UNIVERSITÀ degli STUDI di "ROMA TOR VERGATA"
Facoltà di Scienze Matematiche Fisiche e Naturali

Scuola di Dottorato in Matematica

Projections onto low complexity matrix algebras with applications to regularization, preconditioning and optimization

Candidato:
Stefano Cipolla
0221950
XXX Ciclo

Relatori:
Prof. Carmine Di Fiore
Prof. Paolo Zellini

Coordinatore Scuola di Dottorato:
Prof. Andrea Braides

Anno Accademico 2017/2018

Projections onto low complexity matrix algebras with applications to regularization, preconditioning and optimization

Stefano Cipolla

Data discussione Tesi: 13 Febbraio 2018,
Dipartimento di Matematica, Università di Roma "Tor Vergata".

Ringraziamenti

Guardando indietro nei nostri ricordi d'infanzia, in ognuno di noi è presente il ricordo riguardante la prima volta in cui un conoscente dei nostri genitori o un parente ci ha posto la fatidica domanda "Cosa vuoi fare da grande?". Nel mio caso, devo essere sincero, i primi ricordi riguardanti le risposte fornite a questa domanda hanno a che fare con gli aerei e il desiderio di pilotarli. La cosa singolare – o forse no –, è che riesco a rintracciare nella memoria un punto in cui le mie risposte hanno deviato sempre di più su ambiti legati alla scienza.

Oggi, mentre mi accingo a finire questo percorso di studi, mentre mi preparo alla difesa della mia tesi di dottorato, ad essere sincero, non riesco a pensare ad altro se non a quanto sono stato fortunato: poche persone possono dire di aver realizzato uno dei loro sogni d'infanzia. Anche se nei miei ricordi, a dire la verità, l'immagine di ciò che volevo fare da grande era rappresentata da quella classica del camice e del laboratorio, mi rendo conto, forse, che quello che mi sono ritrovato a fare in questi ultimi tre anni non è molto diverso da quello che fa un chimico in un laboratorio.

Costruire nuove conoscenze usandone di note, a dire il vero, non mi è sembrata essere una cosa semplice. La strada per il risultato finale mi è sembrata essere spesso piuttosto lunga, faticosa e costellata da spasmodiche inquietudini legate all'atto pratico della Ricerca, e insite, probabilmente, in maniera fisiologica in questa.

Fortunatamente, la fatica e l'affanno lasciatomi sotto pelle da questo percorso che sta per concludersi, non riescono a farmi perdere la lucidità per capire quanto le esperienze collezionate lungo questo siano state profondamente formanti dal punto di vista scientifico e umano e, per capire, quanto la persona che sono oggi sia stata plasmata da queste e indissolubilmente legata a queste.

Nel prendere coscienza della persona che sono oggi, sarebbe da pazzi non riconoscere quanto siano state importanti le persone incontrate lungo questo cammino e quanto immensamente grato io debba essere – e sono – verso di loro. Al Professor Di Fiore, per avermi insegnato praticamente tutto e per avermi indirizzato verso le ricerche contenute in questa tesi, al Professor Zellini per avermi sempre spronato a non perdere d'occhio la visione d'insieme, a Francesco, per avermi fatto capire che poi si deve diventare grandi e che forse un giorno la Ricerca camminerà – chissà – anche sulle nostre spalle, vorrei rivolgere il mio più profondo e sentito grazie. Altrettanto sentito e doveroso è il grazie che rivolgo alla Professoressa Redivo-Zaglia e al Professor Tyrtyshnikov per aver significativamente migliorato la qualità di questa tesi con i loro preziosi e puntuali consigli.

Un ultimo sentito, profondo e doveroso grazie va alla mia famiglia, per avermi sempre consigliato, incoraggiato e sostenuto in tutte le cose della mia vita.

Roma, 13 febbraio 2018

Contents

Introduction

Simulations of real world phenomena often require the solution of *large scale* numerical problems and, equally often, the main computational tasks are connected with computational problems involving large scale matrices.

The dimensions of the problems which have to be faced in this context, typically consisting in the solution of systems of equations, make *direct methods* less affordable and hence *iterative methods* become more competitive.

Iterative methods could suffer of an extremely *low rate of convergence* and/or high computational cost per step. For these reasons the employment of *ad hoc efficiency improvement techniques* becomes necessary. The problem of devising a suitable efficiency improvement strategy for a given iterative method can be considered one of the core tasks in numerical analysis. Taking a deeper look, one can consider these strategies as the instruments which have made the simulation of real world phenomena *effective* and have permitted to develop most of nowadays technologies.

The preconditioning of *Krylov* solvers for linear systems by means of *suitable approximations* of the original matrix ([91, 84, 3, 2]) can be considered as one in the paradigm of efficiency improvement techniques and has inspired much of the literature in the last fifty years. For this reason, the problem of accurately approximating – in some sense – a given matrix with a *lower complexity* one has become ubiquitous in numerical linear algebra.

Usually, the matrices of linear systems arising from applications exhibit some kind of *structure* – more or less evident – which must be exploited in order to produce the desired approximations. The task of producing *accurate and low complexity* approximations corresponding to a given structure has been the topic of an intense investigation in recent years and it has produced a constellation of elegant and useful mathematical results.

A class of remarkable results in this context are those connecting *Toeplitz* structure – naturally related with shift-invariant phenomena – with *Circulant* structure: under standard hypotheses, the operation of projecting a

Toeplitz matrix onto the *algebra* of Circulant matrices produces a low complexity approximation which is an ideal preconditioner for Krylov methods. The projected matrix is able to catch the spectral distribution of the original matrix very accurately when the dimension of the Toeplitz matrix grows.

More in detail (see [22] for an exhaustive survey on the topic), consider f in the Banach space of all 2π periodic continuous real-valued functions on $[-\pi, \pi]$ equipped with the $\|\cdot\|_\infty$ norm, i.e.,

$$f(x) = \sum_{k=-\infty}^{+\infty} t_k e^{\mathbf{i}kx}, \;\; t_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-\mathbf{i}kx}dx,$$

and consider the associated Toeplitz matrix

$$[T_n(f)]_{ij} = \{t_{i-j}\} \text{ for } i, j \in \{0, \ldots, n-1\}.$$

Define the algebra of Circulant matrices as

$$\mathcal{C} := \{Fd(\mathbf{z})F^H, \; \mathbf{z} \in \mathbb{C}^n\},$$

where $d(\mathbf{z}) := \mathrm{diag}(z_i, \; i = 1, \ldots, n)$ and F is the Fourier matrix. The Circulant algebra is a *low complexity matrix algebra* since the matrix vector product $C\mathbf{x}$ can be performed in $O(n\log(n))$ FLoating point OPerations (FLOPs) for all $C \in \mathcal{C}$ [84, 3, 22, 2]. Define, finally, $\mathcal{C}_{T_n(f)}$ as the projection of $T_n(f)$ onto $\mathcal{C}$:

$$\mathcal{C}_{T_n(f)} := argmin_{X\in\mathcal{C}}\|X - T_n(f)\|_F \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm. The following theorem holds:

Theorem 0.0.1. *If $f > 0$, for all $\epsilon > 0$ there exist $M, N \in \mathbb{N}$ such that, for all $n > N$, at most M eigenvalues of the matrix $\mathcal{C}_{T_n(f)}^{-1}T_n(f) - I$ have larger absolute value than ϵ, i.e., $\mathcal{C}_{T_n(f)}^{-1}T_n(f)$ have clustered spectra around 1.*

In other words, under the hypothesis of Theorem 0.0.1, we can say that the number of iterations needed to solve the positive definite linear system $T_n(f)\mathbf{x} = \mathbf{b}$ using the *Conjugate Gradient* method preconditioned with $\mathcal{C}_{T_n(f)}$, is bounded from a constant independent from n, the size of the problem.

Projections onto Circulant and other low complexity matrix algebras have been deeply investigated and profitably used in the last two decades not only as preconditioners for linear systems solvers but even in connection with regularization theory for ill–conditioned linear systems [62, 36, 49, 35].

Observe that the operation of projecting a given matrix $A \in \mathbb{C}^{n\times n}$ onto a given algebra $\mathcal{L}$ of matrices simultaneously diagonalized by a fixed matrix U unitary of low complexity, i.e., $\mathcal{L} := \operatorname{sd} U = \{Ud(\mathbf{z})U^H,\ \mathbf{z} \in \mathbb{C}^n\}$, is defined in general. It is then natural to try to understand which are the properties inherited by the projected matrix $\mathcal{L}_A$, how these relate with the choice of the subspace $\mathcal{L}$ – projections are well defined for arbitrary closed convex subsets of a Hilbert space –, and in which applications the projected matrix represents a satisfactory approximation of the original matrix.

More recently, projections onto low complexity matrix algebras have been employed in connection with Quasi-Newton methods in solving the unconstrained minimization problem for large scale smooth functions [40, 7]. The key observation, which justifies their use in this context, can be traced in the remarkable fact that, in general, projecting a given Hermitian matrix onto a generic $\operatorname{sd} U$ algebra, even if it does not provide an accurate approximation of its spectrum as in the Toeplitz–Circulant case, is able to produce, in a precise sense, *global* spectral approximations. In most recent developments of the use of matrix projections in connection with minimization algorithms [42, 27], it has been introduced the idea of *adaptive* low complexity matrix algebras, i.e., low complexity spaces are constructed adaptively – step by step – on the sequence of Hessian approximations produced by the Quasi-Newton updating scheme in order to yield as more as possible accurate approximations of these matrices and maintain a *Quasi-Newton rate of convergence.*

The aim of this dissertation is to provide an up-to-date overview of techniques and results connected with the general theory of projections onto matrix algebras and of some applications where their use produces evident computational benefits in gaining the efficiency of iterative methods. In particular this dissertation collects and presents some new results obtained by the author during his Ph.D. studies.

The Chapters 2,3,4 and 5 are extracted, sometimes verbatim, from the following works:

- **Chapter 2:** *Low complexity matrix projections preserving actions on vectors.* Cipolla S., Di Fiore C., Zellini P. , submitted for publication, 2017.
- **Chapter 3:** *Regularizing properties of a class of matrices including the Optimal and the Superoptimal preconditioners.* Cipolla S., Di Fiore C., Zellini P., submitted for publication, 2017.
- **Chapter 4:** *Euler-Richardson method preconditioned by weakly stochastic matrix algebras: a potential contribution to Pagerank computation.*

Cipolla S., Di Fiore C., Tudisco, F., Electronic Journal of Linear Algebra, 2017(32): 254-272.

- **Chapter 5:** *Updating Broyden Class-type descent directions by Householder adaptive transforms.* Cipolla S., Di Fiore C., Zellini P., submitted for publication, 2017.

More in detail the dissertation is structured as follows:

- Chapter 1: in the first part we review projection onto matrix spaces from a general point of view and its properties. In the second part we focus on $\mathrm{sd}\, U$ algebras and give a survey on some not so well known results concerning the quality of the spectrum of the projected matrix when regarded as a global approximant of the original one.

- Chapter 2: in this chapter we fully solve a problem originally introduced in [27]. Given a symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$ and $\mathbf{v} \in \mathbb{R}^n$, find a low complexity unitary matrix U such that defining $\mathcal{L} = \mathrm{sd}\, U$, we have $\mathcal{L}_A\mathbf{v} = A\mathbf{v}$, where $\mathcal{L}_A$ is the projection of A onto $\mathcal{L}$ in Frobenius norm. Actually, by using the Arnoldi procedure for Block-Krylov subspaces, we solve the above problem in the more general case where $\mathbf{v}$ is replaced by a matrix $V \in \mathbb{R}^{n\times r}$. The solution matrix U turns out to be real and can be written as the product of $2r$ Householder matrices.

- Chapter 3: in this chapter given a positive definite matrix $A \in \mathbb{R}^{n\times n}$, we introduce a class of matrices related to A obtained by suitably combining projections of its powers onto algebras of matrices simultaneously diagonalized by a unitary transform. After a detailed theoretical study of some spectral properties of the matrices of this class – which suggest their use as regularizing preconditioners –, we prove that such matrices can be cheaply computed when the matrix A has Toeplitz structure. We provide numerical evidence of the goodness of the proposed approach in regularizing procedures.

- Chapter 4: in this chapter we address the efficient solution of M-matrix linear system $Mx = y$, where $M = I - \tau A$, being A a column stochastic matrix and τ a positive coefficient smaller than one. The Pagerank centrality index on graphs is a relevant example where this problem appears. Previous investigations have shown that the Euler-Richardson (ER) method can be considered in order to approach the Pagerank computation problem by means of preconditioning strategies. In this work we observe indeed that the classic power method can be embedded into the ER scheme, by means of a suitable simple preconditioner. Therefore, we propose a new preconditioner based on fast Householder transformations and the concept of low complexity weakly stochastic algebras, which gives rise to an effective alternative to the power method for large-scale sparse problems. We give detailed mathematical reasonings for this choice and discuss the convergence properties. Numerical tests performed on real-world datasets are presented, showing the advantages given by the use of the proposed Householder-Richardson method.

- Chapter 5: in this chapter we introduce and study a novel Quasi Newton minimization methods based on a Hessian approximation Broyden Class-*type* updating scheme, where a suitable matrix $\tilde{B}_k$ is updated instead of the current Hessian approximation B_k. We identify conditions which imply the convergence of the algorithm and, if exact line search is chosen, its quadratic termination. By a remarkable connection between the projection operation and Krylov spaces studied in Chapter 2, such conditions can be ensured using low complexity matrices $\tilde{B}_k$ obtained projecting B_k onto algebras of matrices diagonalized by products of a constant number of Householder matrices adaptively chosen step by step. Extended experimental tests show that the introduction of the adaptive criterion considerably improves the performance of the minimization scheme when compared with a non-adaptive choice and confirm that the method could be particularly suitable to solve large scale problems.

Table of Notations

- ":=" means "by definition".
- $\mathbb{C}^{n\times n}$, $\mathbb{R}^{n\times n}$ are the set of $n \times n$ complex or real matrices.
- $\mathbb{C}^n$, $\mathbb{R}^n$ are the set of length n complex or real vectors.
- I_n is the $n \times n$ identity matrix.
- If $A \in \mathbb{C}^{n\times n}$, $\operatorname{tr}(A)$ and $\det(A)$ are the trace and the determinant of A.
- If $A, B \in \mathbb{C}^{n\times n}$, $(A, B) := \operatorname{tr}(A^H B)$.
- If $A \in \mathbb{C}^{n\times n}$, $\|A\|_F := \sqrt{\operatorname{tr}(A^H A)}$ (Frobenius Norm).
- $\mathbf{e}$ is the vector of all ones.
- $\mathbf{e}_j$ is the j-th vector of the canonical basis.
- If $A \in \mathbb{C}^{n\times n}$, we write $A \geq 0$ ($A > 0$) if A is Hermitian positive semi-definite (positive definite).
- If $A \in \mathbb{C}^{n\times n}$, we write A spd if A is real symmetric positive definite.
- If $A \in \mathbb{C}^{n\times n}$, $\boldsymbol{\lambda}(A)$ is the vector of the eigenvalues of A.
- $\lambda_j(A)$ is the j-th entry of $\boldsymbol{\lambda}(A)$.
- $\lambda(A)$ is the generic eigenvalue of A.
- If A is Hermitian positive definite, if not otherwise stated, the eigenvalues are ordered non-increasingly, i.e., $\lambda_1(A) \geq \cdots \geq \lambda_n(A)$.
- If $\mathbf{z} \in \mathbb{C}^n$, $d(\mathbf{z})$ is the diagonal matrix whose diagonal entries are the elements of $\mathbf{z}$.
- If $A \in \mathbb{C}^{n\times n}$, $d(A)$ is the vector of the diagonal elements of A.
- If $A \in \mathbb{C}^{n\times n}$, $\chi(A)$ is the computational cost of the product $A \times \mathit{vector}$.

- If $\mathcal{L} \subset \mathbb{C}^{n\times n}$, $A \in \mathbb{C}^{n\times n}$, we write $\mathcal{L}_A$ for the projection of A onto $\mathcal{L}$ in the Frobenius norm (if it is well defined).
- If $U \in \mathbb{C}^{n\times n}$ unitary, we write $\operatorname{sd} U := \{Ud(\mathbf{z})U^H \;:\; \mathbf{z} \;\in\; \mathbb{C}^n\}$.
- If $A \in \mathbb{C}^{n\times n}$, $\mu_2(A)$ is the condition number in the 2-norm of A.
- If $A, B \in \mathbb{C}^{n\times n}$, we write $A \approx B$ if they are similar.
- If $A \in \mathbb{C}^{n\times n}$, $\rho(A)$ denotes the spectral radius of A.

Chapter 1

Matrix Projections: A survey

The aim of this chapter is to survey important results concerning the properties that the projected matrices inherit from the original one.

1.1 Hilbert Spaces and Projections

In this section we recall some results concerning Hilbert spaces and projections. Even if in the following of this thesis we will use just spaces of finite dimension – for which Pythagoras Theorem is fairly enough to speak about projections –, we prefer to present here the theory in the greatest possible generality in order to ease the work of the interested reader in generalizing results contained in this section for spaces with infinitely many dimensions. We borrow this section from [19] and refer there for the proofs of the results stated in the following.

Theorem 1.1.1 (Hilbert's Projection Theorem)**.** *Let $\mathcal{H}$ be a Hilbert space with respect to a inner product $(\ ,\ )$ and $\mathcal{K} \subset \mathcal{H}$ be a nonempty closed convex set. Then for any $x \in \mathcal{H}$ there is a unique element $\mathcal{K}_x \in \mathcal{K}$, called the orthogonal projection of x onto $\mathcal{K}$, such that*

$$\|x - \mathcal{K}_x\| = \inf_{y \in \mathcal{K}} \|x - y\|, \tag{1.1}$$

where $\|\ \|$ is the norm induced by $(\ ,\ )$. Moreover, $\mathcal{K}_x$ is the unique solution of the problem

$$\begin{cases} y \in \mathcal{K} \\ (x - y, z - y) \leq 0 \text{ for all } z \in \mathcal{K}. \end{cases} \tag{1.2}$$

Corollary 1.1.2. *Let $\mathcal{H}$ be a Hilbert space and $\mathcal{K} \subset \mathcal{H}$ a nonempty closed convex set. Then*

$$(x - y, \mathcal{K}_x - \mathcal{K}_y) \geq \|\mathcal{K}_x - \mathcal{K}_y\|^2. \tag{1.3}$$

Corollary 1.1.3. *Let $\mathcal{M}$ be a nonempty closed subspace of a Hilbert space $\mathcal{H}$. Then, for every $x \in \mathcal{H}$, $\mathcal{M}_x$ is the unique solution of the problem*

$$\begin{cases} y \in \mathcal{M} \\ (x-y, v) = 0 \text{ for all } v \in \mathcal{M}. \end{cases} \tag{1.4}$$

Let us recall the following properties of Hilbert spaces:

- if $\mathcal{M}$ is a subspace of $\mathcal{H}$, then $cl(\mathcal{M})$ is a subspace of $\mathcal{H}$, where the closure is in the topology induced by the scalar product;
- for any subset $\mathcal{A} \subset \mathcal{H}$ let us set

  $$\mathcal{A}^{\perp} = \{x \in \mathcal{H} \text{ s.t. } (x, a) = 0 \text{ for all } a \in \mathcal{A}\};$$

  then, for any $\mathcal{A}, \mathcal{B} \subset \mathcal{H}$ we have

  1. $\mathcal{A}^{\perp}$ is a closed subspace of $\mathcal{H}$ and $cl(\mathcal{A})^{\perp} = \mathcal{A}^{\perp}$;
  2. $\mathcal{A} \subset \mathcal{B} \Rightarrow \mathcal{B}^{\perp} \subset \mathcal{A}^{\perp}$;
  3. $(\mathcal{A} \cup \mathcal{B}) = \mathcal{A}^{\perp} \cap \mathcal{B}^{\perp}$;

Proposition 1.1.4. *Let $\mathcal{M}$ be a nonempty closed subspace of a Hilbert space $\mathcal{H}$. Then the following statements hold:*

1. *For every $x \in \mathcal{H}$ there exists a unique pair $(\mathcal{M}_x, \mathcal{M}_x^{\perp}) \in \mathcal{M} \times \mathcal{M}^{\perp}$ such that $x = \mathcal{M}_x + \mathcal{M}_x^{\perp}$ (Riesz Orthogonal Decomposition);*
2. *$\mathcal{M}_{(\cdot)} : \mathcal{H} \to \mathcal{H}$ is linear and $\|\mathcal{M}_x\| \leq \|x\|$ for all $x \in \mathcal{H}$.*
3. *$\mathcal{M}_{(\cdot)} \circ \mathcal{M}_{(\cdot)} = \mathcal{M}_{(\cdot)}$, $\ker \mathcal{M}_{(\cdot)} = \mathcal{M}_{(\cdot)}^{\perp}$, $\mathcal{M}_{\mathcal{H}} = \mathcal{M}$.*

1.2 Projections onto general spaces of matrices

We borrow this section from [44] and refer there for the proofs of the results stated in the following. Given a square matrix $A \in \mathbb{C}^{n\times n}$ and $\mathcal{L}$ a linear subspace of $\mathbb{C}^{n\times n}$ of dimension m, we are interested in the elements of $\mathcal{L}$ which best approximate A in some given norm. Consider the minimum problem

$$\min_{X\in\mathcal{L}} ||X - A|| \tag{1.5}$$

where $||\cdot||$ is a matrix norm in $\mathbb{C}^{n\times n}$. Let us remember that $\mathbb{C}^{n\times n}$ with the inner product

$$(A, A') = \sum_{r,t=1}^{n} \overline{a}_{rt} a'_{rt} = \operatorname{tr}(A^H A'), \quad A, A' \in \mathbb{C}^{n\times n}$$

is a Hilbert space. This inner product induces the so-called Frobenius norm:

$$||A||_F^2 = \sum_{r,t=1}^{n} \overline{a}_{rt} a_{rt} = \sum_{r,t=1}^{n} |a_{rt}|^2 = \text{tr}\,(A^H A).$$

Note that each linear subspace $\mathcal{L}$ is a closed subspace of $\mathbb{C}^{n\times n}$ with respect to $||\cdot||_F$. We can hence apply results of Section 1.1.

Theorem 1.2.1. *If the norm in (1.5) is the Frobenius norm, there exists a unique matrix $\mathcal{L}_A$ solving problem (1.5). The matrix $\mathcal{L}_A$, which is referred to as the best least-squares (l.s. in short) fit to A from $\mathcal{L}$, is equivalently defined by the condition*

$$(A - \mathcal{L}_A, X) = 0, \ \ \forall X \ \in \ \mathcal{L}, \tag{1.6}$$

i.e., $\mathcal{L}_A$ is the unique element of $\mathcal{L}$ such that $A - \mathcal{L}_A$ is orthogonal to $\mathcal{L}$. If $\mathcal{L}$ is spanned by the matrices J_k, $k = 1, \dots, m$, then

$$\mathcal{L}_A = \sum_{k=1}^{m} [B_{\mathcal{L}}^{-1} \mathbf{c}_{\mathcal{L},A}]_k J_k \tag{1.7}$$

where $B_{\mathcal{L}}$ is the $m \times m$ Hermitian positive definite matrix

$$[B_{\mathcal{L}}]_{ij} = \sum_{r,t=1}^{n} \overline{[J_i]}_{rt} [J_j]_{rt} = (J_i, J_j), \ i, j = 1, \dots, m \tag{1.8}$$

and $\mathbf{c}_{\mathcal{L},A}$ is the $m \times 1$ vector whose entries are

$$[\mathbf{c}_{\mathcal{L},A}]_i = \sum_{r,t=1}^{n} \overline{[J_i]}_{rt} a_{rt} = (J_i, A), \ i = 1, \dots, m \tag{1.9}$$

(notice that $B_{\mathcal{L}}$ and $\mathbf{c}_{\mathcal{L},A}$ depend upon the choice of the J_k's).

Proof. A direct proof for Theorem 1.2.1 is obtained by Corollary 1.1.3 and using the identity:

$$||\sum_{k=1}^{m} z_k J_k - A||_F^2 = \mathbf{z}^H B_{\mathcal{L}} \mathbf{z} - 2\,\text{Re}(\mathbf{z}^H \mathbf{c}_{\mathcal{L},A}) + ||A||_F^2, \ \ \mathbf{z} \ \in \mathbb{C}^m. \tag{1.10}$$

In details, the previous identity with $A = 0$ and the linear independence of the J_k imply that

$$\mathbf{z}^H B_{\mathcal{L}} \mathbf{z} = ||\sum_{k=1}^{m} z_k J_k||_F^2 > 0, \ \ \forall\, \mathbf{z} \ \in \mathbb{C}^m, \ \ \mathbf{z} \neq 0.$$

Thus, $B_{\mathcal{L}}$ is positive definite and the matrix

$$\mathcal{L}_A = \sum_{k=1}^{m} [B_{\mathcal{L}}^{-1}\mathbf{c}_{\mathcal{L},A}]_k J_k$$

is well defined. Moreover, by (1.10), we have, for an "increment" $\sum_{k=1}^{m} z_k J_k$,

$$||\mathcal{L}_A + \sum_{k=1}^{m} z_k J_k - A||_F^2 = ||\mathcal{L}_A - A||_F^2 + \mathbf{z}^H B_{\mathcal{L}} \mathbf{z} > ||\mathcal{L}_A - A||_F^2, \ \ \forall\, \mathbf{z} \in \mathbb{C}^m, \ \mathbf{z} \neq 0.$$

□

$B_{\mathcal{L}}$ is usually known in literature as *Gram* matrix and its rank structure has been investigated for *fast algebras* in [34].
We have, moreover, using Corollary 1.1.2 and Proposition 1.1.4, for every subspace $\mathcal{L} \subset \mathbb{C}^{n\times n}$ and $A, B \subset \mathbb{C}^{n\times n}$, that

- $\mathrm{tr}\,(A^H A) \geq \mathrm{tr}\,(\mathcal{L}_A^H \mathcal{L}_A)$ even if in general $\mathcal{L}_A^H \notin \mathcal{L}$;
- $\mathrm{tr}\,((A-B)^H(\mathcal{L}_A - \mathcal{L}_B)) \geq \mathrm{tr}\,((\mathcal{L}_A - \mathcal{L}_B)^H(\mathcal{L}_A - \mathcal{L}_B)) \geq 0.$

Remark 1. *If A is real and there exist real matrices J_k spanning $\mathcal{L}$, then also $\mathcal{L}_A$ is real. In fact* $\mathrm{Re}\,\mathcal{L}_A \in \mathcal{L}$ *and*

$$||\mathcal{L}_A - A||_F^2 = ||\,\mathrm{Re}\,\mathcal{L}_A - A||_F^2 + ||\,\mathrm{Im}\,\mathcal{L}_A||_F^2.$$

$\mathrm{Im}\,\mathcal{L}_A \neq 0$ *would imply that* $\mathrm{Re}\,\mathcal{L}_A$ *approximates A better than $\mathcal{L}_A$, which is absurd. Observe, moreover, that A real $\Rightarrow$ $\mathcal{L}_A$ real is true in the more general setting where $\overline{\mathcal{L}} \subseteq \mathcal{L}$ being $\mathcal{L}$ a subspace of $\mathbb{C}^{n\times n}$. The proof follows by the uniqueness of the projection.*

In the following, when we refer to minimum problem (1.5), we assume that the norm is the Frobenius norm. The uniqueness result in Theorem 1.2.1 implies that possible symmetries of A are inherited by its best l.s. fit $\mathcal{L}_A$ under suitable assumptions on $\mathcal{L}$, as is stated in the following lemma. The symbol J is used to denote the reversion matrix $[J]_{ij} = \delta_{i,n+1-j}$, $i, j = 1, \ldots, n$.

Lemma 1.2.2. *The following implications hold:*

1. *Assume $X^T \in \mathcal{L}$, $\forall\, X \in \mathcal{L}$ ($\mathcal{L}$ is closed under transposition):*
   $A^T = \pm A \Rightarrow \mathcal{L}_A^T = \pm \mathcal{L}_A$;
2. *Assume $JX^TJ \in \mathcal{L}$, $\forall\, X \in \mathcal{L}$ ($\mathcal{L}$ is closed under transposition through the secondary diagonal):*
   $A^T = \pm JAJ \Rightarrow \mathcal{L}_A^T = \pm J\mathcal{L}_A J$;
3. *Assume $X^H \in \mathcal{L}$, $\forall\, X \in \mathcal{L}$ ($\mathcal{L}$ is closed under conjugate transposition):*
   $A^H = \pm A \Rightarrow \mathcal{L}_A^H = \pm \mathcal{L}_A$;

Proof. see [44] □

1.3 Class $\mathbb{V}$ spaces and *-spaces

Definition 1. *Define a space of class $\mathbb{V}$, a space $\mathcal{L}$ of dimension n such that there exists $\mathbf{v} \in \mathbb{C}^n$ satisfying $\mathbf{v}^T J_k = \mathbf{e}_k^T$, $k = 1, \ldots, n$, for n linearly independent matrices $J_k \in \mathcal{L}$.*

As the J_k's span $\mathcal{L}$, the conditions $\mathbf{v}^T J_k' = \mathbf{e}_k^T$, $J_k' \in \mathcal{L}$, imply $J_k' = J_k$, $\forall\, k$, and the matrices J_k are uniquely determined. The matrix $\mathcal{L}_{\mathbf{v}}(\mathbf{z}) = \sum_{k=1}^n z_k J_k$ for which

$$\mathbf{v}^T \mathcal{L}_{\mathbf{v}}(\mathbf{z}) = \mathbf{z}^T \tag{1.11}$$

is referred to as the matrix of $\mathcal{L}$ whose $\mathbf{v}$-row is $\mathbf{z}^T$. Notice that two matrices of $\mathcal{L}$ with the same $\mathbf{v}$ row are equal and that $\mathcal{L}_{\mathbf{v}}(\mathbf{e}_k) = J_k$. If $\mathbf{v}$ is one of the vectors of the canonical basis of $\mathbb{C}^n$, say $\mathbf{e}_h$, then $\mathcal{L}$ is called an h-space.
In more intuitive terms, in a space of class $\mathbb{V}$, the generic matrix is determined by a linear combination of its rows, whereas only one row (the h row) is sufficient to define the generic matrix of a h-space.

