varepsilon = 0.01\,\text{tr}(\boldsymbol{A})$, $\delta = 0.05$ and $\upsilon = 0$ to obtain $\boldsymbol{A}_{\text{rest}}$ defined in line 16. The x-axis shows the number of matrix-vector products with $\boldsymbol{A}_{\text{rest}}$. The red line shows the evolution of the sequence M_k defined in (7.8), the blue line shows the linear line k against k and the black line is the number of matrix-vector products with $\boldsymbol{A}_{\text{rest}}$ to guarantee an error less than ε with probability at least $1-\delta$. We stop the while loop in Algorithm 11 once the red and blue line cross.

Algorithm 11 A-Hutch++

```
input: Symmetric $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Error tolerance $\varepsilon > 0$. Failure probability $\delta \in (0,1)$. Block-size $b$.
output: An approximation to $\text{tr}(\boldsymbol{A})$ : $\text{tr}_{\mathsf{adap}}(\boldsymbol{A})$.
Perform lines 1–16 in Algorithm 10 to get $\boldsymbol{Q}$, $\text{trest}_1$ and $\boldsymbol{A}_{\text{rest}}$.
Initialize $\boldsymbol{\Psi}^{(0)} = []$ and $\boldsymbol{C}^{(0)} = []$.
Initialize $M_0 = \infty$ and $k = 0$.
while $M_k > k$ do
    $k = k + b$
    $\alpha_k = \sup\left\{\alpha \in (0,1) : \frac{\gamma\left(\frac{k}{2},\alpha\frac{k}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \leq \delta\right\}$
    Generate a random matrix $\widehat{\boldsymbol{\Psi}}^{(k)} \in \mathbb{R}^{n\times b}$ and append $\boldsymbol{\Psi}^{(k)} = \begin{bmatrix}\boldsymbol{\Psi}^{(k-b)} & \widehat{\boldsymbol{\Psi}}^{(k)}\end{bmatrix}$.
    Compute $\widehat{\boldsymbol{C}}^{(k)} = \boldsymbol{A}_{\text{rest}}\widehat{\boldsymbol{\Psi}}^{(k)}$ and append $\boldsymbol{C}^{(k)} = \begin{bmatrix}\boldsymbol{C}^{(k-b)} & \widehat{\boldsymbol{C}}^{(k)}\end{bmatrix}$.
    Over-estimate $\|\boldsymbol{A}_{\text{rest}}\|_{\text{F}}^2$ with $\text{estFrob}_k = \frac{1}{k\alpha_k}\|\boldsymbol{C}^{(k)}\|_{\text{F}}^2$.
    Define $M_k = C(\varepsilon,\delta)\text{estFrob}_k$.
end while
return $\text{tr}_{\mathsf{adap}}(\boldsymbol{A}) = \text{trest}_1 + \frac{1}{k}\text{tr}(\boldsymbol{\Psi}^{(k)T}\boldsymbol{C}^{(k)})$
```

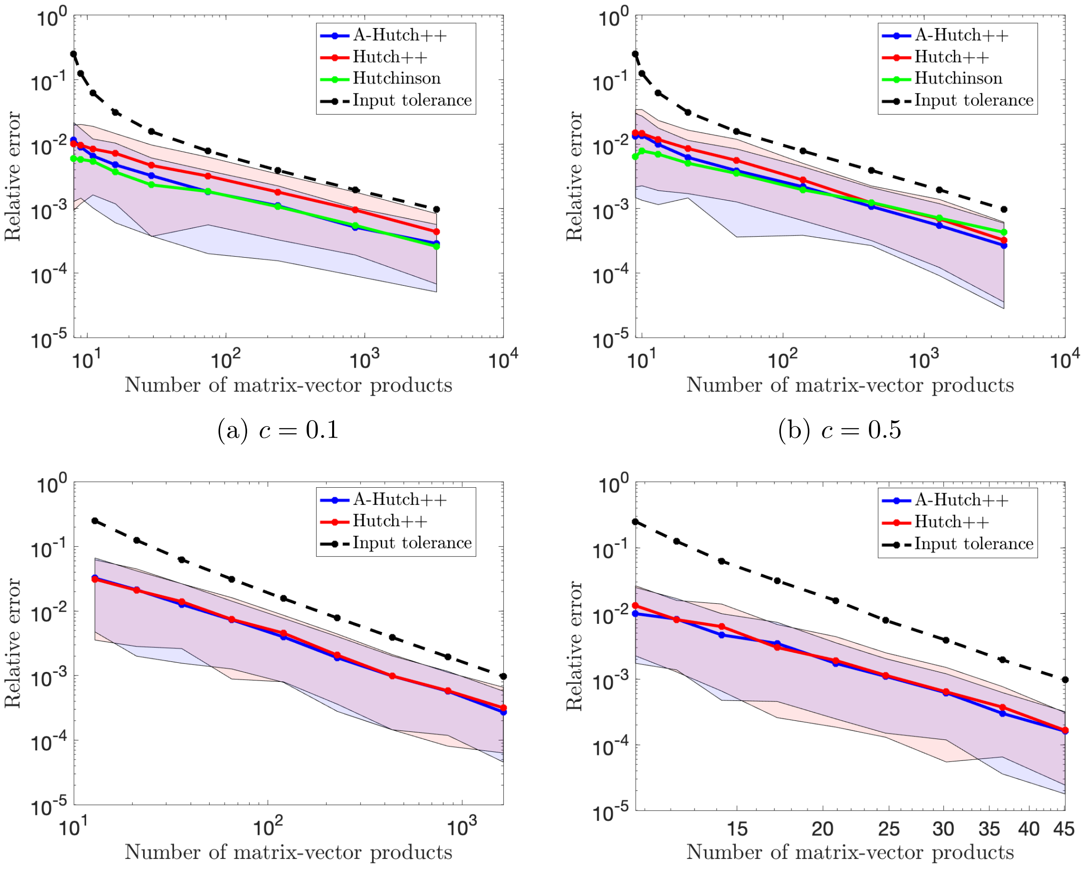
7.3 Numerical experiments

All numerical experiments in this paper have been performed in Matlab, version R2020a; our implementation of Algorithm 11 is available at https://github.com/davpersson/A-Hutch- together with the scripts to reproduce all figures and tables in this paper.

For a variety of matrices from [13, 59, 111, 138], we compare the newly proposed A-Hutch++ algorithm with Hutch++. In A-Hutch++ we fix $\delta = 0.05$ in all our experiments and we let $\varepsilon = \frac{|\operatorname{tr}(\boldsymbol{A})|}{2^t}$ for $t = 2, 3, \dots, 10$, except in Figure 7.7b where we let $t = 3, 4, \dots, 11$. The error of the estimate produced by A-Hutch++ implemented in a block-wise fashion is essentially identical to the unblocked version of A-Hutch++, i.e. $b = 1$, as long as the block-size is small compared to the number of required matrix-vector products. Therefore, for simplicity, we set the block-size to $b = 1$ in all experiments. Furthermore, as discussed in Section 7.1.1 we set $v = 0$ and omit the side condition on m. For each considered matrix, for each value of ε, we first run Algorithm 11 and count the number of matrix-vector products that have been used to obtain the estimate, then we run Algorithm 9 with the same number of matrix-vector products. For each value of ε we repeat this 100 times and plot the average relative error on the y-axis and the average required matrix-vector products on the x-axis. In each figure, the blue line is the average relative error from A-Hutch++, the red line is the average relative error from Hutch++, with the same number of matrix vector products, and the black dashed line is the ε that was used as the input tolerance of A-Hutch++. For matrices with slow eigenvalue decay we have also included the average relative error from the Hutchinson estimator (6.1), see the green line in Figures 7.4a, 7.4b, 7.7b and 7.8a. The shaded blue area shows the 10^{th} to 90^{th} percentiles[2] of the results from A-Hutch++, and the shaded red area shows the 10^{th} to 90^{th} percentiles of the results from Hutch++, see e.g. Figure 7.4.

In the numerical experiments we observe that A-Hutch++ performs better compared to Hutch++ for matrices with slower singular value decay; see e.g. Figure 7.4a, in which A-Hutch++ achieves an average relative error of 0.001827 using an average of 74.41 matrix-vector products (6^{th} blue point in the figure). In comparison, Hutch++ achieves an average relative error of 0.001804 using an average of 237.7 matrix-vector products (7^{th} red point in the figure). Hence, in these cases the adaptivity does improve the performance compared to Hutch++. For faster singular value decay the two algorithms perform similarly. However, in no case does Hutch++ perform noticeably better compared to A-Hutch++.

[2] We show the 90% percentile because, if we did not reuse the matrix-vector products of the Frobenius norm estimation for the Hutchinson trace estimator, Lemma 7.3 would ensure a failure probability of at most $2\delta = 10\%$.



(a) $c = 0.1$

(b) $c = 0.5$

(c) $c = 1$

(d) $c = 3$

Figure 7.4: Comparison of A-Hutch++ and Hutch++ for the estimation of the trace of the synthetic matrices with algebraic decay from Section 7.3.1.

7.3.1 Synthetic matrices

We create matrices with algebraically decaying singular values as in (4.24) with $n = 5000, C = 1$ and $c \in \{0.1, 0.5, 1, 3\}$ and whose eigenvectors are drawn uniformly from random orthogonal matrices. The results are shown in Figure 7.4.

Furthermore, using these example matrices we also estimated the failure probability of A-Hutch++. Table 7.1 demonstrates the empirical failure probabilities from 100000 repeats of A-Hutch++ for different input pairs (ε, δ). In all cases the empirical failure probabilities remain well below the prescribed failure probability.

In addition, to demonstrate that A-Hutch++ allocates more matrix-vector products to the Hutchinson estimator for matrices with slow eigenvalue decay and vice versa for matrices with fast eigenvalue decay, we also include a table displaying the distribution of the matrix-vector products between the two phases. See Table 7.2.

δ ε	0.1	0.05	0.01
$0.1\,\mathrm{tr}(\boldsymbol{A})$	0	0	0
$0.01\,\mathrm{tr}(\boldsymbol{A})$	0.00285	0.00076	0.00005
$0.005\,\mathrm{tr}(\boldsymbol{A})$	0.00686	0.00244	0.00015

(a) $c = 0.1$

δ ε	0.1	0.05	0.01
$0.1\,\mathrm{tr}(\boldsymbol{A})$	0	0	0
$0.01\,\mathrm{tr}(\boldsymbol{A})$	0.00484	0.00126	0.00010
$0.005\,\mathrm{tr}(\boldsymbol{A})$	0.00855	0.00331	0.00032

(b) $c = 0.5$

δ ε	0.1	0.05	0.01
$0.1\,\mathrm{tr}(\boldsymbol{A})$	0.00026	0.00002	0
$0.01\,\mathrm{tr}(\boldsymbol{A})$	0.00607	0.00186	0.00018
$0.005\,\mathrm{tr}(\boldsymbol{A})$	0.00804	0.00250	0.00030

(c) $c = 1$

δ ε	0.1	0.05	0.01
$0.1\,\mathrm{tr}(\boldsymbol{A})$	0	0	0
$0.01\,\mathrm{tr}(\boldsymbol{A})$	0.00002	0	0
$0.005\,\mathrm{tr}(\boldsymbol{A})$	0.00006	0	0

(d) $c = 3$

Table 7.1: Empirical failure probabilities from 100000 repeats of applying A-Hutch++ on the synthetic matrices described in Section 7.3.1.

t	Total	Low rank approx.	Hutchinson est.	Ratio
2	8.00	6.00	2.00	0.25
3	9.00	6.00	3.00	0.33
4	11.00	6.00	5.00	0.45
5	16.00	6.00	10.00	0.63
6	29.04	6.00	23.04	0.79
7	74.41	6.00	68.41	0.92
8	237.66	6.00	231.66	0.97
9	858.13	6.00	852.13	0.99
10	3302.76	6.00	3296.76	1.00

(a) $c = 0.1$

t	Total	Low rank approx.	Hutchinson est.	Ratio
2	9.00	6.00	3.00	0.33
3	10.01	6.00	4.01	0.40
4	13.06	6.00	7.06	0.54
5	21.21	6.00	15.21	0.72
6	46.94	6.02	40.92	0.87
7	138.24	10.14	128.10	0.93
8	424.31	49.18	375.13	0.88
9	1287.60	206.96	1080.64	0.84
10	3688.39	914.18	2774.21	0.75

(b) $c = 0.5$

t	Total	Low rank approx.	Hutchinson est.	Ratio
2	12.86	6.16	6.70	0.52
3	21.07	8.86	12.21	0.58
4	36.02	15.08	20.94	0.58
5	65.15	27.68	37.47	0.58
6	120.04	52.84	67.20	0.56
7	228.02	101.18	126.84	0.56
8	436.75	199.14	237.61	0.54
9	843.98	396.32	447.66	0.53
10	1630.29	793.36	836.93	0.51

(c) $c = 1$

t	Total	Low rank approx.	Hutchinson est.	Ratio
2	10.66	8.20	2.46	0.23
3	12.24	8.88	3.36	0.27
4	14.24	10.76	3.48	0.24
5	17.16	12.44	4.72	0.28
6	20.91	15.22	5.69	0.27
7	24.70	18.28	6.42	0.26
8	30.28	22.50	7.78	0.26
9	36.57	27.68	8.89	0.24
10	45.14	34.50	10.64	0.24

(d) $c = 3$

Table 7.2: The average distribution of matrix-vector products between the low rank approximation phase and stochastic trace esimation phase of A-Hutch++ applied on the synthetic matrices with algebraic decay and input tolerance $\varepsilon = 2^{-t}\operatorname{tr}(\boldsymbol{A})$ for $t = 2, 3, \ldots, 10$. A-Hutch++ requires at least 6 matrix-vector products to detect a minimum of the function $\tilde{m}(r)$ in (7.6).

7.3.2 Triangle counting

As discussed in Section 3.2, for an undirected graph with adjacency matrix $\boldsymbol{C}$, the number of triangles in the graph is equal to $\frac{1}{6}\operatorname{tr}(\boldsymbol{C}^3)$; counting triangles arises for instance in data mining applications [8]. We apply A-Hutch++ and Hutch++ to $\boldsymbol{A} = \boldsymbol{C}^3$, where $\boldsymbol{C}$ is the adjacency matrix of the following graphs:

- a Wikipedia vote network[3] of size 7115 ($\operatorname{tr}(\boldsymbol{C}^3) = 3650334$);
- an arXiv collaboration network[4] of size 5242 ($\operatorname{tr}(\boldsymbol{C}^3) = 289560$).

Note that one matrix-vector product with $\boldsymbol{A}$ corresponds to three matrix-vector products with $\boldsymbol{C}$. The numerical results are shown in Figure 7.5.

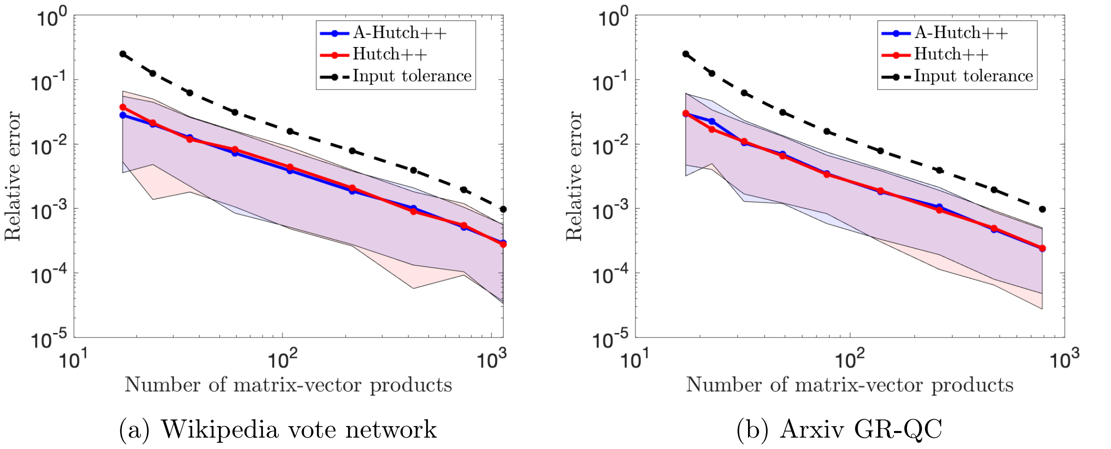
7.3.3 Estrada index

As discussed in Section 3.2, for an undirected graph with adjacency matrix $\boldsymbol{C}$, the Estrada index is defined as $\operatorname{tr}(\exp(\boldsymbol{C}))$ and its applications include measuring the degree of protein protein folding [54] and network analysis [55]. As in [111], we estimate the Estrada index of Roget's Thesaurus semantic graph adjacency matrix[5]. We approximate matrix-vector products with $\boldsymbol{A} = \exp(\boldsymbol{C})$ using 30 iterations of the Lanczos method [84, Chapter 13.2], after which the error from the approximated matrix-vector product is negligible. The results are shown in Figure 7.6.

[3] Accessed from https://snap.stanford.edu/data/wiki-Vote.html

[4] Accessed from https://snap.stanford.edu/data/ca-GrQc.html

[5] Accessed from http://vlado.fmf.uni-lj.si/pub/networks/data/



(a) Wikipedia vote network

(b) Arxiv GR-QC

Figure 7.5: Comparison of A-Hutch++ and Hutch++ for the triangle counting examples from Section 7.3.2.

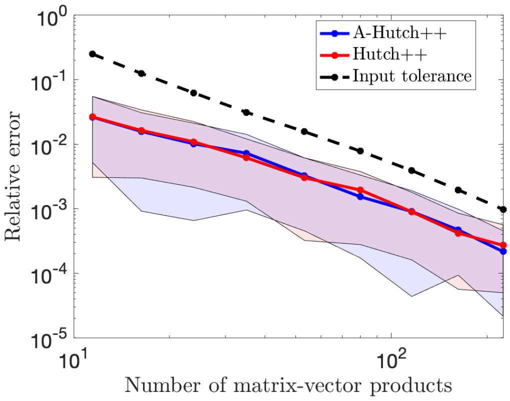
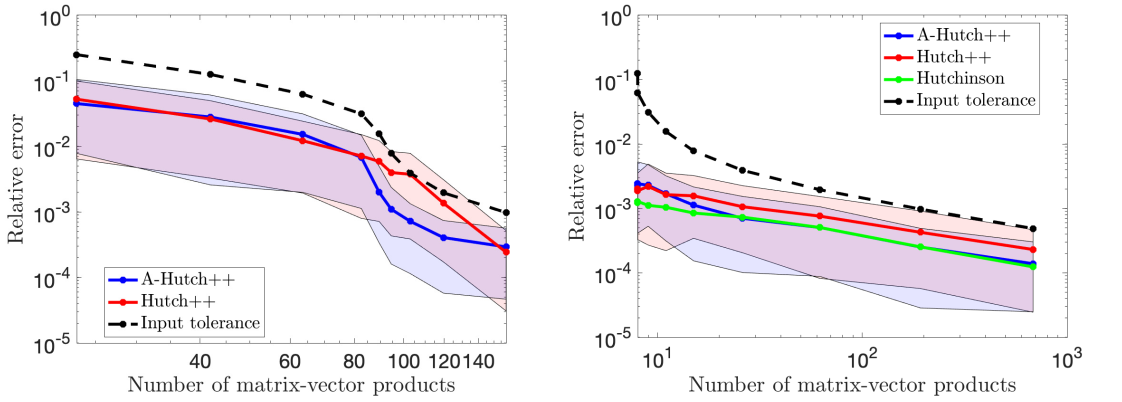


Figure 7.6: Comparison of A-Hutch++ and Hutch++ for the estimation of the Estrada index of the matrix from Section 7.3.3.



(a) Estimating the log-determinant of the matrix from [138].

(b) Estimating the log-determinant of the matrix Thermomech TC.

Figure 7.7: Comparison of A-Hutch++ and Hutch++ for the log determinant estimation of the matrices from Section 7.3.4.

7.3.4 Log-determinant

The computation of the log-determinant of a symmetric positive definite matrix, which arises for instance in statistical learning [2] and Markov random fields models [159], can be addressed by trace estimation exploiting the relation

$$\log\det(\boldsymbol{C}) = \operatorname{tr}(\log(\boldsymbol{C})).$$

In our setting we apply A-Hutch++ and Hutch++ to $\boldsymbol{A} = \log(\boldsymbol{C})$ for the following symmetric positive definite matrices $\boldsymbol{C}$:

- $\boldsymbol{C} = \boldsymbol{I} + \sum\limits_{j=1}^{40} \frac{10}{j^2}\boldsymbol{x}_j\boldsymbol{x}_j^T + \sum\limits_{j=41}^{300} \frac{1}{j^2}\boldsymbol{x}_j\boldsymbol{x}_j^T$ where $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{300} \in \mathbb{R}^{5000}$ are generated in Matlab using `sprandn(5000,1,0.025)`. This example comes from [138, 143]. $\boldsymbol{C}$ has an eigenvalue gap at index 40. Matrix-vector products with $\boldsymbol{A} = \log(\boldsymbol{C})$ are approximated using 25 iterations of Lanczos method.

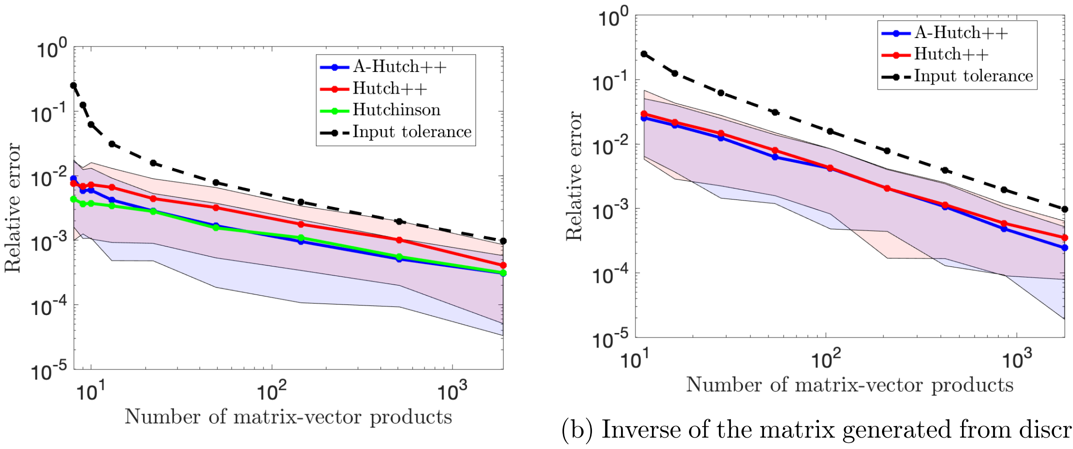
- $\boldsymbol{C}$ is the Thermomech TC matrix[6] from the SuiteSparse Matrix Collection [42]. Matrix-vector products with $\boldsymbol{A} = \log(\boldsymbol{C})$ are approximated using 35 iterations of Lanczos method.

The numerical results are shown in Figure 7.7.

7.3.5 Trace of inverses

We consider $\boldsymbol{A} = \boldsymbol{C}^{-1}$ for the following choices of $\boldsymbol{C}$:

[6] Accessed from https://sparse.tamu.edu/Botonakis/thermomech_TC



(a) Inverse of tridiag$(-1, 4, -1)$.

(b) Inverse of the matrix generated from discretizing Poisson's equation.

Figure 7.8: Comparison of A-Hutch++ and Hutch++ for the estimation of the trace of the inverse of the matrices described in Section 7.3.5.

- $\boldsymbol{C}$ = tridiag$(-1, 4, -1)$ is a 10000×10000 tridiagonal matrix with 4 along the diagonal and -1 along the upper and lower subdiagonal (taken from [59]);
- $\boldsymbol{C}$ a block tridiagonal matrix of size $k^2 \times k^2$ generated from discretizing Poisson's equation with the 5-point operator on a $k \times k$ mesh, with $k = 100$ (taken from [13]).

Matrix-vector products with $\boldsymbol{A} = \boldsymbol{C}^{-1}$ are computed using backslash in Matlab. The results are shown in Figure 7.8.