Theorem 1.3.1. *If $\mathcal{L} = \{Md(\mathbf{z})M^{-1},\ \mathbf{z} \in \mathbb{C}^n\}$ for a non-singular matrix M, then $\mathcal{L} \in \mathbb{V}$. More specifically, for any fixed vector $\mathbf{v}$ such that $[M^T\mathbf{v}]_j \neq 0\ \forall\, j$, the matrix $\mathcal{L}_{\mathbf{v}}(\mathbf{z})$ is well defined and can be represented as:*

$$\mathcal{L}_{\mathbf{v}}(\mathbf{z}) = Md(M^T\mathbf{z})d(M^T\mathbf{v})^{-1}M^{-1}. \tag{1.12}$$

Moreover, $\mathcal{L}$ is a h-space iff $[M]_{hj} \neq 0$, $\forall\, j$.

Proof. The matrices

$$J_k := Md(M^T\mathbf{e}_k)d(M^T\mathbf{v})^{-1}M^{-1}, \ \ k = 1, \ldots, n$$

belong to $\mathcal{L}$, satisfy the identities

$$\begin{aligned} \mathbf{v}^T J_k = \mathbf{v}^T Md(M^T\mathbf{e}_k)d(M^T\mathbf{v})^{-1}M^{-1} = \\ \mathbf{e}_k^T Md(M^T\mathbf{v})d(M^T\mathbf{v})^{-1}M^{-1} = \mathbf{e}_k, \ \ k = 1, \ldots, n, \end{aligned} \tag{1.13}$$

and span $\mathcal{L}$. For the last assertion notice that, if $\mathcal{L}$ is a h-space, then there exists a $\mathbf{z}_k$ such that $\mathbf{e}_k^T = \mathbf{e}_h^T Md(\mathbf{z}_k)M^{-1}$, $k = 1, \ldots, n$, and thus $d(M^T\mathbf{e}_h)$ must be non-singular. □

It is interesting to observe that the concept of non derogatority can be characterized in terms of space of class $\mathbb{V}$. Let $\{p(X)\}$ denote the space of polynomials $p(X)$ in X with complex coefficients. The following theorem holds:

Theorem 1.3.2. *X is non-derogatory iff $\{p(X)\} \in \mathbb{V}$.*

Proof. see [44] □

Now a list of algebraic properties will be given in order to simplify the analysis of the best least square fit $\mathcal{L}_A$ to A from a space $\mathcal{L}$ of class $\mathbb{V}$. In the following let us denote by P_k the $n \times n$ matrices related to the J_k by the identities $\mathbf{e}_i^T P_k = \mathbf{e}_k^T J_i$ (or equivalently $[P_k]_{i,j} = [J_i]_{kj}$), $1 \leq i, k \leq n$. Observe that the matrices P_k can be written as

$$P_k = \sum_{m=1}^{n} \mathbf{e}_m \mathbf{e}_k^T J_m.$$

Lemma 1.3.3. *Let $\mathcal{L} \in \mathbb{V}$. Let $\mathbf{v} \in \mathbb{C}^n$ and $J_k \in \mathcal{L}$ be such that $\mathbf{v}^T J_k = \mathbf{e}_k^T$, $k = 1 \ldots, n$. Then:*

1. $J_i X \in \mathcal{L}, X \in \mathbb{C}^{n\times n} \Rightarrow J_i X = \sum_{k=1}^{n} [X]_{ik} J_k$.

2. *$\mathcal{L}$ is closed (under matrix multiplication) if and only if*

$$J_i J_j = \sum_{k=1}^{n} [J_j]_{ik} J_k, \quad 1 \leq i, j \leq n, \tag{1.14}$$

being the last condition equivalent to $J_i P_k = P_k J_i$ $1 \leq i, k \leq n$.

3. *If $\mathcal{L}$ is closed, then $\mathcal{L}_{\mathbf{v}}(\mathcal{L}_{\mathbf{v}}(\mathbf{z})^T \mathbf{z}') = \mathcal{L}_{\mathbf{v}}(\mathbf{z}')\mathcal{L}_{\mathbf{v}}(\mathbf{z})$, $\mathbf{z}, \mathbf{z}' \in \mathbb{C}^n$.*

4. *If $I \in \mathcal{L}$ ($\mathcal{L}_{\mathbf{v}}(\mathbf{v}) = I$) and $\mathcal{L}$ is closed, then $X \in \mathcal{L}$ is non-singular iff $\exists\, \mathbf{z} \in \mathbb{C}^n$ such that $\mathbf{z}^T X = \mathbf{v}^T$; in this case $X^{-1} = \mathcal{L}_{\mathbf{v}}(\mathbf{z})$.*

5. *If $\mathcal{L}$ is commutative, then $\mathbf{e}_i^T J_j = \mathbf{e}_j^T J_i$ (or $[J_j]_{ik} = [J_i]_{jk}$), $1 \leq i, j \leq n$, $J_i = P_i, 1 \leq i \leq n$, $\mathbf{z}^T \mathcal{L}_{\mathbf{v}}(\mathbf{z}') = \mathbf{z}'^T \mathcal{L}_{\mathbf{v}}(\mathbf{z})$, $\mathbf{z}$, $\mathbf{z}' \in \mathbb{C}^n$, $I \in \mathcal{L}$ and $\mathcal{L}$ is closed.*

Proof. See [44] □

Definition 2. *Call *-space a subspace $\mathcal{L}$ of $\mathbb{C}^{n\times n}$ spanned by J_i, $i = 1, \ldots, n$, linearly independent, subject to the following conditions:*

$$I \in \mathcal{L} \text{ and } J_i^H J_j = \sum_{k=1}^{n} \overline{[J_k]}_{ij} J_k, \ 1 \leq i, j \leq n. \tag{1.15}$$

Lemma 1.3.4. *Let $\mathcal{L} \in \mathbb{V}$ ($\mathbf{v}^T J_k = \mathbf{e}_k^T$). Then $\mathcal{L}$ is a *-space if:*

1. *$\mathcal{L}$ is commutative, $J_i^H \in \mathcal{L}$, or/and*

2. *$\mathcal{L}$ is closed under matrix multiplication, $J_i^H = \alpha_i J_{t_i}$, $|\alpha_i| = 1$,*
   $t : \{1, \ldots, n\} \to \{1, \ldots, n\}$ is a permutation, and $I \in \mathcal{L}$.

Proof. See [44] □

Remark 2. *Observe that if $\mathcal{L}$ is a sdU algebra (see Section 1.4), point 1. of Lemma 1.3.4 holds; instead if $\mathcal{L}$ is a group algebra (see [45]) point 2. of the same lemma holds.*

The following proposition states some important properties of a *-space. In particular, a *-space is a space of class $\mathbb{V}$.

Lemma 1.3.5. *Let $\mathcal{L}$ be a *-space. Then*

1. $\mathbf{v}^T J_k = \mathbf{e}_k^T$, $1 \leq k \leq n$, *where* $\mathbf{v} = [v_1, \ldots, v_n]^T$ *is such that* $I = \sum_{k=1}^n v_k J_k$, *thus* $\mathcal{L} \in \mathbb{V}$.

2. *$\mathcal{L}$ is closed under conjugate transposition* ($J_i^H \in \mathcal{L}$).

3. *$\mathcal{L}$ is closed under matrix multiplication.*

Proof. See [44] □

Lemma 1.3.6. *Let $\mathcal{L}$ be a *-space. Then $\overline{B}_{\mathcal{L}} \in \mathcal{L}$ ($B_{\mathcal{L}}$ defined in Theorem 1.2.1), in fact*

$$\overline{B}_{\mathcal{L}} = \sum_{k=1}^{n} P_k P_k^H = \sum_{k=1}^{n} \overline{trJ_k} P_k = \sum_{k=1}^{n} \overline{trJ_k} J_k. \tag{1.16}$$

Moreover, if $\mathcal{L}_A$ is the best l.s. to $A \in \mathbb{C}^{n \times n}$ from $\mathcal{L}$, then

$$\mathcal{L}_A = \mathcal{L}_{\mathbf{v}}(B_{\mathcal{L}}^{-1}\mathbf{c}_{\mathcal{L},A}) = \mathcal{L}_{\mathbf{v}}(\mathbf{c}_{\mathcal{L},A})\overline{B}_{\mathcal{L}}^{-1} = \overline{B}_{\mathcal{L}}^{-1}\mathcal{L}_{\mathbf{v}}(\mathbf{c}_{\mathcal{L},A}). \tag{1.17}$$

Proof. See [44] □

Lemma 1.3.7. *Let $\mathcal{L}$ satisfy Definition 2. Then $\forall\, \mathbf{z} \in \mathbb{C}^n$,*

$$\mathbf{z}^H \mathcal{L}_{\mathbf{v}}(\mathbf{c}_{\mathcal{L},A})\mathbf{z} = \sum_{k=1}^{n} [P_k^H \mathbf{z}]^H A [P_k^H \mathbf{z}]. \tag{1.18}$$

Proof. See [44] □

We can finally state the following theorem which gives precise information about the spectrum of the projected matrix. In the following Theorem 1.3.8, for a real matrix X, X_s will denote the matrix $\frac{1}{2}(X + X^T)$, i.e., the symmetric part of X.

Theorem 1.3.8. *Let $\mathcal{L}$ be a subspace of $\mathbb{C}^{n\times n}$ satisfying the conditions in Definition 2. Let $A \in \mathbb{C}^{n \times n}$ and $\mathcal{L}_A$ be the best least squares fit to A from $\mathcal{L}$.*

1. *If $A = A^H$, then $\mathcal{L}_A = \mathcal{L}_A^H$ and* $\min \boldsymbol{\lambda}(A) \leq \boldsymbol{\lambda}(\mathcal{L}_A) \leq \max \boldsymbol{\lambda}(A)$. *As a consequence $\mathcal{L}_A$ is positive definite if A is positive definite.*

2. *If A is real, then* $\min \boldsymbol{\lambda}(A_s) \leq \mathrm{Re}\{\boldsymbol{\lambda}\}(\mathcal{L}_A) \leq \max \boldsymbol{\lambda}(A_s)$. *Moreover, if the J_k in Definition 2 are real (in this case $\mathcal{L}_A$ is real), then $(\mathcal{L}_A)_s$ is positive definite if A_s is positive definite.*

Proof. Let M be a Hermitian matrix such that $M^2 = \overline{B}_{\mathcal{L}}^{-1}$ and consider the matrix $M\mathcal{L}_{\mathbf{v}}(\mathbf{c}_{\mathcal{L},A})M$. As a consequence of 1.17, $M\mathcal{L}_{\mathbf{v}}(\mathbf{c}_{\mathcal{L},A})M$ is similar to $\mathcal{L}_A$. Then $\lambda(\mathcal{L}_A)$ is an eigenvalue of $M\mathcal{L}_{\mathbf{v}}(\mathbf{c}_{\mathcal{L},A})M$, i.e. $\exists\ \mathbf{x} \in \mathbb{C}^n$ with $||\mathbf{x}|| = 1$ such that $\lambda(\mathcal{L}_A) = \mathbf{x}^H M\mathcal{L}_{\mathbf{v}}(\mathbf{c}_{\mathcal{L},A})M\mathbf{x}$; thus by Lemma 1.3.7

$$\lambda(\mathcal{L}_A) = \sum_{k=1}^{n} \mathbf{x}_k^H A\mathbf{x}_k, \ \ \mathbf{x}_k = P_k^H M\mathbf{x}. \tag{1.19}$$

Notice that the first identity in (1.16) implies

$$\sum_{k=1}^{n} \mathbf{x}_k^H \mathbf{x}_k = 1 = \sum_{k=1}^{n} [(\mathrm{Re}\,\mathbf{x}_k)^T(\mathrm{Re}\,\mathbf{x}_k) + (\mathrm{Im}\,\mathbf{x}_k)^T(\mathrm{Im}\,\mathbf{x}_k)].$$

So inequalities in points 1. and 2. of Theorem 1.3.8 hold. Moreover, if $A = A^H$, then by Lemma 1.2.2, $\mathcal{L}_A = \mathcal{L}_A^H$.
Now assume that A and the J_k are real and that $\mathbf{z}^T A_s \mathbf{z} > 0\ \forall\ \mathbf{z} \in \mathbb{R}^n$, $\mathbf{z} \neq 0$. Then the matrix M can be chosen real and, by Lemma 1.3.7, we have

$$\mathbf{z}^T M\mathcal{L}_{\mathbf{v}}(\mathbf{c}_{\mathcal{L},A})M\mathbf{z} = \sum_{k=1}^{n} [P_k^T M\mathbf{z}]^T A[P_k^T M\mathbf{z}], \ \ \forall\ \ \mathbf{z} \in \mathbb{R}^n.$$

This identity implies that the matrix $(\mathcal{L}_A)_s$ is positive definite because, by (1.17), $(\mathcal{L}_A)_s = M(M\mathcal{L}_{\mathbf{v}}(\mathbf{c}_{\mathcal{L},A})M)_s M^{-1}$. □

1.4 An insight of projection onto sd U algebras

1.4.1 Majorization and doubly stochastic matrices

We borrow this section from [4] and refer there for the proofs of the results stated in the following. Let us start giving some definitions.

Definition 3. *Let* $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^n$. *We say that* $\mathbf{y}$ *"majorizes"* $\mathbf{x}$ *and we write* $\mathbf{x} \prec \mathbf{y}$ *if*

1. $x_1 \geq x_2 \geq \cdots \geq x_n$, $y_1 \geq y_2 \geq \cdots \geq y_n$;
2. $\sum_{i=1}^{k} x_i \leq \sum_{i=1}^{k} y_i$ *for all* $1 \leq k \leq n$;
3. $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$.

Definition 4. *We will say that* $S \in \mathbb{R}^{n\times n}$ *is " doubly stochastic" if*

1. $S_{ij} \geq 0$ *for all* $i, j \in \{1, \ldots, n\}$;

2. $S\mathbf{e} = \mathbf{e}$ *and* $\mathbf{e}^T S = \mathbf{e}^T$ *being* $\mathbf{e} = (1, \ldots, 1)^T$.

Definition 5. $T : \mathbb{R}^n \to \mathbb{R}^n$ *is a T-transform if there exists* $0 \leq t \leq 1$ *and* $j, k \in \{1, \ldots, n\}$ *such that*

$$T\mathbf{y} = (y_1, \ldots y_{j-1}, ty_j + (1-t)y_k, y_{j+1}, \ldots, y_{k-1}, (1-t)y_j + ty_k, y_{k+1}, \ldots, y_n)^T.$$

Observe that $T = tI + (1-t)Q_{jk}$, *being* Q_{jk} *the permutation matrix which permutes the components* j *and* k *of a given vector. Observe, moreover, that if* T *is a* T*-transform, then it is a doubly stochastic matrix.*

Theorem 1.4.1. $S \in \mathbb{R}^{n \times n}$ *is doubly stochastic if and only if* $S\mathbf{y} \prec \mathbf{y}$ *for all* $\mathbf{y} \in \mathbb{R}^n$.

Proof. Let us suppose that S is doubly stochastic. Define $\mathbf{x} := S\mathbf{y}$. Let us suppose, moreover, $y_1 \geq \cdots \geq y_n$ and $x_1 \geq \cdots \geq x_n$, otherwise we can consider $P_1\mathbf{y}$, $P_2\mathbf{x}$ and $P_2 S P_1^{-1}$ (being P_1 and P_2 permutation matrices chosen such that $P_1\mathbf{x}$ and $P_2\mathbf{y}$ have non increasing components). Let us define, for any fixed k, $t_j := \sum_{i=1}^k s_{ij} \in [0,1]$. Observe that $\sum_{j=1}^n t_j = \sum_{j=1}^n \sum_{i=1}^k s_{ij} = k$ since S is doubly stochastic. We have

$$\sum_{i=1}^k x_i = \sum_{i=1}^k \sum_{j=1}^n s_{ij} y_j = \sum_{j=1}^n t_j y_j$$

and thus

$$\sum_{j=1}^k x_j - \sum_{j=1}^k y_j = \sum_{j=1}^n t_j y_j - \sum_{j=1}^k y_j = \sum_{j=1}^n t_j y_j - \sum_{j=1}^k y_j + y_k(k - \sum_{j=1}^n t_j) =$$
$$\sum_{j=1}^k (y_j - y_k)(t_j - 1) + \sum_{k+1}^n t_j(y_j - y_k) \leq 0,$$

i.e., $\sum_{i=1}^k x_i \leq \sum_{i=1}^k y_i$ for all $1 \leq k \leq n$. Moreover, being $\mathbf{e}^T\mathbf{x} = \mathbf{e}^T S\mathbf{y}$ we have $\mathbf{x} \prec \mathbf{y}$.

For the reverse implication let us suppose $S\mathbf{y} \prec \mathbf{y}$ for all $\mathbf{y} \in \mathbb{R}^n$. Since $S\mathbf{e}_j \prec \mathbf{e}_j$ for all $j = \{1, \ldots, n\}$ we have $s_{ij} \geq 0$ for all $i, j \in \{1, \ldots, n\}$ and $\sum_{i=1}^n s_{ij} = 1$ for all $j \in \{1, \ldots, n\}$. Moreover, being $S\mathbf{e} \prec \mathbf{e}$ we have $\sum_{i=1}^n (S\mathbf{e})_i = 1$ and $0 \leq (S\mathbf{e})_i \leq 1$ for all $i \in \{1, \ldots, n\}$, i.e., $S\mathbf{e} = \mathbf{e}$. □

Lemma 1.4.2. *If* $\mathbf{x} \prec \mathbf{y}$, *then* $\mathbf{x}$ *can be obtained from* $\mathbf{y}$ *by successive applications of a finite number of* T*-transforms.*

Proof. By induction on the dimension. We assume that $\mathbf{x} \neq \mathbf{y}$. If $n = 2$ we have $x_1 \leq y_1$ and $x_1 + x_2 = y_1 + y_2$, thus $y_1 \geq x_1 \geq x_2 \geq y_2$. Therefore, there exists $t \in [0,1]$ s.t. $x_1 = ty_1 + (1-t)y_2$ and hence $x_2 = (1-t)y_1 + ty_2$, i.e., the thesis holds. Let us suppose that the thesis holds for $n - 1$. Since

$\mathbf{x} \prec \mathbf{y}$ we have $x_n \geq y_n$ and hence $y_1 \geq x_1 \geq \cdots \geq x_n \geq y_n$. Choose k such that $y_k \leq x_1 \leq y_{k-1}$ and write $x_1 = ty_1 + (1-t)y_k$ for some $t \in [0,1]$. Let us define moreover $T_1 := tI + (1-t)Q_{1k}$, being Q_{1k} the matrix which permutes the components 1 and k of any given vector, and

$$\mathbf{x}' := (x_2, \dots, x_n), \quad \mathbf{y}' := (y_2, \dots, y_{k-1}, (1-t)y_1 + ty_k, y_{k+1}, \dots, y_n).$$

Since $y_2 \geq \cdots \geq y_{k-1} \geq x_2 \geq \cdots \geq x_{k-1} \geq x_k \geq \cdots \geq x_n$, for every $2 \leq m \leq k-1$, it holds $\sum_{j=2}^m x_j \leq \sum_{j=2}^m y_j$. For $k \leq m \leq n$ we have instead

$$\begin{aligned}
&\sum_{j=2}^m y_j = \sum_{j=2}^{k-1} y_j + [(1-t)y_1 + ty_k] + \sum_{j=k+1}^m y_j = \\
&\sum_{j=1}^m y_j - x_1 \geq \sum_{j=1}^m x_j - x_1 = \sum_{j=2}^m x_j,
\end{aligned}$$

being the last inequality an equality when $m = n$ since $\mathbf{x} \prec \mathbf{y}$. We proved $\mathbf{x}' \prec \mathbf{y}'$. By induction hypothesis there exists a finite number of T-transformations $\tilde{T}_2, \dots, \tilde{T}_r \in \mathbb{R}^{(n-1)\times(n-1)}$ such that $\mathbf{x}' = (\tilde{T}_r \cdots \tilde{T}_2)\mathbf{y}'$. Defining

$$T_i := \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \tilde{T}_i & \\ 0 & & & \end{bmatrix}$$

for every $i = 2, \dots, r$, we have

$$(T_r \cdots T_1)\mathbf{y} = (\tilde{T}_r \cdots \tilde{T}_2)(x_1, \mathbf{y}')^T = (x_1, \mathbf{x}')^T = \mathbf{x}.$$

□

Corollary 1.4.3. *If* $\mathbf{x} \prec \mathbf{y}$*, then there exists a doubly stochastic matrix* S *such that* $\mathbf{x} = S\mathbf{y}$ *for some doubly stochastic matrix* S*.*

Proof. By Lemma 1.4.2 observing that a finite product of T-transforms is doubly stochastic. □

Lemma 1.4.4. *If* $\mathbf{x} \prec \mathbf{y}$*, then* $\mathbf{x}$ *is in the convex hull of all vectors obtained permuting the coordinates of* $\mathbf{y}$*.*

Proof. By Lemma 1.4.2 observing that the product of finite number of T-transforms is a convex combination of permutation matrices. □

As a consequence of the above results, we obtain Theorem 1.4.5 here below:

Theorem 1.4.5. *The following conditions are equivalent:*

1. *$\mathbf{x} \prec \mathbf{y}$;*

2. *$\mathbf{x}$ is obtained from $\mathbf{y}$ by a finite number of T-transforms;*

3. *$\mathbf{x}$ is in the convex hull of all vectors obtained permuting the coordinates of $\mathbf{y}$;*

4. *$\mathbf{x} = S\mathbf{y}$ for some doubly stochastic matrix S.*

Theorem 1.4.6 (Schur's Theorem)**.** *If $B \in \mathbb{C}^{n\times n}$ is a Hermitian matrix, then, defining $\mathbf{b} \in \mathbb{R}^n$ as the vector of diagonal elements of B ordered in non increasing order, we have $\mathbf{b} \prec \boldsymbol{\lambda}(B)$.*

Proof. By spectral theorem we have $B = UD_\lambda U^H$ where $U = (u_{ij})_{i,j\in\{1,\dots,n\}}$ is a unitary matrix. For every $i \in \{1,\dots,n\}$ we have

$$b_{ii} = \mathbf{e}_i^T U d(\boldsymbol{\lambda}) U^H \mathbf{e}_i = \sum_{j=1}^{n} u_{ij}\overline{u}_{ij}\lambda_j = \sum_{j=1}^{n} p_{ij}\lambda_j,$$

being $P = (p_{ij})_{i,j\in\{1,\dots,n\}} := U \circ \overline{U}$ a doubly stochastic matrix, i.e., $\mathbf{b} = P\boldsymbol{\lambda}$. Thesis follows from Theorem 1.4.5. □

Theorem 1.4.7 (Cauchy Interlacing Property, [63])**.** *If $B \in \mathbb{C}^{n\times n}$ is a Hermitian matrix partitioned as follows:*

$$B = \begin{bmatrix} B_{11} & C \\ C^H & B_{22} \end{bmatrix}$$

where $B_{11} \in \mathbb{C}^{k\times k}$ and $B_{22} \in \mathbb{C}^{(n-k)\times(n-k)}$, then we have

$$\lambda_{i+n-k}(B) \le \lambda_i(B_{11}) \le \lambda_i(B) \ \ \textit{for all } i \in \{1,\dots,k\}. \tag{1.20}$$

Proof. Consider $\mathbf{x}_1,\dots,\mathbf{x}_n \in \mathbb{C}^n$ an orthonormal set of eigenvectors of B and $\hat{\mathbf{y}}_1,\dots,\hat{\mathbf{y}}_k \in \mathbb{C}^k$ an orthonormal set of eigenvectors of B_{11}. For every $j \in \{1,\dots,k\}$ define $\mathbf{y}_j := [\hat{\mathbf{y}}_j, \mathbf{0}]^T \in \mathbb{C}^n$. Define, moreover, $S_1^{(i)} = span\{\mathbf{y}_1,\dots,\mathbf{y}_i\}$ and $S_2^{(i)} = span\{\mathbf{x}_i,\dots,\mathbf{x}_n\}$. Since $dim(S_1^{(i)}) + dim(S_2^{(i)}) = i + (n-i+1)$ there exists a unitary vector $\mathbf{y} \neq \mathbf{0} \in S_1^{(i)} \bigcap S_2^{(i)}$ of the form $\mathbf{y} = [\hat{\mathbf{y}}, \mathbf{0}]^T$ where $\hat{\mathbf{y}} \in span\{\hat{\mathbf{y}}_1,\dots,\hat{\mathbf{y}}_i\}$. Since for all unitary $\hat{\mathbf{v}} \in span\{\hat{\mathbf{y}}_1,\dots,\hat{\mathbf{y}}_i\}$ it holds $\lambda_i(B_{11}) \le \hat{\mathbf{v}}^H B_{11}\hat{\mathbf{v}} \le \lambda_1(B_{11})$ and
for all unitary $\mathbf{v} \in span\{\mathbf{x}_i,\dots,\mathbf{x}_n\}$ it holds $\lambda_n(B) \le \mathbf{v}^H B\mathbf{v} \le \lambda_i(B)$, we have

$$\lambda_i(B_{11}) \le \hat{\mathbf{y}}^T B_{11}\hat{\mathbf{y}} = \mathbf{y}^T B\mathbf{y} \le \lambda_i(B). \tag{1.21}$$

Analogously defining $S_1^{(i)} = span\{\mathbf{y}_i,\dots,\mathbf{y}_k\}$ and $S_2^{(i)} = span\{\mathbf{x}_1,\dots,\mathbf{x}_{i+n-k}\}$ we have

$$\lambda_i(B_{11}) \ge \hat{\mathbf{y}}^T B_{11}\hat{\mathbf{y}} = \mathbf{y}^T B\mathbf{y} \ge \lambda_{i+n-k}(B). \tag{1.22}$$

□

Theorem 1.4.8. *Let us suppose* $\mathbf{x} \prec \mathbf{y}$*. Then for any convex function* $\phi : \mathbb{R}^n \to \mathbb{R}^n$ *it holds* $\sum_{i=i}^n \phi(x_i) \leq \sum_{i=i}^n \phi(y_i)$*.*

Proof. Since $\mathbf{x} \prec \mathbf{y}$, using Theorem 1.4.5 there exists a doubly stochastic matrix S s.t. $\mathbf{x} = S\mathbf{y}$. We have that

$$\sum_{i=1}^n \phi(x_i) \leq \sum_{i=1}^n \sum_{j=1}^n s_{ij}\phi(y_j) = \sum_{j=1}^n \phi(y_j).$$

Observe that even the reverse implication holds. See [4] for more details. □

1.4.2 Projection onto sd U algebras

In this section we will summarize some results connected with the projection of Hermitian matrices onto sd U spaces.
Define, for a given unitary matrix $U \in \mathbb{C}^{n\times n}$, the associated sd U algebra as

$$\mathcal{L} := \operatorname{sd} U := \{Ud(\mathbf{z})U^H \text{ s.t. } \mathbf{z} \in \mathbb{C}^n\}. \tag{1.23}$$

As we observed in Remark 2, an sd U algebra is a $*$-space, hence results of Section 1.3 can be applied. Before continuing, observe that one can obtain points 1. and 2. in Theorem 1.3.8, in the more specific case when $\mathcal{L}$ is a sd U algebra, using the identities $\mathbf{u}_k^H \mathcal{L}_A \mathbf{u}_k = \mathbf{u}_k^H A \mathbf{u}_k$, $\mathbf{u}_k = U\mathbf{e}_k$. These identities follow from the equality

$$\mathcal{L}_A = Ud([U^H AU]_{kk},\ k = 1, \dots, n)U^H \tag{1.24}$$

found as a simple consequence of the fact that $||\cdot||_F$ is unitary invariant. In particular the following theorem holds:

Theorem 1.4.9. *Let* $\mathcal{L} = \operatorname{sd} U$ and let $B \in \mathbb{C}^{n\times n}$. Then

1. $\mathcal{L}_B = Ud(\mathbf{z}_B)U^H$ where $[\mathbf{z}_B]_i = [U^H BU]_{ii}$, $i = 1, \dots, n$; in particular $\mathbf{z}_{\mathbf{x}\mathbf{y}^T} = d(U^H \mathbf{x})U^T \mathbf{y}$, where $\mathbf{x},\ \mathbf{y} \in \mathbb{C}^n$.

2. If $B \in \mathbb{R}^{n\times n}$ then $\mathcal{L}_B \in \mathbb{R}^{n\times n}$ provided that $\mathcal{L}$ is spanned by real matrices, or more generally, whenever the conjugate of the space $\mathcal{L}$ is included in $\mathcal{L}$, i.e $\overline{\mathcal{L}} \subset \mathcal{L}$ ($\mathcal{L}$ is closed under conjugation).