8 Nyström++: A single pass trace estimation algorithm

As explained in the Section 6.3.2, Hutch++ requires at least two passes over the matrix $\boldsymbol{A}$. In [111], Algorithm 12 was presented, and its analysis was improved in [92]. It requires only one pass over the input matrix, when computing the matrix vector products in line 3, and we thus call it *Single Pass Hutch++*. It also fits the streaming model because an update $\boldsymbol{A}+\boldsymbol{E}$ of the input matrix trivially translated into an update of the matrix-vector products, without having to revisit $\boldsymbol{A}$. It is similar to Hutch++ since it consists of a randomized low-rank approximation phase and a stochastic trace estimation phase. The low rank approximation phase is performed by computing the low rank approximation $\boldsymbol{A\Psi}(\boldsymbol{\Omega}^T\boldsymbol{A\Psi})^\dagger(\boldsymbol{A\Omega})^T = \boldsymbol{Y}(\boldsymbol{\Omega}^T\boldsymbol{Y})^\dagger\boldsymbol{X}^T$, where $\boldsymbol{X},\boldsymbol{Y}$ and $\boldsymbol{Z}$ are as in line 3 of Single Pass Hutch++. The trace of the low rank approximation equals $\operatorname{tr}((\boldsymbol{\Omega}^T\boldsymbol{Y})^\dagger(\boldsymbol{X}^T\boldsymbol{Y}))$ via the cyclic property of the trace. In the stochastic trace estimation phase the trace of $\boldsymbol{A}-\boldsymbol{Y}(\boldsymbol{\Omega}^T\boldsymbol{Y})^\dagger\boldsymbol{X}^T$ is estimated, which is done by the stochastic trace estimator (6.1). Single Pass Hutch++ satisfies similar guarantees as Hutch++, but is observed to produce a less accurate trace estimate than Hutch++ with the same number of matrix-vector products. More formally, the following result was proved.

```
Algorithm 12 Single Pass Hutch++
input: Symmetric positive semi-definite $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Number of matrix-vector products $m \in \mathbb{N}$.
output: An approximation to $\operatorname{tr}(\boldsymbol{A}) : \operatorname{tr}_m^{\mathsf{sph++}}(\boldsymbol{A})$
1: Fix positive constants $c_1, c_2$ and $c_3$ such that $c_1 < c_2$ and $c_1 + c_2 + c_3 = 1$.
2: Sample $\boldsymbol{\Omega} \in \mathbb{R}^{d\times c_1 m}, \boldsymbol{\Psi} \in \mathbb{R}^{d\times c_2 m}, \boldsymbol{\Phi} \in \mathbb{R}^{d\times c_3 m}$ with i.i.d. $\mathcal{N}(0,1)$ or Rademacher entries.
3: Compute $\begin{bmatrix}\boldsymbol{X} & \boldsymbol{Y} & \boldsymbol{Z}\end{bmatrix} = \boldsymbol{A}\begin{bmatrix}\boldsymbol{\Omega} & \boldsymbol{\Psi} & \boldsymbol{\Phi}\end{bmatrix}$.
4: return $\operatorname{tr}_m^{\mathsf{sph++}}(\boldsymbol{A}) = \operatorname{tr}((\boldsymbol{\Omega}^T\boldsymbol{Y})^\dagger(\boldsymbol{X}^T\boldsymbol{Y})) + \frac{1}{c_3 m}(\operatorname{tr}(\boldsymbol{\Phi}^T\boldsymbol{Z}) - \operatorname{tr}(\boldsymbol{\Phi}^T\boldsymbol{Y}(\boldsymbol{\Omega}^T\boldsymbol{Y})^\dagger\boldsymbol{X}^T\boldsymbol{\Phi}))$
```

Theorem 8.1 ([92, Theorem 1.1]). *If Single Pass Hutch++ is implemented with* $m =$

$O\left(\varepsilon^{-1}\sqrt{\log(\delta^{-1})} + \log(\delta^{-1})\right)$ *matrix-vector products then*

$$\left|\mathrm{tr}_m^{\mathsf{sph++}}(\boldsymbol{A}) - \mathrm{tr}(\boldsymbol{A})\right| \leq \varepsilon\,\mathrm{tr}(\boldsymbol{A})$$

holds with probability at least $1-\delta$*.*

On the other hand, the numerical experiments in [92] demonstrated that due to the single pass property, which allows for performing matrix-vector products in parallel, Single Pass Hutch++ outperforms Hutch++ in terms of wall-clock time.[1]

For SPSD $\boldsymbol{A}$ one can obtain a version of Hutch++ by utilizing the Nyström approximation $\boldsymbol{A\Omega}(\boldsymbol{\Omega}^T\boldsymbol{A\Omega})^\dagger\boldsymbol{\Omega}^T\boldsymbol{A}$, where $\boldsymbol{\Omega}$ is a random matrix; see (2.9).[2] We call this algorithm Nyström++, see Algorithm 13. The idea of using the Nyström approximation in the context of trace estimation had previously been presented in [102, Section 4] in a broader context, but no analysis was presented. A version of Hutch++ using a similar low-rank approximation was also mentioned in [112]. Furthermore, Nyström++ also fits the streaming model. Another possible advantage of Nyström++ over Hutch++ is that while the Nyström approximation is less accurate than the randomized SVD, one can spend more matrix-vector products for both attaining a low-rank approximation of $\boldsymbol{A}$ and on estimating the trace of $\boldsymbol{A} - \boldsymbol{A\Omega}(\boldsymbol{\Omega}^T\boldsymbol{A\Omega})^\dagger\boldsymbol{\Omega}^T\boldsymbol{A}$. Recall that the trace of the Nyström approximation $\boldsymbol{X}(\boldsymbol{\Omega}^T\boldsymbol{X})^\dagger\boldsymbol{X}^T$ equals $\mathrm{tr}\left((\boldsymbol{\Omega}^T\boldsymbol{X})^\dagger(\boldsymbol{X}^T\boldsymbol{X})\right)$ via the cyclic property of the trace.

Algorithm 13 Nyström++

```
input: Symmetric positive semi-definite $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Number of matrix-vector products $m \in \mathbb{N}$ (multiple of 2).
output: An approximation to $\mathrm{tr}(\boldsymbol{A}) : \mathrm{tr}_m^{\mathsf{n++}}(\boldsymbol{A})$.
Sample $\boldsymbol{\Omega} \in \mathbb{R}^{n\times\frac{m}{2}}, \boldsymbol{\Phi} \in \mathbb{R}^{n\times\frac{m}{2}}$ with i.i.d. $N(0,1)$ entries.
Compute $\begin{bmatrix}\boldsymbol{X} & \boldsymbol{Y}\end{bmatrix} = \boldsymbol{A}\begin{bmatrix}\boldsymbol{\Omega} & \boldsymbol{\Phi}\end{bmatrix}$.
return $\mathrm{tr}_m^{\mathsf{n++}}(\boldsymbol{A}) = \mathrm{tr}((\boldsymbol{\Omega}^T\boldsymbol{X})^\dagger(\boldsymbol{X}^T\boldsymbol{X})) + \frac{2}{m}(\mathrm{tr}(\boldsymbol{\Phi}^T\boldsymbol{Y}) - \mathrm{tr}(\boldsymbol{\Phi}^T\boldsymbol{X}(\boldsymbol{\Omega}^T\boldsymbol{X})^\dagger\boldsymbol{X}^T\boldsymbol{\Phi}))$
```

8.1 Analysis of Nyström++

In the following, we show that Algorithm 13 enjoys the same theoretical guarantees as Algorithm 9 [111, Theorem 1.1]. We begin with a result on the Frobenius norm error of the Nyström approximation.

[1] One needs to be careful how to implement the low-rank approximation in Single Pass Hutch++, since it is prone to numerical instabilities due to the pseudoinverse of $\boldsymbol{\Omega}^T\boldsymbol{Y}$. In our implementation we follow the suggestion given in [117, Section 5.1]. We compute a thin QR-decomposition of $(\boldsymbol{\Omega}^T\boldsymbol{Y})^T = \boldsymbol{QR}$ and let $\boldsymbol{S} = \boldsymbol{YQ}$ and $\boldsymbol{Z} = \boldsymbol{XR}^{-1}$. Then $\boldsymbol{Y}(\boldsymbol{\Omega}^T\boldsymbol{Y})^\dagger\boldsymbol{X}^T = \boldsymbol{SZ}^T$.

[2] Recall that the Nyström approximation depends only on the range of $\boldsymbol{\Omega}$; see Section 2.3. Therefore, the Nyström approximation remains unchanged if we replace the orthonormal basis $\boldsymbol{Q}$ with $\boldsymbol{\Omega}$, as long as $\mathrm{range}(\boldsymbol{\Omega}) = \mathrm{range}(\boldsymbol{Q})$.

Lemma 8.2. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be SPSD and let $\boldsymbol{\Omega}$ be a $n \times 2k$ random matrix whose entries are i.i.d. $\mathcal{N}(0,1)$ random variables with $k \geq 5$. Then*

$$\|\boldsymbol{A} - \boldsymbol{A\Omega}(\boldsymbol{\Omega}^T\boldsymbol{A\Omega})^\dagger\boldsymbol{\Omega}^T\boldsymbol{A}\|_{\mathrm{F}} \leq \frac{499}{\sqrt{k}}\,\mathrm{tr}(\boldsymbol{A})$$

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

Proof. Let $\boldsymbol{A}$ have eigenvalue decomposition partitioned as (2.6) and define $\boldsymbol{\Omega}_1 = \boldsymbol{U}_1^T\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_2 = \boldsymbol{U}_2^T\boldsymbol{\Omega}$. By Lemma 4.29 we have

$$\begin{aligned}\|\boldsymbol{A} - \boldsymbol{A\Omega}(\boldsymbol{\Omega}^T\boldsymbol{A\Omega})^\dagger\boldsymbol{\Omega}^T\boldsymbol{A}\|_{\mathrm{F}} \leq &\|\boldsymbol{\Lambda}_2\|_{\mathrm{F}} + \|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{(4)}^2 \leq \\ &\|\boldsymbol{\Lambda}_2\|_{\mathrm{F}} + \|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_2\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}.\end{aligned}$$

By proceeding as in the beginning of the proof of [68, Lemma 7] with probability at least $1 - 3e^{-k}$ we have[3]

$$\begin{aligned}\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_2 &\leq \|\boldsymbol{\Lambda}_2^{1/2}\|_2\left(\sqrt{\frac{3k}{k+1}}e + \frac{2e^2k}{k+1}\right) + \|\boldsymbol{\Lambda}_2^{1/2}\|_{\mathrm{F}}\frac{e^2\sqrt{2k}}{k+1} \\ &\leq \sqrt{\|\boldsymbol{\Lambda}_2\|_2}(\upsilon_1 + \upsilon_2) + \sqrt{\|\boldsymbol{\Lambda}_2\|_*}\frac{\upsilon_3}{\sqrt{k}} \end{aligned} \tag{8.1}$$

by letting $\upsilon_1 := \sqrt{3}e, \upsilon_2 := 2e^2$ and $\upsilon_3 := \sqrt{2}e^2$. Similarly, we have with probability at least $1 - 3e^{-k}$

$$\begin{aligned}\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}} &\leq \|\boldsymbol{\Lambda}_2^{1/2}\|_{\mathrm{F}}\sqrt{\frac{3k}{k+1}}e + \|\boldsymbol{\Lambda}_2^{1/2}\|_2\frac{2e^2k}{k+1} \\ &\leq \sqrt{\|\boldsymbol{\Lambda}_2\|_*}\upsilon_1 + \sqrt{\|\boldsymbol{\Lambda}_2\|_2}\upsilon_2. \end{aligned} \tag{8.2}$$

By the union bound, both (8.1) and (8.2) hold simultaneously with probability at least $1 - 6e^{-k}$.

$\boldsymbol{\Omega}_1 \in \mathbb{R}^{k\times 2k}$ is a standard Gaussian matrix and therefore has full row rank almost surely. We may therefore apply Lemma 4.29 combined with the bounds (8.1) and (8.2). Hence,

[3] In the setting of [68, Lemma 7], set the quantities $p = k, t = e, u = \sqrt{2k}$ and $\boldsymbol{D} = \boldsymbol{\Lambda}_2^{1/2}$ to obtain (8.1) and (8.2).

with probability at least $1 - 6e^{-k}$ we have

$$\begin{aligned}
&\|\boldsymbol{A} - \boldsymbol{A\Omega}(\boldsymbol{\Omega}^T\boldsymbol{A\Omega})^\dagger\boldsymbol{\Omega}^T\boldsymbol{A}\|_{\mathrm{F}} \\
\leq&\|\boldsymbol{\Lambda}_2\|_{\mathrm{F}} + \|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_2\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}} \\
\leq&\|\boldsymbol{\Lambda}_2\|_{\mathrm{F}} + \left(\sqrt{\|\boldsymbol{\Lambda}_2\|_2}(v_1 + v_2) + \sqrt{\|\boldsymbol{\Lambda}_2\|_*}\frac{v_3}{\sqrt{k}}\right)\left(\sqrt{\|\boldsymbol{\Lambda}_2\|_*}v_1 + \sqrt{\|\boldsymbol{\Lambda}_2\|_2}v_2\right) \\
=&\|\boldsymbol{\Lambda}_2\|_{\mathrm{F}} + \sqrt{\|\boldsymbol{\Lambda}_2\|_2\|\boldsymbol{\Lambda}_2\|_*}\tilde{v}_1 + \|\boldsymbol{\Lambda}_2\|_2\tilde{v}_2 + \|\boldsymbol{\Lambda}_2\|_*\frac{\tilde{v}_3}{\sqrt{k}} \\
\leq&\frac{1 + \tilde{v}_1 + \tilde{v}_2 + \tilde{v}_3}{\sqrt{k}}\operatorname{tr}(\boldsymbol{A})
\end{aligned}$$

where we set

$$\tilde{v}_1 := (v_1 + v_2)v_1 + \frac{v_2v_3}{\sqrt{k}},\ \tilde{v}_2 := (v_1 + v_2)v_2,\ \tilde{v}_3 := v_3v_1$$

and use the norm inequalities

$$\begin{aligned}
&\|\boldsymbol{\Lambda}_2\|_2 \leq \|\boldsymbol{\Lambda}_2\|_{\mathrm{F}} \leq \sqrt{\|\boldsymbol{\Lambda}_2\|_2\|\boldsymbol{\Lambda}_2\|_*} \leq \frac{1}{\sqrt{k}}\|\boldsymbol{\Lambda}\|_* = \frac{1}{\sqrt{k}}\operatorname{tr}(\boldsymbol{A}) \\
&\|\boldsymbol{\Lambda}_2\|_* \leq \|\boldsymbol{\Lambda}\|_* = \operatorname{tr}(\boldsymbol{A})
\end{aligned}$$

in the last step. The proof is completed by noting that $1 + \tilde{v}_1 + \tilde{v}_2 + \tilde{v}_3 \leq 499$.

□

We can now proceed to extend the main result on Hutch++ [111, Theorem 1.1] to Nyström++.

Theorem 8.3. *Suppose that Algorithm 13 (Nyström++) is executed with* $m = O\Big(\varepsilon^{-1}\sqrt{\log(\delta^{-1})} + \log(\delta^{-1})\Big)$ *matrix-vector products and* $\delta \in (0, 1/2)$[4]. *Then its output satisfies*

$$|\operatorname{tr}_m^{\mathsf{n++}}(\boldsymbol{A}) - \operatorname{tr}(\boldsymbol{A})| \leq \varepsilon\operatorname{tr}(\boldsymbol{A})$$

with probability at least $1 - \delta$.

Proof. We follow the proof of [111, Theorem 1.1]. Let us first recall that $\boldsymbol{X} = \boldsymbol{A\Omega}$, $\boldsymbol{Y} = \boldsymbol{A\Phi}$ for $d \times m/2$ standard Gaussian random matrices $\boldsymbol{\Omega}, \boldsymbol{\Phi}$ in Algorithm 13. Throughout the proof, we assume that $m \geq c\log(\delta^{-1})$ for some (sufficiently large) constant c.

[4] This condition on δ allows us to bound all $\log(p\delta^{-1})$ terms that would otherwise appear in the proof (see e.g. Lemma 7.1 where the term $\log(2\delta^{-1})$ appears) from above with $c\log(\delta^{-1})$ for some sufficiently large constant c.

By Lemma 8.2, there is a constant C_1 such that

$$\|\boldsymbol{A} - \boldsymbol{X}(\boldsymbol{\Omega}^T\boldsymbol{X})^\dagger\boldsymbol{X}^T\|_{\mathrm{F}} \le C_1 m^{-1/2}\operatorname{tr}(\boldsymbol{A}),$$

with probability at least $1-\delta/2$. By Lemma 7.1 there is a constant C_2 such that

$$\begin{aligned}&|\operatorname{tr}(\boldsymbol{A} - \boldsymbol{X}(\boldsymbol{\Omega}^T\boldsymbol{X})^\dagger\boldsymbol{X}^T) - \operatorname{tr}_{m/2}(\boldsymbol{A} - \boldsymbol{X}(\boldsymbol{\Omega}^T\boldsymbol{X})^\dagger\boldsymbol{X}^T)|\\ &\le C_2 m^{-1/2}\sqrt{\log(\delta^{-1})}\|\boldsymbol{A} - \boldsymbol{X}(\boldsymbol{\Omega}^T\boldsymbol{X})^\dagger\boldsymbol{X}^T\|_{\mathrm{F}}\end{aligned}$$

with probability at least $1-\delta/2$. By the union bound it holds with probability at least $1-\delta$ that

$$\begin{aligned}|\operatorname{tr}_m^{\mathrm{n}++}(\boldsymbol{A}) - \operatorname{tr}(\boldsymbol{A})| &= \left|\operatorname{tr}(\boldsymbol{A} - \boldsymbol{X}(\boldsymbol{\Omega}^T\boldsymbol{X})^\dagger\boldsymbol{X}^T) - \operatorname{tr}_{m/2}(\boldsymbol{A} - \boldsymbol{X}(\boldsymbol{\Omega}^T\boldsymbol{X})^\dagger\boldsymbol{X}^T)\right|\\ &\le C_2 m^{-1/2}\sqrt{\log(\delta^{-1})}\|\boldsymbol{A} - \boldsymbol{X}(\boldsymbol{\Omega}^T\boldsymbol{X})^\dagger\boldsymbol{X}^T\|_{\mathrm{F}}\\ &\le C_1C_2 m^{-1}\sqrt{\log(\delta^{-1})}\operatorname{tr}(\boldsymbol{A}).\end{aligned}$$

Hence, setting $m = O\Big(\varepsilon^{-1}\sqrt{\log(\delta^{-1})} + \log(\delta^{-1})\Big)$ implies the claim. □

8.1.1 Adaptive Nyström++

It is natural to aim at designing an adaptive version of Nyström++. Following A-Hutch++ we would need to find the minimum of

$$m(r) = r + C(\varepsilon,\delta)\|\boldsymbol{A} - \boldsymbol{A}_{\mathrm{n}}^{(r)}\|_{\mathrm{F}}^2,$$

where $\boldsymbol{A}_{\mathrm{n}}^{(r)}$ is the rank-r Nyström approximation. Such an adaptive version clearly does not fit the streaming model. Moreover, we lose another advantage of Nyström++, that it only needs to perform r matrix-vector products with $\boldsymbol{A}$ to get a rank-r approximation, compared to $2r$ for the randomized SVD. Since we cannot compute $\|\boldsymbol{A} - \boldsymbol{A}_{\mathrm{n}}^{(r)}\|_{\mathrm{F}}^2$ we would need to decompose this term as done in (7.5). This yields $\|\boldsymbol{A} - \boldsymbol{A}_{\mathrm{n}}^{(r)}\|_{\mathrm{F}}^2 = \|\boldsymbol{A}\|_{\mathrm{F}}^2 - 2\operatorname{tr}(\boldsymbol{A}\boldsymbol{A}_{\mathrm{n}}^{(r)}) + \|\boldsymbol{A}_{\mathrm{n}}^{(r)}\|_{\mathrm{F}}^2$. However, evaluating the term $-2\operatorname{tr}(\boldsymbol{A}\boldsymbol{A}_{\mathrm{n}}^{(r)}) + \|\boldsymbol{A}_{\mathrm{n}}^{(r)}\|_{\mathrm{F}}^2$ depending on r requires additional matrix-vector products with $\boldsymbol{A}$. In summary, there is little advantage of using such an adaptive version of Nyström.

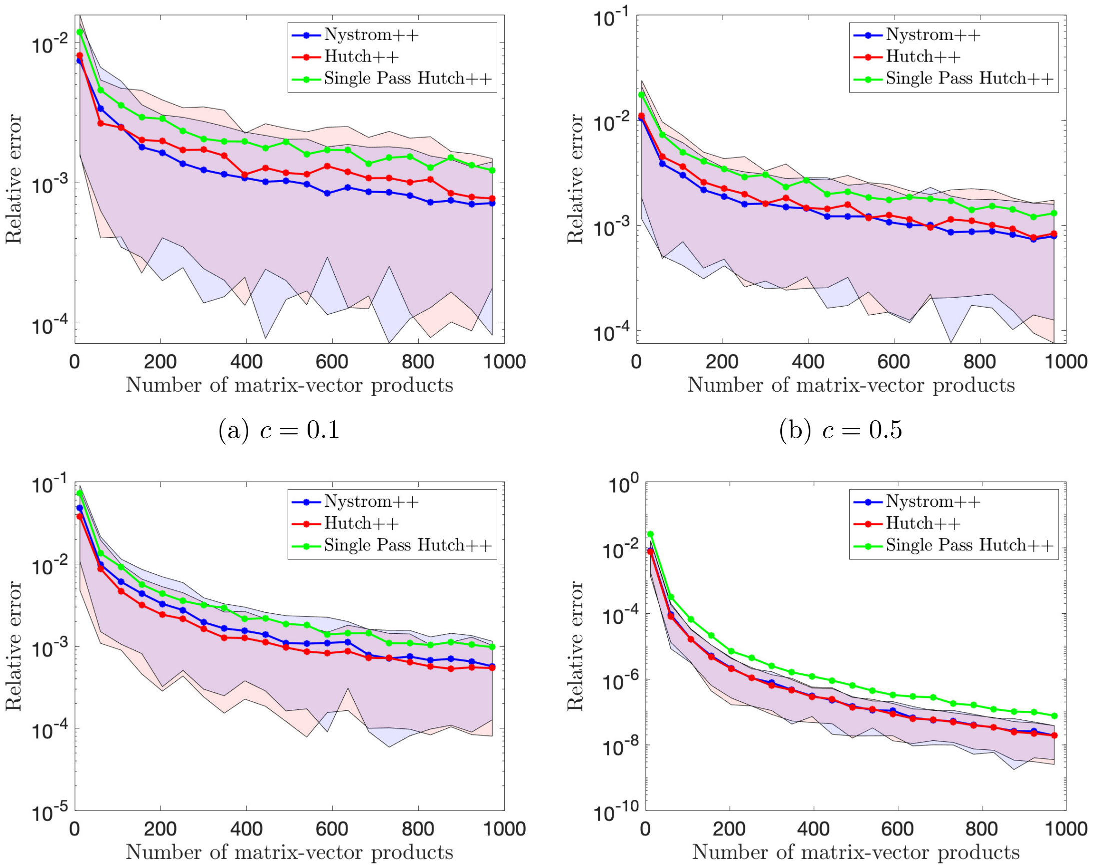
8.2 Numerical results

To deal with potential numerical instabilities due to the appearance of the pseudoinverse in the Nyström approximation in line 3 of Algorithm 13, in our implementation we use [104, Algorithm 16]. This algorithm computes an eigenvalue decomposition $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^T$ of the Nyström approximation of $\boldsymbol{A}+\nu\boldsymbol{I}$, where ν is a small shift, without explicitly forming the Nyström approximation. Once the eigenvalue decomposition is obtained the algorithm removes the shift by setting $\boldsymbol{\Lambda} = \max\{0, \boldsymbol{\Sigma} - \nu\boldsymbol{I}\}$ and returns $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, in factored form, as the stabilized Nyström approximation. The shift is set as $\nu = \sqrt{n}\texttt{eps}(\|\boldsymbol{A\Omega}\|_2)$, where $\texttt{eps}(x)$ returns the distance to the next larger double precision floating point number to $x \in \mathbb{R}$ and $\boldsymbol{\Omega}$ is as in Algorithm 13. For further details, we refer to [104, 151].

We compare Nyström++ with Hutch++ and Single Pass Hutch++. We consider $m = 12 + 48k$ for $k \in \{0, 1, 2, \dots, 20\}$ and for each value of m we run Hutch++, Single Pass Hutch++ and Nyström++ 100 times each. We run the experiments on the matrices from Section 7.3.1, Section 5.3.1, and Section 7.3.5. Moreover, we create two matrices with exponential decay, i.e. $\boldsymbol{A} = \boldsymbol{U\Lambda U}^T \in \mathbb{R}^{5000\times 5000}$ where $\boldsymbol{U}$ is a random orthogonal matrix and $\boldsymbol{\Lambda}$ is the diagonal matrix with entries $\boldsymbol{\Lambda}_{ii} = \gamma^i$ for $i = 1, \dots, 5000$, where γ is a parameter controlling the rate of the decay. We let $\gamma = \exp(-1/10)$ and $\gamma = \exp(-1/100)$.

The results are displayed in Figures 8.1, 8.2, 8.3, and 8.4, respectively. In each figure, the blue line is the average relative error from Nyström++, the red line is the average relative error from Hutch++ and the green line is the average relative error from Single Pass Hutch++. The shaded blue area shows the 10^{th} to 90^{th} percentiles of the results from Nyström++, and the shaded red area shows the 10^{th} to 90^{th} percentiles of the results from Hutch++.

In all cases we observe that Single Pass Hutch++ is the weakest alternative. Moreover, in many cases Hutch++ and Nyström++ have similar performances, and in some cases Nyström++ outperforms Hutch++, see e.g. Figure 8.4.



(a) $c = 0.1$

(b) $c = 0.5$

(c) $c = 1$

(d) $c = 3$

Figure 8.1: Comparison of Hutch++, Single Pass Hutch++ and Nyström++ for the estimation of the trace of the synthetic matrices with algebraic decay described in Section 7.3.1.

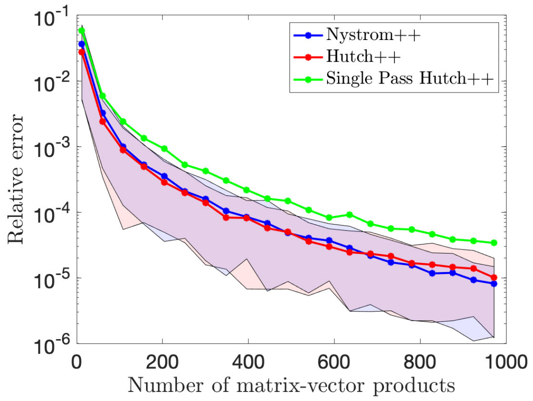
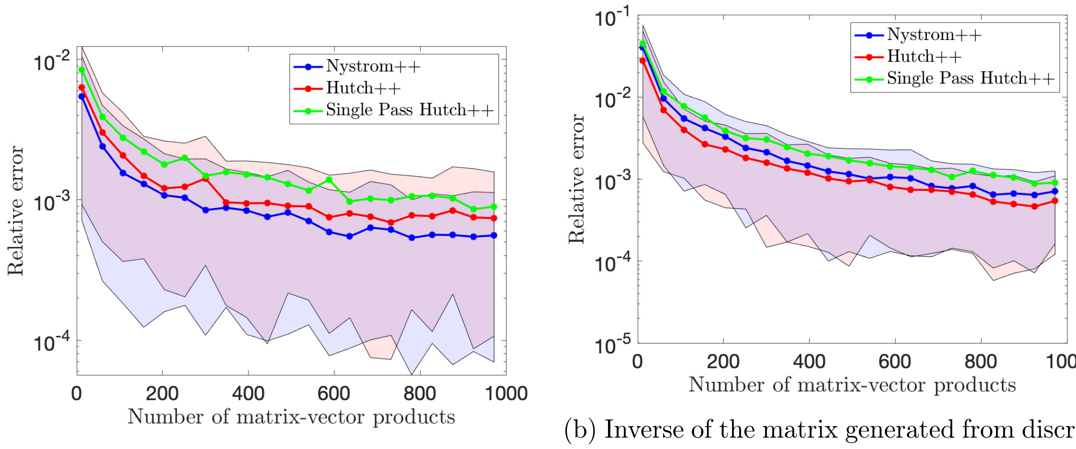


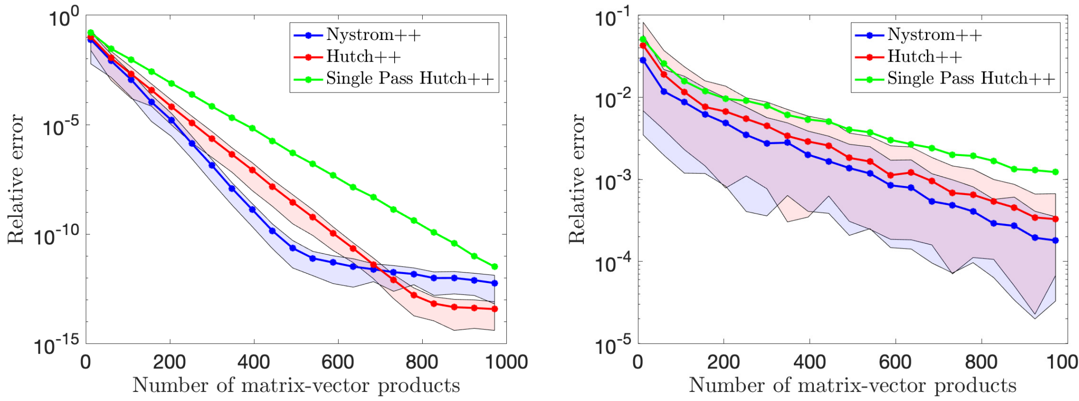
Figure 8.2: Comparison of Hutch++, Single Pass Hutch++ and Nyström++ for the estimation of the Estrada index as described in Section 5.3.1.



(a) Inverse of tridiag$(-1, 4, -1)$.

(b) Inverse of the matrix generated from discretizing Poisson's equation.

Figure 8.3: Comparison of Hutch++, Single Pass Hutch++ and Nyström++ for the estimation of the trace of the inverse of the matrices described in Section 7.3.5.



(a) $\gamma = \exp(-1/10)$

(b) $\gamma = \exp(-1/100)$

Figure 8.4: Comparison of Hutch++, Single Pass Hutch++ and Nyström++ for the estimation of the trace of the synthetic matrices with exponential decay described in Section 8.2.

9 Randomized Nyström approximation of non-negative self-adjoint trace class operators

This chapter is concerned with an infinite-dimensional generalization of the Nyström approximation defined in Section 2.3. In Section 9.1 we motivate this work and explain the difficulties of a generalization of the Nyström approximation to non-negative trace class operators. In Section 9.2 we provide an analysis of the Nyström approximation

$$\widehat{\boldsymbol{A}} := \boldsymbol{A\Omega}(\boldsymbol{\Omega}^T\boldsymbol{A\Omega})^{\dagger}\boldsymbol{\Omega}^T\boldsymbol{A} \approx \boldsymbol{A},$$

where the columns of $\boldsymbol{\Omega}$ are drawn independently from a non-standard Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{K})$ for a SPSD covariance matrix $\boldsymbol{K}$. Recall that this form of the Nyström approximation is equivalent to (2.9) provided $\operatorname{range}(\boldsymbol{Q}) = \operatorname{range}(\boldsymbol{\Omega})$. With the analysis of the finite-dimensional Nyström approximation at hand, we have the necessary tools to analyze an infinite-dimensional generalization of the Nyström approximation, which will be provided in Section 9.3. In Section 9.4 we present the numerical experiments.

This chapter is based on the work in [121].

9.1 Motivation

Recently, Boullé and Townsend [28, 29] generalized the randomized SVD from matrices to Hilbert–Schmidt operators. Subsequent work [27, 26] employed this infinite-dimensional generalization of the randomized SVD to learn Green's functions associated with an elliptic or parabolic partial differential equations (PDE) from a few solutions of the PDE. This approach uses hierarchical low-rank techniques and exploits the fact that Green's functions are smooth away from the diagonal and therefore admit accurate off-diagonal low-rank approximations [17, 18]. Other applications, like Gaussian process regression and Support Vector Machines [53, 65, 119, 157, 162, 163], involve integral operators that feature positive and *globally* smooth kernels. In turn, the operator is not only self-adjoint and positive but it also allows for directly applying low-rank approximation, without the need to resort to hierarchical techniques. Given existing results on matrices, it would be sensible to use an infinite-dimensional extension of the randomized Nyström approximation in such situations.

A difficulty in the design and analysis of our extension is that existing results of the finite-dimensional case (2.9) usually assume that the columns of $\boldsymbol{\Omega}$ are drawn from a *standard* Gaussian multivariate distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, which does not have a practically meaningful infinite-dimensional analog. Following [29], the columns of $\boldsymbol{\Omega}$ are replaced with random fields drawn from a Gaussian process $\mathcal{GP}(0, K)$, which is an infinite-dimensional analog of a *non-standard* multivariate Gaussian distribution. Therefore, to analyze the infinite-dimensional generalization of the Nyström approximation, we will proceed through an analysis of the finite-dimensional Nyström approximation (2.9) when the columns of $\boldsymbol{\Omega}$ are drawn from a non-standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{K})$, for some general symmetric positive semi-definite matrix $\boldsymbol{K}$. Then, we use continuity arguments to obtain bounds on the infinite-dimensional generalization of (2.9). As a byproduct of our analysis, we improve the analysis of the infinite-dimensional analog of the randomized SVD. In particular, unlike the bounds presented in [29], our improved bounds coincide with the bounds presented by Halko, Martinsson, and Tropp in [79] in the finite-dimensional case when $\boldsymbol{K}$ is chosen as the identity matrix.

9.2 The randomized Nyström approximation in finite dimensions with correlated Gaussian sketches

9.2.1 Distribution of Gaussian sketches

In this section we will outline some notation and basic background materian on Gaussian random vectors. We will consider a SPSD matrix $\boldsymbol{A}$ with eigenvalue, and equivalently SVD, partitioned as in (2.6). For a $n \times (k+p)$ sketching matrix $\boldsymbol{\Omega}$ we define

$$\boldsymbol{\Omega}_1 = \boldsymbol{U}_1^T \boldsymbol{\Omega}, \quad \boldsymbol{\Omega}_2 = \boldsymbol{U}_2^T \boldsymbol{\Omega}, \tag{9.1}$$

as done in (2.8). The columns of $\boldsymbol{\Omega}$ will be drawn from $\mathcal{N}(\mathbf{0}, \boldsymbol{K})$ for some SPSD covariance matrix $\boldsymbol{K}$. To analyze the distribution of (9.1) we partition the matrix $\widetilde{\boldsymbol{K}} = \boldsymbol{U}^* \boldsymbol{K} \boldsymbol{U}$ as

$$\widetilde{\boldsymbol{K}} = \begin{bmatrix} \boldsymbol{U}_1^* \boldsymbol{K} \boldsymbol{U}_1 & \boldsymbol{U}_1^* \boldsymbol{K} \boldsymbol{U}_2 \\ \boldsymbol{U}_2^* \boldsymbol{K} \boldsymbol{U}_1 & \boldsymbol{U}_2^* \boldsymbol{K} \boldsymbol{U}_2 \end{bmatrix} =: \begin{bmatrix} \widetilde{\boldsymbol{K}}_{11} & \widetilde{\boldsymbol{K}}_{21}^* \\ \widetilde{\boldsymbol{K}}_{21} & \widetilde{\boldsymbol{K}}_{22} \end{bmatrix}, \quad \widetilde{\boldsymbol{K}}_{11} \in \mathbb{R}^{k \times k}. \tag{9.2}$$

We assume that $\widetilde{\boldsymbol{K}}_{11}$ is invertible, which allows us to define the Schur complement $\widetilde{\boldsymbol{K}}_{22.1} = \widetilde{\boldsymbol{K}}_{22} - \widetilde{\boldsymbol{K}}_{21} \widetilde{\boldsymbol{K}}_{11}^{-1} \widetilde{\boldsymbol{K}}_{21}^*$. By well-known properties of Gaussian random vectors [81, Theorem 5.2], $\boldsymbol{\omega} \sim \mathcal{N}(0, \boldsymbol{K})$ implies $\boldsymbol{U}^* \boldsymbol{\omega} \sim \mathcal{N}(0, \widetilde{\boldsymbol{K}})$ and, thus, the marginal distributions

$$\boldsymbol{\omega}_1 := \boldsymbol{U}_1^* \boldsymbol{\omega} \sim \mathcal{N}(0, \widetilde{\boldsymbol{K}}_{11}), \quad \boldsymbol{\omega}_2 = \boldsymbol{U}_2^* \boldsymbol{\omega} \sim \mathcal{N}(0, \widetilde{\boldsymbol{K}}_{22}).$$

In particular, the columns of $\boldsymbol{\Omega}_1$ are i.i.d. as $N(\mathbf{0}, \widetilde{\boldsymbol{K}}_{11})$ or, in other words, $\boldsymbol{\Omega}_1 = \widetilde{\boldsymbol{K}}_{11}^{1/2} \boldsymbol{X}$ with the standard Gaussian matrix $\boldsymbol{X} \in \mathbb{R}^{k \times (k+p)}$. Because $\widetilde{\boldsymbol{K}}_{11}$ is invertible, this shows that, with probability one, $\boldsymbol{\Omega}_1$ has full rank and possesses a right inverse $\boldsymbol{\Omega}_1^\dagger$, which satisfies $\boldsymbol{\Omega}_1\boldsymbol{\Omega}_1^\dagger = \boldsymbol{I}$. Finally, the *conditional* probability distribution of $\boldsymbol{\omega}_2$ given $\boldsymbol{\omega}_1 = \boldsymbol{x}_1$ is normal with mean $\widetilde{\boldsymbol{K}}_{21}\widetilde{\boldsymbol{K}}_{11}^{-1}\boldsymbol{x}_1$ and covariance matrix $\widetilde{\boldsymbol{K}}_{22.1}$ [81, Theorem 5.3].

9.2.2 Nyström approximation with correlated Gaussian sketches

Similarly to the error analysis for the randomized SVD with correlated input vectors [28, 29], our error bounds for the Nyström approximation depend on prior information contained in the representation (9.2) of $\boldsymbol{K}$ with respect to the eigenvectors of $\boldsymbol{A}$. This is measured through the following two quantities:

$$\beta_k^{(\xi)} = \frac{\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}\boldsymbol{\Lambda}_2^{1/2}\|_\xi}{\|\boldsymbol{\Lambda}_2\|_\xi}\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2, \quad \delta_k^{(\xi)} = \frac{\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{21}\widetilde{\boldsymbol{K}}_{11}^{-1}\widetilde{\boldsymbol{K}}_{21}^*\boldsymbol{\Lambda}_2^{1/2}\|_\xi}{\|\boldsymbol{\Lambda}_2\|_\xi}\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2, \tag{9.3}$$

where $\xi \in \{\mathrm{F}, 2, *\}$ such that $\|\cdot\|_\xi$ dictates the choice of norm. The following theorem states our main result concerning the approximation error of the Nyström approximation in the Frobenius norm. Similar results for the spectral and nuclear norms are presented in Theorems 9.3 and 9.4.

Theorem 9.1 (Frobenius norm)**.** *Let $\boldsymbol{A}$ be an $n \times n$ SPSD matrix, $2 \le k \le \operatorname{rank}(\boldsymbol{A})$ be a target rank, and $p \ge 4$ an oversampling parameter. Let $\boldsymbol{\Omega}$ be a random sketch matrix whose columns are i.i.d. $\mathcal{N}(\boldsymbol{0}, \boldsymbol{K})$ random vectors, where the covariance matrix $\boldsymbol{K}$ is such that the matrix $\widetilde{\boldsymbol{K}}_{11}$ defined in* (9.2) *is invertible. Then, the Nyström approximation $\widehat{\boldsymbol{A}} = \boldsymbol{A\Omega}(\boldsymbol{\Omega}^T\boldsymbol{A\Omega})^\dagger\boldsymbol{\Omega}^T\boldsymbol{A}$ satisfies*

$$\mathbb{E}[\|\boldsymbol{A} - \widehat{\boldsymbol{A}}\|_\mathrm{F}] \le \left(1 + 2\delta_k^{(\mathrm{F})} + 2\sqrt{c_1}\beta_k^{(\mathrm{F})}\right)\|\boldsymbol{\Lambda}_2\|_\mathrm{F} + 2\sqrt{c_2}\beta_k^{(*)}\|\boldsymbol{\Lambda}_2\|_*, \tag{9.4}$$

where $c_1 = \mathcal{O}(k^2/p^2)$, $c_2 = \mathcal{O}(k^2/p^2)$ are constants defined in (9.8) *below. Let $u, t \ge 1$, then*

$$\begin{aligned}\|\boldsymbol{A} - \widehat{\boldsymbol{A}}\|_\mathrm{F} \le \|\boldsymbol{\Lambda}_2\|_\mathrm{F} &+ 4\left(\delta_k^{(\mathrm{F})} + t^2(d_1 + d_3)\beta_k^{(\mathrm{F})}\right)\|\boldsymbol{\Lambda}_2\|_\mathrm{F} + 4t^2 d_3\beta_k^{(*)}\|\boldsymbol{\Lambda}_2\|_* \\ &+ 2t^2u^2d_2\beta_k^{(2)}\|\boldsymbol{\Lambda}_2\|_2\end{aligned} \tag{9.5}$$

holds with probability at least $1 - 3t^{-p} - e^{-u^2/2}$. Here, $d_1 = \mathcal{O}(k/p)$, $d_2 = \mathcal{O}(k/p)$, and $d_3 = \mathcal{O}(k^{3/2}/p)$ are constants defined in (9.11).

The proof of Theorem 9.1 is based on an existing structural bound:

$$\|\boldsymbol{A} - \widehat{\boldsymbol{A}}\| \le \|\boldsymbol{\Lambda}_2\| + \|(\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger)^*\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|, \tag{9.6}$$

where $\boldsymbol{\Omega}_1^\dagger = \boldsymbol{\Omega}_1^*(\boldsymbol{\Omega}_1\boldsymbol{\Omega}_1^*)^{-1}$ is the right inverse of $\boldsymbol{\Omega}_1$, assuming that this matrix has full rank, and $\|\cdot\|$ denotes any unitarily invariant norm. This bound follows from Lemma 4.29 with $q = 1$ and the identity function $f : x \mapsto x$. Obtaining probabilistic bounds from (9.6) requires the analysis of the term $\|(\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger)^*\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|$.

Proof of (9.4). To prove inequality (9.4) of Theorem 9.1, we will make use of the L_q norm of a random variable Z defined as $\mathbb{E}^q(Z) = \mathbb{E}[|Z|^q]^{1/q}$. After taking expectation with respect to $\boldsymbol{\Omega}$ and applying Hölder's inequality, the second term in the bound (9.6) for the Frobenius norm satisfies

$$\mathbb{E}[\|(\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger)^*\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}] = \mathbb{E}[\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{(4)}^2] \leq \left(\mathbb{E}^4[\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{(4)}]\right)^2.$$

To proceed from here, we recall that Section 9.2.1 provides the conditional distribution of $\boldsymbol{\Omega}_2|\boldsymbol{\Omega}_1$ as $\boldsymbol{\Omega}_2|\boldsymbol{\Omega}_1 \sim \widetilde{\boldsymbol{K}}_{21}\widetilde{\boldsymbol{K}}_{11}^{-1}\boldsymbol{\Omega}_1 + \widetilde{\boldsymbol{K}}_{22.1}^{1/2}\boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is a standard Gaussian matrix. Using the triangle inequality for the L_4 norm and $\boldsymbol{\Omega}_1\boldsymbol{\Omega}_1^\dagger = \boldsymbol{I}$ (which holds with probability 1), we obtain that

$$\begin{aligned}\mathbb{E}^4[\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{(4)}] &= \mathbb{E}^4_{\boldsymbol{\Omega}_1,\boldsymbol{\Psi}}[\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{21}\widetilde{\boldsymbol{K}}_{11}^{-1}\boldsymbol{\Omega}_1\boldsymbol{\Omega}_1^\dagger + \boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\boldsymbol{\Psi}\boldsymbol{\Omega}_1^\dagger\|_{(4)}] \\ &\leq \|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{21}\widetilde{\boldsymbol{K}}_{11}^{-1}\|_{(4)} + \left(\mathbb{E}_{\boldsymbol{\Omega}_1}\left[\mathbb{E}_{\boldsymbol{\Psi}}\left[\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\boldsymbol{\Psi}\boldsymbol{\Omega}_1^\dagger\|_{(4)}^4 \mid \boldsymbol{\Omega}_1\right]\right]\right)^{1/4},\end{aligned} \tag{9.7}$$

To bound the second term, we first apply (A.1b),

$$\mathbb{E}_{\boldsymbol{\Psi}}\left[\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\boldsymbol{\Psi}\boldsymbol{\Omega}_1^\dagger\|_{(4)}^4 \mid \boldsymbol{\Omega}_1\right] = \|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\|_{(4)}^4\left(\|\boldsymbol{\Omega}_1^\dagger\|_{(4)}^4+\|\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}^4\right)+\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\|_{\mathrm{F}}^4\|\boldsymbol{\Omega}_1^\dagger\|_{(4)}^4,$$

and then take expectation with respect to $\boldsymbol{\Omega}_1$ using Lemma A.3:

$$\mathbb{E}_{\boldsymbol{\Omega}_1}\left[\mathbb{E}_{\boldsymbol{\Psi}}\left[\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\boldsymbol{\Psi}\boldsymbol{\Omega}_1^\dagger\|_{(4)}^4 \mid \boldsymbol{\Omega}_1\right]\right] \leq c_1\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2^2\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\|_{(4)}^4+c_2\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2^2\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\|_{\mathrm{F}}^4,$$

where

$$c_1 = k\frac{(p-1)(k+1)+2}{p(p-1)(p-3)}, \quad \text{and} \quad c_2 = k\frac{k+p-1}{p(p-1)(p-3)}. \tag{9.8}$$

Inserting the inequality above into (9.7) and using $\|\widetilde{\boldsymbol{K}}_{11}^{-1/2}\|_2 = \|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2^{1/2}$ gives

$$\begin{aligned}&\mathbb{E}^4[\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{(4)}] \leq \\ &\quad\left(\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{21}\widetilde{\boldsymbol{K}}_{11}^{-1/2}\|_{(4)} + \left(c_1\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\|_{(4)}^4 + c_2\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\|_{\mathrm{F}}^4\right)^{1/4}\right)\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2^{1/2},\end{aligned}$$

Finally, inserting the covariance quality factors (see (9.3)) leads to the following inequality,

$$\left(\mathbb{E}^4[\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{(4)}]\right)^2 \leq \left(2\delta_k^{(\mathrm{F})} + 2\sqrt{c_1}\beta_k^{(\mathrm{F})}\right)\|\boldsymbol{\Lambda}_2\|_{\mathrm{F}} + 2\sqrt{c_2}\beta_k^{(*)}\|\boldsymbol{\Lambda}_2\|_*,$$

where we used $(a+b)^2 \leq 2a^2 + 2b^2$ and the subadditivity of the square-root. This concludes the proof of (9.4). □

We now proceed with the proof of the tailbound (9.5) of Theorem 9.1. We begin with a concentration inequality on norms of shifted and rescaled Gaussian matrices.

Lemma 9.2. *Let* $\boldsymbol{\Psi}$ *be a standard Gaussian matrix and let* $\boldsymbol{B}, \boldsymbol{C}$*,* $\boldsymbol{D}$ *be fixed matrices of* *matching sizes. Let $s \geq 2$. Then*

$$\mathbb{P}\left\{\|\boldsymbol{B}+\boldsymbol{C}\boldsymbol{\Psi}\boldsymbol{D}\|_{(s)} \geq \mathbb{E}\left[\|\boldsymbol{B}+\boldsymbol{C}\boldsymbol{\Psi}\boldsymbol{D}\|_{(s)}\right] + \|\boldsymbol{C}\|_2\|\boldsymbol{D}\|_2 u\right\} \leq e^{-u^2/2}.$$

holds for every $u \geq 1$.

Proof. Given $h(\boldsymbol{X}) := \|\boldsymbol{B}+\boldsymbol{C}\boldsymbol{X}\boldsymbol{D}\|_{(s)}$, we have

$$|h(\boldsymbol{X})-h(\boldsymbol{Y})| \leq \|\boldsymbol{C}(\boldsymbol{X}-\boldsymbol{Y})\boldsymbol{D}\|_{(s)} \leq \|\boldsymbol{C}\|_2\|\boldsymbol{D}\|_2\|\boldsymbol{X}-\boldsymbol{Y}\|_{(s)} \leq \|\boldsymbol{C}\|_2\|\boldsymbol{D}\|_2\|\boldsymbol{X}-\boldsymbol{Y}\|_{\mathrm{F}},$$

where we used that the Frobenius norm is larger than any Schatten-s norm for $s \geq 2$. In other words, h is Lipschitz continuous with Lipschitz constant $\|\boldsymbol{C}\|_2\|\boldsymbol{D}\|_2$. This allows us to apply a concentration result for functions of Gaussian matrices [79, Proposition 10.3], which yields the statement of the lemma. □

Proof of (9.5). To obtain (9.5), it suffices to derive a tailbound for the term $\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{(4)}^2$ in the structural bound (9.6) (in the Frobenius norm). Using that $\boldsymbol{\Omega}_2|\boldsymbol{\Omega}_1 \sim \widetilde{\boldsymbol{K}}_{21}\widetilde{\boldsymbol{K}}_{11}^{-1}\boldsymbol{\Omega}_1 + \widetilde{\boldsymbol{K}}_{22.1}^{1/2}\boldsymbol{\Psi}$ with a standard Gaussian matrix $\boldsymbol{\Psi}$, Lemma 9.2 yields the following tailbound for $\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{(4)}^2$ conditioned on $\boldsymbol{\Omega}_1$:

$$\mathbb{P}\Big\{\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{(4)} \geq \mathbb{E}\Big[\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{(4)} \mid \boldsymbol{\Omega}_1\Big] + \|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\|_2\|\boldsymbol{\Omega}_1^\dagger\|_2 u \mid \boldsymbol{\Omega}_1\Big\} \leq e^{-u^2/2}. \tag{9.9}$$

In analogy to (9.7), combining the triangular inequality for the Schatten-4 norm and Jensen's inequality yields

$$\begin{aligned}\mathbb{E}\Big[\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{(4)} \mid \boldsymbol{\Omega}_1\Big] &\leq \|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{21}\widetilde{\boldsymbol{K}}_{11}^{-1}\|_{(4)} + \mathbb{E}_{\boldsymbol{\Psi}}^{4}\Big[\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\boldsymbol{\Psi}\boldsymbol{\Omega}_1^\dagger\|_{(4)} \mid \boldsymbol{\Omega}_1\Big] \\ \leq \|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{21}\widetilde{\boldsymbol{K}}_{11}^{-1}\|_{(4)} &+ \Big(\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\|_{(4)}^4\Big(\|\boldsymbol{\Omega}_1^\dagger\|_{(4)}^4 + \|\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}^4\Big) + \|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\|_{\mathrm{F}}^4\|\boldsymbol{\Omega}_1^\dagger\|_{(4)}^4\Big)^{1/4}.\end{aligned}$$

Using, again, the inequality $(a+b)^2 \leq 2(a^2+b^2)$ and $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$ and the subadditivity of the square-root, we have

$$\begin{aligned}\mathbb{E}\Big[\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{(4)} \mid \boldsymbol{\Omega}_1\Big]^2 \leq 2\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{21}\widetilde{\boldsymbol{K}}_{11}^{-1}\|_{(4)}^2 + 2\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\|_{(4)}^2(\|\boldsymbol{\Omega}_1^\dagger\|_{(4)}^2 + \|\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}^2) \\ + 2\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\|_{\mathrm{F}}^2\|\boldsymbol{\Omega}_1^\dagger\|_{(4)}^2.\end{aligned} \tag{9.10}$$

To control the norms of $\boldsymbol{\Omega}_1^\dagger$ in our bounds, we condition on the event E_t that the following three inequalities are satisfied:

$$\|\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}^2 \leq d_1\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2 t^2, \quad \|\boldsymbol{\Omega}_1^\dagger\|_2^2 \leq d_2\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2 t^2, \quad \|\boldsymbol{\Omega}_1^\dagger\|_{(4)}^2 \leq d_3\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2 t^2,$$

with the constants

$$d_1 = \frac{3k}{p+1}, \quad d_2 = e^2 \frac{k+p}{(p+1)^2}, \quad d_3 = \sqrt{k} d_2. \tag{9.11}$$

Lemma A.4 implies the following bound for $\mathbb{P}(E_t^c)$:

$$\mathbb{P}(E_t^c) \leq 2t^{-(p+1)} + t^{-p} \leq 3t^{-p}. \tag{9.12}$$

Under the event E_t, (9.10) becomes

$$\mathbb{E}\Big[\|\mathbf{\Lambda}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_{(4)} \mid E_t\Big]^2 \leq 2\left(\delta_k^{(\mathrm{F})} + t^2(d_1+d_3)\beta_k^{(\mathrm{F})}\right)\|\mathbf{\Lambda}_2\|_{\mathrm{F}} + 2t^2 d_3 \beta_k^{(*)}\|\mathbf{\Lambda}_2\|_*. \tag{9.13}$$

Conditioning (9.9) on E_t and combining it with (9.13) yields

$$\|\mathbf{\Lambda}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_{(4)}^2 \leq 4\left(\delta_k^{(\mathrm{F})} + t^2(d_1+d_3)\beta_k^{(\mathrm{F})}\right)\|\mathbf{\Lambda}_2\|_{\mathrm{F}} + 4t^2 d_3\beta_k^{(*)}\|\mathbf{\Lambda}_2\|_* + 2d_2\beta_k^{(2)}\|\mathbf{\Lambda}_2\|_2 u^2 t^2,$$

with probability $\geq 1 - e^{-u^2/2}$. Similarly to the proof of [79, Theorem 10.8], by the union bound we remove the conditioning by a union bound using (9.12) and conclude the proof of (9.1). □

The structural bound (9.6) for the Nyström method applies to any unitarily invariant matrix norm. Using properties of Gaussian matrices from Appendix A, this allows us to extend the analysis performed in the proof of Theorem 9.1 to the spectral and nuclear norms.