3. If $B = B^H$, then $\mathcal{L}_B = (\mathcal{L}_B)^H$ and $\min \boldsymbol{\lambda}(B) \leq \boldsymbol{\lambda}(\mathcal{L}_B) \leq \max \boldsymbol{\lambda}(B)$ where $\boldsymbol{\lambda}(X)$ is the spectrum of X. Therefore $\mathcal{L}_B$ is Hermitian positive definite whenever B is Hermitian positive definite.

4. $\operatorname{tr}(\mathcal{L}_B) = \operatorname{tr}(B)$;

5. if B is Hermitian, then $\boldsymbol{\lambda}(\mathcal{L}_B) \prec \boldsymbol{\lambda}(B)$;

6. ([103]) if B is Hermitian and ϕ is convex, then $\sum_{i=1}^n \phi(\lambda_i(\mathcal{L}_B)) \leq \sum_{i=1}^n \phi(\lambda_i(B))$;

7. ([103]) if B is Hermitian and ϕ is convex monotonic non decreasing, then $\sum_{i=1}^{k} \phi(\lambda_i(\mathcal{L}_B)) \leq \sum_{i=1}^{k} \phi(\lambda_i(B))$ for all $k \in \{1, \ldots, n\}$;

8. If B Hermitian positive definite, then $\det(B) \leq \det(\mathcal{L}_B)$ where the equality holds iff U diagonalizes B.

9. ([84]) Define the K-condition number for B as $K(B) = (\operatorname{tr}(B)/n)^n/(\det(B))$. Then

$$K(\mathcal{L}_B^{-1}B) = \min_{X\in\mathcal{L},\, X \text{ Hermitian positive definite}} K(X^{-1}B).$$

Proof. 1. Observe that since the Frobenius norm is unitary invariant we have

$$\min_{X\in\mathcal{L}} \|X - B\|_F = \min_{\mathbf{z}\in\mathbb{C}^n} \|d(\mathbf{z}) - U^H BU\|_F.$$

The second part follows from direct computation.
2. and 3. see Remark 1 and Theorem 1.3.8.
4. Using 1. and observing that $\boldsymbol{\lambda}(\mathcal{L}_B)$ is obtained arranging in a non increasing order the set

$$\{(U^H BU)_{ii} \text{ for } i = 1, \ldots, n\}.$$

5. Define the matrix $\tilde{B} := U^H BU$. From Theorem 1.4.6 we have $\tilde{\mathbf{b}} \prec \boldsymbol{\lambda}(B)$ where $\tilde{\mathbf{b}}$ are the diagonal elements of $\tilde{B}$ arranged in non increasing order. Thesis follows using point 2., i.e., $\boldsymbol{\lambda}(\mathcal{L}_B) = \tilde{\mathbf{b}}$.

6. From Theorem 1.4.8 using point 3.

7. Define U_k as the matrix obtained by by the first k columns of U. Define

$$\tilde{B}_k := U_k^H BU_k \in \mathbb{C}^{k\times k}$$

and $\tilde{\mathbf{b}}_k$ as the diagonal elements of $\tilde{B}_k$ arranged in non increasing order. We have $\tilde{\mathbf{b}}_k \prec \boldsymbol{\lambda}(B_k)$ from Theorem 1.4.6, hence, using Theorem 1.4.8, it holds

$$\sum_{i=1}^{k} \phi((\tilde{\mathbf{b}}_k)_i) \leq \sum_{i=1}^{k} \phi(\lambda_i(B_k)) \leq \sum_{i=1}^{k} \phi(\lambda_i(B)),$$

where the last inequality follows from Theorem 1.4.7 and from the monotonicity of ϕ.

8. Use point 1. and the Hadamard's inequality for the determinant of $U^H BU$:

$$\det(B) = \det\bigl(U^H BU\bigr) \leq \prod_{i=1}^{n} (U^H BU)_{ii} = \det(\mathcal{L}_B). \tag{1.25}$$

Assume that in the above equation (1.25) the equality holds and assume by contradiction that $(U^H BU)_{ts} \neq 0$ for a pair of indexes $t, s \in \{1, \dots, n\}$, where $t \neq s$. Assume, moreover, without loss of generality that $(U^H BU)_{ss} \geq (U^H BU)_{tt}$. Then there exists a Givens transformation Q s.t.

$$((UQ)^H BUQ)_{ii} = \begin{cases} (U^H BU)_{ii} \text{ for } i \neq t, s; \\ (U^H BU)_{ss} + \delta; \\ (U^H BU)_{tt} - \delta; \end{cases} \tag{1.26}$$

with

$$\delta = \sqrt{|(U^H BU)_{ts}|^2 + \Big(\frac{(U^H BU)_{ss} - (U^H BU)_{tt}}{2}\Big)^2} - \frac{|(U^H BU)_{ss} - (U^H BU)_{tt}|}{2} > 0,$$

obtained from the equation $((UQ)^H BUQ)_{st} = 0$. But then we have

$$\det(B) \leq \prod_{i=1}^{n} ((UQ)^H BUQ)_{ii} < \prod_{i=1}^{n} (U^H BU)_{ii},$$

which is a contradiction.

9.

$$K(X^{-1}B) = K(d(\mathbf{z})^{-1} U^H BU) = \Big(\frac{\sum_{i=1}^n (U^H BU)_{ii}/z_i}{n}\Big)^n \frac{\prod_{i=1}^n z_i}{\det(B)}.$$

Thanks to the inequality between the arithmetic and geometric means (which says that for any non-negative real numbers $\{x_i\}_{i=1}^n$ it holds $\Big(\frac{\sum_{i=1}^n x_i}{n}\Big)^n \geq \prod_{i=1}^n x_i$ and equality holds iff $x_1 = \cdots = x_n$), we have

$$\begin{aligned} &\Big(\frac{\sum_{i=1}^n (U^H BU)_{ii}/z_i}{n}\Big)^n \prod_{i=1}^n z_i \geq \prod_{i=1}^n (U^H BU)_{ii} \\ &\Leftrightarrow K(\mathcal{L}_B^{-1} B) \geq \frac{\prod_{i=1}^n (U^H BU)_{ii}}{\det(B)}. \end{aligned} \tag{1.27}$$

In (1.27) equality holds iff $(U^H BU)_{11}/z_1 = \cdots = (U^H BU)_{nn}/z_n$, condition satisfied by $\mathbf{z} = [\dots, (U^H BU)_{ii}, \dots]^T$. □

Chapter 2

Low complexity matrix projections preserving actions on vectors

2.1 Introduction

The projection onto algebras of matrices simultaneously diagonalized by unitary transforms U

$$\mathcal{L} := \text{sd}\, U = \{U d(\mathbf{z}) U^H \ : \ \mathbf{z} \in \mathbb{C}^n\},$$

has been used profitably in the last twenty or thirty years as a core instrument in order to speed up, through preconditioning techniques, iterative methods for linear systems $A\mathbf{x} = \mathbf{b}$ where A is a symmetric positive definite matrix (spd for short) [44, 102, 37, 96], [66, Chap. 5]. The main idea connected with these spaces could be traced in the key observation that very often the matrices corresponding to linear systems arising from applications exhibit some special structures and thus can be naturally approximated in low complexity spaces $\mathcal{L}$ of matrices of the form $\text{sd}\, U$. By a *low complexity space* $\mathcal{L}$ we mean a space such that for any $A \in \mathcal{L}$ the cost of a matrix vector product $A\mathbf{x}$ for any $\mathbf{x} \in \mathbb{R}^n$ and the number of memory allocations sufficient to store A are much less than n^2. In particular $\mathcal{L}_A$, the projection of A onto such $\mathcal{L} = \text{sd}\, U$, has revealed to produce approximations of the spectrum of A good enough (see in particular [102, 101]) to speed up preconditioned iterative solvers of $A\mathbf{x} = \mathbf{b}$ without increasing the time per step and the space complexity.

Recently [40, 7, 43, 39], projection onto low complexity matrix algebras have been also used in order to reduce the computational cost of BFGS method [82] for the unconstrained minimization of a function $f : \mathbb{R}^n \to \mathbb{R}$. In these low complexity BFGS-type methods, a sequence of spd matrices $\{B_k\}_{k\in\mathbb{N}}$ satisfying the $\mathcal{S}$ecant-Equation ($B_k\mathbf{s}_{k-1} = \mathbf{y}_{k-1}$, where $\mathbf{s}_{k-1} =$

$\mathbf{x}_k - \mathbf{x}_{k-1}$, $\mathbf{y}_{k-1} = \nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k-1})$) is iteratively generated defining B_k as a rank-2 correction of $\mathcal{L}_{B_{k-1}}$ (instead of B_{k-1}) and the line search is performed along the corresponding sequence of descent directions $\mathbf{d}_k = -B_k^{-1}\nabla f(\mathbf{x}_k)$. In earlier papers on this subject, the successful use of matrix algebras $\,\mathrm{sd}\, U$ tool was mostly connected with the fact that the approximations $\{\mathcal{L}_{B_k}\}_{k\in\mathbb{N}}$ of the spd matrices $\{B_k\}_{k\in\mathbb{N}}$ generated by BFGS-type scheme, preserve some *global* information about the spectrum of B_k (i.e. trace and determinant, see Theorem 1.4.9). This fact permits to establish global convergence results for a corresponding $\mathcal{N}$on $\mathcal{S}$ecant class of low complexity BFGS-type algorithms where the descent directions are of the form $\mathbf{d}_k = -\mathcal{L}_{B_k}^{-1}\nabla f(\mathbf{x}_k)$. Notice that in such papers, the matrix algebra $\mathcal{L}$ (and hence the unitary matrix U) was fixed during all the execution of the algorithm. More recently, in [41, 39] it has been pointed out, accordingly to some experimental evidences, that in order to improve the efficiency in low complexity BFGS-type methods, an adaptive choice of the matrix algebra $\mathcal{L}$ can be exploited, i.e. at each step a matrix algebra $\mathcal{L}^{(k)} = \mathrm{sd}\, U_k$ (and hence a unitary matrix U_k) can be produced in order to guarantee that the matrix $\mathcal{L}_{B_k}^{(k)}$ retains as much as possible information from B_k. Moreover, in [27], it has been observed that the possibility of changing the algebra $\mathcal{L}$ at each step can be used in order to guarantee the global convergence of Secant low complexity BFGS-type methods. This can be achieved by requiring, apart the global approximation of the spectrum of B_k by $\mathcal{L}_{B_k}$, the coincidence of $\mathcal{L}_{B_k}$ and B_k along the given direction $\mathbf{s}_k$. We are able to obtain such a point-wise property only by changing the space $\mathcal{L}$ at each step. To this aim the following problem was introduced in [27] :

Problem 1. *Given a spd matrix $B \in \mathbb{R}^{n\times n}$ and a vector $\mathbf{v} \in \mathbb{R}^n$, find a low complexity unitary matrix L such that defining $\mathcal{L} = sd\, L$, we have*

$$\mathcal{L}_B\mathbf{v} = B\mathbf{v}. \tag{2.1}$$

In this paper we obtain a general result (Theorem 2.3.1) which in particular gives a full solution of Problem 1 and of its multi-direction generalization :

Problem 2. *Given a symmetric matrix $B \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{n\times r}$ such that $V^TV = I_r$, find a low complexity unitary matrix L such that, defining $\mathcal{L} = sd\, L$, we have*

$$\mathcal{L}_BV = BV. \tag{2.2}$$

Observe that the cost of the construction of the unitary low complexity matrix L and of the projection $\mathcal{L}_B$ (where $\mathcal{L} = \mathrm{sd}\, L$) in Problem 2, is justified by the possibility to exploit $\mathcal{L}_B$ as a low complexity approximation of B. This approximation should be able to capture in $\mathcal{L}_B$ two characteristics of the operator B: its spectral information and its action along a fixed set of directions defined by V. The proof of the main result in Section 2.3, which gives a solution of Problem 2, uses the fact that the Arnoldi method for finding an orthonormal basis of a Krylov subspace permits to highlight the essential informational content of the action of B on a fixed set of directions. In this paper we show how this highlighting leads to the definition of a special low complexity space $\mathcal{L} = \operatorname{sd} L$ such that the projection of B onto $\mathcal{L}$ allows us to gain the two required characteristics. A key instrument to construct the unitary L is given by a remarkable new result concerning the Householder matrices (see Section 2.2.1). More precisely L turns out to be the product of a number of Householder matrices depending on the dimension of V.

Even if Problem 1 has been introduced in connection with an optimization framework (see [27, 30] and Chapter 5), where the aim was to define new efficient minimization procedures, we believe that the general results presented in this chapter have their own theoretical interest and can have applications in other fields of computational mathematics. For the sake of completeness let us point out that a problem similar to Problem 2 has been tackled in [106, 105, 79, 52] in relation to preconditioning problems, but using different techniques.

2.2 Preliminaries: Block-Arnoldi

Given $A \in \mathbb{R}^{n\times n}$, $V_1 \in \mathbb{R}^{n\times r}$ such that $V_1^T V_1 = I_r$ consider the following generalized Krylov space

$$\mathcal{K}_m(A, V_1) := K_m(A, V_1\mathbf{e}_1) + \cdots + K_m(A, V_1\mathbf{e}_r)$$

where $K_m(A, \mathbf{v}) = \mathit{span}\{\mathbf{v}, A\mathbf{v}, \ldots, A^{m-1}\mathbf{v}\}$ denotes the usual Krylov space of order m. Consider, moreover, the following Block-Arnoldi procedure :

```
Input: $A \in \mathbb{R}^{n\times n}$, $V_1 \in \mathbb{R}^{n\times r}$
for $j = 1, \ldots, m$ do
    $W_j := AV_j$;
    for $i = 1, \ldots, j$ do
        $H_{ij} := V_i^T W_j$ ;
        $W_j := W_j - V_i H_{ij}$;
    end
    $W_j := Q_j R_j$ ($Q_j$ orthogonal , $R_j$ upper triangular) ;
    Set $V_{j+1} := Q_j$ and $H_{j+1,j} := R_j$ ;
end
```

Algorithm 1: Block-Arnoldi

The following Lemma 2.2 (see [92, 93, 61]) could be considered as a condensed summary of basic properties on the output of the Block-Arnoldi method.

Lemma 2.2.1. *Let us define*

$$U_m := [V_1, \ldots, V_m],$$

$$H_m := (H_{ij})_{1 \leq i,j \leq m},\ H_{ij} := 0,\ i > j+1$$

constructed considering the output of Algorithm 1, and let E_m be the $mr \times r$ matrix whose columns coincide with the last r columns of I_{mr}. If none of the upper triangular matrices $\{H_{j+1,j}\}_{j=1,\dots,m-1}$ are singular, then the columns of U_m form an orthonormal basis of $\mathcal{K}_m(A, V_1)$. Moreover, the following relations hold:

$$AU_m = U_m H_m + V_{m+1} H_{m+1,m} E_m^T; \tag{2.3}$$

$$H_m = U_m^T A U_m. \tag{2.4}$$

In particular, about the output of the Block-Arnoldi method, a natural block generalization of a well known polynomial vector identity (see Lemma 3.1 in [90]), is stated in the following Lemma 2.3. Note that such block identity (2.5) is related with the approximation of the action on V_1 of the exponential of A [76].

Lemma 2.2.2. *Let U_m and H_m be defined as in Lemma 2.2.1. Then for any polynomial p_j of degree $j \leq m-1$ the following equality holds:*

$$p_j(A)V_1 = U_m p_j(H_m) \begin{bmatrix} I_r \\ 0_{(m-1)r,r} \end{bmatrix}. \tag{2.5}$$

Proof. Analogous of that of Lemma 3.1 in [90], defining the orthogonal projector onto $\mathcal{K}_m(A, V_1)$ as $\pi_m = U_m U_m^T$. □

Before concluding, as already pointed out in the Introduction, observe that thanks to Lemma 2.2.2 the Block-Arnoldi procedure represents a computational strategy for compressing the action of the matrix A on V_1 in a small number of parameters whenever r is small. These parameters, suitably rearranged, will lead to an ad-hoc low complexity space $\mathcal{L} = \operatorname{sd} L$ where the projection $\mathcal{L}_A$ inherits the informational content related to the action A on V_1 (see Theorem 2.3.1).

2.2.1 Householder Matrices

Given a vector $\mathbf{p} \neq 0 \in \mathbb{R}^n$ define

$$\mathcal{H}(\mathbf{p}) := I_n - \frac{2}{\|\mathbf{p}\|^2}\mathbf{p}\mathbf{p}^T.$$

By extension, if $\mathbf{p} = 0$, we will write $\mathcal{H}(\mathbf{p})$ to denote the identity matrix.

Remark 3. *Consider two vectors $\mathbf{v}$, $\mathbf{z} \in \mathbb{R}^n$. From direct computation one can check that defining $\mathbf{p} = \mathbf{v} - \frac{\|\mathbf{v}\|}{\|\mathbf{z}\|}\mathbf{z}$ with $\mathbf{z} \neq 0$, we have*

$$\mathcal{H}(\mathbf{p})\mathbf{v} = \frac{\|\mathbf{v}\|}{\|\mathbf{z}\|}\mathbf{z}.$$

The following Lemma 2.2.3 is a generalization of an analogous in [41].

Lemma 2.2.3. *Consider* $W = [\mathbf{w}_1|\dots|\mathbf{w}_s] \in \mathbb{R}^{n\times s}$ *and* $V = [\mathbf{v}_1|\dots|\mathbf{v}_s] \in \mathbb{R}^{n\times s}$ *such that* $s \leq n$*,* $W^TW = V^TV$*. Then there exist* $\mathbf{h}_1,\dots,\mathbf{h}_s \in \mathbb{R}^n$ *and an orthogonal matrix* $U = \mathcal{H}(\mathbf{h}_s)\cdots\mathcal{H}(\mathbf{h}_1)$ *product of* s *Householder matrices such that*

$$U\mathbf{w}_i = \mathbf{v}_i \textit{ for all } i \in \{1,\dots,s\}.$$

If $s = n$ *we have* $\mathbf{h}_n = 0$*, i.e.* $\mathcal{H}(\mathbf{h}_n) = I$*, or* $\mathcal{H}(\mathbf{h}_n) = \mathcal{H}(\mathbf{v}_n)$*.*

Proof. By induction on s.
For $s = 1$ use Remark 3. Assuming the thesis true for $s - 1$, let us prove it for s. Set $U_i = \mathcal{H}(\mathbf{q}_i)Q$ where Q is an orthogonal matrix and

$$\mathbf{q}_i := Q\mathbf{w}_i - \mathbf{v}_i \text{ for } i \in \{1,\dots,s\}.$$

Since $U_i\mathbf{w}_i = \mathbf{v}_i$ for $i \in \{1,\dots,s\}$ for any choice of Q, thesis will be proved if there exists an orthogonal matrix Q such that $\mathbf{q}_1 = \cdots = \mathbf{q}_s =: \mathbf{q}$, i.e.

$$Q(\mathbf{w}_1 - \mathbf{w}_i) = \mathbf{v}_1 - \mathbf{v}_i \text{ for } i \in \{2,\dots,s\}.$$

In fact in this case $U_1 = \cdots = U_s$ would be the required matrix U. Thesis follows from inductive hypothesis defining

$$W_1 = [\mathbf{w}_1 - \mathbf{w}_2|\dots|\mathbf{w}_1 - \mathbf{w}_s] \in \mathbb{R}^{n\times(s-1)},$$

$$V_1 = [\mathbf{v}_1 - \mathbf{v}_2|\dots|\mathbf{v}_1 - \mathbf{v}_s] \in \mathbb{R}^{n\times(s-1)}$$

and observing that $W_1^TW_1 = V_1^TV_1$. Note that the vector $\mathbf{q}$ can be chosen such that $\|\mathbf{q}\|_2^2 = 2$.
For the second part of the thesis observe that if $Q = \mathcal{H}(\mathbf{h}_{n-1})\cdots\mathcal{H}(\mathbf{h}_1)$ is the orthogonal matrix such that $Q\mathbf{w}_i = \mathbf{v}_i$ for $i = 1,\cdots,n-1$ we have $Q\mathbf{w}_n = k\mathbf{v}_n$, $k = \pm 1$ (V is an orthonormal basis of $\mathbb{R}^n$ and Q orthogonal). If $k = -1$ it is enough to consider $U = \mathcal{H}(\mathbf{v}_n)Q$.

□

Two important observations are in order here:

1. Lemma 2.2.3 can be used to produce an orthogonal matrix $U = \mathcal{H}(\mathbf{h}_s)\cdots\mathcal{H}(\mathbf{h}_1)$ having among its columns s given orthonormal vectors $\mathbf{v}_i$ for $i \in \{1,\dots,s\}$ (just take $\mathbf{w}_i = \mathbf{e}_{k_i}$) and the proof of Lemma 2.2.3 represents a concrete computational procedure of cost $O(s(s-1)n)$ (see Lemma 2.2.5) for finding the vectors $\mathbf{h}_1,\dots,\mathbf{h}_s$. In particular we have a procedure (alternative to the classic one obtained through triangular decomposition by Householder reflections) of cost $O(n^3)$ for finding the vectors $\mathbf{h}_1,\dots,\mathbf{h}_n$ such that

$$V = \mathcal{H}(\mathbf{h}_n)\cdots\mathcal{H}(\mathbf{h}_1),$$

where V is any orthogonal fixed matrix.

2. Matrix algebras of the form $\mathcal{L} = \operatorname{sd} U$ where $U = \mathcal{H}(\mathbf{h}_s)\cdots\mathcal{H}(\mathbf{h}_1)$ is an orthogonal matrix and $s << n$, are low complexity matrix algebras since the structure of U can be exploited in order to perform the matrix-vector product $U\mathbf{d}(\mathbf{z})U^T\mathbf{v}$ in $O(sn)$ FLOPs (whenever the vector $\mathbf{z}$ is given). See also Lemma 2.2.4 below.

Lemma 2.2.4. *Consider an orthogonal matrix U of the form $U = \mathcal{H}(\mathbf{h}_s)\cdots\mathcal{H}(\mathbf{h}_1)$, then for any vector $\mathbf{v} \in \mathbb{R}^n$ the matrix vector product $U\mathbf{v}$ can be written as*

$$\begin{gathered} U\mathbf{v} = \mathbf{v} - \sum_{i=1}^{s} \mathbf{h}_i^T \mathbf{q}_{i-1}\mathbf{h}_i, \textit{ where} \\ \mathbf{q}_0 = \mathbf{v} \textit{ and } \mathbf{q}_i := \mathbf{q}_{i-1} - \mathbf{h}_i^T\mathbf{q}_{i-1}\mathbf{h}_i = \mathbf{v} - \sum_{j=1}^{i} \mathbf{h}_j^T\mathbf{q}_{j-1}\mathbf{h}_j. \end{gathered} \tag{2.6}$$

Proof. Let us preliminarily observe that formula (2.6) for the $\mathbf{q}_i$ can be easily proved. The thesis holds for $s = 1$. Assuming the thesis true for $s-1$, let us prove it for s. Consider $U = \mathcal{H}(\mathbf{h}_s)\cdots\mathcal{H}(\mathbf{h}_1)$. From inductive hypothesis we have

$$\mathcal{H}(\mathbf{h}_{s-1})\cdots\mathcal{H}(\mathbf{h}_1)\mathbf{v} = \mathbf{v} - \sum_{i=1}^{s-1} \mathbf{h}_i^T\mathbf{q}_{i-1}\mathbf{h}_i$$

from which we obtain

$$\mathcal{H}(\mathbf{h}_s)(\mathbf{v} - \sum_{i=1}^{s-1} \mathbf{h}_i^T\mathbf{q}_{i-1}\mathbf{h}_i) = \mathbf{v} - \sum_{i=1}^{s-1} \mathbf{h}_i^T\mathbf{q}_{i-1}\mathbf{h}_i - \mathbf{h}_s^T(\mathbf{v} - \sum_{i=1}^{s-1} \mathbf{h}_i^T\mathbf{q}_{i-1}\mathbf{h}_i)\mathbf{h}_s.$$

The result follows observing that

$$\mathbf{v} - \sum_{i=1}^{s-1} \mathbf{h}_i^T\mathbf{q}_{i-1}\mathbf{h}_i = \mathbf{v} - \sum_{i=1}^{s-2} \mathbf{h}_i^T\mathbf{q}_{i-1}\mathbf{h}_i - \mathbf{h}_{s-1}^T\mathbf{q}_{s-2}\mathbf{h}_{s-1} = \mathbf{q}_{s-1}.$$

□

Remark 4. *Evidently once the coefficients $\mathbf{h}_i^T\mathbf{q}_{i-1}$ for $i \in \{1,\dots,s\}$ are known $O(sn)$ multiplications are sufficient to compute the matrix vector product $U\mathbf{v}$ in (2.6). Observe, moreover, that the coefficients $\mathbf{h}_i^T\mathbf{q}_{i-1}$ for $i \in \{1,\dots,s\}$ in (2.6) can be computed in $O(s^2)$ once the quantities*

$$\mathbf{h}_i^T\mathbf{v} \textit{ for } i \in \{1,\dots,s\}, \textit{ and } \mathbf{h}_i^T\mathbf{h}_j \textit{ for } i \neq j \in \{1,\dots,s\}$$

are known (this last observation will be useful in Proposition 2.3.2).

A non recursive version of Lemma 2.2.3, more implementation-oriented, is stated in the following

Lemma 2.2.5. *Consider $W = [\mathbf{w}_1|\dots|\mathbf{w}_s] \in \mathbb{R}^{n\times s}$ and $V = [\mathbf{v}_1|\dots|\mathbf{v}_s] \in \mathbb{R}^{n\times s}$ such that $s \leq n$, $W^TW = V^TV$. Then there exist $\mathbf{h}_1,\dots,\mathbf{h}_s \in \mathbb{R}^n$ and an orthogonal matrix $U = \mathcal{H}(\mathbf{h}_s)\cdots\mathcal{H}(\mathbf{h}_1)$ product of s Householder matrices such that*

$$U\mathbf{w}_i = \mathbf{v}_i \text{ for all } i \in \{1,\dots,s\}.$$

The vectors $\mathbf{h}_i$ for $i \in \{1,\dots,s\}$ can be obtained by setting :

$$\begin{aligned} &\mathbf{h}_i = (-1)^{s-i}[\mathcal{H}(\mathbf{h}_{i-1})\cdots\mathcal{H}(\mathbf{h}_1)(\mathbf{w}_{s-i+1} - \mathbf{w}_{s-i}) - (\mathbf{v}_{s-i+1} - \mathbf{v}_{s-i})], \\ &\mathbf{h}_i := (\sqrt{2}/\|\mathbf{h}_i\|_2)\mathbf{h}_i \end{aligned} \tag{2.7}$$

(where we set $\mathbf{w}_0 = \mathbf{v}_0 = \mathbf{0}$). The computational cost is :

$[s(s-1)n+s(2n+1)]$ mult. $+[(s(s+2)-2)n+s(n-1)]$ add. $+s$ sq. roots.

Observe that when $\mathbf{w}_i = \mathbf{e}_{k_i}$ for $i = 1,\dots,s$, it is possible to save $(s-1)n$ mult. and $(3s-2)n$ add..

2.3 Main Result

Theorem 2.3.1. *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. For every fixed integers m and r with $1 \leq m \leq n$, $mr \leq n$ and for any $V_1 \in \mathbb{R}^{n\times r}$ such that $V_1^TV_1 = I_r$, there exists an orthogonal matrix $L \in \mathbb{R}^{n\times n}$ such that if $\mathcal{L} = sdL$ and $\mathcal{L}_A$ is the best approximation of A in $\mathcal{L}$, then*

$$p_j(\mathcal{L}_A)V_1 = p_j(A)V_1 \tag{2.8}$$

for any polynomial p_j of degree $j \leq m-1$.