Theorem 9.3 (Expectation bound in spectral and nuclear norms)**.** *Consider the setting of Theorem 9.1 with an oversampling parameter $p \geq 2$, it holds that*

$$\mathbb{E}[\|\boldsymbol{A} - \widehat{\boldsymbol{A}}\|_2] \leq \left(1 + \frac{3k}{p-1}\beta_k^{(2)} + 3\delta_k^{(2)}\right)\|\mathbf{\Lambda}_2\|_2 + \frac{3e^2(k+p)}{p^2-1}\beta_k^{(*)}\|\mathbf{\Lambda}_2\|_*, \tag{9.14a}$$

$$\mathbb{E}[\|\boldsymbol{A} - \widehat{\boldsymbol{A}}\|_*] \leq \left(1 + \frac{k}{p-1}\beta_k^{(*)} + \delta_k^{(*)}\right)\|\mathbf{\Lambda}_2\|_*. \tag{9.14b}$$

Proof. We start by proving (9.14a) and use the structural bound (9.6) for the Nyström method to obtain

$$\mathbb{E}[\|\boldsymbol{A} - \widehat{\boldsymbol{A}}\|_2] \leq \|\mathbf{\Lambda}_2\|_2 + \mathbb{E}[\|\mathbf{\Lambda}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_2^2] = \|\mathbf{\Lambda}_2\|_2 + \mathbb{E}^2[\|\mathbf{\Lambda}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_2]^2.$$

Similarly to the proof of Theorem 9.1, we use the conditional distribution of $\mathbf{\Omega}_2|\mathbf{\Omega}_1$ and the triangle inequality for the L_2 norm twice, to get

$$\begin{aligned}\mathbb{E}^2\Big[\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\|_2\Big] &= \mathbb{E}^2_{\boldsymbol{\Omega}_1}\Big[\mathbb{E}^2_{\boldsymbol{\Psi}}\Big[\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{21}\widetilde{\boldsymbol{K}}_{11}^{-1}\boldsymbol{\Omega}_1\boldsymbol{\Omega}_1^{\dagger} + \boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\boldsymbol{\Psi}\boldsymbol{\Omega}_1^{\dagger}\|_2 \mid \boldsymbol{\Omega}_1\Big]\Big] \\ &\leq \|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{21}\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2 + \Big(\mathbb{E}_{\boldsymbol{\Omega}_1}\Big[\mathbb{E}_{\boldsymbol{\Psi}}\Big[\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\boldsymbol{\Psi}\boldsymbol{\Omega}_1^{\dagger}\|_2^2 \mid \boldsymbol{\Omega}_1\Big]\Big]\Big)^{1/2},\end{aligned}$$

where $\boldsymbol{\Psi}$ is an $(n-k)\times(k+p)$ standard Gaussian matrix. Then (A.1c) leads to

$$\mathbb{E}_{\boldsymbol{\Psi}}\Big[\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\boldsymbol{\Psi}\boldsymbol{\Omega}_1^{\dagger}\|_2^2 \mid \boldsymbol{\Omega}_1\Big] \leq \Big(\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\|_{\mathrm{F}}\|\boldsymbol{\Omega}_1^{\dagger}\|_2 + \|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\|_2\|\boldsymbol{\Omega}_1^{\dagger}\|_{\mathrm{F}}\Big)^2.$$

After taking expectation with respect to $\boldsymbol{\Omega}_1$ and using the triangle inequality for the L_2 norm, we have

$$\begin{aligned}\Big(\mathbb{E}_{\boldsymbol{\Omega}_1}\Big[\mathbb{E}_{\boldsymbol{\Psi}}\Big[\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\boldsymbol{\Psi}\boldsymbol{\Omega}_1^{\dagger}\|_2^2 \mid \boldsymbol{\Omega}_1\Big]\Big]\Big)^{1/2} &\leq \|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\|_{\mathrm{F}}(\mathbb{E}_{\boldsymbol{\Omega}_1}\Big[\|\boldsymbol{\Omega}_1^{\dagger}\|_2^2\Big])^{1/2} \\ &\quad + \|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\|_2(\mathbb{E}_{\boldsymbol{\Omega}_1}\Big[\|\boldsymbol{\Omega}_1^{\dagger}\|_{\mathrm{F}}^2\Big])^{1/2}.\end{aligned}$$

We then apply Lemma A.3 to obtain the following inequality

$$\begin{aligned}\mathbb{E}^2[\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\|_2] &\leq \frac{\|\widetilde{\boldsymbol{K}}_{11}^{-1/2}\|_{\mathrm{F}}}{\sqrt{p-1}}\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\|_2 + \|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{21}\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2 \\ &\quad + \frac{e\sqrt{k+p}}{\sqrt{p^2-1}}\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2^{1/2}\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\|_{\mathrm{F}} \\ &\leq \sqrt{\frac{k}{p-1}\beta_k^{(2)}\|\boldsymbol{\Lambda}_2\|_2} + \sqrt{\delta_k^{(2)}\|\boldsymbol{\Lambda}_2\|_2} + \sqrt{\frac{e^2(k+p)}{p^2-1}\beta_k^{(*)}\|\boldsymbol{\Lambda}_2\|_*}.\end{aligned}$$

We conclude the proof of (9.14a) using the inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$.

The bound for the nuclear norm follows through a similar argument from the structural bound (9.6) in the nuclear norm:

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{A}-\widehat{\boldsymbol{A}}\|_*] &\leq \|\boldsymbol{\Lambda}_2\|_* + \mathbb{E}[\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\|_{\mathrm{F}}^2] \\ &\leq \|\boldsymbol{\Lambda}_2\|_* + \|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{21}\widetilde{\boldsymbol{K}}_{11}^{-1}\|_{\mathrm{F}}^2 + \|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\|_{\mathrm{F}}^2\frac{\mathrm{tr}(\widetilde{\boldsymbol{K}}_{11}^{-1})}{p-1} \\ &\leq \left(1+\frac{k}{p-1}\beta_k^{(*)} + \delta_k^{(*)}\right)\|\boldsymbol{\Lambda}_2\|_*.\end{aligned}$$

□

Theorem 9.4 (Tailbound in spectral and nuclear norms). *Using the notation of theo-*

rem 9.1 with an oversampling parameter $p \geq 4$, *and* $u, t \geq 1$, *it holds that*

$$\|\boldsymbol{A} - \widehat{\boldsymbol{A}}\|_2 \leq \left(1 + 4\delta_k^{(2)} + 4(d_1 + d_2 u^2)t^2 \beta_k^{(2)}\right) \|\boldsymbol{\Lambda}_2\|_2 + 4d_2 t^2 \beta_k^{(*)} \|\boldsymbol{\Lambda}_2\|_*, \tag{9.15a}$$

$$\|\boldsymbol{A} - \widehat{\boldsymbol{A}}\|_* \leq \left(1 + 2\delta_k^{(*)} + d_1 t^2 \beta_k^{(*)}\right) \|\boldsymbol{\Lambda}_2\|_* + 2d_2 t^2 u^2 \beta_k^{(2)} \|\boldsymbol{\Lambda}_2\|_2, \tag{9.15b}$$

where each inequality holds with probability $\geq 1 - 2t^{-p} - e^{u^2/2}$.

Proof. We begin by deriving the tailbound (9.15a) in the spectral norm. Similar to the proof of Equation (9.5), we process the second term of the structural bound (9.6) with the concentration inequality of Lemma 9.2 in the spectral norm:

$$\mathbb{P}\left\{\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_2 \geq \mathbb{E}\Big[\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_2 \mid \boldsymbol{\Omega}_1\Big] + \|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\|_2 \|\boldsymbol{\Omega}_1^\dagger\|_2 u \mid \boldsymbol{\Omega}_1\right\} \leq e^{-u^2/2}. \tag{9.16}$$

Using Lemma A.2, it holds that

$$\mathbb{E}\Big[\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_2 \mid \boldsymbol{\Omega}_1\Big] \leq \|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{21}\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2 + \|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\|_{\mathrm{F}} \|\boldsymbol{\Omega}_1^\dagger\|_2 + \|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}^{1/2}\|_2 \|\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}.$$

We then consider the probability event E_t that the Frobenius and spectral norms of $\boldsymbol{\Omega}_1^\dagger$ are well controlled:

$$E_t = \left\{\|\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}} \leq \sqrt{d_1 \|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2} t, \quad \|\boldsymbol{\Omega}_1^\dagger\|_2 \leq \sqrt{d_2 \|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2} t\right\},$$

where $\mathbb{P}(E_t^c) \leq 2t^{-p}$ by Lemma A.4. We conclude the proof of (9.15a) by conditioning (9.16) on E_t and using the inequality $\|\boldsymbol{x}\|_1 \leq \sqrt{d}\|\boldsymbol{x}\|_2$ for $\boldsymbol{x} \in \mathbb{R}^d$ to obtain:

$$\|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_2^2 \leq \left(4\delta_k^{(2)} + 4(d_1 + d_2 u^2)t^2 \beta_k^{(2)}\right) \|\boldsymbol{\Lambda}_2\|_2 + 4d_2 t^2 \beta_k^{(*)} \|\boldsymbol{\Lambda}_2\|_*,$$

with probability $\geq 1 - 2t^{-p} - e^{-u^2/2}$.

The proof of the tail bound for the nuclear norm (9.15b) follows from a similar argument. □

Remark 9.1 (Connection with the randomized SVD). *Consider an arbitrary matrix* $\boldsymbol{B} \in \mathbb{R}^{m \times n}$ *with singular value decomposition*

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{W}_1 & \boldsymbol{W}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{S}_1 & \\ & \boldsymbol{S}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_1^* \\ \boldsymbol{U}_2^* \end{bmatrix},$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{m \times k}, \boldsymbol{U}_1 \in \mathbb{R}^{n \times k}$, *and* $\boldsymbol{S}_1 \in \mathbb{R}^{k \times k}$ *is a diagonal matrix containing the* k *largest singular values of* $\boldsymbol{B}$. *Letting* $\boldsymbol{Q}$ *be an orthonormal basis for* $\mathrm{range}(\boldsymbol{B}\boldsymbol{\Omega})$, *then* $\boldsymbol{Q}\boldsymbol{Q}^*\boldsymbol{B}$ *is the approximation attained by the (basic) randomized SVD. Setting* $\boldsymbol{A} = \boldsymbol{B}^*\boldsymbol{B}$

and $\widehat{\boldsymbol{A}} = \boldsymbol{A\Omega}(\boldsymbol{\Omega}^*\boldsymbol{A\Omega})^\dagger\boldsymbol{\Omega}^*\boldsymbol{A}$ *we have for any* $s \geq 1$ *that*

$$\|\boldsymbol{B} - \boldsymbol{QQ}^*\boldsymbol{B}\|_{(2s)}^2 = \|\boldsymbol{A} - \widehat{\boldsymbol{A}}\|_{(s)}.$$

Hence, obtaining a bound for the randomized SVD applied to $\boldsymbol{B}$ *is equivalent to obtaining a bound for the Nyström approximation on* $\boldsymbol{A}$*; similar observations have been made in [68, 150]. Therefore, one can apply the results obtained earlier in this section to derive error bounds for the randomized SVD. As an example, Theorem 9.3 implies that*

$$\mathbb{E}\|\boldsymbol{B} - \boldsymbol{QQ}^*\boldsymbol{B}\|_{\mathrm{F}}^2 \leq \left(1 + \frac{k}{p-1}\tilde{\beta}_k + \tilde{\delta}_k\right)\|\boldsymbol{S}_2\|_{\mathrm{F}}^2,$$

where $\tilde{\beta}_k = \mathrm{tr}(\boldsymbol{S}_2\widetilde{\boldsymbol{K}}_{22.1}\boldsymbol{S}_2)\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2/\|\boldsymbol{S}_2\|_{\mathrm{F}}^2$ *and* $\tilde{\delta}_k = \mathrm{tr}(\boldsymbol{S}_2\widetilde{\boldsymbol{K}}_{21}\widetilde{\boldsymbol{K}}_{11}^{-1}\widetilde{\boldsymbol{K}}_{21}^*\boldsymbol{S}_2)\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2/\|\boldsymbol{S}_2\|_{\mathrm{F}}^2$. *This bound coincides with the standard randomized SVD bound [79, Theorem 10.5] when* $\boldsymbol{K} = \boldsymbol{I}$*, unlike the bound proved in [28, Theorem 2].*

9.3 The randomized Nyström approximation in infinite dimensions

This section presents an infinite-dimensional extension of the randomized Nyström approximation. We begin by briefly introducing the concepts of quasimatrices, Hilbert–Schmidt operators, and Gaussian processes, which will be useful to generalize of the bounds of Section 9.2 to operators between function spaces.

9.3.1 Quasimatrices

For a bounded domain $D \subset \mathbb{R}^d$, $d \geq 1$, we consider the Hilbert space $L^2(D)$ of square-integrable functions. Quasimatrices are a convenient way to represent and work with collections of functions or more, generally, elements of infinite-dimensional vector spaces; see, e.g., [148]. In particular, a function $Y : \mathbb{R}^m \to L^2(D)$ is expressed as the quasimatrix

$$Y = \begin{bmatrix} y_1 & \cdots & y_m \end{bmatrix}, \quad y_i \in L^2(D).$$

Similar to matrices, compositions of linear operators can be conveniently expressed by extending the usual matrix multiplication rules to quasimatrices. The adjoint $Y^* : L^2(D) \to \mathbb{R}^m$ can also be viewed as a quasimatrix with the m rows $\langle y_1, \cdot\rangle, \cdots, \langle y_m, \cdot\rangle : L^2(D) \to \mathbb{R}$, where $\langle\cdot,\cdot\rangle$ denotes the standard inner product in $L^2(D)$. If $Z : \mathbb{R}^\ell \to L^2(D)$ is another quasimatrix, then Y^*Z yields the following $m \times \ell$ matrix:

$$Y^*Z = \begin{bmatrix} \langle y_1, z_1\rangle & \cdots & \langle y_1, z_\ell\rangle \\ \vdots & \ddots & \vdots \\ \langle y_m, z_1\rangle & \cdots & \langle y_m, z_\ell\rangle \end{bmatrix}.$$

9.3.2 Non-negative self-adjoint trace-class operators

We consider a non-negative self-adjoint trace-class operator $\mathcal{A} : L^2(D) \to L^2(D)$, i.e., it holds that $\langle \mathcal{A}f, f\rangle \geq 0$ and $\langle \mathcal{A}f, g\rangle = \langle f, \mathcal{A}g\rangle$ for every $f, g \in L^2(D)$, and the trace norm [89, Definition 4.5.1] is finite:

$$\|\mathcal{A}\|_{\mathrm{Tr}} := \sum_{j=1}^{\infty} \langle \mathcal{A}e_j, e_j\rangle < \infty,$$

for any orthonormal basis $\{e_j\}_j$ of $L^2(D)$. Non-negative self-adjoint trace-class operators are Hilbert–Schmidt operators [89, Theorem 4.5.2], and therefore admit an eigenvalue decomposition of the form [89, Theorem 4.3.1]:

$$\mathcal{A} = \sum_{\substack{j=1 \\ \lambda_j > 0}}^{\infty} \lambda_j \langle u_j, \cdot\rangle u_j, \tag{9.17}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are the eigenvalues of $\mathcal{A}$, and $\{u_j\}_j$ are orthonormal eigenfunctions. The eigenvalues allow us to express the trace, Hilbert–Schmidt, and operator norms of $\mathcal{A}$ as

$$\|\mathcal{A}\|_{\mathrm{Tr}} = \sum_{j=1}^{\infty} \lambda_j, \quad \|\mathcal{A}\|_{\mathrm{HS}} = \left(\sum_{j=1}^{\infty} \lambda_j^2\right)^{1/2}, \quad \|\mathcal{A}\|_{\mathrm{op}} = \lambda_1,$$

which are infinite-dimensional analogs of the nuclear, Frobenius, and spectral norms discussed in Chapter 2.

Furthermore, we introduce the operator $U : \ell^2 \to L^2(D)$ defined by $Uf = \sum_{i=1}^{\infty} f_i u_i$ for any f in ℓ^2, the space of square-summable sequences (indexed by positive integers). Then, for a given rank $k \geq 1$, the quasimatrix $U_1 : \mathbb{R}^k \to L^2(D)$ contains the first k eigenfunctions of $\mathcal{A}$, and the quasimatrix $U_2 : \ell^2 \to L^2(D)$ contains the remaining eigenfunctions. Finally, we introduce the diagonal matrix and quasimatrices $\mathbf{\Lambda}_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ and $\mathbf{\Lambda}_2 = \mathrm{diag}(\lambda_{k+1}, \lambda_{k+2}, \ldots)$, respectively, which contain the eigenvalues of $\mathcal{A}$ in descending order.

9.3.3 Gaussian processes

In this section we will give a brief outline of random elements in Hilbert spaces and Gaussian processes.

Let ω be a measurable function from a probability space $(S, \mathcal{S}, \mathbb{P})$ to $(\mathbb{H}, \mathcal{B}(\mathbb{H}))$, where $\mathbb{H}$ is a separable Hilbert space with inner product $\langle \cdot, \cdot\rangle$. If $\mathbb{E}\|\omega\| < \infty$ then we can define the mean of ω as the Bochner integral

$$m = \mathbb{E}\omega = \int_S \omega d\mathbb{P};$$

see [89, Definition 7.2.1]. An alternative way of defining the mean is to let m be the unique representer in $\mathbb{H}$ that has the property $\langle m, g\rangle = \mathbb{E}\langle \omega, g\rangle$; see [89, Section 7.2]. We will assume throughout this thesis that $m = 0$.

Assuming $\mathbb{E}\|\omega\|^2 < \infty$, the covariance operator is the self-adjoint non-negative trace class operator $\mathcal{K} : \mathbb{H} \to \mathbb{H}$ defined as the Bochner integral

$$\mathcal{K} = \mathbb{E}\omega \otimes \omega = \int_S \omega \otimes \omega d\mathbb{P},$$

where $\omega \otimes \omega$ denotes the operator given by $f \mapsto \langle \omega, f\rangle\omega$ for any $f \in \mathbb{H}$. The existence of $\mathcal{K}$ is guaranteed by the assumption $\mathbb{E}\|\omega\|^2 < \infty$ and the separability of the Hilbert space consisting of Hilbert Schmidt operators from $\mathbb{H}$ to $\mathbb{H}$. The trace norm of $\mathcal{K}$ is given by $\|\mathcal{K}\|_{\mathrm{Tr}} = \mathbb{E}\|\omega\|^2 < \infty$. Since $\mathcal{K}$ is a self-adjoint operator it has a spectral decomposition given by $\mathcal{K} = \sum_{i=1}^{\infty} \theta_i \psi_i \otimes \psi_i$, where $\{\psi_i\}_i$ is a complete orthonormal system so that $\overline{\mathrm{Im}(\mathcal{K})} = \overline{\mathrm{span}\{\psi_i\}_i}$. Furthermore, we have $\omega \in \overline{\mathrm{span}\{\psi_i\}_i}$ almost surely; see [89, Theorem 7.2.5, Theorem 7.2.6].

In this thesis we will be concerned with random elements in the Hilbert space $L^2(D)$, for some bounded set $D \subset \mathbb{R}^d$, equipped with the standard inner product. In particular, we are interested in Gaussian processes, which are continuous analogues non-standard Gaussian distributions. Formally speaking, a Gaussian process is a stochastic process $\omega = \{\omega(x) : x \in D\}$ defined on some probability space $(S, \mathcal{S}, \mathbb{P})$. We assume that the process is jointly measurably on $\mathcal{B}(D) \times \mathcal{S}$. We say that the process has mean zero if $\mathbb{E}\omega(x) = 0$ for all $x \in D$. The covariance between two instances of ω, $\omega(x)$ and $\omega(y)$, is described by the symmetric positive semi-definite covariance function

$$K(x, y) = \mathrm{Cov}(\omega(x), \omega(y)),$$

provided $\mathbb{E}\omega(x)^2$ exists for all $x \in D$ [89, Theorem 7.3.1]. The process is said to be a Gaussian process if for any finite collection of points $x_1, \ldots, x_m \in D$, the vector $\begin{bmatrix}\omega(x_1) & \cdots & \omega(x_m)\end{bmatrix}^\top$ follows a multivariate normal distribution with mean $\mathbf{0} \in \mathbb{R}^m$ and covariance matrix $\boldsymbol{K} = (K(x_i, x_j))_{1\leq i,j\leq m}$. In this case we write $\omega \sim \mathcal{GP}(0, K)$.

If we further assume that $K : D \times D \mapsto \mathbb{R}$ is a continuous kernel by Mercer's theorem it admits the following eigenvalue decomposition [89, Theorem 4.6.5]:

$$K(x, y) = \sum_{i=1}^{\infty} \theta_i \psi_i(x)\psi_i(y), \quad \int_D K(x, y)\psi_i(y)dy = \theta_i\psi_i(x), \quad x, y \in D.$$

where the sum converges absolutely and uniformly on D [108]. Here, $\theta_1 \geq \theta_2 \geq \ldots \geq 0$ are the eigenvalues of the integral operator $\mathcal{K}$ induced by K:

$$\mathcal{K}[f](x) = \int_D K(x,y)f(y)dy, \quad f \in L^2(D),\, x \in D,$$

and $\{\psi_j\}_j$ are the corresponding orthonormal eigenfunctions of $\mathcal{K}$ in $L^2(D)$. In the following, we assume that $\mathcal{K}$ is trace-class, i.e., $\sum_{i=1}^{\infty} \theta_i < \infty$. Mean-zero processes with such covariance kernels are mean-square continuous [89, Theorem 7.3.2]. For such processes, the different definitions of the mean function and covariance operator given in this section are equivalent [89, Theorem 7.4.3].

For any two functions $f, g \in L^2(D)$ we have that $\langle \omega, f\rangle \sim \mathcal{N}(0, \langle f, \mathcal{K}f\rangle), \langle \omega, g\rangle \sim \mathcal{N}(0, \langle g, \mathcal{K}g\rangle)$, and $\mathrm{Cov}(\langle \omega, f\rangle, \langle \omega, g\rangle) = \langle g, \mathcal{K}f\rangle$ [89, Theorem 7.3.3 and Theorem 7.4.3]. Furthermore, by the Karhunen-Loève theorem we have that $\omega = \sum\limits_{i=1}^{\infty} \langle \omega, \psi_i\rangle\psi_i$, where the series converges uniformly in D in the mean-square sense; see [89, Theorem 7.3.5 and Theorem 7.4.3]. The Karhunen–Loève expansion of ω is given by

$$\omega = \theta_1^{1/2}\zeta_1\psi_1 + \theta_2^{1/2}\zeta_2\psi_2 + \cdots, \tag{9.18}$$

where $\zeta_1, \zeta_2, \ldots \sim \mathcal{N}(0,1)$ are mutually independent; see [1, Chapter 3.2]. With probability one, a realization of ω is in $L^2(D)$ [89, Theorem 7.2.5].

Recalling that $\{u_j\}_j$ denote the eigenfunctions of $\mathcal{A}$, see (9.17), we define the function $\widetilde{\boldsymbol{K}} : \mathbb{N}^* \times \mathbb{N}^* \to \mathbb{R}$ elementwise as

$$\widetilde{\boldsymbol{K}}(i,j) = \langle u_i, \mathcal{K}u_j\rangle = \sum_{k=1}^{\infty} \theta_k \langle \psi_k, u_i\rangle\langle \psi_k, u_j\rangle, \quad i,j \in \mathbb{N}^*,$$

which is bounded by $\|\mathcal{K}\|_{\mathrm{Tr}}$ using the Cauchy–Schwarz inequality. We also define the following restrictions of $\widetilde{\boldsymbol{K}}$ to different sets of indices:

$$\widetilde{\boldsymbol{K}}_{11} = \widetilde{\boldsymbol{K}}_{[\![1,k]\!]\times[\![1,k]\!]}, \quad \widetilde{\boldsymbol{K}}_{21} = \widetilde{\boldsymbol{K}}_{[\![k+1,\infty)\times[\![1,k]\!]}, \quad \widetilde{\boldsymbol{K}}_{22} = \widetilde{\boldsymbol{K}}_{[\![k+1,\infty)\times[\![k+1,\infty)}.$$

Then, $\widetilde{\boldsymbol{K}}_{11}$ defines a $k \times k$ matrix, which we assume to be of rank k in the rest of this section. In terms of the quasimatrices U_1, U_2 containing the eigenfunctions defined as above, we can write

$$\widetilde{\boldsymbol{K}}_{11} = U_1^*\mathcal{K}U_1, \quad \widetilde{\boldsymbol{K}}_{21} = U_2^*\mathcal{K}U_1, \quad \widetilde{\boldsymbol{K}}_{22} = U_2^*\mathcal{K}U_2, \tag{9.19}$$

in analogy to (9.2).

9.3.4 Infinite-dimensional extension of the Nyström approximation

We are now ready to present the infinite-dimensional extension of the Nyström approximation. Let k be a target rank, p be an oversampling parameter, and $\Omega = \begin{bmatrix}\omega_1 & \cdots & \omega_{k+p}\end{bmatrix}$ be a random quasimatrix with $k+p$ columns, whose columns are i.i.d. from $\mathcal{GP}(0,K)$. The Nyström approximation $\widehat{\mathcal{A}}$ to $\mathcal{A}$ is defined as

$$\widehat{\mathcal{A}} := \mathcal{A}\Omega(\Omega^*\mathcal{A}\Omega)^\dagger(\mathcal{A}\Omega)^*. \tag{9.20}$$

Assuming that the realization of ω_i is in $L^2(D)$, which holds with probability 1, the Nyström approximation is an operator $\widehat{\mathcal{A}} : L^2(D) \to L^2(D)$ of rank at most $k+p$ with the explicit representation

$$\widehat{\mathcal{A}}[f] = \sum_{i,j=1}^{k+p} \mathcal{A}\omega_i \left[(\Omega^*\mathcal{A}\Omega)^\dagger\right]_{ij} \langle \omega_j, \mathcal{A}f\rangle$$

As in the finite-dimensional case, the error bounds for the infinite-dimensional analog of the Nyström approximation depend on the prior information of eigenvectors contained in $\mathcal{K}$, which is measured by the following two quantities:

$$\beta_k^{(\xi)} = \frac{\|\mathbf{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}\mathbf{\Lambda}_2^{1/2}\|_\xi}{\|\mathbf{\Lambda}_2\|_\xi}\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2, \quad \delta_k^{(\xi)} = \frac{\|\mathbf{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{21}\widetilde{\boldsymbol{K}}_{11}^{-1}\widetilde{\boldsymbol{K}}_{21}^*\mathbf{\Lambda}_2^{1/2}\|_\xi}{\|\mathbf{\Lambda}_2\|_\xi}\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2, \tag{9.21}$$

where $\xi \in \{\mathrm{HS}, \mathrm{Tr}, \mathrm{op}\}$. These quantities are the infinite-dimensional analogs of (9.3).