Proof. Consider the matrices U_m and H_m constructed from Algorithm 1 applied to $\mathcal{K}_m(A, V_1)$ (observe that the first r columns of U_m form V_1). From Lemma 2.2.2 with $j = 1$ we have

$$AV_1 = U_mH_mU_m^TU_m \begin{bmatrix} I_r \\ 0_{(m-1)r,r} \end{bmatrix}.$$

From (2.4), the last equality becomes

$$AV_1 = U_mQQ^TU_m^TAU_mQQ^TU_m^TV_1 \tag{2.9}$$

for any orthogonal matrix $Q \in \mathbb{R}^{mr\times mr}$. In particular, being $U_m^TAU_m$ symmetric, in (2.9) we can choose Q as the orthogonal matrix which diagonalizes $U_m^TAU_m$, i.e.

$$AV_1 = U_mQ \begin{bmatrix} x_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & x_{mr} \end{bmatrix} Q^TU_m^TV_1, \tag{2.10}$$

where $x_i = \mathbf{e}_i^T Q^T U_m^T A U_m Q \mathbf{e}_i$ for $i = 1, \ldots, mr$. Consider now the matrix

$$L = [U_m Q\mathbf{e}_1 | \ldots | U_m Q\mathbf{e}_{mr} | \mathbf{g}_{mr+1} | \ldots | \mathbf{g}_n]$$

where $\{\mathbf{g}_{rm+1}, \ldots, \mathbf{g}_n\}$ is an orthonormal basis for

$$< U_m Q\mathbf{e}_1, \ldots, U_m Q\mathbf{e}_{mr} >^{\perp} = < U_m \mathbf{e}_1, \ldots, U_m \mathbf{e}_{mr} >^{\perp}, \tag{2.11}$$

set $\mathcal{L} = \operatorname{sd} L$ and consider $\mathcal{L}_A$ the best approximation of A in $\mathcal{L}$. In order to prove that $\mathcal{L}_A$ satisfies (2.8) it is sufficient to prove that

$$\mathcal{L}_A^j V_1 = A^j V_1 \text{ for } 0 \leq j \leq m-1. \tag{2.12}$$

Of course, (2.12) is true for $j = 0$. The equality $\mathcal{L}_A V_1 = AV_1$ follows observing that using the first formula in Theorem 1.4.9 we have

$$\begin{aligned} \mathcal{L}_A V_1 &= (\sum_i^n (L^T A L)_{ii} L\mathbf{e}_i (L\mathbf{e}_i)^T) V_1 \\ &= (\sum_i^{mr} x_i (U_m Q\mathbf{e}_i)(U_m Q\mathbf{e}_i)^T) V_1 = AV_1 \end{aligned} \tag{2.13}$$

where in the second equality we take into account that $\mathbf{g}_i^T V_1 = \mathbf{0}^T$ for $i \in \{mr+1, \ldots, n\}$ (see (2.11) and (2.10)). Suppose now (2.12) true for all $j \leq m-2$ and let us prove it for $j = m-1$. From inductive hypothesis and Lemma 2.2.2 we have

$$\mathcal{L}_A^{m-1} V_1 = \mathcal{L}_A \mathcal{L}_A^{m-2} V_1 = \mathcal{L}_A A^{m-2} V_1 = \mathcal{L}_A U_m H_m^{m-2} \begin{bmatrix} I_r \\ 0_{(m-1)r,r} \end{bmatrix}.$$

From direct computation, we have $\mathcal{L}_A U_m = U_m H_m$ and thus, using (2.11) and the definition of Q, we have

$$\mathcal{L}_A U_m H_m^{m-2} \begin{bmatrix} I_r \\ 0_{(m-1)r,r} \end{bmatrix} = U_m H_m^{m-1} \begin{bmatrix} I_r \\ 0_{(m-1)r,r} \end{bmatrix} = A^{m-1} V_1$$

where the last equality follows using again Lemma 2.2.2. Hence (2.12) holds also for $j = m-1$. □

Thanks to Lemma 2.2.3, given $A \in \mathbb{R}^{n \times n}$ symmetric and $V_1 \in \mathbb{R}^{n \times r}$, $V_1^T V_1 = I$, one can realize concretely the procedure introduced in the proof of Theorem 2.3.1 to obtain $\mathcal{L}$ and $\mathcal{L}_A$ such that (2.8) holds.

Proposition 2.3.2. *The matrix $\mathcal{L}_A$ of Theorem 2.3.1 can be constructed by the following steps:*

1. *Apply the Block-Arnoldi procedure, Algorithm 1, to $\mathcal{K}_m(A, V_1)$ in order to obtain the matrices U_m and $H_m = U_m^T A U_m$ of Lemma 2.2.1. This step requires $O(mr\chi(A) + m(m+1)r^2 n + mC_1)$ arithmetic operations, where C_1 is the computational cost of a QR decomposition of $n \times r$ matrix.*

2. *Produce sufficiently accurate matrices Q (orthogonal) and D (diagonal) such that $Q^T H_m Q = D$.*

3. *By using Lemma 2.2.3, compute $\mathbf{h}_1, \dots, \mathbf{h}_{mr}$ such that*

$$\mathcal{H}(\mathbf{h}_{mr}) \cdots \mathcal{H}(\mathbf{h}_1) \begin{bmatrix} I_{mr} \\ 0_{n-mr,mr} \end{bmatrix} = U_m Q,$$

   and set $L = \mathcal{H}(\mathbf{h}_{mr}) \cdots \mathcal{H}(\mathbf{h}_1)$. This step requires $O(mr(mr-1)n)$ arithmetic operations.

4. *Set $\mathcal{L} = sdL$ and compute $(L^T AL)_{ii}$ for $i = mr+1, \dots, n$ (for $i = 1, \dots, mr$ we have $(L^T AL)_{ii} = D_{ii}$), and observe that*

$$\mathcal{L}_A = Ld(((L^T AL)_{ii})_{i=1}^n) L^T$$

   satisfies (2.8)*. This last step requires $mr\chi(A) + O(mrn)$ arithmetic operations.*

Proof. For the computational cost needed to perform point 3. consider the first observation after Lemma 2.2.3 and Lemma 2.2.5.
Instead, concerning point 4., observe that using Lemma 2.2.4 and the symmetry of A we have :

$$\mathbf{e}_j^T L^T AL\mathbf{e}_j = \mathbf{e}_j^T A\mathbf{e}_j - 2\sum_{i=1}^{mr} \mathbf{h}_i^T \mathbf{q}_{i-1} \mathbf{e}_j^T A\mathbf{h}_i + \sum_{i,l=1}^{mr} \mathbf{h}_i^T \mathbf{q}_{i-1} \mathbf{h}_l^T \mathbf{q}_{l-1} \mathbf{h}_i^T A\mathbf{h}_l$$

for all $j \in \{1, \dots, n\}$. The computational cost estimation follows from Remark 4 observing that $\mathbf{h}_i^T \mathbf{e}_j$ for $i \in \{1, \dots, rm\}$ and $j \in \{1, \dots, n\}$ do not require FLOPs to be computed. □

The following Corollary states the cost of solving the original Problem 2 by using the procedure indicated in Theorems 2.3.1 and Proposition 2.3.2 in the cases $r = 1$ and $r = 2$.

Corollary 2.3.3. *Consider $A \in \mathbb{R}^{n\times n}$ and $V_1 \in \mathbb{R}^{n\times r}$. The computational cost to produce an orthogonal matrix U and $\mathcal{L}_A$ such that $\mathcal{L}_A V_1 = AV_1$, where $\mathcal{L} = sdU$ is : $O(n) + 2\chi(A)$ when $m = 2$, $r = 1$ and $O(n) + 4\chi(A)$ when $m = 2$, $r = 2$.*

Remark 5. *Observe that since $U_m U_m^T \in \mathcal{L}$ we have that*

$$\|A - \mathcal{L}_A\|_F \leq \|A - U_m U_m^T\|_F.$$

Before concluding this section let us stress once more that Problem 2 with $r = 1$ has been introduced in [27] in order to guarantee the convergence of $\mathcal{S}$ecant low complexity BFGS-type minimization schemes originally introduced in [40] (where it was proved the convergence of a $\mathcal{N}$on $\mathcal{S}$ecant version). This result has been recently generalized to Broyden-Class-type methods, see [30] and Chapter 5 for more details. Interestingly enough, the above mentioned papers [40] and [27] suggest that the Secant Equation is not a necessary ingredient for the convergence – and perhaps for the definition – of quasi-Newton methods, in striking contrast to [82, p.24] [33, p.54] and [13, p.223] where the Secant Equation seems to be a necessary requirement of Quasi-Newton methods (and is even called quasi-Newton Equation).

2.4 Numerical Results

In the following Table 2.1 we present some numerical results confirming the effectiveness of the construction presented in Proposition 2.3.2. In particular we will focus our attention on the case $m = 3$, $r = 1$, using some matrices from [31] and $V_1 \in \mathbb{R}^n$ generated randomly.

Table 2.1: Numerical results for $m = 3$, $r = 1$

Name	Dimension	$\mu_2(A)$	spd	$\|A^j V_1 - \mathcal{L}_A^j V_1\|_2$	
plat362	362	$2.178223e + 11$	yes	$j = 1$	$6.0561e - 15$
				$j = 2$	$2.6162e - 15$
1138_bus	1138	$8.572646e + 06$	yes	$j = 1$	$1.4353e - 10$
				$j = 2$	$2.4170e - 06$
ex4	1601	$2.386583e + 03$	no	$j = 1$	$8.0444e - 14$
				$j = 2$	$8.0444e - 14$
can_1072	1072	$1.214866e + 35$	no	$j = 1$	$1.2807e - 12$
				$j = 2$	$3.5074e - 11$

Observe that, as the experiment 1138_bus shows, further investigation should be devoted in order to analyze the numerical stability issues connected to the proposed construction, to better understand, for example, why the accuracy deteriorates so consistently in this particular example.

2.5 Useful remarks and future works

In this section we present some useful applications and possible research problems related with the result stated in Theorem 2.3.1. An application to numerical optimization is described more in detail in the next Chapter 5.

1)[Generic V_1] Observe that if $V_1 \in \mathbb{R}^{n\times r}$ has maximum rank but $V_1^T V_1 \neq I_r$, in order to find the matrix algebra $\mathcal{L} = \mathrm{sd}\, U$ such that $\mathcal{L}_A^j V_1 = A^j V_1$ for $0 \le j \le m-1$, it is sufficient to consider the matrix algebra $\mathcal{L} = \mathrm{sd}\, U$ such that $\mathcal{L}_A^j Q = A^j Q$ for $0 \le j \le m-1$, being $V_1 = QR$ (QR decomposition).

2) [Projections which preserve eigenvectors] Consider $V_1 \in \mathbb{R}^{n\times r}$ such that $V_1^T V_1 = I_r$ and $AV_1 = V_1 d(\boldsymbol{\lambda})$ where $\boldsymbol{\lambda} \in \mathbb{R}^r$. Consider now any

orthogonal L having r columns coinciding with those of V_1, then, defining $\mathcal{L} = \operatorname{sd} L$, it is easy to verify that we have

$$\mathcal{L}_A V_1 = V_1 d(\boldsymbol{\lambda}) \tag{2.14}$$

(such L can be constructed as the product of r Householder matrices). Note that (2.14) represents a block generalization of an analogous formula used in [26] in order to precondition Euler-Richardson method for solving stochastic linear systems where the matrix algebra $\mathcal{L} = \operatorname{sd} U$ has been chosen such that $\mathbf{e}^T \mathcal{L}_A = \mathbf{e}^T A = \mathbf{e}^T$ (being A a column stochastic matrix).

3) [The generated ideal] Given $B \in \mathbb{R}^{n \times n}$ and $V_1 \in \mathbb{R}^{n \times r}$ such that $V_1^T V_1 = I_r$, adopting an algebraic-geometric point of view, if we try to tackle directly Problem 2 (i.e. find $\mathcal{L} = \operatorname{sd} U$ such that $\mathcal{L}_B V_1 = B V_1$), we should seek a matrix U of the form

$$U = (I - \mathbf{x}_s \mathbf{x}_s^T) \dots (I - \mathbf{x}_1 \mathbf{x}_1^T),$$

satisfying the following system of $nr + s$ polynomial equations :

$$\begin{cases} U d((U^T B U)_{ii}) U^T V_1 - B V_1 = 0 \\ \|\mathbf{x}_1\|^2 = 2 \\ \vdots \\ \|\mathbf{x}_s\|^2 = 2 \end{cases} \tag{2.15}$$

If we denote now by I the ideal generated by such system and by $Z_a(I)$ the corresponding affine algebraic set, from Theorem 2.3.1 with $m = 2$ we have $Z_a(I) \neq \varnothing$ when $s = 2r$ and hence, from Hilbert Nullstellenstaz-weak form ([78]), we can conclude $I \neq (1)$ in $\mathbb{C}[\mathbf{x}_1, \dots, \mathbf{x}_s]$. Observe that if we consider, instead of an orthogonal matrix a unitary matrix, (2.15) is not longer an algebraic variety. Nevertheless, we suspect that an analogous result could hold if the construction presented in Section 2.3 could be extended to the complex case and if we suitably increase the number of unknowns in (2.15).

4) [Projections with fixed columns] Observe that if we form $V_1 \in \mathbb{R}^{n \times r}$ using a subset of the canonical basis in Theorem 2.3.1, then projections will have the same columns of the original matrix. Of course, in order to obtain a good approximation in Frobenius norm of the original matrix, a straightforward choice of the columns to be preserved in the projection, would be to choose those of maximum 2-norm. Nevertheless, we believe that, in order to make choices which produce more accurate approximations, in this context, ideas from the theory of pseudo-skeleton approximation and maximal volume from [57, 58, 59], would be extremely useful.

5) [Eigenvalues approximation] In the particular case of A spd matrix, the Kantorovich inequality [63] could represent a measure of the accuracy of the extremal eigenvalues of the projection $\mathcal{L}_A$ as approximations of the

extremal eigenvalues of A. To fix ideas consider $m = 2$ in Theorem 2.3.1. From the Kantorovich inequality we have that

$$\frac{(\lambda_1(A)+\lambda_n(A))^2}{4\lambda_1(A)\lambda_n(A)} \geq \frac{\mathbf{e}_i^T V_1^T A^2 V_1 \mathbf{e}_i}{(\mathbf{e}_i^T V_1^T A V_1 \mathbf{e}_i)^2} \text{ for all } i \in \{1,\ldots,r\}.$$

Considering $\mathcal{L} = \operatorname{sd} L$ such that $\mathcal{L}_A V_1 = AV_1$, we have

$$\frac{(\lambda_1(A)+\lambda_n(A))^2}{4\lambda_1(A)\lambda_n(A)} \geq \frac{(\lambda_1(\mathcal{L}_A)+\lambda_n(\mathcal{L}_A))^2}{4\lambda_1(\mathcal{L}_A)\lambda_n(\mathcal{L}_A)} \geq \frac{\mathbf{e}_i^T V_1^T \mathcal{L}_A^2 V_1 \mathbf{e}_i}{(\mathbf{e}_i^T V_1^T \mathcal{L}_A V_1 \mathbf{e}_i)^2} = \frac{\mathbf{e}_i^T V_1^T A^2 V_1 \mathbf{e}_i}{(\mathbf{e}_i^T V_1^T A V_1 \mathbf{e}_i)^2}$$

for all $i \in \{1,\ldots,r\}$ which, of course, is not guaranteed to hold for the projection on a generic matrix algebra $\mathcal{L} = \operatorname{sd} U$ for the vectors in V_1 (the first inequality follows using Theorem 1.4.9). A suitable choice of V_1 could maximize the last ratio producing in this way a good approximation of the extremal eigenvalues of A. Finally, observe that if one form V_1 using the eigenvectors corresponding to the maximal and minimal eigenvalues of A we obtain, for all $\mathbf{v} \in \mathbb{R}^n$,

$$\frac{\mathbf{v}^T A^2 \mathbf{v}}{(\mathbf{v}^T A \mathbf{v})} \leq \frac{(\lambda_1(A)+\lambda_n(A))^2}{4\lambda_1(A)\lambda_n(A)} = \frac{(\lambda_1(\mathcal{L}_A)+\lambda_n(\mathcal{L}_A))^2}{4\lambda_1(\mathcal{L}_A)\lambda_n(\mathcal{L}_A)}.$$

Chapter 3

Regularizing properties of a class of matrices including the optimal and the superoptimal preconditioners

3.1 Introduction

Optimal circulant preconditioners have been introduced in [20] and studied in [21, 102]. Superoptimal circulant preconditioners have been introduced in [102] and studied [24, 36]. In this paper we introduce a class of matrices which include the optimal and the superoptimal preconditioners. We prove that the matrices in such a class share, in an enhanced form, some good properties of the superoptimal matrix [36, 49, 35], particularly useful when exploited as regularizing preconditioners [62]. We prove moreover, that the proposed preconditioners can be computed cheaply when the coefficient matrix of the linear system has Toeplitz structure. Finally, we exhibit experimental results confirming the goodness of the proposed preconditioners when employed as regularizing preconditioners.

3.2 Main Results

In this section we introduce and study a class of matrices parametrized by the natural numbers. Theorem 3.2.3 represents the main result in this section and generalize an analogous result obtained in [25] involving the optimal and the superoptimal preconditioners. Let us introduce the following theorem by Ostrowski (Theorem 4.5.9 [63]):

Proposition 3.2.1. *Let A, $S \in \mathbb{C}^{n\times n}$ with A Hermitian and S non singular. Let the eigenvalues of A, SAS^* and SS^* be arranged in nondecreas-*

ing order. Let $\sigma_1 \geq \cdots \geq \sigma_n > 0$ be the singular values of S. For each $k = 1, \ldots, n$, there is a positive real number $\theta_k \in [\sigma_n^2, \sigma_1^2]$ such that

$$\lambda_k(SAS^*) = \theta_k \lambda_k(A).$$

Lemma 3.2.2. *For any Hermitian positive definite $A \in \mathbb{C}^{n\times n}$ and unitary $U \in \mathbb{C}^{n\times n}$ we have :*

$$d(U^* A^{2^i} U) \geq d(U^* A^{2^{i-1}} U)^2, \quad i \in \mathbb{N}\backslash\{0\}. \tag{3.1}$$

Proof. By direct computation exploiting the equality

$$U^* A^{2^i} U = U^* A^{2^{i-1}} U U^* A^{2^{i-1}} U.$$

□

Definition 6. *For any $\mathcal{L} =$ sd U and $i \in \mathbb{N}$ define*

$$P_{(i)}(A) := \mathcal{L}_{A^{2^i}}^{\frac{1}{2^{i-1}}} \mathcal{L}_A^{-1}. \tag{3.2}$$

Observe that choosing $i = 0$ in equation (3.2) we obtain the optimal preconditioner introduced in [20], choosing instead $i = 1$ we obtain the superoptimal preconditoner introduced in [102]. We can consider for this reason $P_{(i)}(A)$ as a possible extension of the above mentioned preconditioners.

Remark 6. *From Lemma 3.2.2 we have*

$$d(U^* A^{2^i} U)^{\frac{1}{2^{i-1}}} \geq d(U^* A^{2^{i-1}} U)^{\frac{1}{2^{i-2}}}$$

or equivalently

$$P_{(i)}(A) \geq P_{(i-1)}(A) \text{ for } i \in \mathbb{N}\backslash\{0\}, \tag{3.3}$$

and hence, for $i \in \mathbb{N}$ it holds

$$\begin{aligned} \max_k \lambda_k(P_{(i)}(A)) &\geq \max_k \lambda_k(P_{(i-1)}(A)), \\ \min_k \lambda_k(P_{(i)}(A)) &\geq \min_k \lambda_k(P_{(i-1)}(A)). \end{aligned} \tag{3.4}$$

Observe, moreover, that applying repeatedly Lemma 3.2.2 it follows that

$$d(U^* A^{2^i} U) \geq \cdots \geq d(U^* A U)^{2^i}$$

and hence

$$d(U^* A U)^{-1} d(U^* A^{2^i} U)^{\frac{1}{2^i}} \geq I_{n\times n}. \tag{3.5}$$

Remark 7. *Observe that using Theorem 3.1 in [49], $P_{(i)}(A)$ are the solutions of the following optimization problem:*

$$P_{(i)}(A) := \arg\min_{X\in\mathcal{L}} \|AX - \mathcal{L}_{A^2} \mathcal{L}_{A^{2^i}}^{-\frac{1}{2^{i-1}}}\|. \tag{3.6}$$

The choice of the matrices $\mathcal{L}_{A^2}\mathcal{L}_{A^{2^i}}^{-\frac{1}{2^{i-1}}}$ in the optimization problem (3.6) is justified by the following theorem:

Theorem 3.2.3. *Given an Hermitian positive definite $A \in \mathbb{C}^{n\times n}$, for any $i \in \mathbb{N}\backslash\{0\}$ we have that*

$$\lambda_k((P_{(i)}(A))^{-1}A) \leq \lambda_k((P_{(i-1)}(A))^{-1}A) \quad k = 1, \dots, n. \tag{3.7}$$

Proof.

$$\begin{aligned}
&(P_{(i)}(A))^{-1}A = (P_{(i)}(A))^{-1}P_{(i-1)}(A)(P_{(i-1)}(A))^{-1}A \\
&= \mathcal{L}_{A^{2^i}}^{-\frac{1}{2^{i-1}}}\mathcal{L}_{A^{2^{i-1}}}^{\frac{1}{2^{i-2}}}\mathcal{L}_{A^{2^{i-1}}}^{-\frac{1}{2^{i-2}}}\mathcal{L}_A A \\
&\approx d(U^*A^{2^i}U)^{-\frac{1}{2^{i-1}}}d(U^*A^{2^{i-1}}U)^{\frac{1}{2^{i-2}}}d(U^*A^{2^{i-1}}U)^{-\frac{1}{2^{i-2}}}d(U^*AU)U^*AU \\
&\approx DMD
\end{aligned}$$

where

$$D = d(U^*A^{2^i}U)^{-\frac{1}{2^i}}d(U^*A^{2^{i-1}}U)^{\frac{1}{2^{i-1}}}$$

and

$$M = d(U^*A^{2^{i-1}}U)^{-\frac{1}{2^{i-1}}}d(U^*AU)^{\frac{1}{2}}U^*AUd(U^*AU)^{\frac{1}{2}}d(U^*A^{2^{i-1}}U)^{-\frac{1}{2^{i-1}}}.$$

From (3.1) we have $D_{ii} \in (0, 1]$ and hence thesis follows from Theorem 3.2.1 where $S = D$ and $A = M$, observing that

$$M \approx (P_{(i-1)}(A))^{-1}A.$$

□

Corollary 3.2.4. *For every $k = 1, \dots, n$, there exist $\lambda_k^{\infty\downarrow}$ and $\lambda_k^{\infty\uparrow}$ such that*

$$\lim_{i\to\infty}\lambda_k((P_{(i)}(A))^{-1}A) = \lambda_k^{\infty\downarrow}, \tag{3.8}$$

$$\lim_{i\to\infty}\lambda_k(P_{(i)}(A^{-1})A) = \lambda_k^{\infty\uparrow}. \tag{3.9}$$

Proof. Observe that using Theorem 1.4.9 we have

$$\lambda(\mathcal{L}_{A^{2^i}}^{\frac{1}{2^{i-1}}}) \in [\lambda_n(A^{2^i})^{\frac{1}{2^{i-1}}}, \lambda_1(A^{2^i})^{\frac{1}{2^{i-1}}}] = [\lambda_n(A)^2, \lambda_1(A)^2]$$

and hence

$$\lambda(P_{(i)}(A)) \in [\frac{\lambda_n(A)^2}{\lambda_1(A)}, \frac{\lambda_1(A)^2}{\lambda_n(A)}]. \tag{3.10}$$

Thesis follows from Bolzano-Weierstrass theorem observing that

$$\lambda((P_{(i)}(A))^{-1}A) \in [\frac{\lambda_n(A)^2}{\lambda_1(A)^2}, \frac{\lambda_1(A)^2}{\lambda_n(A)^2}]$$

and observing that, from Theorem 3.2.3, $\lambda_k((P_{(i)}(A))^{-1}A)$ is a monotonically decreasing sequence for each $k = 1, \dots, n$.

For the second part, applying Theorem 3.2.3 to the inverse matrix A^{-1}, we obtain

$$\lambda_k((P_{(i)}(A^{-1}))^{-1}A^{-1}) \leq \lambda_k((P_{(i-1)}(A^{-1}))^{-1}A^{-1}) \quad k = 1, \dots, n$$

and hence

$$\lambda_k(P_{(i)}(A^{-1})A) \geq \lambda_k(P_{(i-1)}(A^{-1})A) \quad k = 1, \dots, n. \tag{3.11}$$

Observe that (3.11) is now a monotonic increasing sequence and thesis follows as in the previous case. □

Remark 8. *Observe that using Proposition 3.2.1, Lemma 3.2.2 and analogous techniques to those in Theorem 3.2.3 and Corollary 3.2.4, it is possible to prove that also the sequences $\{\lambda_k(P_{(i)}(A)A)\}_{i\in\mathbb{N}}$ are monotonic increasing and convergent for every $k = 1, \dots, n$. Interestingly enough, using $P_{(i)}(A)$ and $(P_{(i)}(A))^{-1}$ it is possible to produce, in some cases, a better approximation of A^{-1}. To this extent observe that*

$$\min_{a,b\in\mathbb{R}} \|(aP_{(i)}(A) + b(P_{(i)}(A))^{-1})A - I\|_F$$

has a non trivial solution, namely a, b are obtained solving, under suitable hypotheses, the following linear system

$$\begin{bmatrix} tr(A^2(P_{(i)}(A))^2) & tr(A^2) \\ tr(A^2) & tr(A^2(P_{(i)}(A))^{-2}) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} tr(A(P_{(i)}(A))) \\ tr(A(P_{(i)}(A))^{-1}) \end{bmatrix}. \tag{3.12}$$

The following Theorem 3.2.6 gives a more accurate bound of $\lambda(P_{(i)}(A))$ if compared to the bound contained in (3.10). First let us introduce the following lemma :

Lemma 3.2.5. *If $A \in \mathbb{C}^{n\times n}$ is an Hermitian positive definite matrix then, defining $M = \max_k A_{kk}$, we have*

$$|\operatorname{Re}\{A_{hk}\}| \leq M \text{ and } |\operatorname{Im}\{A_{hk}\}| \leq M \text{ for all } h, k \in \{1, \dots, n\}.$$

Proof. For $h \neq k$ chose the vectors $\mathbf{x}_1 = \mathbf{e}_h - \mathbf{e}_k$, $\mathbf{x}_2 = \mathbf{e}_h + \mathbf{e}_k$, $\mathbf{x}_3 = \imath\mathbf{e}_h - \mathbf{e}_k$, $\mathbf{x}_4 = \imath\mathbf{e}_h + \mathbf{e}_k$ and use the fact that $\mathbf{x}_l^* A \mathbf{x}_l$ for $l = 1, 2, 3, 4$ are strictly positive scalars. □

Theorem 3.2.6. *For every $i \in \mathbb{N}$ we have*

$$\sigma(P_{(i)}(A)) \subset [\beta_{\min}, C_{(i)}\beta_{\max}], \tag{3.13a}$$

$$\mu_2(P_{(i)}(A)) \leq \min\{\mu_2(A)^3, (2n-1)^{\frac{1}{2^{i-1}}}\mu_2(A)^{\frac{1}{2^{i-2}}}\mu_2(\mathcal{L}_A)\}, \tag{3.13b}$$

being $\beta_{\min} = \min_k(U^*AU)_{kk}$, $\beta_{\max} = \max_k(U^*AU)_{kk}$ and $C_{(i)} \geq 1$ a suitable constant.

Proof. For the first part, to ease the notation, define

$$B := U^*AU \quad \text{and} \quad X_{(i)} := d(B)^{-2^{i-2}}(B^{2^{i-1}})d(B)^{-2^{i-2}}.$$

It can be easily checked that $P_{(i)}^{2^{i-1}}(A) = Ud(X_{(i)}d(B)^{2^{i-1}}X_{(i)})U^*$ and that

$$\beta_{\min}^{2^{i-1}}X_{(i)}^2 \leq X_{(i)}d(B)^{2^{i-1}}X_{(i)} \leq \beta_{\max}^{2^{i-1}}X_{(i)}^2. \tag{3.14}$$

From (3.14) it follows that

$$\beta_{\min}d(X_{(i)}^2)^{\frac{1}{2^{i-1}}} \leq d(X_{(i)}d(B)^{2^{i-1}}X_{(i)})^{\frac{1}{2^{i-1}}} \leq \beta_{\max}d(X_{(i)}^2)^{\frac{1}{2^{i-1}}}$$

where $d(X_{(i)}d(B)^{2^{i-1}}X_{(i)})^{\frac{1}{2^{i-1}}}$ are the eigenvalues of $P_{(i)}(A)$. To complete the proof define $M_{(i)} := \max_k(X_{(i)})_{kk}$. We have that

$$1 \leq (X_{(i)}^2)_{kk}^{\frac{1}{2^{i-1}}} \leq (M_{(i)}^2 + 2M_{(i)}^2(n-1))^{\frac{1}{2^{i-1}}} \text{ for all } k \in 1,\ldots,n, \tag{3.15}$$

where the first inequality follows observing that $(X_{(i)}^2)_{kk} \geq (X_{(i)})_{kk}^2$ and (3.5), the second inequality follows instead from Lemma 3.2.5. Define

$$C_{(i)} := (M_{(i)}^2(2n-1))^{\frac{1}{2^{i-1}}}. \tag{3.16}$$

For the second part let us bound the constant $M_{(i)}$. Observe that if we consider the spectral norm $\|\cdot\|_2$ we have for $i \in \mathbb{N}$ that

$$\begin{aligned} M_{(i)} \leq \rho(X_{(i)}) = \rho(d(B)^{-2^{i-1}}B^{2^{i-1}}) \leq \|d(B)^{-2^{i-1}}\|_2\|B^{2^{i-1}}\|_2 \Rightarrow \\ (M_{(i)})^{\frac{1}{2^{i-1}}} \leq \|d(B)^{-2^{i-1}}\|_2^{\frac{1}{2^{i-1}}}\|B^{2^{i-1}}\|_2^{\frac{1}{2^{i-1}}} = \frac{\lambda_1(A)}{\beta_{\min}}, \end{aligned} \tag{3.17}$$

and hence

$$1 \leq C_{(i)} \leq (\frac{\lambda_1(A)}{\beta_{\min}})^{\frac{1}{2^{i-2}}}(2n-1)^{\frac{1}{2^{i-1}}}. \tag{3.18}$$

Observe finally that from (3.18) we have $\lim_{i\to\infty} C_{(i)} = 1$. □

Before concluding this section, let us observe that an analogous bound to (3.13a) was derived in [36] for the superoptimal preconditioner, and hence Theorem 3.2.6 could be considered as an extension of Theorem 3.4 in [36].

Observe, moreover, that (3.13b) could be particularly relevant if the matrix algebra $\mathcal{L} = \operatorname{sd} U$ is chosen such that $\mu_2(\mathcal{L}_A) << \mu_2(A)$ (see Appendix for a way to construct such $\mathcal{L}$). In fact, once such $\mathcal{L}$ is available, the corresponding sequence of preconditioners $P_{(i)}(A) = \mathcal{L}_{A^{2^i}}^{\frac{1}{2^{i-1}}}\mathcal{L}_A^{-1}$ must be, by (3.13b), such that $\mu_2(P_{(i)}(A)) \approx \mu_2(\mathcal{L}_A)$ when $i \in \mathbb{N}$ is sufficiently large. Thus it is possible to introduce and use preconditioners $P_{(i)}(A)$ which satisfy property (3.7) – a property which favors regularizing properties (see Section 3.4) –, without resulting in a significant deterioration of the condition number.

Remark 9. *It is possible to compute directly* $\lim_{i\to\infty} P_{(i)}(A)$. *To this end let us define* $\mathbf{u}_k := U\mathbf{e}_k$ *for* $k = 1, \dots, n$ *and write* $\mathbf{u}_k = \sum_{j=1}^{n} \alpha_j^{(k)} \mathbf{v}_j$ *where* $A\mathbf{v}_j = \lambda_j(A)\mathbf{v}_j$. *We have*

$$\lim_{s\to+\infty} (\mathbf{u}_k A^s \mathbf{u}_k)^{\frac{1}{s}} = \lim_{s\to+\infty} (\sum_{j=1}^{n} (\alpha_j^{(k)})^2 \lambda_j(A)^s)^{\frac{1}{s}} = \lambda_{(1)}(A),$$

and hence,

$$\lim_{i\to+\infty} P_{(i)}(A) = Ud(\mathbf{b})U^* \text{ where } \mathbf{b} = [\frac{\lambda_1(A)^2}{(U^*AU)_1}, \dots, \frac{\lambda_1(A)^2}{(U^*AU)_n}]. \qquad (3.19)$$

3.3 The Toeplitz Case

"When using superoptimal preconditioners in practice, one obviously should be assured that there is a way to compute them sufficiently quickly ([102])".

In this section we will prove that, when T is a Toeplitz matrix and $\mathcal{C} = \text{sd } F$ where F is the Fourier matrix (i.e. $\mathcal{C}$ is the algebra of circulant matrices), the preconditioners $P_{(i)}(T)$ can be computed cheaply for moderate values of $i \in \mathbb{N}$.

3.3.1 An insight into Toeplitz and Toeplitz-like structures

Let us start introducing some notations, definitions and results. What follows in this subsection is entirely borrowed from [71] and hence we refer there for further details and proofs.

Definition 7. *Define the displacement* $\nabla_Z(A)$ *of* $A \in \mathbb{C}^{n\times n}$ *with respect to* $Z \in \mathbb{C}^{n\times n}$ *as*

$$\nabla_Z(A) := A - ZAZ^*,$$

and the rank of $\nabla_Z(A)$ *the displacement rank of* A. *Define, moreover, Toeplitz-like matrices as those matrices with small displacement rank.*

Definition 8. *For* $r \geq rank(\nabla_Z(A))$ *we call a pair of matrices* (G, B) *where* G, $B \in \mathbb{C}^{n\times r}$ *generator for* A *if*

$$\nabla_Z(A) := A - ZAZ^* = GB^*.$$

If we consider

$$Z = \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix},$$

then the matrix A can be reconstructed from generators as

$$A = \mathcal{T}(G, B) := \sum_{k=0}^{n-1} Z^k G B^* (Z^*)^k. \tag{3.20}$$

If we denote by $\mathbf{g}_j$, $\mathbf{b}_j \in \mathbb{C}^n$ respectively the columns of G, B for $j = 1, \ldots, r$, we can rewrite (3.20) as

$$A = \mathcal{T}(G, B) = \sum_{k=1}^{r} L(\mathbf{g}_j) U(\mathbf{b}_j^*), \tag{3.21}$$

where $U(\mathbf{x})$ and $L(\mathbf{x})$ are the triangular Topelitz matrices

$$L(\mathbf{x}) = \begin{pmatrix} x_0 & 0 & \cdots & 0 \\ x_1 & x_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ x_{n-1} & \cdots & x_1 & x_0 \end{pmatrix}, \quad U(\mathbf{x}) = \begin{pmatrix} x_0 & x_1 & \cdots & x_{n-1} \\ 0 & x_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & x_1 \\ 0 & \cdots & 0 & x_0 \end{pmatrix}.$$

Observe, moreover, that if $T \in \mathbb{C}^{n \times n}$ is the Toeplitz matrix $T_{ij} = t_{i-j}$ for $i, j \in \{0, \ldots, n-1\}$ we have that $rank(\nabla_Z(T)) = 2$ and a set of minimal generator is

$$G = \begin{pmatrix} t_0 & 1 \\ t_1 & 0 \\ \vdots & \vdots \\ t_{n-1} & 0 \end{pmatrix}, B = \begin{pmatrix} 1 & 0 \\ 0 & \bar{t}_{-1} \\ \vdots & \vdots \\ 0 & \bar{t}_{-(n-1)} \end{pmatrix}.$$

Lemma 3.3.1. *Let T be a Toeplitz matrix. Then T^s is a Toeplitz-like matrix of displacement rank at most $2s$ for any integer $s \geq 1$. Letting (G, B) denote the generators for T, a sequence of (non-minimal) generators $(G_1, B_1), \ldots, (G_s, B_s)$ for $T, \ldots, T^s$ is given by*

$$G_1 = G, \ G_{i+1} = [P_G^i G \ \ P_G^{i-1} G \ \ \ldots \ \ G \ \ -P_G \mathbf{e}_1 \ \ \ldots \ \ -P_G^i \mathbf{e}_1], \tag{3.22}$$

$$B_1 = B, \ B_{i+1} = [B \ \ P_B B \ \ \ldots \ \ P_B^i B \ \ P_B^i \mathbf{e}_1 \ \ \ldots \ \ P_B \mathbf{e}_1], \tag{3.23}$$

for $i = 1, \ldots, s-1$, where $P_G := (Z-I)T(Z-I)^{-1}$ and $P_B := (Z-I)T^*(Z-I)^{-1}$. Moreover

$$\mathbf{e}_1 \in range(G_1) \subset \ldots range(G_s) \text{ and } \mathbf{e}_1 \in range(B_1) \subset \ldots range(B_s).$$

Corollary 3.3.2. *Let $T \in \mathbb{C}^{n \times n}$ be a Toeplitz matrix, then a set of generators for the monomial $T, \ldots, T^s$ can be computed with $O(sn \log_2(n))$ operations.*

Proof. Applying $(Z - I)^{-1}$ to a vector amounts simply to computing the vector of its cumulative sums and the application of $Z - I$ to a vector can be evaluated with $n - 1$ subtractions. The multiplication of Toeplitz matrix by a vector can be performed in $O(n \log(n))$. □

3.3.2 Projecting the powers of Toeplitz matrices onto the Circulant Algebra

Let us start recalling Theorem 5.1 and Corollary 1 in [102]:

Theorem 3.3.3. *Let $M = LR \in \mathbb{C}^{n\times n}$, where*

$$L = \begin{pmatrix} l_0 & 0 & \cdots & 0 \\ l_1 & l_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ l_{n-1} & \cdots & l_1 & l_0 \end{pmatrix}, \quad R = \begin{pmatrix} r_0 & r_1 & \cdots & r_{n-1} \\ 0 & r_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & r_1 \\ 0 & \cdots & 0 & r_0 \end{pmatrix}. \tag{3.24}$$

Then, defining

$$s_k(M) := \sum_{i-j=k\, mod\, n} m_{ij} \quad where \quad i,j,k \in \{0,\dots,n-1\}, \tag{3.25}$$

we have

$$\begin{pmatrix} s_0 \\ s_1 \\ \vdots \\ s_{n-1} \end{pmatrix} = P \begin{pmatrix} r_{n-1} \\ r_{n-2} \\ \vdots \\ r_0 \end{pmatrix} + L \begin{pmatrix} 0 \\ r_{n-1} \\ \vdots \\ (n-1)r_1 \end{pmatrix} \tag{3.26}$$

where

$$P = \begin{pmatrix} l_{n-1} & 2l_{n-2} & \cdots & (n-1)l_1 & nl_0 \\ 0 & l_{n-1} & \ddots & \ddots & (n-1)l_1 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 2l_{n-2} \\ 0 & \cdots & \cdots & 0 & l_{n-1} \end{pmatrix}. \tag{3.27}$$

Remark 10. *From direct computation $s_k(M) = tr(Q^k M)$.*

Corollary 3.3.4. *If $M = LR$ as in Theorem 3.3.3, then the values s_k for $0 \le k \le n-1$, can be computed in $O(n\log_2(n))$ operations.*

Theorem 3.3.5. *Given $T \in \mathbb{C}^{n\times n}$ a Toeplitz matrix, then*

$$argmin_{C\in\mathcal{C}} \|T^s - C\|_F \tag{3.28}$$

can be computed in $O(2sn\log_2(n))$ operations.

Proof. It is well known that any matrix $C \in \mathcal{C}$, can be written in the form

$$C = \sum_{k=0}^{n-1} c_j Q^j \tag{3.29}$$

where

$$Q = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 0 & 0 \end{pmatrix},$$

and that $\mathcal{C}_{T^s} = \sum_{j=0}^{n-1} \alpha_j Q^j$ where $\alpha_j = [B^{-1}\mathbf{c}]_j$, $B_{i,j} = (Q^i, Q^j)$ and $c_j = (Q^j, T^s)$ for $i, j = 1, \ldots, n-1$.

Observe, moreover, that from Lemma 3.3.1 we have

$$T^s = \sum_{i=1}^{2s} L(\mathbf{g}_i^s) U(\mathbf{b}_i^s)$$

being $\mathbf{g}_i^s$, $\mathbf{b}_i^s$ for $i = 1, \ldots, 2s$, the columns of the generators (G_s, B_s) of T^s. Thesis follows from the linearity of the trace, Theorem 3.3.3 and Remark 10. □

3.4 Experimental results and Conclusions

3.4.1 Experimental Results

In [36, 35, 49] it has been proved that when $\mathcal{C} = \mathrm{sd}\, F$ (F Fourier matrix) and A is a Toeplitz matrix, under suitable hypotheses, the spectrum of $P_{(1)}(A)$ stays bounded from below. In these papers, this particular feature of the superoptimal preconditioner, has been profitably exploited in order to use $P_{(1)}(A)$ as a regularizing preconditioner for A since it provides an approximation of the matrix which ignores some "bad" frequencies corresponding to small eigenvalues. From Remark 6 and Theorem 3.2.3 it is easy to understand that the same good properties hold for the class of preconditioners $P_{(i)}(A)$ presented in this chapter, allowing us to infer that such class of preconditioners presents the same regularizing behaviour of the superoptimal preconditioner.

In this section some preliminary numerical experiences are carried on. Such experiments confirm that the preconditioners proposed in Section 3.2 could be suitably employed as regularizing preconditioners for the conjugate gradient method (see [36, 62, 49, 35]).

We focus on the solution of the system $A\mathbf{x} = \mathbf{g}$ where $A \in \mathbb{R}^{n \times n}$ is severely ill conditioned and $\mathbf{g} \in \mathbb{R}^n$ is contaminated by noise. With the same choices made in [49], the matrix algebra chosen is $\mathcal{C} = \mathrm{sd}\, F$, the dimension of the system A is set to $n = 50$ and the solution vector is the sum of two different "impulses": the $j-$th component of the solution vector $\mathbf{f}$ is

$$f_j = 0.5\, k_{0.1}(x_j + 0.9) + k_{0.05}(x_j - 0.8), \quad x_j = -2 + \frac{4}{n+1}(j+1),$$

where the points x_j are equally distributed in $[-2, 2]$ and $k_\sigma(t)$ denotes the Gaussian distribution

$$k_\sigma(t) := \frac{1}{2\sqrt{\pi\sigma}} e^{-\frac{t^2}{4\sigma}}.$$

The right hand side vector $\mathbf{g}$ is the sum of $\overline{\mathbf{g}} = A\mathbf{f}$ and the noise component η, that is $\mathbf{g} = \overline{\mathbf{g}} + \eta$, where η comes from a normal distribution with zero mean and deviation $\alpha\|\overline{\mathbf{g}}\|$. The matrix A is the symmetric real Toeplitz matrix

$$A_{r,s} = a_{r-s} = \begin{cases} \frac{4}{51}\, k_{0.15}(x_r - x_s) \text{ if } |r-s| \le b, & b \le n, \\ 0, \text{ otherwise }. \end{cases} \tag{3.30}$$

Table 3.1: Experimental Results

	$b = 8,\ \mu_2(A) = 8.36 \times 10^5$			$b = 30,\ \mu_2(A) = 2.98 \times 10^{14}$		
	i	$k_{m.e.}$	$\|\mathbf{f} - \mathbf{x}_{k_{m.e.}}\|/\|\mathbf{f}\|$	i	$k_{m.e.}$	$\|\mathbf{f} - \mathbf{x}_{k_{m.e.}}\|/\|\mathbf{f}\|$
$\alpha = 10^{-3}$	1	3	0.339173	1	4	0.328979
	2	11	**0.323669**	2	11	**0.318053**
	3	16	0.330664	3	16	0.327823
	4	16	0.336834	4	16	0.334599
	5	16	0.339157	5	16	0.337059
$\alpha = 10^{-2}$	1	1	0.427361	1	1	0.424058
	2	3	0.408945	2	3	0.408333
	3	7	**0.405855**	3	7	**0.404537**
	4	10	0.407286	4	11	0.405856
	5	12	0.409811	5	12	0.409198
$\alpha = 10^{-1}$	1	1	1.000000	1	1	1.000000
	2	1	0.609964	2	1	0.607666
	3	1	0.634762	3	1	0.634365
	4	2	**0.599344**	4	2	**0.598570**
	5	2	0.608385	5	2	0.608331

We remark that we have not addressed the important problem of deciding when to stop the PCG method. We perform 50 iterations of the PCG for every preconditioner and, in Table 3.1, we report the iteration $k_{m.e}$ such that the corresponding Relative Restoration Error (RRE), i.e., $\|\mathbf{f} - \mathbf{x}_{k_{m.e.}}\|/\|\mathbf{f}\|$, is minimal. In Figure 3.1 we plot the input signal, the noisy signal and the best reconstructed solution.

3.4.2 Conclusions

As Table 3.1 shows, a higher regularization level and better filtering capabilities for the noise space are obtained in correspondence of higher values of i for the preconditioner $P_{(i)}(A)$. We can hence infer that the class of the regularizing preconditioners proposed in this chapter could be suitably employed even when critical conditions are registered, i.e., high noise level or excessive ill-conditioning. In fact, even in these unfavorable conditions, satisfactory reconstruction performances can be obtained, as it is clear in Figure 3.1.

3.5 Appendix

Given A positive definite, for any $\mathcal{L} = \mathrm{sd}\, U$ we have that $\mu_2(\mathcal{L}_A) \leq \mu_2(A)$. In this Appendix we show that, performing no more than $O(n^2)$ FLOPs, one can define a matrix algebra $\mathcal{L} = \mathrm{sd}\, U$ such that $\mu_2(\mathcal{L}_A)$ is as small as desired, completing hence the observation started after Theorem 3.2.6. Moreover, even if we are aware that in literature such kind of result already exists, we prefer to repeat here the proof in order to keep explicit track of the connections with matrix algebras framework. It can be easily proved that, given

$$A = \begin{bmatrix} a & c \\ c & b \end{bmatrix}, \quad a, b, c \in \mathbb{R}, \quad \text{with} \quad a < b, \quad \text{and any} \quad z \in (a, b),$$

there exist $\alpha, \beta \in \mathbb{R}$ such that $\alpha^2 + \beta^2 = 1$ and

$$\begin{bmatrix} \alpha & -\beta \\ \beta & \alpha \end{bmatrix} \begin{bmatrix} a & c \\ c & b \end{bmatrix} \begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix} = \begin{bmatrix} a_z & c_z \\ c_z & b_z \end{bmatrix}, \text{ with } |c_z| > |c|, \text{ and } a_z,\, b_z \text{ such that}$$

$$a < a_z \leq b_z < b, \; a_z - a = b - b_z, \text{ and } a_z = z\,(\,b_z = z\,) \text{ if } z \in \left(a, \frac{a+b}{2}\right] \left(\, z \in \left[\frac{a+b}{2}, b\right)\,\right)$$

(note that $a_{a+t} = a_{b-t}$, $b_{a+t} = b_{b-t}$, $c_{a+t} = c_{b-t}$, $\forall\, t \in (0, b-a)$). Of course, an analogous result can be stated if $a > b$.

Thus, if $A = \begin{bmatrix} a & c \\ c & b \end{bmatrix}$, $a, b, c \in \mathbb{R}$, $a \neq b$, then, by a Givens similarity transformation, we can cluster the diagonal entries a and b of A as much as we want, maintaining their order; or, equivalently, given any $z \in (\min\{a, b\}, \max\{a, b\})$, we can cluster the diagonal entries a and b of A so to make equal to z the one of them nearer to z.

Let A be an arbitrary $n \times n$ symmetric matrix with real entries. Assume that $[A]_{ii} \neq [A]_{jj}$ for some $i \neq j$ (i.e. we suppose that the equalities $[A]_{11} = [A]_{22} = \ldots = [A]_{nn}$ are not all verified).

Set $A_0 = A$ and, for $k = 0, 1, \ldots$, define the $n \times n$ matrix A_{k+1} from the $n \times n$ matrix A_k as follows:

(1) Choose i, j such that $[A_k]_{ii} < \frac{\operatorname{tr}(A)}{n} < [A_k]_{jj}$.

(2) If $\frac{\operatorname{tr}(A)}{n} - [A_k]_{ii} \leq [A_k]_{jj} - \frac{\operatorname{tr}(A)}{n}$, then introduce the $n \times n$ Givens matrix G_{k+1} such that

$$[G_{k+1}^T A_k G_{k+1}]_{ii} = [A_k]_{ii} + \Big(\frac{\operatorname{tr}(A)}{n} - [A_k]_{ii}\Big) \text{ and } [G_{k+1}^T A_k G_{k+1}]_{jj} = [A_k]_{jj} - \Big(\frac{\operatorname{tr}(A)}{n} - [A_k]_{ii}\Big),$$

otherwise, introduce the $n \times n$ Givens matrix G_{k+1} such that

$$[G_{k+1}^T A_k G_{k+1}]_{jj} = [A_k]_{jj} - \Big([A_k]_{jj} - \frac{\operatorname{tr}(A)}{n}\Big) \text{ and } [G_{k+1}^T A_k G_{k+1}]_{ii} = [A_k]_{ii} + \Big([A_k]_{jj} - \frac{\operatorname{tr}(A)}{n}\Big).$$

After $\hat{k}$ steps (note that $\hat{k} \leq n-1-d$ where $d \geq 0$ is the number of diagonal entries of A equal to $\frac{\operatorname{tr}(A)}{n}$), this procedure yields a matrix

$$G_{\hat{k}}^T \cdots G_2^T\, G_1^T\, A\, G_1\, G_2 \cdots G_{\hat{k}} = \begin{bmatrix} \frac{\operatorname{tr}(A)}{n} & * & \cdot & * \\ * & \frac{\operatorname{tr}(A)}{n} & \cdot & \cdot \\ \cdot & \cdot & \ddots & * \\ * & \cdot & * & \frac{\operatorname{tr}(A)}{n} \end{bmatrix}$$

with diagonal entries all equal to $\frac{\operatorname{tr}(A)}{n}$.

If $U_k = G_1 G_2 \cdots G_k$ and $\mathcal{L}_k = \operatorname{sd} U_k$, then the eigenvalues of $(\mathcal{L}_k)_A = U_k \operatorname{diag}([U_k^T A U_k]_{ss}) U_k^T$ satisfy the following inequalities

$$\min_s [U_k^T A U_k]_{ss} \leq \min_s [U_{k+1}^T A U_{k+1}]_{ss}, \;\; \max_s [U_{k+1}^T A U_{k+1}]_{ss} \leq \max_s [U_k^T A U_k]_{ss},$$

$$l_{k+1} \leq l_k := \max_s [U_k^T A U_k]_{ss} - \min_s [U_k^T A U_k]_{ss},$$

$$l_{\hat{k}} = 0, \;\; (\mathcal{L}_{\hat{k}})_A = \frac{\operatorname{tr}(A)}{n} I.$$

In particular, if A is positive definite, we have

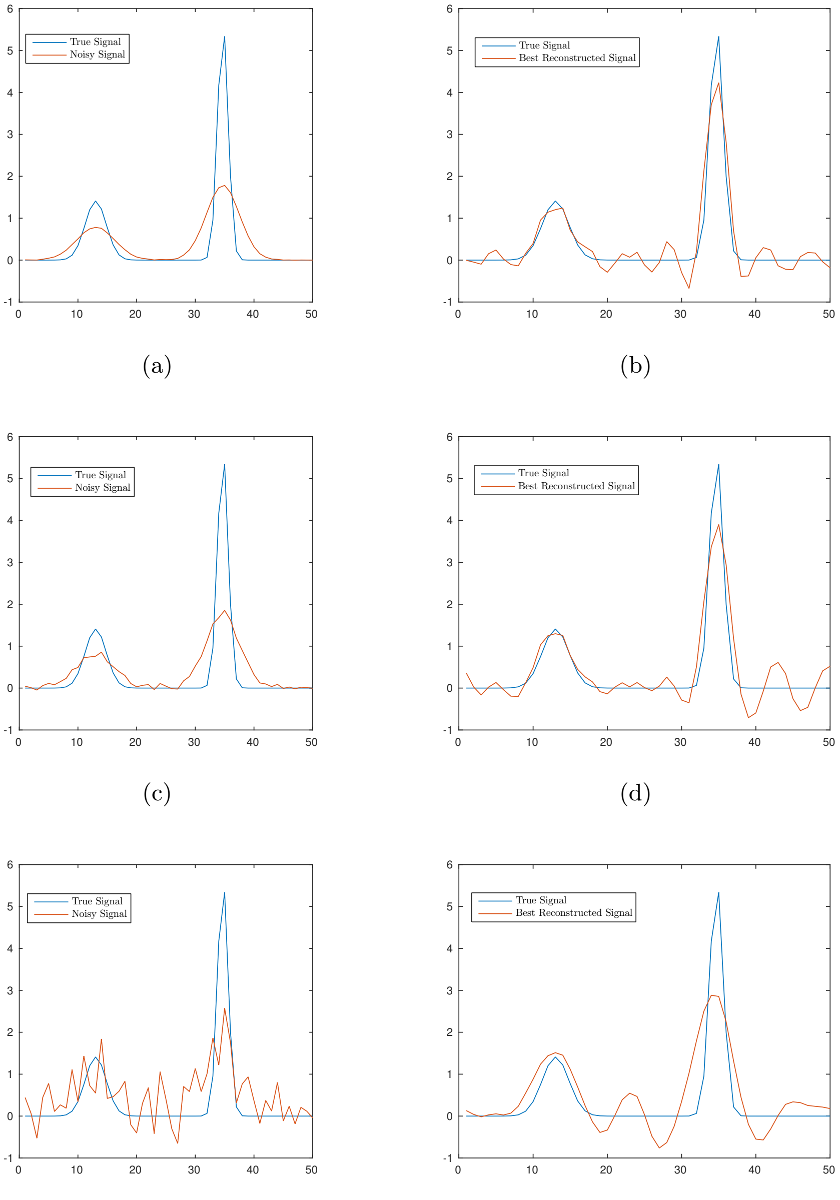
$$\mu_2((\mathcal{L}_{\hat{k}})_A) = 1 \leq \mu_2((\mathcal{L}_{k+1})_A) \leq \mu_2((\mathcal{L}_k)_A) \leq \mu_2((\mathcal{L}_0)_A) \leq \mu_2(A),$$

where $\mathcal{L}_0 = \operatorname{sd} I$.

Of course, if in step (1) we require that $[A_k]_{ii}$ is the smallest among the $[A_k]_{ss}$ smaller than $\frac{\operatorname{tr}(A)}{n}$, and $[A_k]_{jj}$ is the greatest among the $[A_k]_{ss}$ greater than $\frac{\operatorname{tr}(A)}{n}$, then

(i) the sequence l_k decreases as fast as possible;

(ii) if A is positive definite, the sequences l_k and $\mu_2((\mathcal{L}_k)_A)$ decrease as fast as possible.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3.1: $b = 30$. Left column : true and noisy signals for $\alpha = 10^{-3}$, $\alpha = 10^{-2}$ and $\alpha = 10^{-1}$.
Right column : best reconstructed signals.

Chapter 4

Euler-Richardson method preconditioned by weakly stochastic matrix algebras: a potential contribution to Pagerank computation

4.1 Introduction

Markov chains are used to model many different real world systems which evolve in time. When the total number of states which the system may occupy is finite, the chain is typically well represented by a column stochastic matrix S. The state of equilibrium is described by the ergodic distribution p, defined as the solution of the eigenproblem $Sp = p$. Under suitable hypotheses on S, as for instance irreducibility, the solution p is unique and entry-wise positive. The problem of computing such p is one of the crucial issues in Markov processes analysis.

The power method is one of the simplest iterative schemes that converges to the solution p (provided that the eigenvalues of S different from one have absolute value smaller than one). The rate of convergence of such method is well known to be proportional to the magnitude of the subdominant eigenvalue of S. Due to its simplicity and its well understood limit behavior, this method is often used in practice, especially for large-scale unstructured problems.

Examples of growing interest in recent literature are connected with the analysis of complex networks, where, the pattern of the edges of the network is used for localizing important nodes or group of nodes. Many important models, based on matrices or functions of matrices and describing certain features of the network, are related with a random walk defined on the graph and thus exploit extremal eigenvectors and eigenvalues of such matrices (see f.i. [50, 51, 81, 83, 86, 88]). A popular example to which we are particularly interested in is the centrality index problem on graphs known as Pagerank problem ([1] for instance). In that case the web surfer moves randomly on the web graph $W = (V, E)$ and the importance of each node in V is given by the ergodic distribution $w = Gw$ of the random walk defined on W by the Google engine web matrix G (see Section 4.1.2 for more details). The dimension of w in that case is the number of web pages that populate the World Wide Web, thus w roughly has 10^9 entries. The power method can be performed on G in a relatively cheap way by means of the transition matrix of the graph, which is typically sparse. On the other hand, the original formula by Brin and Page [11] defines the same Pagerank vector w as the solution of a linear system whose coefficient matrix is a M-matrix and, as a consequence, the ergodic Pagerank distribution w can be computed either by solving the eigenproblem or by solving such linear system. Thus one can use any linear system solver to approximate w, and several approaches have been investigated and compared to the power method, e.g. [32, 53, 54, 99]. Although such methods sometimes have a convergence rate greater than the one achieved by the power method, they are often more demanding in terms of memory storage and number of operations per step.

The equivalence between the eigenproblem $Sp = p$ and a linear system problem holds in general for a large set of stochastic matrices, not only the Google matrix. Indeed, it has been observed in [98] that, if S is a column stochastic matrix having at least one full row, then 1 is a simple and dominant eigenvalue of S, the ergodic distribution $Sp = p$ is well defined and p is also solution of a M-matrix linear system problem associated to S. In this work we propose a class of simple iterative schemes, named preconditioned Euler-Richardson, to solve such linear system. These methods can be seen as a subset of the class of stationary iterative methods often introduced in terms of a splitting of the coefficient matrix, [77, 104] e.g. Here we observe that this kind of methods provides a natural generalization of the power method and of the well known Jacobi iterative scheme, which correspond to two particular choices of the preconditioner. Then we introduce the concept of weakly stochastic matrix algebra in order to define a new fast and efficient preconditioner, based on Householder unitary transformations. We discuss the relation among the new preconditioned method, the original power method and the Jacobi iterations by providing, in particular, an analysis of the convergence and a number of results on the spectral radius of the respective iteration matrices. Finally we present several numerical tests on synthetic datasets and matrices coming from real-world models. Although the proposed Householder preconditioner does not preserve the nonnegativity of the entries of the original matrix and despite we cannot provide an exhaustive convergence theorem when the coefficient matrix is not assumed symmetric, the analysis made in Section 4.5 and the experiments proposed in Section 4.6 show that the Householder preconditioner reduces significantly the number of iterations without significantly affecting the computational cost nor the memory storage. Thus it stands as a preconditioned version of the power method, well suited for large-scale stochastic M-matrix problems with sparsity structure.

4.1.1 Notational Remarks

In this section we change slightly the notation as follows: for an integer n, the linear space of square $n \times n$ real matrices is denoted by $\mathbb{M}_n$. The symbols O and I denote the zero and the identity matrices, respectively. A matrix is called nonnegative (resp. positive), if its entries are nonnegative (resp. positive) numbers, in symbols $A \geq O$ (resp. $A > O$); for real matrices A, B we write $A \geq B$ if $A - B \geq O$; the cone of nonnegative matrices is denoted by $\mathbb{M}_n^+$, the one of nonnegative vectors by $\mathbb{R}_+^n$.

We use the reverse magnitude ordering

$$|\lambda_1(A)| \geq |\lambda_2(A)| \geq \cdots \geq |\lambda_n(A)| \,.$$

When a matrix A satisfies the equality $A^T\boldsymbol{e} = \boldsymbol{e}$, we say that A is a weakly (column) stochastic matrix. If both A and A^T are weakly stochastic we say that A is doubly weakly stochastic. Note that a nonnegative weakly stochastic matrix is a stochastic matrix in the standard sense, that is a matrix having the set of discrete probability distributions as an invariant. Finally we will not denote anymore vectors by bold letters.