For each $1 \le j \le k+p$, we consider the stochastic process $\boldsymbol{\omega}_j = \{\langle u_i, \omega_j\rangle,\ i \in \mathbb{N}^*\}$ whose trajectories are in ℓ^2, and denote by $\mathbf{\Omega} = U^*\Omega = \begin{bmatrix}\boldsymbol{\omega}_1 & \cdots & \boldsymbol{\omega}_{k+p}\end{bmatrix}$ the random quasimatrix whose columns are i.i.d. from $\mathcal{GP}(0,\widetilde{\boldsymbol{K}})$. Then, we introduce the random $k \times (k+p)$ matrix $\mathbf{\Omega}_1$ as

$$\mathbf{\Omega}_1 := U_1^*\Omega = \begin{bmatrix} \langle u_1, \omega_1\rangle & \cdots & \langle u_1, \omega_{k+p}\rangle \\ \vdots & \ddots & \vdots \\ \langle u_k, \omega_1\rangle & \cdots & \langle u_k, \omega_{k+p}\rangle \end{bmatrix}, \quad \omega_j \sim \mathcal{GP}(0,K), \quad 1 \le j \le k+p,$$

whose columns are i.i.d. from $\mathcal{N}(0,\widetilde{\boldsymbol{K}}_{11})$ [29, Lemma 1]. Since we assume that $\mathrm{rank}(\widetilde{\boldsymbol{K}}_{11}) = k$ we know that $\mathbf{\Omega}_1$ has full row-rank with probability one. Therefore, $\mathbf{\Omega}_1$ has almost surely a right inverse $\mathbf{\Omega}_1^\dagger = \mathbf{\Omega}_1^*(\mathbf{\Omega}_1\mathbf{\Omega}_1^*)^{-1}$. We define $\mathbf{\Omega}_2 = U_2^*\Omega$ similarly. Note that the i^{th} entry of the columns of $\mathbf{\Omega}_2$ are distributed as $\mathcal{N}(0,(\widetilde{\boldsymbol{K}}_{22})_{ii})$ and the covariance between the i^{th} and j^{th} entry is $(\widetilde{\boldsymbol{K}}_{22})_{ij}$ [29, Section 3.3].

9.3.5 Structural bound

In this section, we prove an infinite-dimensional analog of the structural bound (9.6). We begin by stating some basic but useful results on norms of finite sections of ℓ^2 operators.

Lemma 9.5. *Let $\boldsymbol{B}, \boldsymbol{C} : \ell^2 \to \ell^2$ be non-negative self-adjoint trace-class operators such that $\boldsymbol{B} - \boldsymbol{C}$ is non-negative. For $n \in \mathbb{N}^*$, consider the restriction $\boldsymbol{B}_{[\![1,n]\!] \times [\![1,n]\!]}$ of $\boldsymbol{B}$ to its first n rows and columns. Then, for $\xi \in \{\mathrm{Tr}, \mathrm{HS}, \mathrm{op}\}$, we have*

$$\lim_{n\to\infty} \|\boldsymbol{B}_{[\![1,n]\!]\times[\![1,n]\!]}\|_\xi = \|\boldsymbol{B}\|_\xi, \tag{9.22a}$$

$$\|\boldsymbol{C}\|_\xi \le \|\boldsymbol{B}\|_\xi \le \|\boldsymbol{B}_{[\![1,n]\!]\times[\![1,n]\!]}\|_\xi + \|\boldsymbol{B}_{[\![n+1,\infty)\times[\![n+1,\infty)}\|_\xi. \tag{9.22b}$$

For $\xi \in \{\mathrm{HS}, \mathrm{Tr}\}$, the property (9.22a) follows from the absolute convergence of the involved series. For $\xi = \mathrm{op}$, the triangle inequality yields

$$|\|\boldsymbol{B}\|_{\mathrm{op}} - \|\boldsymbol{B}_{[\![1,n]\!]\times[\![1,n]\!]}\|_{\mathrm{op}}| \le \sqrt{\|\boldsymbol{B}\|_{\mathrm{HS}}^2 - \|\boldsymbol{B}_{[\![1,n]\!]\times[\![1,n]\!]}\|_{\mathrm{HS}}^2} \to 0, \quad \text{as } n \to \infty.$$

(9.22b) is an infinite-dimensional analog of Lemmas 4.27 and 4.28, which can be proven using (9.22a). We can now proceed with an infinite-dimensional analog of the finite-dimensional structural bound (9.6).

Lemma 9.6 (Infinite-dimensional structural bound)**.** *Let $\mathcal{A} : L^2(D) \to L^2(D)$ be a self-adjoint non-negative trace-class operator, $k, p \ge 1$, and $\Omega : \mathbb{R}^{k+p} \to L^2(D)$ be a quasimatrix with $k+p$ columns such that the matrix $\boldsymbol{\Omega}_1 = U_1^*\Omega \in \mathbb{R}^{k\times(k+p)}$ is full rank. Then,*

$$\|\mathcal{A} - \mathcal{A}\Omega(\Omega^*\mathcal{A}\Omega)^\dagger(\mathcal{A}\Omega)^*\|_\xi \le \|\boldsymbol{\Lambda}_2\|_\xi + \|(\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger)^*\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_\xi,$$

where $\xi \in \{\mathrm{op}, \mathrm{HS}, \mathrm{Tr}\}$.

Proof. Let $\boldsymbol{\Omega} = U^*\Omega$ be defined as in Section 9.3.4 and $\mathcal{P}_U = UU^* : L^2(D) \to L^2(D)$ denote the orthogonal projection onto the range of $\mathcal{A}$. Since $\mathcal{A}$ is a self-adjoint operator, we have $\mathcal{A} = \mathcal{P}_U\mathcal{A} = \mathcal{A}\mathcal{P}_U = \mathcal{P}_U\mathcal{A}\mathcal{P}_U$. Therefore,

$$\begin{aligned}\|\mathcal{A} - \mathcal{A}\Omega(\Omega^*\mathcal{A}\Omega)^\dagger(\mathcal{A}\Omega)^*\|_\xi &= \|\mathcal{P}_U\mathcal{A}\mathcal{P}_U - \mathcal{P}_U\mathcal{A}\mathcal{P}_U\Omega(\Omega^*\mathcal{P}_U\mathcal{A}\mathcal{P}_U\Omega)^\dagger(\mathcal{P}_U\mathcal{A}\mathcal{P}_U\Omega)^*\|_\xi \\ &= \|U(U^*\mathcal{A}U - U^*\mathcal{A}U\boldsymbol{\Omega}(\boldsymbol{\Omega}^*U^*\mathcal{A}U\boldsymbol{\Omega})^\dagger(U^*\mathcal{A}U\boldsymbol{\Omega})^*)U^*\|_\xi \\ &= \|U^*\mathcal{A}U - U^*\mathcal{A}U\boldsymbol{\Omega}(\boldsymbol{\Omega}^*U^*\mathcal{A}U\boldsymbol{\Omega})^\dagger(U^*\mathcal{A}U\boldsymbol{\Omega})^*\|_\xi = \|\boldsymbol{\Lambda} - \boldsymbol{\Lambda}\boldsymbol{\Omega}(\boldsymbol{\Omega}^*\boldsymbol{\Lambda}\boldsymbol{\Omega})^\dagger(\boldsymbol{\Lambda}\boldsymbol{\Omega})^*\|_\xi,\end{aligned}$$

where the third equality follows from the unitary invariance of the norms and the fourth equality is due to the relation $U^*\mathcal{A}U = \boldsymbol{\Lambda}$. As in the finite-dimensional case, the rest of the proof follows the argument of the proof of Lemma 4.29 for the operator monotone function $f : x \mapsto x$ and $q = 1$, now using the fact that the inequalities in (9.22b) are infinite-dimensional analogs of Lemmas 4.27 and 4.28 with $f : x \mapsto x$ and $\|\cdot\| := \|\cdot\|_\xi$. □

9.3.6 Probabilistic bounds

With the structural bound in place, we proceed to derive probabilistic bounds for the infinite-dimensional Nyström approximation (9.20).

Theorem 9.7 (Infinite-dimensional Nyström approximation). *Let $\mathcal{A} : L^2(D) \to L^2(D)$ be a non-negative self-adjoint trace-class operator, $2 \leq k \leq \operatorname{rank}(\mathcal{A})$ be a target rank, and $p \geq 4$ be an oversampling parameter. Let Ω be a quasimatrix with $k+p$ columns i.i.d. from $\mathcal{GP}(0, K)$, with a kernel K such that the matrix $\widetilde{\boldsymbol{K}}_{11}$ defined in* (9.19) *is invertible. Then, the Nyström approximation $\widehat{\mathcal{A}} : L^2(D) \to L^2(D)$ defined in* (9.20) *satisfies*

$$\mathbb{E}[\|\mathcal{A} - \widehat{\mathcal{A}}\|_{\mathrm{HS}}] \leq \left(1 + 2\delta_k^{(\mathrm{HS})} + 2\sqrt{c_1}\beta_k^{(\mathrm{HS})}\right) \|\boldsymbol{\Lambda}_2\|_{\mathrm{HS}} + 2\sqrt{c_2}\beta_k^{(\mathrm{Tr})}\|\boldsymbol{\Lambda}_2\|_{\mathrm{Tr}}, \tag{9.23a}$$

$$\mathbb{E}[\|\mathcal{A} - \widehat{\mathcal{A}}\|_{\mathrm{op}}] \leq \left(1 + \frac{3k}{p-1}\beta_k^{(\mathrm{op})} + 3\delta_k^{(\mathrm{op})}\right) \|\boldsymbol{\Lambda}_2\|_{\mathrm{op}} + \frac{3e^2(k+p)}{p^2-1}\beta_k^{(\mathrm{Tr})}\|\boldsymbol{\Lambda}_2\|_{\mathrm{Tr}}, \tag{9.23b}$$

$$\mathbb{E}[\|\mathcal{A} - \widehat{\mathcal{A}}\|_{\mathrm{Tr}}] \leq \left(1 + \frac{k}{p-1}\beta_k^{(\mathrm{Tr})} + \delta_k^{(\mathrm{Tr})}\right) \|\boldsymbol{\Lambda}_2\|_{\mathrm{Tr}}, \tag{9.23c}$$

where $c_1 = \mathcal{O}(k^2/p^2)$, $c_2 = \mathcal{O}(k^2/p^2)$ are the constants defined in (9.8)*. Let $u, t \geq 1$, then*

$$\begin{aligned} \|\mathcal{A} - \widehat{\mathcal{A}}\|_{\mathrm{HS}} \leq \|\boldsymbol{\Lambda}_2\|_{\mathrm{HS}} + 4\left(\delta_k^{(\mathrm{HS})} + t^2(d_1+d_3)\beta_k^{(\mathrm{HS})}\right) \|\boldsymbol{\Lambda}_2\|_{\mathrm{HS}} + 4t^2 d_3 \beta_k^{(\mathrm{Tr})}\|\boldsymbol{\Lambda}_2\|_{\mathrm{Tr}} \\ + 2t^2u^2 d_2 \beta_k^{(2)}\|\boldsymbol{\Lambda}_2\|_{\mathrm{op}}, \end{aligned} \tag{9.24a}$$

$$\|\mathcal{A} - \widehat{\mathcal{A}}\|_{\mathrm{op}} \leq \left(1 + 4\delta_k^{(\mathrm{op})} + 4(d_1 + d_2u^2)t^2\beta_k^{(\mathrm{op})}\right) \|\boldsymbol{\Lambda}_2\|_{\mathrm{op}} + 4d_2t^2\beta_k^{(\mathrm{Tr})}\|\boldsymbol{\Lambda}_2\|_{\mathrm{Tr}}, \tag{9.24b}$$

$$\|\mathcal{A} - \widehat{\mathcal{A}}\|_{\mathrm{Tr}} \leq \left(1 + 2\delta_k^{(\mathrm{Tr})} + d_1t^2\beta_k^{(\mathrm{Tr})}\right) \|\boldsymbol{\Lambda}_2\|_{\mathrm{Tr}} + 2d_2t^2u^2\beta_k^{(\mathrm{op})}\|\boldsymbol{\Lambda}_2\|_{\mathrm{op}}, \tag{9.24c}$$

with probability $\geq 1 - 3t^{-p} - e^{-u^2/2}$. Here, $d_1 = \mathcal{O}(k/p)$, $d_2 = \mathcal{O}(k/p)$, and $d_3 = \mathcal{O}(k^{3/2}/p)$ are the constants defined in (9.11)*.*

The proof of Theorem 9.7 occupies the rest of this section and follows from a continuity argument on the generalization of the Nyström approximation to correlated Gaussian vectors analyzed earlier in Section 9.2. Let $n \geq 1$ and $s \in [1, \infty]$, we first define the following random variables:

$$X_{s,n} = \|(\boldsymbol{\Lambda}_2^{1/2})_{[\![1,n]\!]\times[\![1,n]\!]}(\boldsymbol{\Omega}_2)_{[\![1,n]\!]\times[\![1,k+p]\!]}\boldsymbol{\Omega}_1^\dagger\|_{(2s)}, \quad X_s = \|\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{(2s)}, \tag{9.25}$$

where $\|\cdot\|_{(2s)}$ denotes the Schatten-$2s$ norm. We aim to show the convergence of $X_{s,n}$ to X_s as $n \to \infty$, and begin with a preliminary result on the finiteness of the expectation of X_s.

Lemma 9.8 (Expectation of X_s). *For $s \in [1, \infty]$, let X_s be the random variable defined in* (9.25)*. Then, if $p \geq 4$, we have $\mathbb{E}_\Omega^2[X_s] < \infty$.*

Proof. We first notice that $2s \geq 2$ implies

$$\|\mathbf{\Lambda}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_{(2s)}^2 \leq \|\mathbf{\Lambda}_2^{1/2}\|_{\mathrm{op}}^2\|\mathbf{\Omega}_2\|_{\mathrm{HS}}^2\|\mathbf{\Omega}_1^\dagger\|_{\mathrm{F}}^2.$$

Noting that $\mathbf{\Omega}_1$, $\mathbf{\Omega}_2$ are not independent, we need to establish $\mathbb{E}[\|\mathbf{\Omega}_2\|_{\mathrm{HS}}^4] < \infty$ and $\mathbb{E}[\|\mathbf{\Omega}_1^\dagger\|_{\mathrm{F}}^4] < \infty$ in order to conclude the result from Hölder's inequality. First, Lemma A.3 ensures that $\mathbb{E}[\|\mathbf{\Omega}_1^\dagger\|_{\mathrm{F}}^4] < \infty$ since $\mathbf{\Omega}_1$ is a $k \times (k+p)$ matrix whose columns are i.i.d. from $\mathcal{N}(\mathbf{0}, \widetilde{\boldsymbol{K}}_{11})$ [29, Lemma 1]. By the Karhunen–Loève expansion (9.18), an arbitrary column ω of Ω satisfies

$$\mathbb{E}\Big[\|\omega\|_{L^2(D)}^4\Big] = \sum_{i,j=1}^{\infty} \mathbb{E}[\zeta_i^2\zeta_j^2]\theta_i\theta_j = \sum_i \mathbb{E}[\zeta_i^4]\theta_i^2 + \sum_{i\neq j}\mathbb{E}[\zeta_i^2\zeta_j^2]\theta_i\theta_j =$$
$$3\sum_i \theta_i^2 + \sum_{i\neq j}\theta_i\theta_j \leq 3\|\mathcal{K}\|_{\mathrm{Tr}}^2 < \infty.$$

In turn,

$$\mathbb{E}\|\mathbf{\Omega}_2\|_{\mathrm{HS}}^4 \leq \mathbb{E}\|\Omega\|_{\mathrm{HS}}^4 \leq \mathbb{E}\Big(\|\omega_1\|_{L^2(D)}^2 + \cdots + \|\omega_{k+p}\|_{L^2(D)}^2\Big)^2 \leq (k+p)^2\mathbb{E}\Big[\|\omega\|_{L^2(D)}^4\Big] < \infty.$$

□

We proceed by showing that $\lim_{n\to\infty}\mathbb{E}^2[X_s - X_{s,n}] = 0$.

Lemma 9.9 (Convergence of $X_{s,n}$ to X_s)**.** *For $s \in [1,\infty]$, let X_s, $X_{s,n}$ be the random variables defined in* (9.25) *for $n \geq 1$. Then, if $p \geq 4$, we have* $\lim_{n\to\infty}\mathbb{E}^2[X_s - X_{s,n}] = 0$.

Proof. For $n \geq 1$, we define the quasimatrix $\mathbf{\Omega}_2^{(n)}$ whose first n rows are equal to the first n rows of $\mathbf{\Omega}_2$ and the remaining rows are zero. Then,

$$X_{s,n} = \|(\mathbf{\Lambda}_2^{1/2})_{[\![1,n]\!]\times[\![1,n]\!]}(\mathbf{\Omega}_2)_{[\![1,n]\!]\times[\![1,k+p]\!]}\mathbf{\Omega}_1^\dagger\|_{(2s)} = \|\mathbf{\Lambda}_2^{1/2}\mathbf{\Omega}_2^{(n)}\mathbf{\Omega}_1^\dagger\|_{(2s)}.$$

Combining the triangle inequality and sub-multiplicativity of the Schatten-s norm, and using the fact that $2s \geq 2$, we have

$$(X_s - X_{s,n})^2 \leq \|\mathbf{\Lambda}_2^{1/2}\|_{\mathrm{op}}^2\|\mathbf{\Omega}_1^\dagger\|_{\mathrm{F}}^2\|\mathbf{\Omega}_2 - \mathbf{\Omega}_2^{(n)}\|_{\mathrm{HS}}^2.$$

By Hölder's inequality it suffices to show that $\lim_{n\to\infty}\mathbb{E}[\|\mathbf{\Omega}_2 - \mathbf{\Omega}_2^{(n)}\|_{\mathrm{HS}}^4] = 0$, since $\mathbb{E}[\|\mathbf{\Omega}_1^\dagger\|_{\mathrm{F}}^4] < \infty$ by Lemma A.3. Let $\boldsymbol{\omega}_i$ and $\boldsymbol{\omega}_i^{(n)}$ denote the i^{th} columns of $\mathbf{\Omega}_2$ and $\mathbf{\Omega}_2^{(n)}$, respectively. Using the monotonicity and triangle inequality of L^p-norms, we have

$$\begin{aligned}\mathbb{E}^4[\|\boldsymbol{\Omega}_2 - \boldsymbol{\Omega}_2^{(n)}\|_{\mathrm{HS}}] \leq \mathbb{E}^4\left[\sum_{i=1}^{k+p} \|\boldsymbol{\omega}_i - \boldsymbol{\omega}_i^{(n)}\|_2\right] &\leq \sum_{i=1}^{k+p} \mathbb{E}^4\left[\|\boldsymbol{\omega}_i - \boldsymbol{\omega}_i^{(n)}\|_2\right] \\ &= (k+p)\mathbb{E}^4\left[\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_1^{(n)}\|_2\right],\end{aligned}$$

since the columns of $\boldsymbol{\Omega}_2$ are identically distributed.

We are now going to verify that

$$\lim_{n\to\infty} \mathbb{E}\left[\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_1^{(n)}\|_2^4\right] = 0.$$

Following [29, Section 3.3], we know that the entries of $\boldsymbol{\omega}_1$ satisfy $(\boldsymbol{\omega}_1)_i \sim \mathcal{N}(0, (\widetilde{\boldsymbol{K}}_{22})_{ii})$ for $i \in \mathbb{N}^*$. Let $Y = \sum_{i=n+1}^{\infty} (\boldsymbol{\omega}_1)_i^2 = \|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_1^{(n)}\|_2^2$. Combining the non-negativity of the summands and the Fubini–Tonelli theorem, we can interchange the summation and expectation to obtain

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_1^{(n)}\|_2^4\right] = \mathbb{E}[Y^2] = \mathbb{E}\left[\sum_{i=n+1}^{\infty} (\boldsymbol{\omega}_1)_i^2 Y\right] &= \sum_{i=n+1}^{\infty} \mathbb{E}\left[(\boldsymbol{\omega}_1)_i^2 Y\right] \\ &\leq \sum_{i=n+1}^{\infty} \sqrt{\mathbb{E}[(\boldsymbol{\omega}_1)_i^4]}\sqrt{\mathbb{E}[Y^2]},\end{aligned}$$

where the last inequality follows from the Cauchy–Schwarz inequality. Hence,

$$\mathbb{E}\left[\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_1^{(n)}\|_2^4\right] \leq \left(\sum_{i=n+1}^{\infty} \sqrt{\mathbb{E}[(\boldsymbol{\omega}_1)_i^4]}\right)^2 = 3\left(\sum_{i=n+1}^{\infty} (\widetilde{\boldsymbol{K}}_{22})_{ii}\right)^2 \to 0, \quad \text{as } n \to \infty,$$

since $\operatorname{tr}(\widetilde{\boldsymbol{K}}_{22}) \leq \operatorname{tr}(\mathcal{K}) < \infty$. □

Combining Lemmas 9.8 and 9.9, we obtain that $\lim_{n\to\infty} \mathbb{E}^2[X_{s,n}] = \mathbb{E}^2[X_s]$. Hence, if we have a family of bounds $\mathbb{E}^2[X_{s,n}] \leq y_{s,n}$ with $y_{s,n} \to y_s$ as $n \to \infty$, then $\mathbb{E}^2[X_s] \leq y_s$. Furthermore, combining the continuous mapping theorem and the fact that L^2 convergence implies convergence in distribution, for any positive sequence $z_{s,n} \to z_s < \infty$, we have $\lim_{n\to\infty} \mathbb{P}(X_{s,n}^2 > z_{s,n}) = \mathbb{P}(X_s^2 > z_s)$. We can now proceed with the proof of Theorem 9.7, which uses results from Section 9.2 to derive expressions for $y_{s,n}$ and $z_{s,n}$ and show that, for $s \in \{2, \infty, 1\}$, they converge to the right-hand sides of (9.23) and (9.24).