4.1.2 A generalization of the Pagerank linear system formulation

We say that $M \in \mathbb{M}_n$ is a (column) stochastic M-matrix if it can be decomposed as $M = I - \tau A$, with $A \geq O$, $A^T\boldsymbol{e} = \boldsymbol{e}$ and $0 < \tau < 1$. We let SK_n denote the set of such matrices, namely

$$\mathrm{SK}_n = \{I - \tau A \mid \tau \in (0,1),\ A \geq O,\ A^T\boldsymbol{e} = \boldsymbol{e}\}.$$

If $S \in \mathbb{M}_n$ is any stochastic matrix having at least one full row we say that S belongs to Σ_n,

$$\Sigma_n = \{A \in \mathbb{M}_n : A \geq O, A^T\boldsymbol{e} = \boldsymbol{e}, \max_i \min_j a_{ij} > 0\}\,.$$

The following theorem is a collection of results proved in [89, 98]. It shows that the two sets of matrices SK_n and Σ_n are strictly related.

Given S stochastic (nonnegative, weakly stochastic) define $\tau(S) \in \mathbb{R}_+$ and $y_S \in \mathbb{R}_+^n$ as

$$\tau(S) = 1 - \sum_{i=1}^{n} \min_{j=1,\ldots,n} s_{ij}, \quad (y_S)_i = \min_{j=1,\ldots,n} s_{ij}$$

and, for nonzero $\tau(S)$, let $A_S \in \mathbb{M}_n$ be

$$A_S = \tau(S)^{-1}(S - y_S\, \boldsymbol{e}^T)\,.$$

Theorem 4.1.1.

- *Let S be a stochastic matrix. The quantity $\tau(S)$ belongs to the interval $[0,1]$ and $\tau(S) \geq |\lambda|$, $\forall \lambda \in \boldsymbol{\lambda}(S) \setminus \{1\}$.*
- *Let $S \in \Sigma_n$ and p be the ergodic distribution of S (i.e. $p \geq 0$, $p \neq 0$, $Sp = p$, $p^T\boldsymbol{e} = 1$). Then $\tau(S) \in [0,1)$, $y_S \neq 0$, $(I-(S-y_S\boldsymbol{e}^T))p = y_S$. If moreover $\tau(S) > 0$ then $I - (S - y_S\boldsymbol{e}^T) = I - \tau(S)A_S$ with $\tau(S) \in (0,1)$, $A_S \geq 0$ and A_S stochastic (by columns), that is, $I-(S-y_S\boldsymbol{e}^T) \in \mathrm{SK}_n$.*
- *Let $M = I - \tau A \in \mathrm{SK}_n$, $y \geq 0$ nonzero and let x be such that $Mx = y$. Define $\tilde{y} = (1-\tau)/(y^T\boldsymbol{e})y$ and $\tilde{x} = (1-\tau)/(y^T\boldsymbol{e})x$. Then*

$$S := \tilde{y}\boldsymbol{e}^T + \tau A \in \Sigma_n, \tag{4.1}$$

$\tau(S) \in [0,\tau] \subset [0,1)$, and $\tilde{x}$ is the ergodic distribution of S (i.e $\tilde{x} \geq 0$, $\tilde{x} \neq 0$, $S\tilde{x} = \tilde{x}$, $\tilde{x}^T\boldsymbol{e} = 1$).

If $S \in \Sigma_n$, then the theorem above shows that the eigenproblem $Sp = p$ can be solved by solving the linear system $(I - \tau(S)A_S)x = y_S$, and vice-versa, if $M \in \mathrm{SK}_n$, then the solution of $Mx = y$ is a multiple of the ergodic distribution $p = Sp$ (where S is obtained from $M = I - \tau A$ through (4.1)).

It is worth noting that this generalizes to any matrix $S \in \Sigma_n$ a famous property of the Google's Pagerank index, where the particular structure of the problem allows to recast the stationary distribution problem in terms of a linear system problem [72]. Let us observe to conclude, that in general, if an eigenvalue is known, then computing its eigenvector always reduces to solving some consistent linear system; what is interesting in the connection with the Pagerank, as the theorem above shows, is the possibility to find such a system with a nonsingular matrix.

Let $W = (V, E)$ be the direct graph where nodes correspond to web-pages and edges to hyperlinks between pages. The Pagerank index vector p of W is the solution of the equation

$$Gp = p \tag{4.2}$$

where $G = \alpha T^T + (1-\alpha)v\boldsymbol{e}^T$ is the Google engine web matrix, T is the row stochastic transition matrix of W, v is a real positive personalization vector such that $v^T\boldsymbol{e} = 1$ and $0 < \alpha < 1$. Due to the huge dimension of G, several algorithms essentially based on the power method have been proposed to compute the stationary distribution of (4.2). However the original formula by Brin and Page [72] defines the Pagerank vector p as the solution of a M-matrix linear system of the type

$$\gamma(I-\alpha T)^{\mathsf{T}}p = v \qquad \gamma \in \mathbb{R}\,. \tag{4.3}$$

In fact, such system follows immediately from (4.2), by the particular form of G, but can be also recovered by means of Theorem 4.1.1, as we show now:

Since $\max_i \min_j (G)_{ij} \geq (1-\alpha)\max_i v_i > 0$ we deduce that $G \in \Sigma_n$ and $p = (I - \tau(G)A_G)^{-1}y_G$. For the sake of simplicity, suppose that each column of T has at least one zero entry. Then $\tau(G) = 1 - \sum_i \min_j (G)_{ij} = \alpha$, $y_G = (1-\alpha)v$ and $A_G = \alpha^{-1}(G - y_G\boldsymbol{e}^{\mathsf{T}}) = T^{\mathsf{T}}$. This shows, indeed, that p is both the solution of the eigenvector problem (4.2) and of the Pagerank linear system (4.3), with $\gamma = (1-\alpha)^{-1}$.

4.1.3 The Euler-Richardson method

We assume from now on that any random walk considered is described by a stochastic matrix $S \in \Sigma_n$. We discuss a method which computes the solution of the eigenproblem $p = Sp$ by solving the associated stochastic M-matrix linear system. By virtue of Theorem 4.1.1 the two problems are equivalent, so for the sake of clarity and generality we always assume that a stochastic M-matrix $M \in \mathrm{SK}_n$ and a nonnegative vector $y \geq 0$ are given, and we are interested in the solution of the equation $Mx = y$.

The preconditioned Euler-Richardson method (briefly, PER method) for the solution of $Mx = y$ is the stationary iterative scheme based on the splitting $M = P - (P - M)$ and defined by the following sequence ([104] e.g.)

$$x_{k+1} = P^{-1}y + (I - P^{-1}M)x_k, \qquad k = 0, 1, 2, 3, \ldots \tag{4.4}$$

where P is a suitable nonsingular preconditioner. The iteration matrix of such method is evidently $I - P^{-1}M$, thus we write

$$H(P) = I - P^{-1}M$$

to denote such matrix, underlining the dependence upon the chosen preconditioner P. Since the eigenvalues of any $M \in \mathrm{SK}_n$ have positive real part, the standard Euler-Richardson method (ER), obtained by setting $P^{-1} = \omega I$ inside (4.4), is convergent for all $\omega \in (0, \min \frac{2\,\mathrm{Re}(\lambda_i)}{|\lambda_i^2|})$ and its rate of convergence is optimized by setting $\omega_0 = \arg\min_{\omega\in\mathbb{R}} \rho(H(\omega I))$. This is the simplest iterative method and it may not be the best choice in terms of efficiency. However, its simplicity allows its easy implementation for problems that are unstructured and have huge dimension, as for instance the Pagerank problem. In particular, as for the power method, the ER scheme requires only one real vector to store the data. Moreover, the analysis made throughout this paper shows that (4.4) can be seen as a preconditioned power method.

This opens the way to a number of further investigations and improvements, as, for instance, the variety of techniques proposed to speed-up the power method for Pagerank computation can be potentially applied to (4.4) (e.g. extrapolation [9, 10, 12, 69] or structural adaptive mathods [60, 67, 68]). More precisely, when $M \in \mathrm{SK}_n$, one can show that $\omega_0 = 1$ (see e.g. [99, Thm. 4.2]). It is therefore easy to realize that, if $M = I - \tau(S)A_S$ is defined as in Theorem 4.1.1, then the ER, with $\omega = 1$, and the power methods are very close. In particular, we show in the sequel that there exists a simple choice P_{pm} for the preconditioner P in (4.4) that gives rise exactly to the same convergent sequence as the one defined by the power method applied to the original eigenproblem $Sp = p$. To this aim we initially devote Section 4.2 to define and investigate the concept of weakly stochastic matrix algebra, then we show that the power method preconditioner P_{pm} is indeed defined in terms of such algebras. In Section 3 we consider a new preconditioner chosen in a suitable weakly stochastic and low complexity Householder algebra, giving rise to a competitive method that can be implemented with linear memory storage allocation (two real vectors) and with the same order of operations per step of the power method. Besides its direct application to the iterative scheme (4.4), the analysis made in Sections 4.2 and 4.4 provides a number of interesting relations among matrix algebras, stochastic and nonnegative matrices, and in our opinion it is of self interest.

4.2 Low complexity matrix subspaces

Let us start recalling some facts in order to simplify the reading of the following sections.
Given $J_1, \dots, J_m \in \mathbb{M}_n$ linearly independent, the subspace $\mathcal{L} = \mathrm{span}(J_1, \dots, J_m)$ is said to be of low complexity if for any $L \in \mathcal{L}$, the order of complexity required to multiply L times a vector or to solve a linear system with L as coefficient matrix, is much less then $O(n^2)$ (tipically $O(n \log n)$, see examples in Section 4.2.2). A preconditioner for (4.4) can thus be chosen inside $\mathcal{L}$. A popular choice for $P \in \mathcal{L}$ is the so called optimal fit preconditioner obtained by projecting the coefficient matrix over $\mathcal{L}$. For any given matrix $X \in \mathbb{M}_n$, we write $\mathcal{L}_X$ to denote its projection over $\mathcal{L}$. Note that, by definition of projection, one has that $\|\mathcal{L}_X - X\|_F \leq \|Y - X\|_F$, for any $Y \in \mathcal{L}$, being $\|\cdot\|_F$ the Frobenius norm. A possible representation of $\mathcal{L}_X$ is as follows

$$\mathcal{L}_X = \textstyle\sum_{i=1}^{m} (B^{-1}c)_i J_i \tag{4.5}$$

where B is the Gram matrix $b_{ij} = (J_i, J_j)$ and c is the vector $c_i = (J_i, X)$. Of course the number of arithmetic operations required to identify such $\mathcal{L}_X$ in $\mathcal{L}$ should be “not too large”, that is the linear system $Bx = c$ should be easily solvable. It is shown in [34, 44] that, under suitable hypotheses on $\mathcal{L}$, the matrix B is in $\mathcal{L}$ itself or belongs to other low complexity classes, and thus the projection $\mathcal{L}_X$ can be obtained with a small amount of computations whenever $\mathcal{L}$ is of low complexity. Typical examples of such spaces are the algebras of matrices simultaneously diagonalized by a unitary transform, henceforth briefly called *sd$U$ spaces*. Fixed any unitary matrix $U \in \mathbb{M}_n$, any such a space is denoted by $\mathrm{sd}\, U$ and is defined by

$$\mathcal{U} = \mathrm{sd}\, U = \{Ud(\boldsymbol{\lambda})U^H \mid \lambda \in \mathbb{C}^n\}\,.$$

Any $\mathcal{U} = \mathrm{sd}\, U$ is an n-dimensional matrix algebra, thus a commutative set of matrices closed under addition, multiplication and inversion. Moreover, the following further representation for $\mathcal{U}_X$ holds for $\mathcal{U} = \mathrm{sd}\, U$:

$$\mathcal{U}_X = Ud((U^H XU)_{ii}\, i = 1, \dots, n)U^H \tag{4.6}$$

where, for a matrix M, $d(M)$ denotes the diagonal matrix with diagonal entries $m_{11}, \dots, m_{nn}$. As $I \in \mathrm{sd}\, U$, the linearity of the projection operator implies that $\mathcal{U}_M = I - \tau\mathcal{U}_A$, for any $M \in \mathrm{SK}_n$. Therefore the problem of defining a preconditioner for (4.4) reduces to the problem of identifying the projection of the nonnegative and weakly stochastic matrix A, and solving low complexity systems with $I - \tau\mathcal{U}_A$ as coefficient matrix.

Note that in many cases, if $\mathcal{U}_A$ is a weakly stochastic matrix, then $\mathcal{U}_M$ is invertible. Indeed if $\|A\|_2 \leq 1$, and this is true at least for all $A \geq O$ which are stochastic and normal, then, using the Cauchy-Schwartz inequality, we get

$$\rho(\mathcal{U}_A) = \max_{i=1,\dots,n} |u_i^{\mathsf{T}} Au_i| \leq \max_{\|x\|_2=1} |x^{\mathsf{T}} Ax| \leq \max_{\|x\|_2=1} \|Ax\|_2 = \|A\|_2 \leq 1\,,$$

where u_i are the columns of U, defining $\mathcal{U}$. Thus $\mathcal{U}_M = I - \tau\mathcal{U}_A$ is evidently invertible.

Next Section 4.2.1 contains a theorem characterizing the $\mathrm{sd}\, U$ spaces $\mathcal{U}$ such that A weakly stochastic implies $\mathcal{U}_A$ weakly stochastic as well. Then in Section 4.4.1 we show that P_{pm} can be defined in terms of such spaces, and finally we introduce a new $\mathrm{sd}\, U$ space where to select a different preconditioner for (4.4).

4.2.1 Weakly stochastic matrix algebras

This subsection is devoted to characterize the $\mathrm{sd}\, U$ matrix algebras $\mathcal{U}$ which preserve the weakly stochasticity of A, when projecting A on them. For a vector u such that $u^{\mathsf{T}}U$ has no zero entries, define the map $L_u : \mathbb{C}^n \longrightarrow \mathcal{U}$ that associates to a vector x the matrix $L_u(x)$ of $\mathcal{U}$ such that $u^{\mathsf{T}}L_u(x) = x^{\mathsf{T}}$. As $u^{\mathsf{T}}U$ has no zero entries, it is not difficult to see that L_u is a well defined bijection, for any $\mathrm{sd}\, U$ matrix algebra $\mathcal{U} = \mathrm{sd}\, U$. However, it is worth pointing out that the class of spaces for which the operator L_u is a well defined bijection contains properly the set of $\mathrm{sd}\, U$ spaces (see [44]). A direct computation shows that the following representation of $L_u(x) \in \mathcal{U} = \operatorname{sd} U$ holds

$$L_u(x) = Ud(U^T x)d(U^T u)^{-1}U^H \tag{4.7}$$

Definition 9. *If there exists a column of U which has all constant entries then we call $\mathcal{U} = sd U$ a* weakly stochastic $\operatorname{sd} U$ matrix algebra.

The reason of such name is made evident by the following Theorem 4.2.1 which completely characterizes those $\operatorname{sd} U$ spaces $\mathcal{U}$ with the property that the projection over $\mathcal{U}$ of a weakly stochastic matrix, is still weakly stochastic.

Theorem 4.2.1. *Let $\mathcal{U} = sd U$ for some unitary U, and let u be any vector such that $u^T U$ has no zero entries. The following statements are equivalent*

1. *There exists an index k s.t. the k^{th} column of U has constant entries*
2. $\boldsymbol{e}\boldsymbol{e}^T \in \mathcal{U}$
3. *For any vector $x \in \mathbb{C}^n$, it holds $L_u(x)\boldsymbol{e} = L_u(x)^T\boldsymbol{e} = \left(x^T\boldsymbol{e}/u^T\boldsymbol{e}\right)\boldsymbol{e}$*
4. *For any matrix $A \in \mathbb{M}_n$, it holds $\mathcal{U}_A\boldsymbol{e} = \mathcal{U}_A^T\boldsymbol{e} = \frac{1}{n}(\boldsymbol{e}^T A\boldsymbol{e})\boldsymbol{e}$*

In particular if A or A^T are weakly stochastic then $\mathcal{U}_A$ is doubly weakly stochastic.

Proof. (1)$\Longrightarrow$(2) Let D_i be the diagonal rank one matrix whose only nonzero entry is $(D_i)_{ii} = \boldsymbol{e}^T\boldsymbol{e}$. The rank one matrices $R_i = UD_iU^H = (\boldsymbol{e}^T\boldsymbol{e})(Ue_i)(Ue_i)^H$ clearly all belong to $\mathcal{U}$, and in particular $R_k = \boldsymbol{e}\boldsymbol{e}^T \in \mathcal{U}$. (2) $\Longrightarrow$ (3) Since L_u is a bijection and since $\boldsymbol{e}\boldsymbol{e}^T \in \mathcal{U}$, we have $L_u(\boldsymbol{e}) = \boldsymbol{e}\boldsymbol{e}^T/\boldsymbol{e}^T u$. Thus formula (4.7) implies

$$\boldsymbol{e}^T\left(\tfrac{\boldsymbol{e}^T u}{\boldsymbol{e}^T x}\right)L_u(x) = \left(\tfrac{\boldsymbol{e}^T u}{\boldsymbol{e}^T x}\right)x^T L_u(\boldsymbol{e}) = \left(\tfrac{\boldsymbol{e}^T u}{\boldsymbol{e}^T x}\right)\left(\tfrac{x^T\boldsymbol{e}}{\boldsymbol{e}^T u}\right)\boldsymbol{e}^T = \boldsymbol{e}^T,$$

that is $L_u(x)^T\boldsymbol{e} = (x^T\boldsymbol{e}/u^T\boldsymbol{e})\boldsymbol{e}$, for any $x \in \mathbb{C}^n$. Now using the hypothesis $\boldsymbol{e}\boldsymbol{e}^T \in \mathcal{U}$ and the fact that matrices in $\mathcal{U}$ commute, we have the equality $L_u(x)\boldsymbol{e}\boldsymbol{e}^T - \boldsymbol{e}\boldsymbol{e}^T L_u(x) = O$ implying that $L_u(x)\boldsymbol{e} = \left(x^T\boldsymbol{e}/\boldsymbol{e}^T u\right)\boldsymbol{e}$. (3) $\Longrightarrow$ (4) Since $\mathcal{U}_A \in \mathcal{U}$, there exists a vector $z_A \in \mathbb{C}^n$ such that $\mathcal{U}_A = L_u(z_A)$. It is enough to show that $\boldsymbol{e}^T u/\boldsymbol{e}^T z_A = n/\boldsymbol{e}^T A\boldsymbol{e}$. Matching the representations (4.7) and $\mathcal{U}_A = Ud(U^H AU)U^H$ we get $z_A^T = u^T Ud(U^H AU)U^H$. Thus since $U^H\boldsymbol{e} = \overline{\alpha}e_k$, with $|\alpha|^2 = 1/n$, it holds

$$z_A^T\boldsymbol{e} = \overline{\alpha}u^T Ue_k(U^H AU)_{kk} = u^T\boldsymbol{e}(U^H AU)_{kk} = u^T\boldsymbol{e}\left(\boldsymbol{e}^T A\boldsymbol{e}/n\right).$$

(4)$\Longrightarrow$(1) The eigenvectors of any matrix $L \in \mathcal{U} = \operatorname{sd} U$ are the columns of U. The fact that $\boldsymbol{e}$ is an eigenvector of $\mathcal{U}_A$ for any $A \in \mathbb{M}_n$, implies that there exists k such that $Ue_k = \alpha\boldsymbol{e}$, with $|\alpha|^2 = 1/n$. □

4.2.2 Examples

Low complexity matrix algebras have been studied extensively in relatively recent years in the context of preconditioning, displacement and optimization, see e.g. [23, 34, 38, 37, 44, 100, 101] and the references therein. Among the best known matrix algebras developed in past literature, we recognize several weakly stochastic matrix algebras. For the sake of completeness, we briefly discuss some relevant examples in the following. Other examples can be found among the Hartley-type algebras and the matrix algebras associated with trigonometric transforms, see [5, 6, 8] e.g. In particular, it is not difficult to observe that the Hartley [5] and the τ_{11} [8] algebras are weakly stochastic.

Circulants

The discrete Fourier transform is realized through the action of the Fourier matrix $F_n = 1/\sqrt{n}\left(e^{(}2\pi\mathbf{i}(ij)/n)\right)_{i,j=0}^{n-1}$. It is easy to check that F_n is unitary and that $F_n e_1 = \boldsymbol{e}/\sqrt{n}$, that is, the first column of F_n is constant. The matrix algebra $\mathcal{C} = \mathrm{sd}\,(F_n)$ is usually referred to as the *circulant* algebra. If $\varSigma$ is the modulo-n shift backward matrix

$$\varSigma = \begin{bmatrix} & 1 & & \\ & & \ddots & \\ & & & 1 \\ 1 & & & \end{bmatrix}$$

then $\{I, \varSigma, \varSigma^2, \dots, \varSigma^{n-1}\}$ is a basis for $\mathcal{C}$ and this basis is made by nonnegative and mutually orthogonal matrices. It follows that the computation of the projection $\mathcal{C}_A$ only requires additive operations among the entries of A.

Haar

The discrete Haar transform is realized through the action of the Haar matrix W_n, which can be described recursively in terms of two matrices Q_n and D_n. Let $S_1 = Q_1 = D_1 = 1$, then for $n = 2, 4, 8, \dots, 2^m$ let $S_n = d\left(1/\sqrt{2}, 1, \dots, 1\right)$ and

$$Q_n = \begin{bmatrix} \boldsymbol{e}e_1^T & Q_{n/2} \\ Q_{n/2} & -\boldsymbol{e}e_1^T \end{bmatrix}, \quad D_n = \begin{bmatrix} D_{n/2}S_{n/2} & \mathbf{0} \\ \mathbf{0} & D_{n/2}S_{n/2} \end{bmatrix}.$$

Thus the Haar matrix is given by

$$W_n = Q_n D_n = \begin{bmatrix} \frac{1}{\sqrt{n}}\boldsymbol{e}e_1^T & W_{n/2}S_{n/2} \\ W_{n/2}S_{n/2} & -\frac{1}{\sqrt{n}}\boldsymbol{e}e_1^T \end{bmatrix}$$

Other equivalent definitions of this matrix can be found in the literature ([55] e.g.) obtained by permuting rows and columns of W_n. Using the proposed construction, it is not difficult to check that such W_n is unitary, that its first column is $(1/\sqrt{n})\boldsymbol{e}$ and that its multiplication times a generic vector can be performed very cheaply. The generated algebra $\mathcal{W} = \mathrm{sd}\,(W_n)$ is called Haar matrix algebra.

Eta

Consider the matrix U defined as follows:

$$\begin{cases} U_{i,1} = 1/\sqrt{n}, & \\ U_{ij} = \sqrt{\frac{2}{n}}\left\{\cos\left(\frac{(2i-1)(j-1)\pi}{n}\right) + \sin\left(\frac{(2i-1)(j-1)\pi}{n}\right)\right\}, & j = 2, \dots, \lceil \frac{n}{2} \rceil \\ U_{i,\frac{n}{2}+1} = \frac{(-1)^{i-1}}{\sqrt{n}}, & \text{if } n \text{ is even} \\ U_{ij} = \sqrt{\frac{2}{n}}\left\{\cos\left(\frac{(2i-1)(j-1)\pi}{n}\right) + \sin\left(\frac{(2i-1)(j-1)\pi}{n}\right)\right\}, & j = \lfloor \frac{n}{2} + 2 \rfloor, \dots, n \end{cases}$$

for $i = 1, \dots, n$. The matrix U is unitary and real, moreover $Ue_1 = \frac{1}{\sqrt{n}}\boldsymbol{e}$. As a consequence, setting $\eta = \mathrm{sd}\,U$, we obtain a weakly stochastic $\mathrm{sd}\,U$ low complexity matrix algebra (see [6, 44, 94, 95] for the fast sine-cosine transforms used in the computations involving U). For the sake of completeness let us recall the following representation [6]: $\eta = \mathcal{C}_s + J\mathcal{C}_s$ where $\mathcal{C}_s$ is the algebra of symmetric circulant matrices and J is the reverse identity matrix, $(J)_{ij} = 1$ if $i+j = n+1$, and $(J)_{ij} = 0$ otherwise. It follows that any matrix A in η is symmetric and persymmetric and satisfies the cross-sum rule

$$a_{i-1,j} + a_{i+1,j} = a_{i,j-1} + a_{i,j+1}, \quad i, j = 1, \dots, n$$

with border conditions $a_{0,i} = a_{1,n+1-i}$, $i = 1, \dots, n$. See also [44, 34, 100] and the references therein.

Hadamard

The Sylvester-Hadamard orthogonal matrix of order $n = 2^m$ is defined recursively by the rule $H_m = H_1 \otimes H_{m-1}$, where

$$H_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

and $\otimes$ is the Kronecker product. As the first column of H_1 has all ones entries, we immediately see that the first column of H_m has constant entries as well. The associated unitary matrix is $\frac{1}{\sqrt{2}^m}H_m$ and the associated matrix algebra $\mathrm{sd}\,(\frac{1}{\sqrt{2}^m}H_m)$ is therefore a weakly stochastic $\mathrm{sd}\,U$ algebra. The Hadamard matrix has relevant applications both in statistics, where it is used f.i. to uncover the dependencies in a multivariable data set, and in error correcting codes theory. We refer to [107] for a detailed description of applications and properties of the Hadamard matrix.

4.3 Preserving the nonnegativity of the entries

As the original matrix A is both nonnegative and weakly stochastic, one ideally would like to have $\mathcal{L}_A$ both nonnegative and weakly stochastic. The following theorem characterizes the subspaces $\mathcal{L}$ which preserve the nonnegativity of a matrix A, when projecting A onto them.

Definition 10. *We say that $\mathcal{L}$ is a* nonnegative matrix space *if $\mathcal{L}_A \geq O$ for any $A \geq O$, i.e. if the projection $A \mapsto \mathcal{L}_A$ preserves the cone of nonnegative matrices.*

Theorem 4.3.1. *$\mathcal{L}$ is a nonnegative matrix space if and only if $\mathcal{L}$ has a basis of orthonormal nonegative matrices.*

Proof. It is straightforward to see that if $\mathcal{L}$ has a basis of orthogonal nonnegative matrices, then $\mathcal{L}$ is a nonnegative space. Let $\pi : \mathbb{M}_n \to \mathcal{L}$ be the projection operator. If $\mathcal{L}$ is a nonnegative matrix space, we have that $\pi(\mathbb{M}_n^+) \subseteq \mathbb{M}_n^+$, that is the projection leaves the cone of nonnegative matrices invariant. For the sake of simplicity let us consider the vectorization operator which realizes the standard isomorphism between $\mathbb{M}_n$ and $\mathbb{R}^{n^2}$. We have $\mathrm{vec}(\mathbb{M}_n^+) \equiv \mathbb{R}_+^{n^2}$, $\Pi = \mathrm{vec}(\pi) \in \mathbb{M}_{n^2}^+$ and $\boldsymbol{\lambda}(\Pi) = \{0, 1\}$. The multiplicity of $1 = \rho(\Pi)$ is the dimension of $\mathcal{L}$. If $\dim(\mathcal{L}) = 1$ then the proof is trivial. If $\dim(\mathcal{L}) = k > 1$ then $\rho(\Pi)$ is not simple; thus, due to the Perron-Frobenius theorem, Π is reducible and there exists a permutation matrix $Q \in \mathbb{M}_{n^2}^+$ such that

$$Q\Pi Q^T = \Pi_1 \oplus \cdots \oplus \Pi_k \oplus N$$

where each Π_i is irreducible and N is nilpotent. Therefore there exist k positive vectors $x_1, \ldots, x_k$ such that $\Pi_i x_i = x_i$. Let $\tilde{x}_i$ be the embedding of x_i into $\mathbb{R}^{n^2}$, obtained by filling x_i with zero entries, and set $y_i = Q^T \tilde{x}_i$. Then $\Pi y_i = y_i$ and $y_i^T y_j = 0$ for $i = 1, \ldots, k$ and any $j \neq i$. Finally note that, if $Y_i \in \mathbb{M}_n$ is the matrix such that $y_i = \mathrm{vec}(Y_i)$, we have $Y_i \geq O$, $Y_i \in \mathcal{L}$ and $(Y_i, Y_j) = y_i^T y_j$ so that $Y_1, \ldots, Y_k$ is a basis for $\mathcal{L}$ made by orthogonal nonnegative matrices. □

It is immediate to note that the algebra of diagonal matrices $\mathcal{D} = \mathrm{sd}\, I$ is a nonnegative matrix space. If we define the preconditioner for PER as the projection of M on $\mathcal{D}$ then we have $P = \mathcal{D}_M = d(M)$ and the method in (4.4) coincides with the Jacobi iterative scheme. The convergence properties of such method are well known (see for instance [77, 104]). We analyze them in more details in Section 4.5, taking into account the structure of the matrices $M \in \mathrm{SK}_n$. Despite its simple formulation and cheap implementation, the diagonal preconditioner does not preserve the weakly stochasticity of the original matrix. This has two drawbacks: on the one hand the use of $P = \mathcal{D}_M$ in (4.4) has a less clear relation to the power method, since the power method preconditioner P_{pm} is weakly stochastic, as we will show in the next section; on the other hand, the numerical implementations in Section 4.6 show that the property of being weakly stochastic ensures faster convergence.

To our knowledge, if we exclude the multilevel generalizations, the only low complexity matrix algebra satisfying the hypothesis of both Theorems 4.2.1 and 4.3.1 is the circulant algebra. The use of a circulant preconditioner in this context has been analyzed in details in [99]. Although the analysis in [99] shows a reduction of the number of iterations with respect the classic power method, the use of the circulant algebra requires one Fourier transform and two complex vectors to be stored per each step. In Section 4.4.3 we will introduce a new preconditioner based on a weakly stochastic matrix algebra diagonalized by a suitable Householder transform. We propose a convergence analysis of PER with such novel preconditioner under the assumption that A is symmetric. Although an exhaustive convergence analysis for the more realistic case where A is generic is still missing, we point out that the low memory storage and the linear order of operations per step required by the new technique make it effectively applicable also when the dimension of the problem is huge. This is further highlighted by the numerical tests proposed in Section 4.6.

4.4 The choice of the preconditioner

In this section we show that the power method preconditioner P_{pm} for the PER scheme is a matrix belonging to a class of weakly stochastic $\mathrm{sd}\, U$ matrix algebras. Then we develop a new matrix algebra defined in terms of Householder unitary transformations, leading to a new cheap preconditioner for PER. In the subsequent sections we analyze the convergence of PER method and we provide numerical evidences of the advantages obtained by using the new Householder preconditioner.

4.4.1 The power method embedded into a PER iterative scheme

Recall that the unpreconditioned Euler-Richadson method is obtained by chosing $P = I$. Given $M = I - \tau A \in \mathrm{SK}_n$, the power method preconditioner P_{pm} for the solution of $Mx = y$, instead, is the following rank-one correction of the identity matrix

$$P_{\mathrm{pm}} = I - \frac{\tau}{n}\boldsymbol{e}\boldsymbol{e}^T.$$

We observe that P_{pm} belongs to any weakly stochastic $\mathrm{sd}\, U$ matrix algebra. In other words such preconditioner belongs to the intersection $\cap\{\,\mathrm{sd}\, U \mid U \text{ has a constant column}\}$. Indeed, let U be any unitary matrix such that Ue_k has constant entries, and consider the diagonal matrix $D = d(1, \dots, 1, 1-\tau, 1, \dots, 1)$, where $1-\tau$ lies in the k-th diagonal position, then $P_{\mathrm{pm}} = UDU^H$. It is worth noting that this is somehow analogous to the property shown in point 2 of Theorem 4.2.1. Indeed observe that, as for the projection $\mathcal{U}_M = I - \tau \mathcal{L}_A$, the matrix $P_{\rm pm}$ has the structure $P_{\rm pm} = I - \tau E$, where E is the weakly stochastic matrix $E = \boldsymbol{e}\boldsymbol{e}^T/n$ which indeed belongs to any weakly stochastic $\mathrm{sd}\, U$ algebra. Note moreover that

$$P_{\rm pm}^{-1} = I + \left(\frac{\tau}{1-\tau}\right)\frac{\boldsymbol{e}\boldsymbol{e}^T}{n}\,. \tag{4.8}$$

In view of Theorem 4.1.1 we can show the connection between the PER method applied to $M \in \mathrm{SK}_n$ and the power method for the ergodic distribution of a Markov chain described by $S \in \Sigma_n$.

Theorem 4.4.1. *Given $S \in \Sigma_n$ let $\tau(S)$, A_S and y_S be defined as before Theorem 4.1.1. When the preconditioner is $P = P_{\rm pm}$, the PER method for the solution of $Mx = (I - \tau(S)A_S)x = y_S$ coincides with the power method applied to S.*

Proof. Let $\{x_k\}$ be the sequence defined by the PER method (4.4). By Theorem 4.1.1 and the formula above, it follows that $y_S^T\boldsymbol{e} = 1 - \tau(S)$ and $P_{\rm pm}^{-1}\boldsymbol{e} = (1-\tau(S))^{-1}\boldsymbol{e}$. As a consequence we observe that $\boldsymbol{e}^T x_k = 1$ implies

$$\boldsymbol{e}^T x_{k+1} = \boldsymbol{e}^T P_{\rm pm}^{-1} y_S + \boldsymbol{e}^T (I - P_{\rm pm}^{-1}M)x_k = 1\,.$$

Therefore we can assume that the entries of the initial x_0 sum up to 1, and that $\boldsymbol{e}^T x_k = 1$ for all $k \geq 0$. We have

$$\begin{aligned} x_{k+1} &= P_{\rm pm}^{-1} y_S + (I - P_{\rm pm}^{-1}M)x_k \\ &= y_S + \frac{\tau(S)}{n}\boldsymbol{e} + x_k - Mx_k - \frac{\boldsymbol{e}^T M x_k}{1-\tau(S)}\frac{\tau(S)}{n}\boldsymbol{e} \\ &= y_S + \tau(S)A_S x_k \end{aligned}$$

and, by Theorem 4.1.1, $S = \tau(S)A_S + y_S\boldsymbol{e}^T$, therefore $x_{k+1} = Sx_k$, and the proof is complete. □

Let $|\lambda_1(X)| \geq \cdots \geq |\lambda_n(X)|$ be the eigenvalues of a matrix X. As $S \in \Sigma_n$, by Theorem 4.1.1, we have $|\lambda_2(S)| < 1$ and, by the well known behavior of the power method, we have that x_k converges to the solution of $Mx = y_S$ as $O(|\lambda_2(S)|^k)$. However, a different bound can be observed by using the equivalence shown in Theorem 4.4.1 as indeed we will show in Section 4.5 that x_k converges to x as $O(\tau(S)^k|\lambda_2(A_S)|^k)$. Note that $S \in \Sigma_n$ implies, by Theorem 4.1.1, that $\tau(S) < 1$ and A_S is stochastic too, thus $|\lambda_2(S)| < 1$ and $\tau(S)|\lambda_2(A_S)| < 1$. However in several cases (for instance if A_S is primitive, i.e. $A_S^k > O$ for some integer power $k > 0$), one has $|\lambda_2(A_S)| < 1$, thus $\tau(S)|\lambda_2(A_S)| < \tau(S)$ whereas $|\lambda_2(S)| \leq \tau(S)$, thus suggesting that $\tau(S)|\lambda_2(A_S)|$ could provide a better bound on the convergence rate.

4.4.2 The Householder weakly stochastic matrix algebra

Let $\mathcal{U}$ be any weakly stochastic $\operatorname{sd} U$ algebra. The power method is obtained by applying PER and choosing P inside $\mathcal{U}$ as the matrix with the following eigenvalues: $1-\tau$, with multiplicity one, and 1 with multiplicity $n-1$. To improve the performances of the power method, we define a new preconditioner by replacing the eigenvalues 1 with the spectrum of the projection $\mathcal{U}_M$ of M onto $\mathcal{U}$. Note that the eigenvalue $1-\tau$ is an invariant, that is $1-\tau \in \boldsymbol{\lambda}(\mathcal{U}_M)$ for any weakly stochastic $\operatorname{sd} U$ algebra $\mathcal{U}$. In fact, as $A^{\mathsf{T}}\boldsymbol{e} = \boldsymbol{e}$ implies $\mathcal{U}_A^{\mathsf{T}}\boldsymbol{e} = \boldsymbol{e}$, we have $\mathcal{U}_M^{\mathsf{T}}\boldsymbol{e} = (I - \tau\mathcal{U}_A)^{\mathsf{T}}\boldsymbol{e} = (1-\tau)\boldsymbol{e}$. Also note that $\mathcal{U}_M$ minimizes the distance $\|X - M\|_F$ among the matrices $X \in \mathcal{U}$, being $\|X\|_F = \sqrt{(X,X)}$ the Frobenius norm. Therefore $\|\mathcal{U}_M - M\|_F \leq \|P_{\mathrm{pm}} - M\|_F$ for any weakly stochastic $\operatorname{sd} U$ space $\mathcal{U}$, and the inequality is strict up to trivial cases. This motivates the choice $P = \mathcal{U}_M$, in place of the classic $P = P_{\mathrm{pm}}$, to improve the performances of the method.

The example spaces shown in Section 4.2.2 are defined in terms of fast transformations U whose space and time complexities are $O(n \log n)$. In order to keep the complexity of the PER iterations as low as possible, we define a weakly stochastic $\operatorname{sd} U$ algebra diagonalized by a Householder transformation. As we prove in the next section this allows us to keep the time and space complexity per step linear in n. It is worth mentioning that, due to their linear computational complexity, matrix algebras diagonalized by Householder unitary transforms have been already involved in a number of applications. In particular they have been recently used to define competitive iterative optimization algorithms, whose space and time per step complexity is $O(n)$, c.f. [27, 41]. Let us introduce a linear space $\mathcal{H}$ of the form

$$\mathcal{H} = \{H(w)d(z)H(w) \mid z \in \mathbb{C}^n\}, \qquad H(w) = I - 2ww^H, \qquad \|w\| = 1$$

where $H(w)$ is a Householder unitary matrix such that $H(w)e_k = \left(e^{\mathbf{i}\theta}/\sqrt{n}\right)\boldsymbol{e}$, for some $\theta \in \mathbb{R}$. We shall observe that all the Householder matrices of this kind are of the form $H(w^{\pm})$, where w^+ and w^-, are two suitable vectors in $\mathbb{R}^n$. We firstly look at the k-th column of $H(w)$, and we get $(I - 2ww^H)e_k = e_k - 2w(w^H e_k) = \left(e^{\mathbf{i}\theta}/\sqrt{n}\right)\boldsymbol{e}$. Therefore

$$2\overline{w_k}\, w = e_k - \frac{e^{\mathbf{i}\theta}}{\sqrt{n}}\boldsymbol{e}. \tag{4.9}$$

The k-th component of the above equality implies $2\overline{w}_k w_k = 1 - \left(e^{\mathbf{i}\theta}/\sqrt{n}\right)$, so that $w_k \neq 0$, $\theta \in \{0, \pi\}$, and thus $|w_k|^2 = (1 \pm 1/\sqrt{n})/2$. As a consequence the k-th entry of w is given by either of the two following formulas, corresponding to $\theta = 0$ and $\theta = \pi$, respectively:

$$w_k = \left(\frac{\sqrt{n}-1}{2\sqrt{n}}\right)^{1/2} e^{\mathbf{i}\phi}, \quad \text{or} \quad w_k = \left(\frac{\sqrt{n}+1}{2\sqrt{n}}\right)^{1/2} e^{\mathbf{i}\phi}, \quad \phi \in \mathbb{R}\,.$$

Writing now the j-th component of (4.9) for $j \neq k$ we obtain an analogous formula also for the other entries of w:

$$w_j = -\frac{e^{\mathbf{i}\phi}}{\sqrt{2}\sqrt[4]{n}\sqrt{\sqrt{n}-1}}, \qquad \text{or} \qquad w_j = \frac{e^{\mathbf{i}\phi}}{\sqrt{2}\sqrt[4]{n}\sqrt{\sqrt{n}+1}}, \qquad \phi \in \mathbb{R}\,.$$

The previous relations can be written in compact form, showing that any vector w such that $H(w)$ defines a weakly stochastic algebra is either $w = w_\phi^+$ or $w = w_\phi^-$, where

$$w_\phi^- = e^{\mathbf{i}\phi}\cdot\beta_n^-(\sqrt{n}e_k + \boldsymbol{e}), \qquad w_\phi^+ = e^{\mathbf{i}\phi}\cdot\beta_n^+(\sqrt{n}e_k - \boldsymbol{e}), \qquad \beta_n^\pm = \frac{1}{\sqrt{2}\sqrt[4]{n}\sqrt{\sqrt{n}\mp 1}}\,.$$

This finally shows an explicit formula for all the possible weakly stochastic Householder algebras. Note indeed that the Householder matrices $H(w_\phi^+)$ and $H(w_\phi^-)$ do not depend on $\phi \in \mathbb{R}$ therefore, setting $w^\pm = w_0^\pm$, we see that $H(w^+)$ and $H(w^-)$ are the only two Householder matrices which define a weakly stochastic $\mathrm{sd}\,U$ algebra. They are both real unitary matrices and such that $H(w^\pm)e_k = (\pm 1/\sqrt{n})\boldsymbol{e}$.

4.4.3 The Householder PER method

We use the notation

$$\chi(B) = \text{ computational cost of the product } B \times \textit{vector}\,.$$

It is not difficult to check that, when $P = \mathcal{U}_M$ and $\mathcal{U} = \mathrm{sd}\,U$, the overall computational cost of the PER method for the solution of $Mx = y$, $M = I - \tau A \in \mathrm{SK}_n$, is

$$\chi(A) + \chi(U) + \chi(U^H) + O(n) \tag{4.10}$$

for each step, plus a preprocessing phase which is required essentially for the computation of the eigenvalues $\lambda_1, \ldots, \lambda_n$ of $\mathcal{U}_A$, and whose computational complexity highly depends on the chosen U as, by (4.6), $\lambda_i = (U^H A U)_{ii}$.

The Householder PER method (HPER in short) is obtained by projecting M over one of the two Householder $\mathrm{sd}\,U$ algebras introduced above. As we can freely choose either w^+ or w^-, in what follows we set $w = w^+$ and assume for simplicity that the constant column of $H(w)$ is the first one (i.e. $k = 1$ in the construction of Section 4.4.2). Then we let $\mathcal{H} = \mathrm{sd}\,H(w)$. Let us briefly analyze the computational cost of HPER.

Set $H(w) = I - 2ww^{\boldsymbol{T}}$, where $w = \beta_n(\sqrt{n}e_1 - \boldsymbol{e})$ and $\beta_n^2 = \frac{1}{2}\frac{1}{\sqrt{n}(\sqrt{n}-1)}$. We immediately see that $\chi(H(w)) = O(n)$. Therefore for this choice, even if A is a sparse or a strongly structured matrix, the complexity per step of HPER is dominated by $\chi(A)$ as the estimate (4.10) becomes $\chi(A) + O(n)$ when $U = H(w)$. Note that this is the same complexity required by the standard power method iterations.

Note that a preprocessing phase is required for the computation of the diagonal entries of $H(w)AH(w)$ (that is the eigenvalues of $\mathcal{H}_A$), as well as to compute Aw, A^Tw and $H(w)y = y + 2\beta_n(n^{-1/2} - y_1)w$. Observe that

$$H(w)AH(w) = A - 2(Aww^T + ww^TA - 2\gamma_n ww^T) \tag{4.11}$$

where

$$\gamma_n = w^TAw = n\beta_n^2\left((A)_{11} - \frac{1}{\sqrt{n}} - \frac{1}{\sqrt{n}}(A\boldsymbol{e})_1 + 1\right). \tag{4.12}$$

Therefore, from (4.11), we obtain the equalities

$$d(H(w)AH(w))_{ii} = (A)_{ii} - 2w_i\Big((Aw)_i + (w^TA)_i - 2\gamma_n w_i\Big),$$

$i = 1, \dots, n$, which, together with (4.12), show that $\chi(A)$ operations are sufficient to compute the diagonal entries of $H(w)AH(w)$. We conclude that the overall cost of the preprocessing phase is $\chi(A) + \chi(A^T) + O(n)$. Let us point out that even in the worst case, when A is a general, non structured and dense matrix, by the particular form of w, $O(n^2)$ additive operations and $O(n)$ multiplicative operations are sufficient to compute Aw and w^TA.

4.5 Convergence analysis

In this section we analyze the convergence of the preconditioned Euler-Richardson method applied to the linear system $Mx = y$ when $M \in \mathrm{SK}_n$ and the preconditioner P is the optimal fit of M onto a matrix algebra $\mathrm{sd}\,U$. First of all we state the following simple but somewhat general theorem. We shortly outline a possible proof.

Theorem 4.5.1. *Let $M \in \mathrm{SK}_n$ and $\mathcal{L}$ be a subspace of $\mathbb{M}_n$ such that $I \in \mathcal{L}$ and $A \geq \mathcal{L}_A \geq O$, then $\rho(H(\mathcal{L}_M)) < 1$. That is the PER scheme* (4.4) *with $P = \mathcal{L}_M$ converges.*

Proof. The linearity of the projection and the fact that $I \in \mathcal{L}$ imply that $\mathcal{L}_M = I - \tau\mathcal{L}_A$. Since $A \geq \mathcal{L}_A \geq O$, the Perron-Frobenius theorem implies $\rho(\mathcal{L}_A) \leq \rho(A) = 1$. Thus $\mathcal{L}_M$ is a M-matrix as well, and $\mathcal{L}_M^{-1} \geq O$. Moreover clearly $\mathcal{L}_M - M = \tau(A - \mathcal{L}_A) \geq O$. We have as a consequence that both $H(\mathcal{L}_M)$ and $(I - H(\mathcal{L}_M))^{-1}$ are nonnegative matrices. Indeed $H(\mathcal{L}_M) = \mathcal{L}_M^{-1}(\mathcal{L}_M - M) \geq O$ and $(I - H(\mathcal{L}_M))^{-1} = M^{-1}(\mathcal{L}_M - M + M) = M^{-1}(\mathcal{L}_M - M) + I \geq O$. Let $\rho = \rho(H(\mathcal{L}_M))$ and let $z \geq 0$, $z \neq 0$ be such that $H(\mathcal{L}_M)z = \rho z$. We have that $y = (I - H(\mathcal{L}_M))^{-1}z$ is nonzero and nonnegative, and $z = (1-\rho)(I - H(\mathcal{L}_M))^{-1}z = (1-\rho)y$. But z is nonzero and nonnegative, that is $(1-\rho) > 0$. □

It is worth noting that the theorem above is related with the concept of regular splitting of M-matrices, see [104] e.g. In fact, under the hypothesis of Theorem 4.5.1, letting $N = \mathcal{L}_M - M$, we observe that $M = \mathcal{L}_M - N$ is a regular splitting of M, that is $N \geq 0$ and $\mathcal{L}_M^{-1} \geq 0$.

Corollary 4.5.2. *Let $M = I - \tau A \in \mathrm{SK}_n$ and let $\{J_1, \dots, J_m\}$ be a set of nonnegative mutually orthogonal matrices such that*

1. *$\#\{$nonzero entries of $J_k\} = 1$ for any $k \in \{1, \dots, m\}$*
2. *$I \in \mathcal{L} = \mathrm{span}\{J_1, \dots, J_m\}$*

Then $\mathcal{L}_M$ is invertible and the PER method with $P = \mathcal{L}_M$ is convergent.

Proof. To prove this corollary we simply show that the hypothesis of the previous theorem are all satisfied. First of all, since $J_1, \dots, J_m$ are orthogonal and nonnegative, $\mathcal{L}$ satisfies the hypothesis of Theorem 4.3.1 by construction, thus $\mathcal{L}_A \geq O$.

Now, for any $k \in \{1, \dots, m\}$, let $(J_k)_{i_k,j_k}$ be the unique nonzero element of J_k. We can obviously assume $(J_k)_{i_k,j_k} = 1$ without loosing generalities. Then, in the notation of (4.5), we have $B = I$, $c_k = a_{i_k,j_k}$ and $\mathcal{L}_A = \sum_k a_{i_k,j_k} J_k$, implying that for any $i \in \{1, \dots, n\}$ it holds

$$\sum_{j=1}^{n} |(\mathcal{L}_A)_{ij}| \leq \sum_{j=1}^{n} |(\sum_{k=1}^{m} a_{i_k,j_k} J_k)_{ij}| = \sum_{j,k:(i_k,j_k)=(i,j)} |a_{i_k,j_k}| \leq \sum_{j=1}^{n} |a_{i,j}|$$

and hence

$$\rho(\mathcal{L}_A) \leq \|\mathcal{L}_A\|_\infty = \max_i \sum_{j=1}^{n} |(\mathcal{L}_A)_{ij}| \leq \max_i \sum_{j=1}^{n} |a_{i,j}| \leq \|A\|_\infty = 1 = \rho(A).$$

Therefore $\rho(\mathcal{L}_A) \leq \rho(A)$. It follows that $\mathcal{L}_M = I - \tau \mathcal{L}_A$ is invertible and $\mathcal{L}_M^{-1} \geq O$. Finally note that

$$(A - \mathcal{L}_A)_{ij} = \begin{cases} a_{ij} & ij \notin \{i_1 j_1, \dots, i_m j_m\} \\ 0 & otherwise \end{cases}$$

which implies $A \geq \mathcal{L}_A$. The thesis follows. □

The set $\{e_1 e_1^T, e_2 e_2^T, \dots, e_n e_n^T\}$ is a simple example of nonnegative matrices satisfying the hypothesis of the corollary above. Their linear span is the algebra of diagonal matrices $\mathcal{D} = \mathrm{sd}\, I$, the PER method applied with $P = \mathcal{D}_M$ coincides with the classic Jacobi method and it is well known to be convergent (see [104] e.g.). Nonetheless the next Theorem 4.5.3 shows that a more precise control on the rate of convergence can be achieved. Recall that the classic unpreconditioned ER scheme is obtained for $P = I$ and one has $\rho(H(I)) \leq \tau$.

Theorem 4.5.3. *Let $\mathcal{D} = \text{sd}\, I$ be the algebra of diagonal matrices and let $M = I - \tau A \in \text{SK}_n$. If $\varepsilon = \arg\max \left\{ \lambda \geq 0 \mid \min_i a_{ii} \geq \frac{1-\tau^\lambda}{1-\tau^{1+\lambda}} \right\}$, then $\rho(H(\mathcal{D}_M)) \leq \tau^{1+\varepsilon}$. In particular $\rho(H(\mathcal{D}_M)) \leq \tau$ for any $M \in \text{SK}_n$ and if $\min_i a_{ii} \geq (1+\tau)^{-1}$ then $\rho(H(\mathcal{D}_M)) \leq \tau^2$.*

Proof. By exploiting the entries of $H = I - \mathcal{D}_M^{-1} M$ we have

$$(H)_{ij} = \begin{cases} \frac{\tau a_{ij}}{1-\tau a_{ii}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}.$$

By Gershgorin localization theorem, the eigenvalues of H are contained inside the ball in $\mathbb{C}$ centered over the origin and with radius $R = \max_i \sum_{j \neq i} (H)_{ij}$. By using the identity $\sum_j a_{ij} = 1$ and by observing that $a_{ii} \geq \frac{1-\tau^\varepsilon}{1-\tau^{1+\varepsilon}}$ if and only if $\frac{1-a_{ii}}{1-\tau a_{ii}} \leq \tau^\varepsilon$, we get $\sum_{j \neq i} (H)_{ij} = \tau \frac{1-a_{ii}}{1-\tau a_{ii}} \leq \tau^{1+\varepsilon}$, thus $\rho(H) \leq R \leq \tau^{1+\varepsilon}$. □

Note that the preceding theorem shows that the larger are the diagonal entries of A, the more likely $\rho(H(\mathcal{D}_M))$ is small.

For the sake of completeness let us point out that other examples of set of matrices satisfying the hypothesis of Corollary 4.5.2 have been considered in literature. For example the authors of [36] define the set of matrix spaces (and matrix algebras) $\{U \Delta U^H : \Delta \in \mathcal{M}(E)\}$, where U is a unitary matrix, $E \in \{0,1\}^{n \times n}$ is a non-degenerate mask matrix, $\mathcal{M}(E) = \{E \circ A \text{ s.t. } A \in \mathbb{M}_n\}$ and $\circ$ is the Hadamard entry-wise product. It is not difficult to observe that any space $\mathcal{L} = \mathcal{M}(E)$ indeed admits a basis of matrices $\{J_1, \dots, J_m\}$ satisfying the hypothesis of the corollary. Of course, for such space $\mathcal{L}$, the matrix $\mathcal{L}_M$ is $\mathcal{L}_M = E \circ M$.

4.5.1 Projecting over a weakly stochastic algebra

The results presented in this section give a further and more detailed intuition on why, when $M \in \text{SK}_n$, a preconditioner for (4.4) based on weakly stochastic matrix algebras behaves well. We assume for the remaining part of this section that any stochastic matrix A, defining the given stochastic M-matrix $M = I - \tau A \in \text{SK}_n$, has a simple dominant eigenvalue $\rho(A) = 1$.

Let $\mathcal{U} = \text{sd}\, U$ be a weakly stochastic matrix algebra. For a matrix $M = I - \tau A \in \text{SK}_n$ we have $\mathcal{U}_M = I - \tau \mathcal{U}_A = I - \tau U d(z_A) U^H$ (which we assume invertible) and

$$H(\mathcal{U}_M) = I - (I - \tau U d(z_A) U^H)^{-1} (I - \tau A),$$

where z_A is the vector whose entries are the eigenvalues of $\mathcal{U}_A$, ordered as usual. Recall that the components of z_A are the diagonal entries of $U^H A U$.

We claim that, for any such algebra $\mathcal{U}$, the spectrum of the iteration matrix $H(\mathcal{U}_M)$ only depends on the eigenvalues $\lambda_i(A)$ and $\lambda_i(\mathcal{U}_A)$, for $i \neq 1$.

In other words when the preconditioner is chosen projecting M over a weakly stochastic algebra, the leading eigenvalues $\lambda_1(A) = \lambda_1(\mathcal{U}_A) = 1$ of A and its projection, are not involved in the analysis of the convergence.

To this end let us observe that, since $\mathcal{U}$ is weakly stochastic, there exists an index $k \in \{1, \dots, n\}$ such that Ue_k has constant elements, that is $Ue_k = \alpha \boldsymbol{e}$ for some $\alpha \in \mathbb{C}$ such that $n\alpha\overline{\alpha} = 1$. Therefore the k-th entry of z_A is

$$(z_A)_k = (U^H AU)_{kk} = e_k^T U^H AUe_k = \overline{\alpha}\boldsymbol{e}^T AUe_k = \overline{\alpha}\boldsymbol{e}^T Ue_k = \overline{\alpha}\alpha n = 1$$

Observe analogously that $e_k^T U^H(I - \tau A)U = (1-\tau)e_k^T$, that is the k-th entry of the k-th row of $U^H MU$ is $1-\tau$ and the remaining components are all zeros. The same holds for $U^H \mathcal{U}_M U$. It follows that the two matrices $U^H MU$ and $U^H \mathcal{U}_M U$ have the same block structure, which we represent here when $k = 1$ for easy of notation:

$$U^H MU = \begin{pmatrix} 1-\tau & 0^T \\ f & I - \tau V^H AV \end{pmatrix}, \quad U^H \mathcal{U}_M U = \begin{pmatrix} 1-\tau & 0^T \\ 0 & I - \tau d(V^H AV) \end{pmatrix}, \tag{4.13}$$

where V is the partial isometry given by the last $n-1$ columns of U and f is a suitable $n-1$ vector. Of course, also $U^H H(\mathcal{U}_M)U = I - (U^H \mathcal{U}_M U)^{-1}(U^H MU)$ has the same structure.

Now consider the matrix $A_2 = A - q\boldsymbol{e}^T/\sqrt{n}$, where q is any vector such that $q \geq 0$, $q^T\boldsymbol{e} = \sqrt{n}$. Also, let $A = XJX^{-1}$ be the Jordan decomposition of A. As $\rho(A) = 1$ is a simple eigenvalue of A, the k-th row of X^{-1} is constant. That is $e_k^T X^{-1} = \boldsymbol{e}^T/\sqrt{n}$ and $Je_k = (J^T e_k) = e_k$. Note moreover that $e_k^T X^{-1} q = 1$. Then

$$X^{-1}A_2 X = X^{-1}\left(A - qe_k^T X^{-1}\right)X = J - X^{-1}qe_k^T = \begin{pmatrix} 1 & 0^T \\ 0 & \tilde{J} \end{pmatrix} - \begin{pmatrix} 1 & 0^T \\ \tilde{q} & O \end{pmatrix} \tag{4.14}$$

where $\tilde{J}$ is the Jordan form of A besides the 1×1 block associated with $\rho(A)$, $\tilde{q}$ is the $n-1$ vector made by the entries of $X^{-1}q$ besides the k-th one, and the right most block representation has been shown for the case where $k = 1$, for notational convenience.

We deduce that, for any $q \geq 0$ with $q^T\boldsymbol{e} = \sqrt{n}$, we have $\boldsymbol{\lambda}(A_2) = \boldsymbol{\lambda}(X^{-1}A_2X) = \boldsymbol{\lambda}(\tilde{J}) \cup \{0\} = \{\boldsymbol{\lambda}(A) \setminus \{1\}\} \cup \{0\}$. Moreover note that this shows that the eigenvalues of A_2 and $V^H AV$ coincide, besides 0. In fact $U^H MU$ and $X^{-1}MX$ are similar and thus, comparing (4.13) and (4.14), the blocks $I - \tau V^H AV$ and $I - \tau\tilde{J}$ have same eigenvalues. The same holds for the eigenvalues of $\mathcal{U}_{A_2}$ and $d(V^H AV)$. Therefore, for any weakly stochastic algebra $\mathcal{U}$, the spectrum of the iteration matrix $H(\mathcal{U}_M)$ of the PER method (4.4), can be decomposed as

$$\boldsymbol{\lambda}(H(\mathcal{U}_M)) = \boldsymbol{\lambda}(I - (I - \tau\mathcal{U}_{A_2})^{-1}(I - \tau A_2))$$

which finally shows our claim.