Proof of Theorem 9.7. Let $\|\boldsymbol{\Lambda}_2\|_{\mathrm{HS}} + y_2$, $\|\boldsymbol{\Lambda}_2\|_{\mathrm{op}} + y_\infty$, and $\|\boldsymbol{\Lambda}_2\|_{\mathrm{Tr}} + y_1$ be the respective right-hand sides in (9.23), and $\|\boldsymbol{\Lambda}_2\|_{\mathrm{HS}} + z_2$, $\|\boldsymbol{\Lambda}_2\|_{\mathrm{op}} + z_\infty$, and $\|\boldsymbol{\Lambda}_2\|_{\mathrm{Tr}} + z_1$ be the right-hand sides in (9.24). For $s \in \{2, \infty, 1\}$, we aim to prove that $\mathbb{E}[X_s^2] \leq y_s$ and

$\mathbb{P}(X_s^2 \geq z_s) \leq 3t^{-p} + e^{-u^2/2}$. Following Theorem 9.1, we have

$$\begin{aligned}
\mathbb{E}[X_{2,n}^2] \leq {} & 2\|(\boldsymbol{\Lambda}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}(\widetilde{\boldsymbol{K}}_{21})_{\llbracket 1,n\rrbracket\times\llbracket 1,k\rrbracket}\widetilde{\boldsymbol{K}}_{11}^{-1}(\widetilde{\boldsymbol{K}}_{21})^*_{\llbracket 1,n\rrbracket\times\llbracket 1,k\rrbracket}(\boldsymbol{\Lambda}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}\|_{\mathrm{F}} \\
& + 2\sqrt{c_1}\|(\boldsymbol{\Lambda}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}(\widetilde{\boldsymbol{K}}_{22.1})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}(\boldsymbol{\Sigma}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}\|_{\mathrm{F}} \\
& + 2\sqrt{c_2}\|(\boldsymbol{\Sigma}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}(\widetilde{\boldsymbol{K}}_{22.1})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}(\boldsymbol{\Sigma}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}\|_* := y_{2,n},
\end{aligned}$$

and with probability greater than $1 - 3^{-p} - e^{-u^2/2}$, we have

$$\begin{aligned}
X_{2,n}^2 \geq {} & 4\|(\boldsymbol{\Lambda}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}(\widetilde{\boldsymbol{K}}_{21})_{\llbracket 1,n\rrbracket\times\llbracket 1,k\rrbracket}\widetilde{\boldsymbol{K}}_{11}^{-1}(\widetilde{\boldsymbol{K}}_{21})^*_{\llbracket 1,n\rrbracket\times\llbracket 1,k\rrbracket}(\boldsymbol{\Lambda}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}\|_{\mathrm{F}} \\
& + 4t^2(d_1 + d_3)\|(\boldsymbol{\Lambda}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}(\widetilde{\boldsymbol{K}}_{22.1})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}(\boldsymbol{\Lambda}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}\|_{\mathrm{F}} \\
& + 4t^2 d_3\|(\boldsymbol{\Lambda}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}(\widetilde{\boldsymbol{K}}_{22.1})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}(\boldsymbol{\Lambda}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}\|_* \\
& + 2t^2u^2 d_2\|(\boldsymbol{\Lambda}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}(\widetilde{\boldsymbol{K}}_{22.1})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}(\boldsymbol{\Lambda}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}\|_2 := z_{2,n}.
\end{aligned}$$

Following Theorems 9.3 and 9.4 we know that

$$\begin{aligned}
&\mathbb{E}[X_{\infty,n}] \leq y_{\infty,n}, && \mathbb{E}[X_{1,n}] \leq y_{1,n}, \\
&\mathbb{P}\left(X_{\infty,n} > z_{\infty,n}\right) \leq 3^{-p} + e^{-u^2/2}, && \mathbb{P}\left(X_{1,n} > z_{1,n}\right) \leq 3^{-p} + e^{-u^2/2},
\end{aligned}$$

where $y_{\infty,n}$, $y_{1,n}$, $z_{\infty,n}$, and $z_{1,n}$ can be defined analogously to $y_{2,n}$ and $z_{2,n}$ using (9.14) and (9.15). Moreover, Lemma 9.9 implies that $\lim_{n\to\infty} \mathbb{E}[X_{s,n}^2] = \mathbb{E}[X_s^2]$ and convergence in distribution, which implies $\lim_{n\to\infty} \mathbb{P}(X_{s,n} > z_{s,n}) = \mathbb{P}(X_s > \lim_{n\to\infty} z_{s,n})$. Hence, it is sufficient to show that $\lim_{n\to\infty} y_{s,n} = y_s$ and $\lim_{n\to\infty} z_{s,n} = z_s$ for $s \in \{2, \infty, 1\}$. For this purpose, it is sufficient to show that

$$\begin{aligned}
&\lim_{n\to\infty} \|(\boldsymbol{\Lambda}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}(\widetilde{\boldsymbol{K}}_{22.1})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}(\boldsymbol{\Lambda}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}\|_\xi = \|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}\boldsymbol{\Lambda}_2^{1/2}\|_\xi, \\
&\lim_{n\to\infty} \|(\boldsymbol{\Lambda}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}(\widetilde{\boldsymbol{K}}_{21})_{\llbracket 1,n\rrbracket\times\llbracket 1,k\rrbracket}\widetilde{\boldsymbol{K}}_{11}^{-1}(\widetilde{\boldsymbol{K}}_{21})^*_{\llbracket 1,n\rrbracket\times\llbracket 1,k\rrbracket}(\boldsymbol{\Lambda}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}\|_\xi = \\
&\qquad\qquad \|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{21}\widetilde{\boldsymbol{K}}_{11}^{-1}\widetilde{\boldsymbol{K}}_{21}^*\boldsymbol{\Lambda}_2^{1/2}\|_\xi,
\end{aligned} \tag{9.26}$$

for $\xi \in \{\mathrm{HS}, \mathrm{Tr}, \mathrm{op}\}$, as inserting the definitions of $\beta_k^{(\xi)}$ and $\delta_k^{(\xi)}$ would then give the result. Note that since $\boldsymbol{\Lambda}_2$ is diagonal we have

$$\begin{aligned}
&(\boldsymbol{\Lambda}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}(\widetilde{\boldsymbol{K}}_{22.1})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}(\boldsymbol{\Lambda}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket} = (\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}\boldsymbol{\Lambda}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}, \\
&(\boldsymbol{\Lambda}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}(\widetilde{\boldsymbol{K}}_{21})_{\llbracket 1,n\rrbracket\times\llbracket 1,k\rrbracket}\widetilde{\boldsymbol{K}}_{11}^{-1}(\widetilde{\boldsymbol{K}}_{21})^*_{\llbracket 1,n\rrbracket\times\llbracket 1,k\rrbracket}(\boldsymbol{\Lambda}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket} = \\
&\qquad\qquad (\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{21}\widetilde{\boldsymbol{K}}_{11}^{-1}\widetilde{\boldsymbol{K}}_{21}^*\boldsymbol{\Lambda}_2^{1/2})_{\llbracket 1,n\rrbracket\times\llbracket 1,n\rrbracket}.
\end{aligned}$$

Since $\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{22.1}\boldsymbol{\Lambda}_2^{1/2}$ and $\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{K}}_{21}\widetilde{\boldsymbol{K}}_{11}^{-1}\widetilde{\boldsymbol{K}}_{21}^*\boldsymbol{\Lambda}_2^{1/2}$ are non-negative trace-class operators, applying (9.22a) yields (9.26), as desired. □

9.4 Numerical experiments

In this section, we test the infinite-dimensional Nyström approximation proposed in this work. Algorithm 14 presents the pseudocode of a suitable "implementation", a variant of [104, Algorithm 16] for non-negative self-adjoint trace-class operators.

```
Algorithm 14 Nyström approximation
input: Non-negative self-adjoint trace-class $\mathcal{A} : L^2(D) \to L^2(D)$, covariance kernel
    $K : D \times D \to \mathbb{R}$. Target rank $k$, oversampling parameter $p$.
output: Rank $k+p$ Nyström approximation $\widehat{\mathcal{A}}$ to $\mathcal{A}$ in factored form.
 1: Draw a random quasimatrix $\Omega = \begin{bmatrix}\omega_1 & \cdots & \omega_{k+p}\end{bmatrix}$ with columns i.i.d. from $\mathcal{GP}(0, K)$.
 2: Orthonormalize columns of $\Omega$: $Q = \text{orth}(\Omega) = \begin{bmatrix}q_1 & \cdots & q_{k+p}\end{bmatrix}$.
 3: Apply operator $\mathcal{A}$ to $Q$: $Y = \mathcal{A}Q = \begin{bmatrix}\mathcal{A}q_1 & \cdots & \mathcal{A}q_{k+p}\end{bmatrix}$.
 4: $\nu = \epsilon\|Y\|_{\text{HS}}$ where $\epsilon$ is equal to the machine precision.
 5: Compute shifted $Y_\nu = Y + \nu\Omega$.
 6: Compute Cholesky factorization of $\boldsymbol{\Omega}^*\boldsymbol{Y}_\nu = \boldsymbol{R}^T\boldsymbol{R}$.   ▷ First compute the symmetric
    part of $\boldsymbol{\Omega}^*\boldsymbol{Y}_\nu$ if needed.
 7: Perform a triangular solve to compute $B = Y_\nu\boldsymbol{R}^{-1}$.
 8: Compute the Hilbert-Schmidt decomposition of $B = \widehat{U}\boldsymbol{S}\boldsymbol{V}^*$.
 9: Remove shift $\widehat{\boldsymbol{\Lambda}} = \max\{\boldsymbol{S}^2 - \nu\boldsymbol{I}, 0\}$, where the maximum is taken entry-wise.
10: return $\widehat{\mathcal{A}} = \widehat{U}\widehat{\boldsymbol{\Lambda}}\widehat{U}^*$ in factored form.
```

Some remarks:

- The orthonormalization in line 2 and the shift in line 5 improves the numerical stability of the algorithm; see e.g. [151].

- In our implementation, we compute the Cholesky factorization of the symmetric part of $\boldsymbol{\Omega}^*\boldsymbol{Y}_\nu$ in line 6.

- In exact arithmetic with $\nu = 0$ in line 4 the approximation returned by Algorithm 14 is mathematically equivalent to (9.20).

The purpose of our experiments is to validate Algorithm 14. For this purpose, we consider an interval $D = [a, b]$ and carry out all operations on functions on D (approximately) using the Chebfun software package [47].

In all experiments, we choose the covariance kernel as a squared-exponential kernel

$$K(x, y) = \exp\left(-\frac{2(x-y)^2}{(b-a)^2\ell^2}\right), \quad x, y \in [a, b],$$

where we vary the length-scale parameter across three values: $\ell = 1, 0.1, 0.01$. A smaller value for ℓ results in smoother Gaussian processes that are more biased towards certain spatial directions. Conversely, a larger value of ℓ results in rougher Gaussian processes that are less spatially biased. We vary ℓ to investigate the effect of the smoothness of the

Gaussian process on the result of the low-rank approximation.

9.4.1 A pretty function

In this example, we compute an approximation to the integral operator defined by the kernel [147]

$$G(x,y) = \frac{1}{1+100(x^2-y^2)^2}, \quad x,y \in [-1,1]. \tag{9.27}$$

We display the results in Figure 9.1. As can be seen from the figures, setting $\ell = 1$ yields a poor approximation to the kernel, and $\ell = 0.01$ is better to obtain high-accuracy approximations.

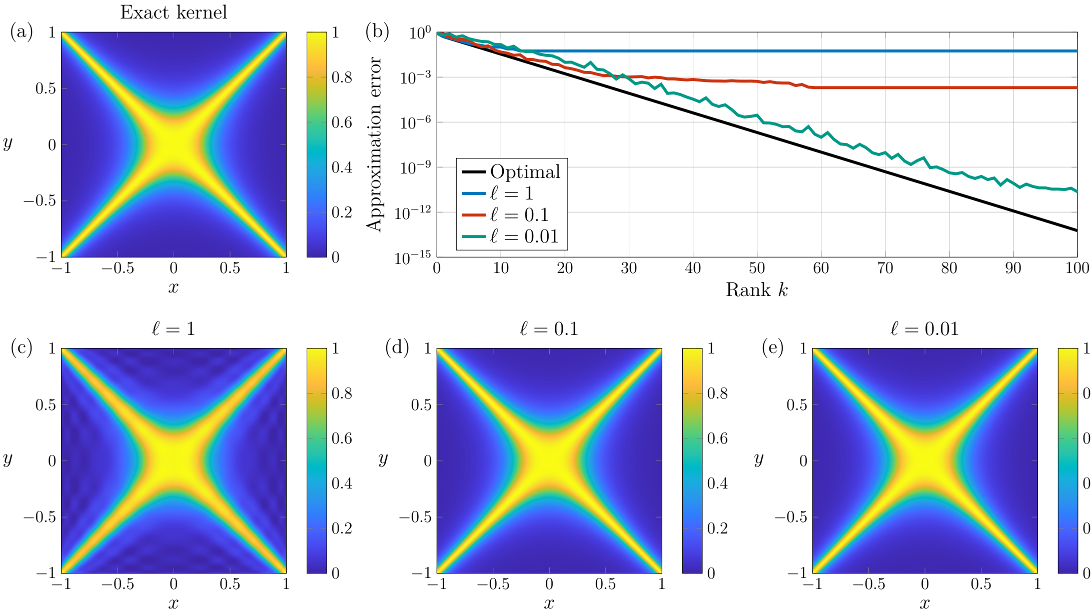


Figure 9.1: (a) Exact kernel defined by (9.27) along with convergence of the Nyström approximation for different values of ℓ (b). (c)-(e) Rank-40 Nyström approximations of the kernel for $\ell = 1, 0.1, 0.01$, respectively.

9.4.2 Matérn Kernels

In this second example, we approximate the integral operator defined by the Matérn-1/2, 3/2, and 5/2 kernels [129, Chapter 4]

$$G_{1/2}(x,y) = \exp(-|x-y|), \quad x,y \in [-1,1]; \tag{9.28}$$

$$G_{3/2}(x,y) = (1+\sqrt{3}|x-y|)\exp\left(-\sqrt{3}|x-y|\right), \quad x,y \in [-1,1]; \tag{9.29}$$

$$G_{5/2}(x,y) = \left(1+\sqrt{5}|x-y|+\frac{5}{3}(x-y)^2\right)\exp\left(-\sqrt{5}|x-y|\right), \quad x,y \in [-1,1]. \tag{9.30}$$

The Matérn class is an important class of covariance kernels, frequently appearing in machine learning. The parameter $\nu = 1/2, 3/2, 5/2$ determines the spectral decay of the kernel and thus the smoothness of the Gaussian process, with higher ν implying faster decay and smoother Gaussian processes. The results are presented in Figure 9.2 for $G_{1/2}, G_{3/2}$, and $G_{5/2}$, respectively.

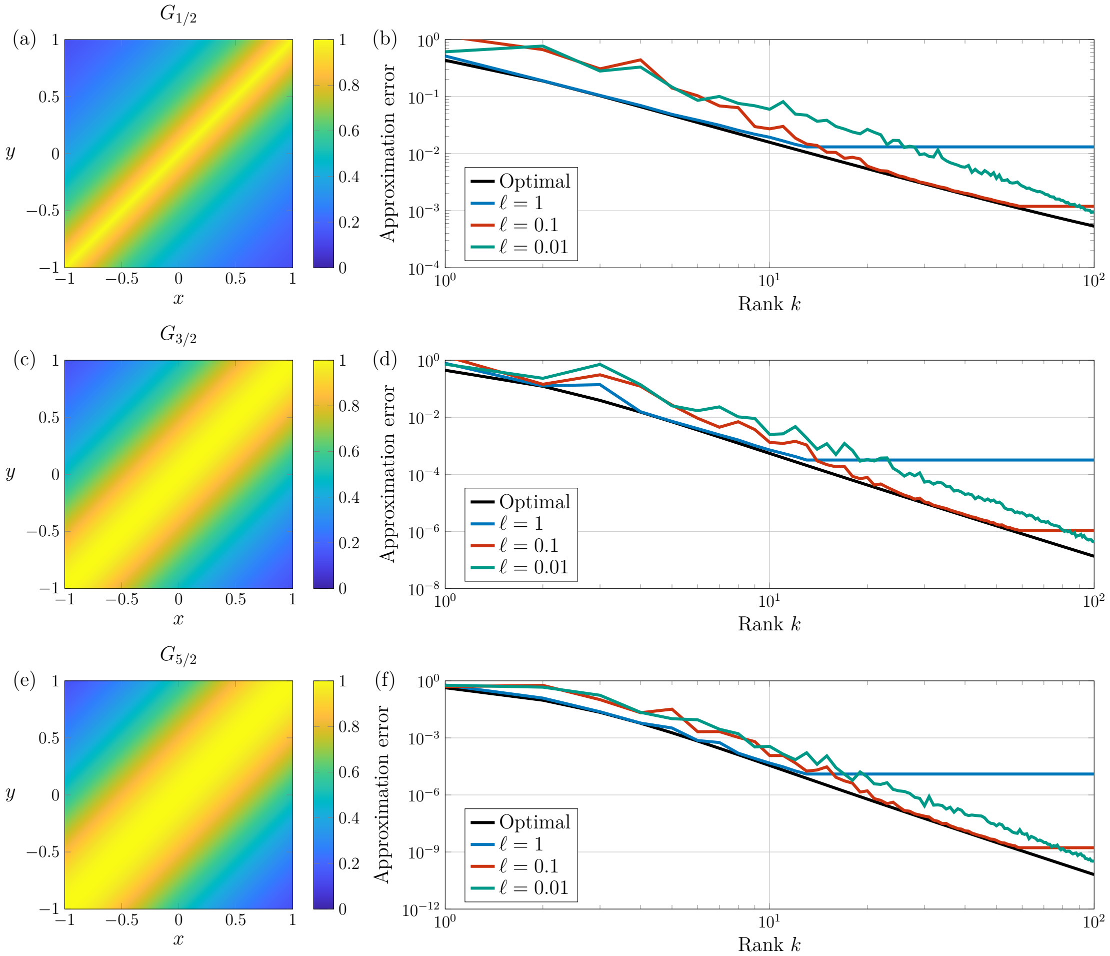


Figure 9.2: The contour plots show the exact kernel defined by (9.28)-(9.30) and the approximated kernels using rank 100 Nyström approximations. The error plots show the relative error in the Hilbert-Schmidt norm. *Optimal* denotes the best low-rank approximation error.

9.4.3 Green's function for an elliptic differential operator

In this example we consider the operator $\mathcal{A}_\eta$ so that $u = \mathcal{A}_\eta f$ solves the equation

$$-\Delta u(\boldsymbol{x}) + \eta u(\boldsymbol{x}) = f(\boldsymbol{x}), \quad \boldsymbol{x} \in [0, 2\pi]^d,$$

where $\eta \geq 0$, with Direchlet boundary conditions. Let $\mathcal{A}_0$ have Hilbert-Schmidt decomposition

$$[\mathcal{A}_0 f](\boldsymbol{x}) = \sum_{i=1}^{\infty} \lambda_i u_i(\boldsymbol{x}) \langle u_i, f \rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the standard $L^2([0, 2\pi]^d)$ inner product. Then $\mathcal{A}_\eta$ has Hilbert-Schmidt decomposition

$$[\mathcal{A}_\eta f](\boldsymbol{x}) = \sum_{i=1}^{\infty} g_\eta(\lambda_i) u_i(\boldsymbol{x}) \langle u_i, f \rangle = [g_\eta(\mathcal{A}_0) f](\boldsymbol{x}), \quad g_\eta(\lambda) = \frac{\lambda}{\eta\lambda + 1}.$$

For each $\eta \geq 0$ the function g_η is operator monotone [25, Section V]. In Chapter 4 we showed that if $\widehat{\mathcal{A}}$ is a near-optimal low-rank approximation to $\mathcal{A}$, then $g(\widehat{\mathcal{A}}_0)$ is a near-optimal low-rank approximation to $g(\mathcal{A})$, for any non-negative continuous operator monotone function g. Hence, to obtain a low-rank approximation to $\mathcal{A}_\eta$ we compute a Nyström approximation $\widehat{\mathcal{A}}_0$ to $\mathcal{A}_0$ and then approximate $\mathcal{A}_\eta$ with $g_\eta(\widehat{\mathcal{A}}_0)$. In our experiments, we set $\eta = 1$ and $d = 1$. In this case, the Green's function is given by

$$G(x, y) = \min(x, y) - \frac{xy}{2\pi}. \tag{9.31}$$

The results are displayed in Figure 9.3.

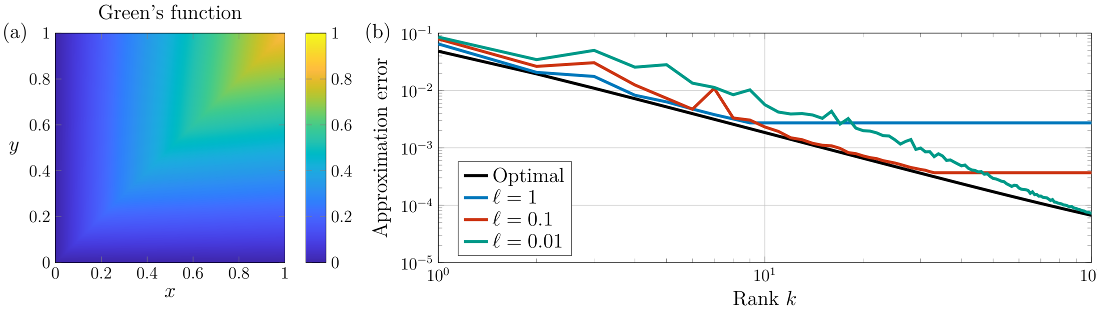


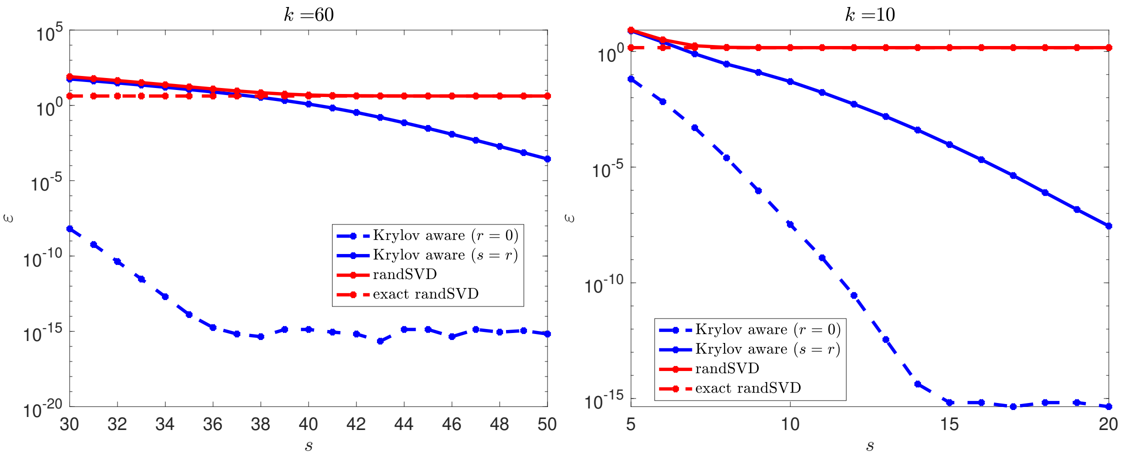
Figure 9.3: The contour plots show the exact kernel defined by (9.31) and the approximated kernels using rank 100 Nyström approximations. The error plots show the relative error in the Hilbert-Schmidt norm. *Optimal* denotes the best low-rank approximation error.

10 Conclusions and outlook

In Chapter 4 we presented funNyström: a simple and effective method to compute low-rank approximations of non-negative operator monotone matrix functions. A significant advantage of funNyström is that it does not require any access to the matrix function $f(\boldsymbol{A})$. Instead, it requires only that we are able to compute a Nyström approximation to the matrix $\boldsymbol{A}$, a task that usually is significantly cheaper. This dramatically reduces the computational cost compared to methods that require access to the matrix function $f(\boldsymbol{A})$. We showed that any near-optimal Nyström approximation to $\boldsymbol{A}$ can be used to compute a near-optimal funNyström approximation to $f(\boldsymbol{A})$. Furthermore, we showed that if $\boldsymbol{Q} \in \mathbb{R}^{n \times \ell}$ is an orthonormal basis so that $\|\boldsymbol{A} - (\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A})_{(k)}\| \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|$, then if $\widehat{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q})^{\dagger}\boldsymbol{Q}^T\boldsymbol{A}$ we have $\|\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}\| \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|$, where $\|\cdot\|$ is the nuclear or Frobenius norm. Moreover, we showed that such a result is impossible in the operator norm. We can use these results to immediately extend results in the literature to obtain results for the funNyström approximation. We also provided bounds for the funNyström approximation in general unitarily invariant norms.

A number of open questions remain. Firstly, we would like to weaken the assumption on Theorem 4.6 so that we only require $\|\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}\|_{\mathrm{F}}^2 \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|_{\mathrm{F}}^2$. Secondly, it would be desirable to extend our results to general non-negative monotonically increasing functions. In particular, we would like to explore if it is possible to show that if $\widehat{\boldsymbol{A}}$ is a Nyström approximation (or another low-rank approximation) so that $\|\boldsymbol{A} - \widehat{\boldsymbol{A}}_{(k)}\| \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_{(k)}\|$ then $\|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}})_{(k)}\| \leq (1+C_f\varepsilon)\|f(\boldsymbol{A}) - f(\boldsymbol{A})_{(k)}\|$ for some constant C_f depending only on f.

In Chapter 5 we analysed the Krylov-aware low-rank approximation method suggested by Chen and Hallman in [37]. We proved that this method will return a good approximation to $f(\boldsymbol{A})$ if there is a low-degree polynomial that can denoise the small eigenvalues of $f(\boldsymbol{A})$. Furthermore, numerical experiments demonstrate that the Krylov-aware algorithm is significantly more efficient than naively implementing the randomized SVD on $f(\boldsymbol{A})$ using approximate matrix-vector products.



(a) Exponential integrator example outlined in Section 5.3.1

(b) Estrada index example outlined in Section 5.3.1

Figure 10.1: Comparing the subobtimality factor $\varepsilon = \frac{\|f(\boldsymbol{A})-\boldsymbol{B}\|_{\mathrm{F}}}{\|f(\boldsymbol{A})-f(\boldsymbol{A})_{(k)}\|_{\mathrm{F}}} - 1$ for a rank k approximations $\boldsymbol{B}$. We consider the rank k approximations given by (10.1) with $q = 2s$, the approximation returned by Algorithm 8 with $s = r$ with (with truncation), Algorithm 7, and Algorithm 1. The sketch matrix is chosen as a $n \times k$ standard Gaussian random matrix. The rank parameter k is visible as titles in the figures.

A fundamental question remains: numerical experiments suggest that the approximation

$$f(\boldsymbol{A}) \approx \boldsymbol{Q}_q f(\boldsymbol{T}_q)_{(k)} \boldsymbol{Q}_q^T, \tag{10.1}$$

where $\boldsymbol{Q}_q$ and $\boldsymbol{T}_q$ are as in Algorithm 4, is a more accurate low-rank approximation compared to the one returned by Algorithm 8; see Figure 10.1. The approximation in (10.1) corresponds to setting $r = 0$ in Algorithm 8 and our theory does not justify this choice of r. Fortunately, our analysis from Chapter 5 is a stepping stone towards an analysis of (10.1), since by Theorem 5.2 and Lemma 5.3 we only need a bound for $\|f(\boldsymbol{T}_q) - \boldsymbol{Q}_q^T f(\boldsymbol{A}) \boldsymbol{Q}_q\|_{\mathrm{F}}$ to obtain a bound for $\|f(\boldsymbol{A}) - \boldsymbol{Q}_q f(\boldsymbol{T}_q)_{(k)} \boldsymbol{Q}_q^T\|_{\mathrm{F}}$.

In Chapter 7 we presented an adaptive version of Hutch++, A-Hutch++, that will estimate the trace of a symmetric matrix $\boldsymbol{A}$ while attempting to minimize the number of matrix-vector products with $\boldsymbol{A}$ used overall. This algorithm also comes with the advantage that the user does not need to determine the number of matrix-vector products required to output an estimate of the trace that is within the prescribed error tolerance.

An important question remains for future study. Recall that an important application of trace estimation is to estimate the trace of a matrix function $f(\boldsymbol{A})$. Approximating matrix-vector products with $f(\boldsymbol{A})$ requires repeated products with $\boldsymbol{A}$, and it would therefore be beneficial to develop an algorithm that estimates $\mathrm{tr}(f(\boldsymbol{A}))$, while minimizing the number of matrix-vector products with $\boldsymbol{A}$.

In Chapter 8 we presented a version of Hutch++ utilizing the Nyström approximation, which requires only one pass over the matrix. We proved that this algorithm satisfies the same theoretical guarantees of Hutch++. While this algorithm offers a similar performance as Hutch++, it performs significantly better than the previously proposed single pass algorithm Single Pass Hutch++.

A future research direction is to develop other variants of the Hutch++ algorithm. For example, [31, 110] studied the Hutchinson-Girard trace estimator in (6.1) where the random test vectors are random Khatri-Rao products. Studying a version of Hutch++ using random Khatri-Rao products is of interest in applications where the matrix $\boldsymbol{A}$ has a special structure that allows for very fast matrix-vector products with structured vectors. Such matrices arise in, for example, stochastic automata networks [97] and in low-rank tensor formats [73]. Developing a version of the Hutch++ algorithm with random Khatri-Rao products would require a crisp analysis of the randomized SVD (Algorithm 1) with these random test vectors. Such results would also be of independent interest. A starting point to establish error bounds for the randomized SVD would be to use the Johnson-Lindenstrauss property of random Kronecker vectors established in, for example, [3].

In Chapter 9 we presented an infinite-dimensional analogue of the Nyström approximation for non-negative self-adjoint trace class operators. We first established bounds for the finite dimensional Nyström approximation when the columns of the sketch matrix $\boldsymbol{\Omega}$ are drawn independently from a non-standard Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{K})$. Subsequently, through a continuity argument we provided analogous bounds for the infinite-dimensional extension of the Nyström approximation for non-negative trace class operators. Additionally, in the process of analyzing the Nyström approximation for trace class operators, we have also improved the existing bounds for the randomized SVD for Hilbert-Schmidt operators.

An important issue remain for future study. Unfortunately, the Nyström approximation presented in Chapter 9 cannot be used to approximate the off-diagonal parts of Green's functions of elliptic differential equations, since the off-diagonal parts of Green's functions are not even self-adjoint. Approximating Green's functions of differential equations was the motivating application for the development of the randomized SVD for Hilbert-Schmidt operators. Therefore, it would be of interest to develop an infinite-dimensional analogue of the generalized Nyström approximation [117, 152], which would be valid for non-self-adjoint operators. A first step towards such generalization is to analyze the finite-dimensional generalized Nyström approximation applied with non-standard Gaussian random vectors. However, such analysis is complicated due to the fact that the existing analysis of the generalized Nyström approximation makes heavy use of the rotational invariance of the random test vectors, a property that is not satisfied by random vectors drawn from non-standard Gaussian distributions.

A Properties of Gaussian matrices

The following lemma is a consequence of the symmetry of standard normal random variables.

Lemma A.1. *Let $\boldsymbol{C} \in \mathbb{R}^{m_1 \times m_2}$, $\boldsymbol{D} \in \mathbb{R}^{n_1 \times n_2}$ be two matrices, consider an $m_2 \times n_1$ standard Gaussian matrix $\boldsymbol{\Psi}$, and define $\boldsymbol{\Phi} = \boldsymbol{C\Psi D}$. Then, $\mathbb{E}_{\boldsymbol{\Psi}}[\boldsymbol{\Phi\Phi}^*\boldsymbol{\Phi}] = \mathbf{0}$.*

Proof. First note that the expectation exists, since each entry of $\boldsymbol{\Phi\Phi}^*\boldsymbol{\Phi}$ is a linear combination of products of Gaussian random variables, which always has a finite expectation. The distribution of $\boldsymbol{\Psi}$ and, in turn, $\boldsymbol{\Phi}$ is symmetric, which gives

$$-\mathbb{E}_{\boldsymbol{\Psi}}[\boldsymbol{\Phi\Phi}^*\boldsymbol{\Phi}] = \mathbb{E}_{\boldsymbol{\Psi}}[-\boldsymbol{\Phi\Phi}^*\boldsymbol{\Phi}] = \mathbb{E}_{\boldsymbol{\Psi}}[(-\boldsymbol{\Phi})(-\boldsymbol{\Phi})^*(-\boldsymbol{\Phi})] = \mathbb{E}_{\boldsymbol{\Psi}}[\boldsymbol{\Phi\Phi}^*\boldsymbol{\Phi}],$$

implying the result of the lemma. □

The following lemma summarizes results on the expected norms of scaled and shifted Gaussian matrices for the Frobenius, Schatten-4, and spectral norms.

Lemma A.2 (Expected norm of shifted Gaussian matrices)**.** *Let $\boldsymbol{C} \in \mathbb{R}^{m_1 \times m_2}, \boldsymbol{D} \in \mathbb{R}^{n_1 \times n_2}$, and $\boldsymbol{B} \in \mathbb{R}^{m_1 \times n_2}$ be three arbitrary matrices, and consider an $m_2 \times n_1$ standard Gaussian matrix $\boldsymbol{\Psi}$. Then, the following relations hold:*

$$\mathbb{E}\Big[\|\boldsymbol{B} + \boldsymbol{C\Psi D}\|_{\mathrm{F}}^2\Big] = \|\boldsymbol{B}\|_{\mathrm{F}}^2 + \|\boldsymbol{C}\|_{\mathrm{F}}^2\|\boldsymbol{D}\|_{\mathrm{F}}^2, \tag{A.1a}$$

$$\mathbb{E}\Big[\|\boldsymbol{C\Psi D}\|_{(4)}^4\Big] = \|\boldsymbol{C}\|_{(4)}^4\|\boldsymbol{D}\|_{(4)}^4 + \|\boldsymbol{C}\|_{\mathrm{F}}^4\|\boldsymbol{D}\|_{(4)}^4 + \|\boldsymbol{C}\|_{(4)}^4\|\boldsymbol{D}\|_{\mathrm{F}}^4, \tag{A.1b}$$

$$\mathbb{E}\Big[\|\boldsymbol{C\Psi D}\|_2^2\Big] \le \left(\|\boldsymbol{C}\|_{\mathrm{F}}\|\boldsymbol{D}\|_2 + \|\boldsymbol{C}\|_2\|\boldsymbol{D}\|_{\mathrm{F}}\right)^2, \tag{A.1c}$$

$$\mathbb{E}\Big[\|\boldsymbol{C\Psi D}\|_{\mathrm{F}}^4] = 2\|\boldsymbol{C}\|_{(4)}^4\|\boldsymbol{D}\|_{(4)}^4 + \|\boldsymbol{C}\|_{\mathrm{F}}^4\|\boldsymbol{D}\|_{\mathrm{F}}^4. \tag{A.1d}$$

Proof. We introduce the matrix $\boldsymbol{\Phi} = \boldsymbol{C\Psi D}$ and begin by proving (A.1a). Using the linearity of trace and expectation we have:

$$\mathbb{E}[\|\boldsymbol{B}+\boldsymbol{\Phi}\|_{\mathrm{F}}^2] = \|\boldsymbol{B}\|_{\mathrm{F}}^2 + \mathbb{E}[\|\boldsymbol{\Phi}\|_{\mathrm{F}}^2] + 2\mathbb{E}[\operatorname{tr}(\boldsymbol{B}^*\boldsymbol{\Phi})] = \|\boldsymbol{B}\|_{\mathrm{F}}^2 + \|\boldsymbol{C}\|_{\mathrm{F}}^2\|\boldsymbol{D}\|_{\mathrm{F}}^2,$$

where we combined [79, Proposition 10.1] with the equality $\mathbb{E}[\operatorname{tr}(\boldsymbol{B}^*\boldsymbol{\Phi})] = 0$ since $\boldsymbol{\Psi}$ has zero mean. A similar result is found in [45, Lemma 3.11]. The equality (A.1b) and inequality (A.1c) can be found in [150, Lemma B.1] and [56, Proposition B.3], respectively. We now conclude with the proof of (A.1d). Let $\boldsymbol{E} = \boldsymbol{D}^* \otimes \boldsymbol{C}$ and $\boldsymbol{\psi} = \operatorname{vec}(\boldsymbol{\Psi})$. Then,

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{C\Psi D}\|_{\mathrm{F}}^4] &= \mathbb{E}[(\boldsymbol{\psi}^*\boldsymbol{E}^*\boldsymbol{E}\boldsymbol{\psi})^2] = \operatorname{Var}(\boldsymbol{\psi}^*\boldsymbol{E}^*\boldsymbol{E}\boldsymbol{\psi}) + (\mathbb{E}[\boldsymbol{\psi}^*\boldsymbol{E}^*\boldsymbol{E}\boldsymbol{\psi}])^2\\ &= 2\|\boldsymbol{E}^*\boldsymbol{E}\|_{\mathrm{F}}^2 + \operatorname{tr}(\boldsymbol{E}^*\boldsymbol{E})^2 = 2\|\boldsymbol{C}\|_{(4)}^4\|\boldsymbol{D}\|_{(4)}^4 + \|\boldsymbol{C}\|_{\mathrm{F}}^4\|\boldsymbol{D}\|_{\mathrm{F}}^4,\end{aligned}$$

where we used a standard result that the variance of a quadratic form with Gaussian random vectors is $2\|\boldsymbol{E}^*\boldsymbol{E}\|_{\mathrm{F}}^2$. □

Lemma A.3. *Let $\boldsymbol{\Omega}_1 \in \mathbb{R}^{k\times(k+p)}$ be a random matrix whose columns are i.i.d. $\mathcal{N}(0, \widetilde{\boldsymbol{K}}_{11})$ random vectors and let $\boldsymbol{B} \in \mathbb{R}^{k\times n}$ be an arbitrary matrix. Then, the following relation hold for $k \geq 1$ and $p \geq 2$:*

$$\mathbb{E}[\|\boldsymbol{\Omega}_1^\dagger\boldsymbol{B}\|_{\mathrm{F}}^2] = \frac{\operatorname{tr}(\boldsymbol{B}^*\widetilde{\boldsymbol{K}}_{11}^{-1}\boldsymbol{B})}{p-1} = \frac{\|\boldsymbol{B}^*\widetilde{\boldsymbol{K}}_{11}^{-1/2}\|_{\mathrm{F}}^2}{p-1}, \tag{A.2a}$$

Additionally, for $p, k \geq 2$ we have:

$$\mathbb{E}[\|\boldsymbol{\Omega}_1^\dagger\|_2^2] \leq \frac{e^2(k+p)}{(p-1)(p+1)}\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2. \tag{A.3a}$$

Finally, for $k \geq 1$ and $p \geq 4$ we have

$$\mathbb{E}[\|\boldsymbol{\Omega}_1^\dagger\|_{(4)}^4] = \frac{(p-1)\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_{\mathrm{F}}^2 + \operatorname{tr}(\widetilde{\boldsymbol{K}}_{11}^{-1})^2}{p(p-1)(p-3)} \leq k\frac{k+p-1}{p(p-1)(p-3)}\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2^2, \tag{A.4a}$$

$$\mathbb{E}[\|\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}^4] = \frac{(p-2)\operatorname{tr}(\widetilde{\boldsymbol{K}}_{11}^{-1})^2 + 2\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_{\mathrm{F}}^2}{p(p-1)(p-3)} \leq k\frac{kp-2k+2}{p(p-1)(p-3)}\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2^2. \tag{A.4b}$$

Proof. First, (A.2a) is proven in [45, Lemma 3.12] and (A.3a) follows from in [117, Lemma 3.1] and the fact $\mathbb{E}[\|\boldsymbol{\Omega}_1^\dagger\|_2^2] \leq \mathbb{E}[\|\boldsymbol{\Psi}^\dagger\|_2^2]\|\boldsymbol{K}_{11}^{-1}\|_2$ where $\boldsymbol{\Psi}$ is a $k\times(k+p)$ standard Gaussian matrix. We proceed to prove the equality in (A.4a) and introduce the random variable $\boldsymbol{X} = \boldsymbol{\Omega}_1\boldsymbol{\Omega}_1^* \sim \mathcal{W}_k(\widetilde{\boldsymbol{K}}_{11}, k+p)$, such that

$$\mathbb{E}\Big[\|\boldsymbol{\Omega}_1^\dagger\|_{(4)}^4\Big] = \mathbb{E}\Big[\|\boldsymbol{X}^{-1}\|_{\mathrm{F}}^2\Big] = \mathbb{E}\Big[\operatorname{tr}(\boldsymbol{X}^{-2})\Big] = \frac{\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_{\mathrm{F}}^2}{p(p-3)} + \frac{\operatorname{tr}(\widetilde{\boldsymbol{K}}_{11}^{-1})^2}{p(p-1)(p-3)},$$

where the last equality follows from [94, Theorem. 2.4.14]. The equality in (A.4b) follows from the fact that $\mathbb{E}[\|\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}^4] = \mathbb{E}[\operatorname{tr}(\boldsymbol{X}^{-1})^2] = \mathbb{E}[\operatorname{tr}\left(\boldsymbol{X}^{-1}\otimes\boldsymbol{X}^{-1}\right)]$. We then exploit a relation on $\mathbb{E}[\boldsymbol{X}^{-1} \otimes \boldsymbol{X}^{-1}]$ [94, Theorem 2.4.14] to obtain

$$\begin{aligned}\mathbb{E}\Big[\|\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}^4\Big] &= \frac{(p-2)\operatorname{tr}(\widetilde{\boldsymbol{K}}_{11}^{-1} \otimes \widetilde{\boldsymbol{K}}_{11}^{-1}) + \operatorname{tr}(\operatorname{vec}(\widetilde{\boldsymbol{K}}_{11}^{-1})\operatorname{vec}(\widetilde{\boldsymbol{K}}_{11}^{-1})^*) + \operatorname{tr}(\boldsymbol{C}_{k\times k}(\widetilde{\boldsymbol{K}}_{11}^{-1} \otimes \widetilde{\boldsymbol{K}}_{11}^{-1}))}{p(p-1)(p-3)} \\ &= \frac{(p-2)\operatorname{tr}(\widetilde{\boldsymbol{K}}_{11}^{-1})^2 + 2\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_{\mathrm{F}}^2}{p(p-1)(p-3)},\end{aligned}$$

where the second equality comes from the relation $\operatorname{tr}(\boldsymbol{C}_{k\times k}(\widetilde{\boldsymbol{K}}_{11}^{-1} \otimes \widetilde{\boldsymbol{K}}_{11}^{-1})) = \|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_{\mathrm{F}}^2$, where $\boldsymbol{C}_{k\times k}$ is the commutation matrix.[1] Finally, the inequalities in (A.4a) and (A.4b) follow from standard norm inequalities. □

The next lemma is a generalization and a consequence of [79, Proposition 10.4], which provides tailbounds on the Frobenius and spectral norms of pseudoinverted standard Gaussian matrices.

Lemma A.4 (Norm bounds for a pseudoinverted scaled Gaussian matrix)**.** *Let* $\boldsymbol{\Omega}_1 \in \mathbb{R}^{k\times(k+p)}$ *be a random matrix whose columns are i.i.d.* $\mathcal{N}(0, \widetilde{\boldsymbol{K}}_{11})$ *random vectors. Then, the following relations hold for* $p \geq 4$ *and all* $t \geq 1$*:*

$$\mathbb{P}\left\{\|\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}} > \sqrt{\frac{3\operatorname{tr}(\widetilde{\boldsymbol{K}}_{11}^{-1})}{p+1}}\,t\right\} \leq t^{-p}, \tag{A.5a}$$

$$\mathbb{P}\left\{\|\boldsymbol{\Omega}_1^\dagger\|_2 > \frac{e\sqrt{(k+p)\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2}}{p+1}t\right\} \leq t^{-(p+1)}, \tag{A.5b}$$

$$\mathbb{P}\left\{\|\boldsymbol{\Omega}_1^\dagger\|_{(4)} > \frac{e\sqrt{(k+p)\|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_{\mathrm{F}}}}{p+1}t\right\} \leq t^{-(p+1)}. \tag{A.5c}$$

Proof. Note that (A.5a) is a restatement of [29, Lemma 3]. Because $\boldsymbol{\Omega}_1 = \widetilde{\boldsymbol{K}}_{11}^{1/2}\boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is a standard Gaussian matrix, it follows that $\|\boldsymbol{\Psi}^\dagger\widetilde{\boldsymbol{K}}_{11}^{-1/2}\|_{(s)} \leq \|\boldsymbol{\Psi}^\dagger\|_2\|\widetilde{\boldsymbol{K}}_{11}^{-1/2}\|_{(s)}$ for any Schatten-s norm. Moreover, the combination with [79, Proposition 10.4] implies

$$\mathbb{P}\left\{\|\boldsymbol{\Psi}^\dagger\|_2 > \frac{e\sqrt{(k+p)}}{p+1}t\right\} \leq t^{-(p+1)},$$

which yield the bounds (A.5b) and (A.5c) using $\|\widetilde{\boldsymbol{K}}_{11}^{-1/2}\|_2^2 = \|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_2$ and $\|\widetilde{\boldsymbol{K}}_{11}^{-1/2}\|_{(4)}^2 = \|\widetilde{\boldsymbol{K}}_{11}^{-1}\|_{\mathrm{F}}$, respectively. □

[1] The commutation matrix $\boldsymbol{C}_{k\times k} \in \mathbb{R}^{k^2\times k^2}$ is a $k \times k$ block matrix, with blocks of size $k \times k$. The (i,j)-block of $\boldsymbol{C}_{k\times k}$ is the matrix $\boldsymbol{E}_{ji}$ with entries $(\boldsymbol{E}_{ji})_{k\ell} = \delta_{jk}\delta_{i\ell}$.

Bibliography

[1] R. J. Adler and J. E. Taylor. *Random fields and geometry.* Springer Monographs in Mathematics. Springer, New York, 2007, pp. xviii+448.

[2] R. H. Affandi, E. Fox, R. Adams, and B. Taskar. “Learning the parameters of determinantal point process kernels”. In: *International Conference on Machine Learning.* PMLR. 2014, pp. 1224–1232.

[3] T. D. Ahle, M. Kapralov, J. B. T. Knudsen, R. Pagh, A. Velingker, D. P. Woodruff, and A. Zandieh. “Oblivious sketching of high-degree polynomial kernels”. In: *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms.* SIAM, Philadelphia, PA, 2020, pp. 141–160.

[4] A. Alaoui and M. W. Mahoney. “Fast Randomized Kernel Ridge Regression with Statistical Guarantees”. In: *Advances in Neural Information Processing Systems.* Vol. 28. 2015.

[5] A. Alexanderian, N. Petra, G. Stadler, and O. Ghattas. “A-optimal design of experiments for infinite-dimensional Bayesian linear inverse problems with regularized ℓ_0-sparsification”. *SIAM J. Sci. Comput.* 36.5 (2014), A2122–A2148.

[6] T. Ando. “Comparison of norms $|||f(A) - f(B)|||$ and $|||f(|A - B|)|||$”. *Math. Z.* 197.3 (1988), pp. 403–409.

[7] M. Aprahamian, D. J. Higham, and N. J. Higham. “Matching exponential-based and resolvent-based centrality measures”. *J. Complex Netw.* 4.2 (2016), pp. 157–176.

[8] H. Avron. “Counting triangles in large graphs using randomized matrix trace estimation”. In: *Workshop on Large-scale Data Mining: Theory and Applications.* Vol. 10. 2010, pp. 10–9.

[9] H. Avron, K. L. Clarkson, and D. P. Woodruff. “Faster kernel ridge regression using sketching and preconditioning”. *SIAM J. Matrix Anal. Appl.* 38.4 (2017), pp. 1116–1138.

[10] H. Avron, K. L. Clarkson, and D. P. Woodruff. “Sharper bounds for regularized data fitting”. In: *Approximation, randomization, and combinatorial optimization. Algorithms and techniques.* Vol. 81. LIPIcs. Leibniz Int. Proc. Inform. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2017, Art. No. 27, 22.

[11] H. Avron and S. Toledo. “Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix”. *J. ACM* 58.2 (2011), Art. 8, 17.

[12] F. R. Bach and M. I. Jordan. “Kernel independent component analysis”. *J. Mach. Learn. Res.* 3.Jul (2002), pp. 1–48.

[13] Z. Bai, M. Fahey, and G. Golub. “Some large-scale matrix computation problems”. *J. Comput. Appl. Math.* 74.1-2 (1996), pp. 71–89.

[14] A. Bakshi, K. L. Clarkson, and D. P. Woodruf. “Low-rank approximation with $1/\epsilon^{1/3}$ matrix-vector products”. In: *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing (STOC).* 2022, pp. 1130–1143.

[15] A. Bakshi and S. Narayanan. “Krylov Methods are (nearly) Optimal for Low-Rank Approximation”. In: *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS).* 2023, pp. 2093–2101.

[16] R. A. Baston and Y. Nakatsukasa. “Stochastic diagonal estimation: probabilistic bounds and an improved algorithm”. *arXiv preprint arXiv:2201.10684* (2022).

[17] M. Bebendorf. *Hierarchical matrices.* Springer, 2008.

[18] M. Bebendorf and W. Hackbusch. “Existence of $\mathcal{H}$-matrix approximants to the inverse FE-matrix of elliptic operators with L^∞-coefficients”. *Numer. Math.* 95 (2003), pp. 1–28.

[19] B. Beckermann and A. Townsend. “Bounds on the singular values of matrices with displacement structure”. *SIAM Rev.* 61.2 (2019), pp. 319–344.

[20] C. Bekas, E. Kokiopoulou, and Y. Saad. “An estimator for the diagonal of a matrix”. *Appl. Numer. Math.* 57.11-12 (2007), pp. 1214–1229.

[21] C. Bekas, A. Curioni, and I. Fedulova. “Low cost high performance uncertainty quantification”. In: *Proceedings of the 2nd Workshop on High Performance Computational Finance.* 2009, pp. 1–8.

[22] R. M. Bell and Y. Koren. “Lessons from the Netflix prize challenge”. *Acm Sigkdd Explorations Newsletter* 9.2 (2007), pp. 75–79.

[23] M. Benzi and P. Boito. “Matrix functions in network analysis”. *GAMM-Mitt.* 43.3 (2020), e202000012, 36.

[24] M. Benzi and C. Klymko. “A matrix analysis of different centrality measures”. *arXiv preprint arXiv:1312.6722* (2014).

[25] R. Bhatia. *Matrix analysis.* Vol. 169. Graduate Texts in Mathematics. Springer-Verlag, New York, 1997, pp. xii+347.

[26] N. Boullé, D. Halikias, and A. Townsend. “Elliptic PDE learning is provably data-efficient”. *Proc. Natl. Acad. Sci. U.S.A.* 120.39 (2023), e2303904120.

[27] N. Boullé, S. Kim, T. Shi, and A. Townsend. “Learning Green’s functions associated with time-dependent partial differential equations”. *J. Mach. Learn. Res.* 23.1 (2022), pp. 9797–9830.

[28] N. Boullé and A. Townsend. “A generalization of the randomized singular value decomposition”. In: *International Conference on Learning Representations.* 2022.

[29] N. Boullé and A. Townsend. “Learning elliptic partial differential equations with randomized linear algebra”. *Found. Comput. Math.* (2022), pp. 1–31.

[30] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. “Metagenes and molecular pattern discovery using matrix factorization”. *Proceedings of the national academy of sciences* 101.12 (2004), pp. 4164–4169.

[31] Z. Bujanovic and D. Kressner. “Norm and trace estimation with random rank-one vectors”. *SIAM J. Matrix Anal. Appl.* 42.1 (2021), pp. 202–223.

[32] D. Burt, C. E. Rasmussen, and M. Van Der Wilk. “Rates of convergence for sparse variational Gaussian process regression”. In: *ICLR.* 2019, pp. 862–871.

[33] J. Calhoun, F. Cappello, L. N. Olson, M. Snir, and W. D. Gropp. “Exploring the feasibility of lossy compression for pde simulations”. *The International Journal of High Performance Computing Applications* 33.2 (2019), pp. 397–410.

[34] S. Cambanis, S. Huang, and G. Simons. “On the theory of elliptically contoured distributions”. *J. Multivariate Anal.* 11.3 (1981), pp. 368–385.

[35] T. Chen, A. Greenbaum, C. Musco, and C. Musco. “Error bounds for Lanczos-based matrix function approximation”. *SIAM J. Matrix Anal. Appl.* 43.2 (2022), pp. 787–811.

[36] T. Chen and E. Hallman. “Krylov-Aware Stochastic Trace Estimation”. *SIAM J. Matrix Anal. Appl.* 44.3 (2023), pp. 1218–1244.

[37] T. Chen and E. Hallman. “Krylov-Aware Stochastic Trace Estimation”. *SIAM J. Matrix Anal. Appl.* 44.3 (2023), pp. 1218–1244.

[38] Y. Chen, E. N. Epperly, J. A. Tropp, and R. J. Webber. “Randomly pivoted Cholesky: Practical approximation of a kernel matrix with few entry evaluations”. *arXiv preprint arXiv:2207.06503* (2022).

[39] M. B. Cohen, C. Musco, and C. Musco. “Input sparsity time low-rank approximation via ridge leverage score sampling”. In: *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms.* SIAM, Philadelphia, PA, 2017, pp. 1758–1777.

[40] A. Cortinovis and D. Kressner. “Low-rank approximation in the Frobenius norm by column and row subset selection”. *SIAM J. Matrix Anal. Appl.* 41.4 (2020), pp. 1651–1673.

[41] A. Cortinovis and D. Kressner. “On randomized trace estimates for indefinite matrices with an application to determinants”. *Found. Comput. Math.* 22.3 (2022), pp. 875–903.

[42] T. A. Davis and Y. Hu. “The University of Florida sparse matrix collection”. *ACM Trans. Math. Software* 38.1 (2011), Art. 1, 25.

[43] N. Deniskin and M. Benzi. “New results and open problems on subgraph centrality”. *J. Comb.* 14.4 (2023), pp. 425–444.

[44] P. Dharangutte and C. Musco. “A tight analysis of Hutchinson’s diagonal estimator”. In: *2023 Symposium on Simplicity in Algorithms (SOSA).* SIAM, Philadelphia, PA, 2023, pp. 353–364.

[45] Y. Diouane, S. Gürol, A. S. Di Perrotolo, and X. Vasseur. “A general error analysis for randomized low-rank approximation methods”. *arXiv preprint arXiv:2206.08793* (2022).

[46] P. Drineas and M. W. Mahoney. “On the Nyström method for approximating a Gram matrix for improved kernel-based learning”. *Journal of Machine Learning Research* 6 (2005), pp. 2153–2175.

[47] H. Driscoll T.A., N. Hale, and L. Trefethen. *Chebfun Guide*. Pafnuty Publications, 2014.

[48] V. Druskin, S. Güttel, and L. Knizhnerman. “Near-optimal perfectly matched layers for indefinite Helmholtz problems”. *SIAM Rev.* 58.1 (2016), pp. 90–116.

[49] E. Dudley, A. K. Saibaba, and A. Alexanderian. “Monte Carlo estimators for the Schatten p-norm of symmetric positive semidefinite matrices”. *Electron. Trans. Numer. Anal.* 55 (2022), pp. 213–241.

[50] C. Eckart and G. Young. “The approximation of one matrix by another of lower rank”. *Psychometrika* 1.3 (1936), pp. 211–218.

[51] E. Epperly. *Stochastic trace estimation.* https://www.ethanepperly.com/index.php/2023/01/26/stochastic-trace-estimation/. Accessed: 17 March 2024.

[52] E. N. Epperly, J. A. Tropp, and R. J. Webber. “XTrace: making the most of every sample in stochastic trace estimation”. *SIAM J. Matrix Anal. Appl.* 45.1 (2024), pp. 1–23.

[53] D. Eriksson, K. Dong, E. Lee, D. Bindel, and A. G. Wilson. “Scaling Gaussian Process Regression with Derivatives”. In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018.

[54] E. Estrada. “Characterization of 3D molecular structure”. *Chemical Physics Letters* 319.5-6 (2000), pp. 713–718.

[55] E. Estrada and D. J. Higham. “Network properties revealed through matrix functions”. *SIAM Rev.* 52.4 (2010), pp. 696–714.

[56] Z. Frangella, J. A. Tropp, and M. Udell. “Randomized Nyström preconditioning”. *SIAM J. Matrix Anal. Appl.* 44.2 (2023), pp. 718–752.

[57] A. Frommer, S. Güttel, and M. Schweitzer. “Convergence of restarted Krylov subspace methods for Stieltjes functions of matrices”. *SIAM J. Matrix Anal. Appl.* 35.4 (2014), pp. 1602–1624.

[58] A. Frommer, K. Lund, and D. B. Szyld. “Block Krylov Subspace Methods for Functions of Matrices II: Modified Block FOM”. *SIAM Journal on Matrix Analysis and Applications* 41.2 (Jan. 2020), pp. 804–837.

[59] A. Frommer, C. Schimmel, and M. Schweitzer. “Analysis of Probing Techniques for Sparse Approximation and Trace Estimation of Decaying Matrix Functions”. *SIAM J. Matrix Anal. Appl.* 42.3 (2021), pp. 1290–1318.

[60] A. Frommer and M. Schweitzer. “Error bounds and estimates for Krylov subspace approximations of Stieltjes matrix functions”. *BIT* 56.3 (2016), pp. 865–892.

[61] A. Frommer and V. Simoncini. “Stopping criteria for rational matrix functions of Hermitian and symmetric matrices”. *SIAM J. Sci. Comput.* 30.3 (2008), pp. 1387–1412.

[62] K. J. Galinsky, G. Bhatia, P.-R. Loh, S. Georgiev, S. Mukherjee, N. J. Patterson, and A. L. Price. “Fast principal-component analysis reveals convergent evolution

of ADH1B in Europe and East Asia". *The American Journal of Human Genetics* 98.3 (2016), pp. 456–472.

[63] A. S. Gambhir, A. Stathopoulos, and K. Orginos. "Deflation as a method of variance reduction for estimating the trace of a matrix inverse". *SIAM J. Sci. Comput.* 39.2 (2017), A532–A558.

[64] Y. Gao and G. Church. "Improving molecular cancer class discovery through sparse non-negative matrix factorization". *Bioinformatics* 21.21 (2005), pp. 3970–3975.

[65] J. Gardner, G. Pleiss, K. Q. Weinberger, D. Bindel, and A. G. Wilson. "GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration". In: *Advances in Neural Information Processing Systems.* Vol. 31. 2018.

[66] M. Gelbrich. "On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces". *Math. Nachr.* 147 (1990), pp. 185–203.

[67] D. A. Girard. "A fast "Monte Carlo cross-validation" procedure for large least squares problems with noisy data". *Numer. Math.* 56.1 (1989), pp. 1–23.

[68] A. Gittens and M. W. Mahoney. "Revisiting the Nyström method for improved large-scale machine learning". *J. Mach. Learn. Res.* 17 (2016), Paper No. 117, 65.

[69] G. H. Golub and R. Underwood. "The block Lanczos method for computing eigenvalues". In: *Mathematical software, III (Proc. Sympos., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1977).* Vol. 39. Publication of the Mathematics Research Center, University of Wisconsin. Academic Press, New York-London, 1977, pp. 361–377.

[70] G. H. Golub and C. F. Van Loan. *Matrix computations.* Fourth. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, 2013, pp. xiv+756.

[71] G. H. Golub and Q. Ye. "An inverse free preconditioned Krylov subspace method for symmetric generalized eigenvalue problems". *SIAM J. Sci. Comput.* 24.1 (2002), pp. 312–334.

[72] L. Grasedyck and W. Hackbusch. "Construction and arithmetics of $\mathcal{H}$-matrices". *Computing* 70.4 (2003), pp. 295–334.

[73] L. Grasedyck, D. Kressner, and C. Tobler. "A literature survey of low-rank tensor approximation techniques". *GAMM-Mitt.* 36.1 (2013), pp. 53–78.

[74] S. Gratton and D. Titley-Peloquin. "Improved bounds for small-sample estimation". *SIAM J. Matrix Anal. Appl.* 39.2 (2018), pp. 922–931.

[75] L. Greengard and V. Rokhlin. "A new version of the fast multipole method for the Laplace equation in three dimensions". In: *Acta numerica, 1997.* Vol. 6. Acta Numer. Cambridge Univ. Press, Cambridge, 1997, pp. 229–269.

[76] M. Gu. "Subspace iteration randomization and singular value problems". *SIAM J. Sci. Comput.* 37.3 (2015), A1139–A1173.

[77] S. Güttel. "Rational Krylov Methods for Operator Functions". PhD thesis. Germany: Institut für Numerische Mathematik und Optimierung, Technische Universität Bergakademie Freiberg, 2010.

[78] S. Güttel. “Rational Krylov approximation of matrix functions: numerical methods and optimal pole selection”. *GAMM-Mitt.* 36.1 (2013), pp. 8–31.

[79] N. Halko, P. G. Martinsson, and J. A. Tropp. “Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions”. *SIAM Rev.* 53.2 (2011), pp. 217–288.

[80] E. Hallman, I. C. F. Ipsen, and A. K. Saibaba. “Monte Carlo methods for estimating the diagonal of a real symmetric matrix”. *SIAM J. Matrix Anal. Appl.* 44.1 (2023), pp. 240–269.

[81] W. K. Härdle and L. Simar. *Applied multivariate statistical analysis.* 4th. Springer, 2015, pp. xiv+580.

[82] T. F. Havel, I. Najfeld, and J.-X. Yang. “Matrix decompositions of two-dimensional nuclear magnetic resonance spectra”. *Proceedings of the National Academy of Sciences* 91.17 (1994), pp. 7962–7966.

[83] E. Herman, A. Alexanderian, and A. K. Saibaba. “Randomization and reweighted ℓ_1-minimization for A-optimal design of linear inverse problems”. *SIAM J. Sci. Comput.* 42.3 (2020), A1714–A1740.

[84] N. J. Higham. *Functions of matrices.* SIAM, Philadelphia, PA, 2008, pp. xx+425.

[85] N. J. Higham. “The scaling and squaring method for the matrix exponential revisited”. *SIAM Rev.* 51.4 (2009), pp. 747–764.

[86] M. Hochbruck and C. Lubich. “On Krylov subspace approximations to the matrix exponential operator”. *SIAM J. Numer. Anal.* 34.5 (1997), pp. 1911–1925.

[87] M. Hochbruck and A. Ostermann. “Exponential integrators”. *Acta Numer.* 19 (2010), pp. 209–286.

[88] R. A. Horn and C. R. Johnson. *Matrix analysis.* Second. Cambridge University Press, Cambridge, 2013, pp. xviii+643.

[89] T. Hsing and R. Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators.* Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2015, pp. xiv+334.

[90] M. F. Hutchinson. “A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines”. *Comm. Statist. Simulation Comput.* 18.3 (1989), pp. 1059–1076.

[91] C. Iyer, A. Gittens, C. Carothers, and P. Drineas. “Iterative Randomized Algorithms for Low Rank Approximation of Tera-scale Matrices with Small Spectral Gaps”. In: *2018 IEEE/ACM 9th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems (scalA).* 2018, pp. 33–40.

[92] S. Jiang, H. Pham, D. Woodruff, and R. Zhang. “Optimal Sketching for Trace Estimation”. In: *Advances in Neural Information Processing Systems.* Vol. 34. 2021, pp. 23741–23753.

[93] P. Kacham and D. Woodruff. “Reduced-Rank Regression with Operator Norm Error”. In: *Conference on Learning Theory.* PMLR. 2021, pp. 2679–2716.

[94] T. Kollo and D. von Rosen. *Advanced multivariate statistics with matrices.* Springer, 2005.

[95] D. Kressner, J. Latz, S. Massei, and E. Ullmann. "Certified and fast computations with shallow covariance kernels". *Found. Data Sci.* 2.4 (2020), pp. 487–512.

[96] C. Lanczos. "An iteration method for the solution of the eigenvalue problem of linear differential and integral operators". *J. Research Nat. Bur. Standards* 45 (1950), pp. 255–282.

[97] A. N. Langville and W. J. Stewart. "The Kronecker product and stochastic automata networks". *J. Comput. Appl. Math.* 167.2 (2004), pp. 429–447.

[98] E.-Y. Lee. "Extension of Rotfel'd theorem". *Linear Algebra Appl.* 435.4 (2011), pp. 735–741.

[99] H. Li and Y. Zhu. "Randomized block Krylov subspace methods for trace and log-determinant estimators". *BIT* 61.3 (2021), pp. 911–939.

[100] 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* 104.51 (2007), pp. 20167–20172.

[101] J. Liesen and Z. Strakoš. *Krylov subspace methods.* Numerical Mathematics and Scientific Computation. Principles and analysis. Oxford University Press, Oxford, 2013, pp. xvi+391.

[102] L. Lin. "Randomized estimation of spectral densities of large matrices made accurate". *Numer. Math.* 136.1 (2017), pp. 183–213.

[103] A. W. Marshall, I. Olkin, and B. C. Arnold. *Inequalities: theory of majorization and its applications.* Second. Springer Series in Statistics. Springer, New York, 2011, pp. xxviii+909.

[104] P.-G. Martinsson and J. A. Tropp. "Randomized numerical linear algebra: foundations and algorithms". *Acta Numer.* 29 (2020), pp. 403–572.

[105] C. A. McCarthy. "c_p". *Israel J. Math.* 5 (1967), pp. 249–271.

[106] A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative risk management.* Revised. Princeton Series in Finance. Princeton University Press, Princeton, NJ, 2015, pp. xix+699.

[107] M. Meier and Y. Nakatsukasa. "Randomized algorithms for Tikhonov regularization in linear least squares". *arXiv preprint arXiv:2203.07329* (2022).

[108] J. Mercer. "Functions of positive and negative type, and their connection the theory of integral equations". *Philos. T. R. Soc. A* 209.441-458 (1909), pp. 415–446.

[109] R. Meyer, C. Musco, and C. Musco. "On the Unreasonable Effectiveness of Single Vector Krylov Methods for Low-Rank Approximation". In: *ACM-SIAM Symposium on Discrete Algorithms (SODA24).* 2024.

[110] R. A. Meyer and H. Avron. "Hutchinson's Estimator is Bad at Kronecker-Trace-Estimation". *arXiv preprint arXiv:2309.04952* (2023).

[111] R. A. Meyer, C. Musco, C. Musco, and D. P. Woodruff. "Hutch++: Optimal Stochastic Trace Estimation". In: *Symposium on Simplicity in Algorithms (SOSA).* SIAM. 2021, pp. 142–155.

[112] R. A. Meyer. *Updates for Hutch++.* https://ram900.hosting.nyu.edu/hutchplusplus/. Accessed: 3 February 2022.

[113] L. Mirsky. “Symmetric gauge functions and unitarily invariant norms”. *Quart. J. Math. Oxford Ser. (2)* 11 (1960), pp. 50–59.

[114] R. J. Muirhead. *Aspects of multivariate statistical theory*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1982, pp. xix+673.

[115] C. Musco and C. Musco. “Randomized Block Krylov Methods for Stronger and Faster Approximate Singular Value Decomposition”. In: *Advances in Neural Information Processing Systems*. Vol. 28. 2015.

[116] C. Musco and C. Musco. “Recursive Sampling for the Nyström Method”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 3833–3845.

[117] Y. Nakatsukasa. “Fast and stable randomized low-rank matrix approximation”. *arXiv preprint arXiv:2009.11392* (2020).

[118] *Pajek datasets*. http://vlado.fmf.uni-lj.si/pub/networks/data/. Accessed: 2023-09-08.

[119] G. Pang, L. Yang, and G. E. Karniadakis. “Neural-net-induced Gaussian process regression for function approximation and PDE solution”. *J. Comput. Phys.* 384 (2019), pp. 270–288.

[120] J. A. de la Peña, I. Gutman, and J. Rada. “Estimating the Estrada index”. *Linear Algebra Appl.* 427.1 (2007), pp. 70–76.

[121] D. Persson, N. Boullé, and D. Kressner. “Randomized Nyström approximation of non-negative self-adjoint operators”. *arXiv preprint arXiv:2404.00960* (2024).

[122] D. Persson, T. Chen, and C. Musco. *Randomized block-Krylov subspace methods for low-rank approximation of matrix functions*. In preparation.

[123] D. Persson, A. Cortinovis, and D. Kressner. “Improved variants of the Hutch++ algorithm for trace estimation”. *SIAM J. Matrix Anal. Appl.* 43.3 (2022), pp. 1162–1185.

[124] D. Persson and D. Kressner. “Randomized low-rank approximation of monotone matrix functions”. *SIAM J. Matrix Anal. Appl.* 44.2 (2023), pp. 894–918.

[125] D. Persson, R. A. Meyer, and C. Musco. “Algorithm-agnostic low-rank approximation of operator monotone matrix functions”. *arXiv preprint arXiv:2311.14023* (2023).

[126] P. Pfeuty. “The one-dimensional Ising model with a transverse field”. *ANNALS of Physics* 57.1 (1970), pp. 79–90.

[127] G. Pleiss, M. Jankowiak, D. Eriksson, A. Damle, and J. Gardner. “Fast matrix square roots with applications to Gaussian processes and Bayesian optimization”. *Advances in Neural Information Processing Systems* 33 (2020), pp. 22268–22281.

[128] F. Pourkamali-Anaraki, S. Becker, and M. Wakin. “Randomized clustered Nystrom for large-scale kernel machines”. *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (2018).

[129] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2006, pp. xviii+248.

[130] W. Reese, A. K. Saibaba, and J. Lee. “Bayesian Level Set Approach for Inverse Problems with Piecewise Constant Reconstructions”. *arXiv preprint arXiv:2111.15620* (2021).

[131] H. Robbins. “A remark on Stirling’s formula”. *Amer. Math. Monthly* 62 (1955), pp. 26–29.

[132] F. Roosta-Khorasani and U. Ascher. “Improved bounds on sample size for implicit matrix trace estimators”. *Found. Comput. Math.* 15.5 (2015), pp. 1187–1212.

[133] F. Roosta-Khorasani, G. J. Székely, and U. M. Ascher. “Assessing stochastic algorithms for large scale nonlinear least squares problems using extremal probabilities of linear combinations of gamma random variables”. *SIAM/ASA J. Uncertain. Quantif.* 3.1 (2015), pp. 61–90.

[134] Y. Saad. “Analysis of some Krylov subspace approximations to the matrix exponential operator”. *SIAM J. Numer. Anal.* 29.1 (1992), pp. 209–228.

[135] Y. Saad. *Iterative methods for sparse linear systems.* Second. SIAM, Philadelphia, PA, 2003, pp. xviii+528.

[136] A. K. Saibaba and A. Międlar. “Randomized low-rank approximations beyond Gaussian random matrices”. *arXiv preprint arXiv:2308.05814* (2023).

[137] A. K. Saibaba. “Randomized subspace iteration: analysis of canonical angles and unitarily invariant norms”. *SIAM J. Matrix Anal. Appl.* 40.1 (2019), pp. 23–48.

[138] A. K. Saibaba, A. Alexanderian, and I. C. F. Ipsen. “Randomized matrix-free trace and log-determinant estimators”. *Numer. Math.* 137.2 (2017), pp. 353–395.

[139] T. Sarlos. “Improved approximation algorithms for large matrices via random projections”. In: *2006 47th annual IEEE symposium on foundations of computer science (FOCS’06).* IEEE. 2006, pp. 143–152.

[140] E. Schmidt. “Zur Theorie der linearen und nichtlinearen Integralgleichungen”. *Math. Ann.* 63.4 (1907), pp. 433–476.

[141] J. Schnack, J. Richter, and R. Steinigeweg. “Accuracy of the finite-temperature Lanczos method compared to simple typicality-based estimates”. *Physical Review Research* 2.1 (2020), p. 013186.

[142] M. Seeger. “Gaussian processes for machine learning”. *International journal of neural systems* 14.02 (2004), pp. 69–106.

[143] D. C. Sorensen and M. Embree. “A DEIM induced CUR factorization”. *SIAM J. Sci. Comput.* 38.3 (2016), A1454–A1482.

[144] G. W. Stewart. *Afternotes on numerical analysis.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996, pp. x+200.

[145] A. M. Stuart. “Inverse problems: a Bayesian perspective”. *Acta Numer.* 19 (2010), pp. 451–559.

[146] C. Thron, S. J. Dong, K. F. Liu, and H. P. Ying. “Padé-Z_2 estimator of determinants”. *Phys. Rev. D* 57 (3 1998), pp. 1642–1653.

[147] A. Townsend. *Pretty functions approximated by Chebfun2.* https://www.chebfun.org/examples/approx2/PrettyFunctions.html. Accessed: 19 December 2023.

[148] A. Townsend and L. N. Trefethen. “Continuous analogues of matrix factorizations”. *Proc. A.* 471.2173 (2015), p. 20140585.

[149] L. N. Trefethen and D. Bau III. *Numerical linear algebra.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997, pp. xii+361.

[150] J. A. Tropp and R. J. Webber. “Randomized algorithms for low-rank matrix approximation: Design, analysis, and applications”. *arXiv preprint arXiv:2306.12418* (2023).

[151] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher. “Fixed-Rank Approximation of a Positive-Semidefinite Matrix from Streaming Data”. In: *Advances in Neural Information Processing Systems.* Vol. 30. 2017.

[152] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher. “Practical sketching algorithms for low-rank matrix approximation”. *SIAM J. Matrix Anal. Appl.* 38.4 (2017), pp. 1454–1485.

[153] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher. “Randomized single-view algorithms for low-rank matrix approximation”. *arXiv preprint arXiv:1609.00048v1* (2016).

[154] S. Ubaru, J. Chen, and Y. Saad. “Fast estimation of $\mathtt{tr}(f(A))$ via stochastic Lanczos quadrature”. *SIAM J. Matrix Anal. Appl.* 38.4 (2017), pp. 1075–1099.

[155] S. Ubaru and Y. Saad. “Applications of trace estimation techniques”. In: *International Conference on High Performance Computing in Science and Engineering.* Springer. 2017, pp. 19–33.

[156] M. Udell and A. Townsend. “Why are big data matrices approximately low rank?” *SIAM J. Math. Data Sci.* 1.1 (2019), pp. 144–160.

[157] V. N. Vapnik. *The nature of statistical learning theory.* 2nd. Springer-Verlag, 2000.

[158] L. N. Vasershtein. “Markov processes over denumerable products of spaces describing large system of automata”. *Problems Inform. Transmission* 5.3 (1969), pp. 47–52.

[159] M. J. Wainwright and M. I. Jordan. “Log-determinant relaxation for approximate inference in discrete Markov random fields”. *IEEE Trans. Signal Process.* 54.6 (2006), pp. 2099–2109.

[160] A. Weiße, G. Wellein, A. Alvermann, and H. Fehske. “The kernel polynomial method”. *Reviews of modern physics* 78.1 (2006), p. 275.

[161] J. Wenger, G. Pleiss, P. Hennig, J. Cunningham, and J. Gardner. “Preconditioning for Scalable Gaussian Process Hyperparameter Optimization”. In: *Proceedings of the 39th International Conference on Machine Learning.* Proceedings of Machine Learning Research. PMLR, 2022, pp. 23751–23780.

[162] C. Williams and D. Barber. “Bayesian classification with Gaussian processes”. *IEEE Trans. Pattern Anal. Mach. Intell.* 20.12 (1998), pp. 1342–1351.

[163] C. Williams and M. Seeger. “Using the Nyström Method to Speed Up Kernel Machines”. In: *Advances in Neural Information Processing Systems.* Vol. 13. 2000.

[164] D. M. Witten, R. Tibshirani, and T. Hastie. “A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis”. *Biostatistics* 10.3 (2009), pp. 515–534.

[165] D. P. Woodruff. “Sketching as a tool for numerical linear algebra”. *Found. Trends Theor. Comput. Sci.* 10.1-2 (2014), pp. iv+157.

[166] L. Wu, J. Laeuchli, V. Kalantzis, A. Stathopoulos, and E. Gallopoulos. “Estimating the trace of the matrix inverse by interpolating from the diagonal of an approximate inverse”. *J. Comput. Phys.* 326 (2016), pp. 828–844.

[167] K. Zhang, I. W. Tsang, and J. T. Kwok. “Improved Nyström Low-rank Approximation and Error Analysis”. In: *International Conference on Machine Learning.* 2008, pp. 1232–1239.

[168] Y. Zhang and W. E. Leithead. “Approximate implementation of the logarithm of the matrix determinant in Gaussian process regression”. *J. Stat. Comput. Simul.* 77.3-4 (2007), pp. 329–348.

Curriculum Vitae

Personal details

Name: Ulf David Persson
Date of birth: 16 December 1996
Place of birth: Linköping, Sweden
Nationality: Swedish

Education

École Polytechnique Fédérale de Lausanne September 2020 - July 2024
Ph.D. Mathematics
Advisor: Prof. Daniel Kressner

New York University February 2023 - July 2023
Visiting research scholar
Advisor: Prof. Christopher Musco

University College London October 2016 - August 2020
MSci Mathematics with Economics
MSci thesis advisor: Prof. Timo Betcke
Awarded First Class Honours

National University of Singapore August 2018 - May 2019
Exchange student
CAP: 4.85/5

Publications and current work

Journal articles

- D. Persson and D. Kressner, *Randomized low-rank approximation of monotone matrix functions*, SIAM Journal on Matrix Analysis and Applications (2023). https://epubs.siam.org/doi/abs/10.1137/22M1523923

- D. Persson, A. Cortinovis, and D. Kressner, *Improved variants of the Hutch++ algorithm for trace estimation*, SIAM Journal on Matrix Analysis and Applications (2022). https://epubs.siam.org/doi/abs/10.1137/21M1447623

In submission

- D. Persson, N. Boullé, and D. Kressner, *Randomized Nyström approximation of non-negative self-adjoint operators*, arXiv preprint, (2024), https://arxiv.org/pdf/2404.00960
- D. Persson, R. A. Meyer, and C. Musco, *Algorithm-agnostic low-rank approximation of operator monotone matrix functions*, arXiv preprint, (2023), https://arxiv.org/pdf/2311.14023

In preparation

- D. Persson, T. Chen, and C. Musco, *Randomized block-Krylov subspace methods for low-rank approximations of matrix functions*, (2024)

Awards

Susan N. Brown Price (UCL) August 2020
Awarded for the best performance in applied mathematics.

UCL Mathematical & Physical Sciences Faculty Dean's List August 2020
For being in the top 5% of graduating students.

Erasmus+ Traineeship Grant May 2019
Received funding to conduct research at Karolinska Institutet.

EPSRC Vacation Bursary May 2018
Received funding to conduct research at UCL.

UCL Dept. of Mathematics 1st Year Undergraduate Prize August 2017
Awarded for excellent exam results.

Teaching experience

École Polytechnique Fédérale de Lausanne

- MSc Thesis co-supervision, Viacheslav Karpii (*Trace estimation of integral operators*), Spring 2024
- TA, MATH-105 (b) Advanced Analysis II, Spring 2024

- Organiser and lecturer, MATH-646 Reading group in Quantum Computing, Fall 2023
- Principal TA, MATH-110 (a) Advanced Linear Algebra I, Fall 2023
- Principal TA, MATH-403 Low-rank approximation techniques, Fall 2022
- Semester project co-supervision, Matthias Zeller (*Randomized algorithms for Gaussian process regression*), Spring 2022
- Principal TA, MATH-202 (c) Analysis III, Spring 2022
- MSc Thesis co-supervision, Tingting Ni (*On the approximation of vector-valued functions by samples*), Fall 2021
- Principal TA, MATH-458 Programming concepts in scientific computing, Fall 2021
- Principal TA, MATH-250 Numerical Analysis, Spring 2021
- Semester project co-supervision, Claudio Boscolo Cegion (*Randomized methods for compressing matrices with hierarchical low-rank structure*), Fall 2020
- Principal TA, MATH-101 (en) Analysis I, Fall 2020

Conference contributions

The $f(A)b$ulous workshop on matrix functions September 2023
Randomized low-rank approximation of monotone matrix functions (talk)
Magdeburg, Germany

Perspectives on Matrix Computations March 2023
Randomized low-rank approximation of monotone matrix functions (talk)
Banff, Canada

Swiss Numerics Day September 2022
Randomized low-rank approximation of monotone matrix functions (poster)
Zurich, Switzerland

ApplMath22 September 2022
Randomized low-rank approximation of monotone matrix functions (poster)
Brijuni, Croatia

Gene Golub SIAM Summer School on Financial Analytics August 2022
Improved variants of the Hutch++ algorithm for trace estimation (poster)
L'Aquila, Italy

EPFL MATHICSE retreat June 2022
Improved variants of the Hutch++ algorithm for trace estimation (talk)
Villars, Switzerland

Conference on random matrix theory and numerical linear algebra June 2022
Improved variants of the Hutch++ algorithm for trace estimation (poster)
Seattle, USA

17th Copper Mountain Conference on Iterative Methods March 2022
Improved variants of the Hutch++ algorithm for trace estimation (talk)
Copper Mountain, USA

Matrix equations and tensor techniques IX September 2021
Improved variants of the Hutch++ algorithm for trace estimation (talk)
Perugia, Italy

Professional experience

Karolinska Institutet May 2019 - September 2019
Visiting undergraduate research
Supervisor: Prof. Roland Nilsson

- Investigated optimization methods to determine metabolic fluxes from measurement data.
- Developed GAMS software to determine metabolic fluxes from measurement data.

University College London June 2018 - August 2018
Undergraduate research
Supervisor: Prof. Erik Burman

- Investigated a numerical method to solve the obstacle problem.

Programming languages

MATLAB, Python, Julia, R, GAMS, STATA.

Languages

Swedish (native), English (fluent), German (C1 level).