It is worth noting that the observations did so far apply to the choice $P = P_{\mathrm{pm}}$. Precisely, from (4.8) we get

$$H(P_{\mathrm{pm}}) = I - P_{\mathrm{pm}}^{-1}M = \tau(A - \boldsymbol{e}\boldsymbol{e}^{\mathsf{T}}/n)$$

and therefore, choosing $q = \boldsymbol{e}/\sqrt{n}$,

$$\boldsymbol{\lambda}(H(P_{pm})) = \boldsymbol{\lambda}(\tau A - \frac{\tau}{n}\boldsymbol{e}\boldsymbol{e}^{\mathsf{T}}) = \boldsymbol{\lambda}(\tau X^{-1}AX - \frac{\tau}{n}X^{-1}\boldsymbol{e}\boldsymbol{e}^{\mathsf{T}}X) = \boldsymbol{\lambda}(\tau A_2).$$

By using Theorems 4.1.1 and 4.4.1, we deduce a new upper-bound on the convergence rate of the power method applied to $S \in \Sigma_n$. Namely we have that the sequence Sx_k converges to the ergodic distribution of S with a rate of convergence bounded by $O(\tau(S)|\lambda_2(A_S)|)$, where $\tau(S)$ and A_S are defined in terms of S as in Theorem 4.1.1.

As a matter of fact, when the preconditioner in $\mathcal{U}$ is not P_{pm}, but instead is chosen in $\mathcal{U}$ as the matrix with smaller Euclidean distance from M, we cannot provide a theoretical control on the eigenvalues of $I-(I-\tau\mathcal{U}_{A_2})^{-1}(I-\tau A_2)$. However both intuition and numerical tests shown in Section 4.6 suggest that the spectral radius of $H(\mathcal{U}_M)$ is significantly smaller than the one of $H(P_{\mathrm{pm}})$.

For the sake of completeness, we observe that, when A is stochastic nonnegative and symmetric, further results hold as stated in the following Theorem 4.5.4, where the eigenvalues of W (symmetric) are ordered as $\lambda_1(W) \geq \cdots \geq \lambda_n(W)$.

Theorem 4.5.4. *Let $M = I - \tau A \in \mathrm{SK}_n$ be such that A is symmetric and $\rho(A)$ is simple. Let $\mathcal{U} = sdU$ be a weakly stochastic matrix algebra. Then*

$$\rho(H(\mathcal{U}_M)) \leq \tau \max\left\{\frac{\lambda_2(\mathcal{U}_A) - \lambda_n(A)}{1 - \tau\lambda_2(\mathcal{U}_A)}, \frac{\lambda_2(A) - \lambda_n(\mathcal{U}_A)}{1 - \tau\lambda_n(\mathcal{U}_A)}\right\}$$

Proof. To lighten the notation, we denote with M_2 the matrix $I - \tau A_2$, where A_2 is the matrix such that $A = \boldsymbol{e}\boldsymbol{e}^{\mathsf{T}}/n + A_2$. Note that the symmetry of A implies that both M and $\mathcal{U}_M$ are positive definite matrices. In fact, M is clearly real symmetric and $\boldsymbol{\lambda}(M) \in \mathbb{R}_+$, whereas $\mathcal{U}_M$ has the form $\mathcal{U}_M = Ud(U^HMU)U^H$. Therefore the eigenvalues of $\mathcal{U}_M$ are inside the convex hull of $\boldsymbol{\lambda}(M)$, so they are real and positive, implying that $\mathcal{U}_M$ is positive definite. A known consequence of the Weyl's inequalities states that, for any two positive definite matrices X and Y, the following inequalities hold (see for instance [4])

$$\lambda_n(X)\lambda_n(Y) \leq \lambda_n(XY) \leq \lambda_1(XY) \leq \lambda_1(X)\lambda_1(Y).$$

Collecting such inequalities, the considerations we did shortly above the statement of this theorem, and the fact that $\lambda_i(\mathcal{U}_M^{-1}) = \lambda_i(\mathcal{U}_M)^{-1}$ for any

$i \in \{1, \ldots, n\}$, we get

$$\begin{aligned}\rho(H(\mathcal{U}_M)) &= \max\Big\{\lambda_1(\mathcal{U}_{M_2}^{-1}M_2) - 1,\ 1 - \lambda_n(\mathcal{U}_{M_2}^{-1}M_2)\Big\} \\ &\leq \max\left\{\frac{\lambda_2(M)}{\lambda_n(\mathcal{U}_M)} - 1,\ 1 - \frac{\lambda_n(M)}{\lambda_2(\mathcal{U}_M)}\right\} \qquad (4.15)\\ &= \tau \max\left\{\frac{\lambda_2(\mathcal{U}_A) - \lambda_n(A)}{1 - \tau\lambda_2(\mathcal{U}_A)},\ \frac{\lambda_2(A) - \lambda_n(\mathcal{U}_A)}{1 - \tau\lambda_n(\mathcal{U}_A)}\right\}\end{aligned}$$

and the thesis follows. □

In particular, under the same hypothesis of the theorem above, we have

$$\rho(H(\mathcal{U}_M)) < \frac{2\tau}{1-\tau},$$

and hence if τ is small enough, precisely $\tau \leq 1/3$, then PER with $P = \mathcal{U}_M$ and A symmetric, converges for any choice of the weakly stochastic algebra $\mathcal{U}$.

4.6 Numerical comparisons

In this final section we present a number of numerical tests comparing the behavior of three methods on several synthetic and real-world datasets. The linear system solved is $Mx = (I - \tau A)x = y$ where y is a random vector with entries in $[0, 1]$, $\tau = 0.9$ and A is a stochastic matrix defined as follows. We consider X, the adjacency matrix of the dataset, then we normalize it into the associated transition matrix $T = D^{-1}X$, being D the diagonal matrix $d_{ii} = e_i^T X \boldsymbol{e}$. In order to force a lower bound on the diagonal entries, we introduce a further parameter $0 < \beta < 1$, and finally define the matrix A as the convex combination $A = \beta I + (1 - \beta)T^T$. The standard Pagerank random walk is retrieved for $\beta = 0$. The methods are defined by different choices of the preconditioner P in (4.4):

HPER. The Euler-Richardson method preconditioned via the optimal fit $P = \mathcal{H}_M$, where $\mathcal{H} = \operatorname{sd} H(w)$ is the Householder weakly stochastic matrix algebra discussed in Section 4.4.3. This method consists of a preprocessing phase in which the quantities β_n, γ_n, Aw, $H(w)y$ and $d(H(w)AH(w))$ must be computed. The overall cost of this initial computation is $O(\chi(A) + \chi(A^T)) + O(n)$. Then each step of the method is performed by the recursive computations of $x_{k+1} = \mathcal{H}_M^{-1}y + (I - \mathcal{H}_M^{-1}M)x_k$, each step requires $O(\chi(A)) + O(n)$ operations.

Jacobi. The Euler-Richardson method preconditioned with the diagonal optimal fit matrix $P = \mathcal{D}_M = I - \tau d(A)$. The rate of convergence is given here by Theorem 4.5.3. Note that $\mathcal{D}$ can be seen as the span of n orthogonal nonnegative rank-one matrices, and thus, accordingly with Theorem 4.3.1, both $\mathcal{D}_A$ and $\mathcal{D}_M^{-1}$ maintain the nonnegativity of the entries. On the other hand this choice for P does not ensure any stochasticity property of P^{-1}, in the general case. We recall that this method coincides with the standard Jacobi iterations.

Power method. The Euler-Richardson method preconditioned with the power method matrix P_{pm} defined as the following rank-one correction of the identity $P_{\mathrm{pm}} = I - \frac{\tau}{n}\boldsymbol{e}\boldsymbol{e}^T$. As discussed in Section 4.4.1, this method coincides with the power method applied to the stochastic eigenproblem $Sp = p$, where S is obtained from A as discussed in Theorem 4.1.1.

It is worth mentioning that both the Jacobi and the power methods can count on a convergence theorem with an explicit upper bound on the convergence rate. The spectral radius of the iteration matrix for the Jacobi method is upper-bounded by $\tau^{1+\varepsilon}$, where ε is defined as in Theorem 4.5.3 and increases with the magnitude of the diagonal entries of A. We have introduced the parameter β to appreciate the acceleration gained by this method when the a_{ii} are close to 1.

Similarly the spectral radius of the iteration matrix for the power method is upper-bounded by $\tau|\lambda_2(A)|$. Note that this convergence rate is linear in τ, but the method can be sensibly faster than the Jacobi one, when the second eigenvalue of A is small. This property is essentially given by the use of a weakly stochastic preconditioner.

Unfortunately we do not have an explicit convergence theorem for HPER for non-symmetric problems. However note that the use of a weakly stochastic preconditioner combines somehow the two previous convergence properties. In fact, on the one hand, as for the power method, the dominant eigenvalue of A is deflated and does not influence the spectral radius of the iteration matrix, on the other hand, as for the Jacobi scheme, the preconditioner is related with the diagonal entries of the matrix U^HAU, that is similar to A. The tests that we present in the following show that HPER goes faster than the Jacobi and power methods and, in particular, its convergence rate increases with β (as the Jacobi iterations do) and increases when the magnitude of the subdominant eigenvalue of A decreases (as for the power method).

We point out that the choice $\tau = 0.9$ has been done accordingly with typical network applications, as for instance the Google's Pagerank centrality. It is worth pointing out that the smaller is τ, the simpler is the problem, thus we do not consider small values of τ in the following.

Tables 4.1, 4.2 show the results for a randomly generated binary matrix X. The eigenvalues of X in this case cluster around the origin ([97, 99] e.g.). As the value of β increases, the number of iterations shown is the median over 10 tests. The HPER method significantly outperforms the other ones. The subsequent Table 4.3 shows, instead, the behavior of the three methods on a number of real world datasets. The test matrices considered are part

	Random matrix of order $n = 10^7$			
	β	HPER	Jacobi	power method
# *iterations*	0.1	11	104	37
	0.2	8	99	32
	0.5	6	60	54
	0.9	4	19	141

Table 4.1: The table shows the number of iterations required by the three methods to achieve a precision of 10^{-7} on the residual $\|Mx - y\|$, when β ranges from 0.1 to 0.9. The coefficient matrix here is defined in terms of a random binary matrix of order 10^7. The number of iterations shown is the median over 10 tests.

	Random matrix of order $n = 10^3$, precision 10^{-7}			
	β	HPER	Jacobi	power method
# *iterations*	0.1	6	157	9
	0.2	6	141	12
	0.5	6	91	24
	0.9	5	25	89

	Random matrix of order $n = 10^3$, precision 10^{-10}			
	β	HPER	Jacobi	power method
# *iterations*	0.1	9	216	13
	0.2	8	194	17
	0.5	8	126	33
	0.9	7	34	122

	Random matrix of order $n = 10^3$, precision 10^{-13}			
	β	HPER	Jacobi	power method
# *iterations*	0.1	11	276	16
	0.2	11	247	21
	0.5	10	160	43
	0.9	9	43	155

Table 4.2: The table shows the number of iterations required by the three methods to achieve a precision of 10^{-7}, 10^{-10}, 10^{-13} on the residual $\|Mx - y\|$, when β ranges from 0.1 to 0.9. The coefficient matrix here is defined in terms of a random binary matrix of order 10^3. The number of iterations shown is the median over 10 tests.

of the University of Florida sparse matrix collection [31]. The matrices considered are both symmetric (undirected) and unsymmetric (directed).

	Network	n	β	# iterations		
				HPER	Jacobi	power method
Undirected	Delaunay19	524 288	0.1	156	184	172
	Delaunay21	2 097 152	0.2	178	206	221
	Italy OSM	6 686 493	0.2	201	214	241
	Europe OSM	50 912 018	0.2	159	167	186
Directed	Indian web crawl	1 382 908	0.2	187	209	235
	Wikipedia 2006	3 148 440	0.1	159	177	175
	Wikipedia 2007	3 566 907	0.1	157	178	173
	LJournal 2008	5 363 260	0.1	154	192	185

Table 4.3: The table shows the number of iterations required by the three methods to achieve a precision of 10^{-7} on the residual $\|Mx - y\|$. Tests here have been made on real world matrices of different sizes, and the value of β has been chosen between 0.1 and 0.2. Tests for larger values of β (here omitted) show a significant acceleration of HPER over the other two methods.

Chapter 5

Updating Broyden Class-type descent directions by Householder adaptive transforms

5.1 Introduction

It is well known that the *BFGS* minimization method is competitive with modified Newton-Raphson-type method, since each iteration of *BFGS* can be performed with only $O(n^2)$ FLOPs and no Hessian evaluation is required. In *BFGS* the search direction $\mathbf{d}_{k+1}$ is defined as $-B_{k+1}^{-1}\mathbf{g}_{k+1}$ where $\mathbf{g}_{k+1} = \nabla f(\mathbf{x}_{k+1})$ and B_{k+1} is a rank-2 correction $\Phi(B_k, \mathbf{s}_k, \mathbf{y}_k)$ of the previous real positive definite Hessian (spd) approximation B_k, defined in terms of the two current difference vectors $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ and $\mathbf{y}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$, and such that

$$B_{k+1}\mathbf{s}_k = \mathbf{y}_k. \tag{5.1}$$

In order to minimize the computational complexity per iteration and the memory required for implementation, it is proposed in [40, 43, 41, 42, 39] to use a *BFGS*-type updating Hessian approximation formula of the form

$$B_{k+1} = \Phi(\tilde{B}_k, \mathbf{s}_k, \mathbf{y}_k) \tag{5.2}$$

where $\tilde{B}_k$ is a suitable approximation of B_k. This scheme is named $\mathcal{L}$QN in the particular case the matrix $\tilde{B}_k$ is chosen to be the projection $\mathcal{L}_{B_k}$ of the matrix B_k in a matrix algebra $\mathcal{L}$. Two possible classes of $\mathcal{L}$QN are considered, the $\mathcal{S}$ecant $\mathcal{L}$QN, satisfying the secant equation (5.1), and the $\mathcal{N}$on $\mathcal{S}$ecant $\mathcal{L}$QN where $\mathbf{d}_{k+1} = -\tilde{B}_{k+1}^{-1}\mathbf{g}_{k+1}$ with $\tilde{B}_{k+1}$ not necessarily satisfying (5.1). The convergence property of $\mathcal{NS}$ $\mathcal{L}$QN [40] and the experimental observed gain of efficiency of the $\mathcal{S}$ $\mathcal{L}$QN with respect to *BFGS* and to its limited memory version *L-BFGS* on some specific class of problems [7, 17, 48], are due, essentially, to the simple fact that the projection $\mathcal{L}_{B_k}$ of the matrix B_k in the space $\mathcal{L}$ approximate in a sufficiently accurate way – under suitable hypotheses on the space $\mathcal{L}$ –, the spectral information of the matrix B_k (see [102]). In [39, 41] it is observed that an adaptive choice of $\mathcal{L}$, i.e, using different algebras $\mathcal{L}^{(k)}$ for each iteration k, could allow to preserve more information from the original matrix B_k, and thus improve the efficiency of $\mathcal{L}$QN techniques. In [27] it is introduced a convergent $\mathcal{L}^{(k)}$QN scheme and preliminary numerical experiences confirm the goodness of the proposed approach when compared with $\mathcal{L}QN$ where $\mathcal{L}$ is fixed.

The main contribution of this work is twofold. On the one hand we extend the theoretical framework and the convergence theory developed in [27] for $\mathcal{S}$ and $\mathcal{NS}$ *BFGS*-type techniques to the restricted Broyden Class-type of quasi Newton methods (introduced here as a generalization of Broyden Class [16]); in this extension it emerges that basic conditions for the convergence are

$$\operatorname{tr}\tilde{B}_k \leq \operatorname{tr}B_k, \quad \det\tilde{B}_k \geq \det B_k, \quad f \in C^2, \text{ convex.}$$

In fact these conditions are sufficient to ensure the convergence of $\mathcal{NS}$ and, with a further condition (see (5.17)), allow to identify a class of convergent $\mathcal{S}$ methods which has nonempty intersection with $\mathcal{NS}$ class. On the other hand, considering a subset of this intersection characterized by the equality $\tilde{B}_k\mathbf{s}_k = B_k\mathbf{s}_k$ and showing that such equality can be imposed for $\tilde{B}_k = \mathcal{L}^{(k)}_{B_k}$, at a low cost (linear in n) and without any assumption on $\mathbf{s}_k$, we introduce a basic class of $\mathcal{L}^{(k)}QN$ methods (see Algorithm 4) which turn out to be a refinement of the class of $\mathcal{L}^{(k)}QN$ considered in [27], where instead the weaker condition $\sigma_k\mathcal{L}^{(k)}\mathbf{s}_k = B_k\mathbf{s}_k$ was considered and it was not yet clear if such condition could be always imposed (for any possible $\mathbf{s}_k$). Moreover, developing a further adaptive criterion (see (5.57)) we produce a low complexity convergent $\mathcal{L}^{(k)}QN$ with quadratic termination property (see Algorithm 6). We show that the used adaptive criteria can be satisfied by low complexity algebras $\mathcal{L}^{(k)}$ defined as the set $\operatorname{sd}U_k$ of all the matrices simultaneously diagonalized by $U_k =$ product of a constant number of Householder matrices, depending only on the number of vectors on which we want that the action of B_k is preserved by $\mathcal{L}^{(k)}_{B_k}$. This implies, in particular, that the memory required to implement Algorithm 6 (or other possible low complexity versions of Algorithm 4) can be considerably smaller that the memory required to implement the $L-BFGS$ method, which is a limited memory modification of BFGS – where only a limited number m of pairs $\mathbf{s}_j,\ \mathbf{y}_j,\ j = k, \dots, k-m+1$ is used to define B_{k+1} – suitable to solve large scale minimization problems [82]. Note moreover that, in contrast with $L-BFGS$, in any Algorithm 4 at each step it is stored, in an approximate way, the second order information generated in all the previous steps of the algorithm.

In detail an outline of the paper is as follows. In Section 5.2.1 we introduce the Broyden Class-type methods based on the parametric Hessian approximation updating formula $B_{k+1} = \Phi(\tilde{B}_k, \mathbf{s}_k, \mathbf{y}_k, \phi)$ with Φ as in [16], see (5.3) (observe that the $BFGS$-type methods are recovered choosing $\phi = 0$). In Section 5.3 we study conditions on the matrices $\tilde{B}_k$ that guarantee the convergence of the restricted Broyden Class-type methods ($\phi \in [0, 1)$) in the $\mathcal{S}$ecant and $\mathcal{N}$on $\mathcal{S}$ecant case (see Algorithm 2). Then, we focus our attention on BFGS-type schemes since a more transparent analysis is possible in this case. In particular, in Section 5.4 we show that convergence conditions on $\tilde{B}_k$ help $BFGS$-type to mimic the $BFGS$ self correction properties. In Section 5.5 we show that such convergence conditions can be imposed for $\tilde{B}_k = \mathcal{L}^{(k)}_{B_k}$ using low complexity algebras $\mathcal{L}^{(k)}$. This result is a consequence of a remarkable connection between the projection operation, Krylov spaces and Householder matrices (see also [28]). In Section 5.6 we analyze conditions on $\tilde{B}_k$ which guarantee the quadratic finite termination property for $BFGS$-type schemes and we show that they can be satisfied for $\tilde{B}_k = \mathcal{L}^{(k)}_{B_k}$. Finally in Section 5.7, using results from previous sections, we introduce a convergent $\mathcal{S}$ecant $\mathcal{L}^{(k)}QN$ method – of linear complexity per step –, that coincides with $BFGS$ on quadratic problems if exact line search is used, i.e., it converges in a finite number of steps. In the same section, using performance profiles [46] based on iterations and function evaluations, are provided the results of numerical experiences on a large set of problems from CUTEst [56]. These experiences indicate that the proposed Algorithm 6 can outperform the previous $\mathcal{L}QN$ and $\mathcal{L}^{(k)}QN$ algorithms studied in literature, but that, in general, with respect to the probability of win, the $L - BFGS$ method can perform better. Nevertheless, we believe that the flexibility of the adaptive $\mathcal{L}^{(k)}QN$ methods in conjunction with the encouraging results obtained in this chapter and further studies, can certainly make $\mathcal{L}^{(k)}QN$ methods a competitive alternative to $L - BFGS$.

5.2 Preliminaries

Let us notice, that in this chapter we will consider just real vectors and matrices, so we will exchange the word ‘unitary’ with the word ‘orthogonal’ and the superscript ‘H’ (Hermitian) with the superscript ‘T’ (transpose).

5.2.1 Broyden Class-type methods

Let us consider a function $f : \mathbb{R}^n \to \mathbb{R}$ where $n \geq 2$.

In this paper we will study the following class of minimization methods obtained generalizing the Broyden Class methods considered in [16]:

```
Data: $\mathbf{x}_0 \in \mathbb{R}^n$, $\mathbf{g}_0 = \nabla f(\mathbf{x}_0)$, $B_0 = \tilde{B}_0$ spd, $\mathbf{d}_0 = -B_0^{-1}\mathbf{g}_0$, k=0;
while $\mathbf{g}_k \neq 0$ do
    $\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{d}_k$ ;        /* $\lambda_k$ verifies conditions (5.4), (5.5) */
    $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$;
    $\mathbf{g}_{k+1} = \nabla f(\mathbf{x}_{k+1})$;
    $\mathbf{y}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$;
    $B_{k+1} = \Phi(\tilde{B}_k, \mathbf{s}_k, \mathbf{y}_k, \phi)$ ;
    $\left\{ \begin{array}{l} \text{Define } \tilde{B}_{k+1} \text{ spd, set } \mathbf{d}_{k+1} = -\tilde{B}_{k+1}^{-1}\mathbf{g}_{k+1} \quad (\mathcal{NS}) ; \\ \text{Set } \mathbf{d}_{k+1} = -B_{k+1}^{-1}\mathbf{g}_{k+1}, \text{ define } \tilde{B}_{k+1} \text{ spd} \quad (\mathcal{S}) ; \end{array} \right.$
    Set $k := k + 1$ ;
end
```

Algorithm 2: Broyden Class-type

where $\tilde{B}_k$ is an approximation of B_k and the updating formula is a generalization of the Broyden's one, i.e.

$$\Phi(\tilde{B}_k, \mathbf{s}_k, \mathbf{y}_k, \phi) = \tilde{B}_k - \frac{\tilde{B}_k \mathbf{s}_k \mathbf{s}_k^T \tilde{B}_k}{\mathbf{s}_k^T \tilde{B}_k \mathbf{s}_k} + \frac{\mathbf{y}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k} + \phi\, \mathbf{s}_k^T \tilde{B}_k \mathbf{s}_k \mathbf{v}_k \mathbf{v}_k^T. \tag{5.3}$$

In (5.3) the vector $\mathbf{v}_k$ is defined by

$$\mathbf{v}_k = \frac{\mathbf{y}_k}{\mathbf{y}_k^T \mathbf{s}_k} - \frac{\tilde{B}_k \mathbf{s}_k}{\mathbf{s}_k^T \tilde{B}_k \mathbf{s}_k}$$

and ϕ is a non negative parameter so that $\Phi(\tilde{B}_k, \mathbf{s}_k, \mathbf{y}_k, \phi)$ is spd whenever $\tilde{B}_k$ is spd and $\mathbf{y}_k^T \mathbf{s}_k > 0$. We assume that the step-length parameter λ_k is chosen by an inexact line search satisfying the two Wolfe conditions

$$f(\mathbf{x}_k + \lambda_k \mathbf{d}_k) \leq f(\mathbf{x}_k) + \alpha \lambda_k \mathbf{g}_k^T \mathbf{d}_k \tag{5.4}$$

$$g(\mathbf{x}_k + \lambda_k \mathbf{d}_k)^T \mathbf{d}_k \geq \beta \mathbf{g}_k^T \mathbf{d}_k \tag{5.5}$$

where $0 < \alpha < 1/2$ and $\alpha < \beta < 1$. Condition (5.5) implies $\mathbf{y}_k^T \mathbf{s}_k > 0$.

Let us observe that in Algorithm 2 the matrices generating the descent search directions $\mathbf{d}_{k+1}$ satisfy the Secant Equation in the $\mathcal{S}$ case. Instead in the $\mathcal{NS}$ case such property is not necessarily fulfilled, i.e.

$$B_{k+1}\mathbf{s}_k = \mathbf{y}_k$$

whereas, in general

$$\tilde{B}_{k+1}\mathbf{s}_k \neq \mathbf{y}_k.$$

In the following three remarks we collect some useful properties we will use in Section 5.3.

Remark 11. *Observe that*

$$\begin{aligned} tr(B_{k+1}) = tr(\Phi(\tilde{B}_k, \mathbf{s}_k, \mathbf{y}_k, \phi)) = tr(\tilde{B}_k) + \frac{\|\mathbf{y}_k\|^2}{\mathbf{y}_k^T\mathbf{s}_k} + \phi\frac{\|\mathbf{y}_k\|^2}{\mathbf{y}_k^T\mathbf{s}_k}\frac{\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k}{\mathbf{y}_k^T\mathbf{s}_k} \\ -(1-\phi)\frac{\|\tilde{B}_k\mathbf{s}_k\|^2}{\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k} - 2\phi\frac{\mathbf{y}_k^T\tilde{B}_k\mathbf{s}_k}{\mathbf{y}_k^T\mathbf{s}_k}. \end{aligned} \tag{5.6}$$

Since $\phi\,\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k \geq 0$, *the last term in* (5.3) *increases the eigenvalues, and hence*

$$\det(B_{k+1}) \geq \det\left(\tilde{B}_k - \frac{\tilde{B}_k\mathbf{s}_k\mathbf{s}_k^T\tilde{B}_k}{\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k} + \frac{\mathbf{y}_k\mathbf{y}_k^T}{\mathbf{y}_k^T\mathbf{s}_k}\right) = \det(\tilde{B}_k)\frac{\mathbf{y}_k^T\mathbf{s}_k}{\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k}$$

(for the last equality see [82]).

Remark 12. *From* (5.5) *it follows that, using definitions in Algorithm 2,*

$$\mathbf{y}_k^T\mathbf{s}_k = \mathbf{g}_{k+1}^T\mathbf{s}_k - \mathbf{g}_k^T\mathbf{s}_k \geq -(1-\beta)\mathbf{g}_k^T\mathbf{s}_k \tag{5.7}$$

from which we obtain

$$\frac{\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k}{\mathbf{y}_k^T\mathbf{s}_k} \leq \frac{\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k}{(1-\beta)(-\mathbf{g}_k^T\mathbf{s}_k)} = \frac{\lambda_k}{1-\beta} \tag{5.8}$$

($\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k = \mathbf{s}_k^T(-\lambda_k\mathbf{g}_k)$ *in the* $\mathcal{NS}$ *case) and*

$$\frac{\mathbf{s}_k^TB_k\mathbf{s}_k}{\mathbf{y}_k^T\mathbf{s}_k} \leq \frac{\mathbf{s}_k^TB_k\mathbf{s}_k}{(1-\beta)(-\mathbf{g}_k^T\mathbf{s}_k)} = \frac{\lambda_k}{1-\beta} \tag{5.9}$$

($\mathbf{s}_k^TB_k\mathbf{s}_k = \mathbf{s}_k^T(-\lambda_k\mathbf{g}_k)$ *in the* $\mathcal{S}$ *case).*

Remark 13. *Let us define* f_* *to be the infimum of* f. *Using* (5.4) *we have (in both* $\mathcal{NS}$ *and* $\mathcal{S}$ *methods)*

$$\begin{aligned} \sum_{k=0}^{N}\mathbf{s}_k^T(-\mathbf{g}_k) &= \sum_{k=0}^{N} -\lambda_k\mathbf{d}_k^T\mathbf{g}_k \\ &\leq \frac{1}{\alpha}\sum_{k=0}^{N}[f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})] \\ &\leq \frac{1}{\alpha}[f(\mathbf{x}_0) - f_*] < \infty. \end{aligned} \tag{5.10}$$

Then the sum converges for $n \to +\infty$, *from which we obtain*

$$\lim_{k\to+\infty}\mathbf{s}_k^T(-\mathbf{g}_k) = 0.$$

Notice that for $\phi \in [0,1]$ we call the Broyden Class-type family “restricted”. If $\tilde{B}_k = B_k$ for all k, then for $\phi = 0$ and $\phi = 1$ one obtains, respectively, the *BFGS* and the DFP method [82].

5.2.2 Assumptions for the function f

In Section 5.3, in order to obtain a convergence result for the Broyden Class-type, we will do the following:

Assumption 1. *The level set*

$$D = \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$$

is convex, the function $f(\mathbf{x})$ is twice continuously differentiable, convex and bounded below in D and the Hessian matrix is bounded in D, i.e.

$$\|G(\mathbf{x})\| \leq M. \tag{5.11}$$

Remark 14. *Observe that if Assumption 1 if fulfilled then the following inequality* (5.12) *holds:*

$$\frac{\|\mathbf{y}_k\|^2}{\mathbf{s}_k^T \mathbf{y}_k} \leq M, \tag{5.12}$$

where $\mathbf{s}_k$ and $\mathbf{y}_k$ are the difference vectors produced by Algorithm 2. In fact, if we define (see [16], [82]) the spd matrix

$$\overline{G} = \int_0^1 G(\mathbf{x}_k + \tau \mathbf{s}_k) d\tau,$$

then we have from standard analysis results,

$$\mathbf{y}_k = \overline{G}\mathbf{s}_k \tag{5.13}$$

and hence if $\mathbf{z}_k = \overline{G}^{\frac{1}{2}}\mathbf{s}_k$,

$$\frac{\|\mathbf{y}_k\|^2}{\mathbf{s}_k^T \mathbf{y}_k} = \frac{\mathbf{s}_k^T \overline{G}^2 \mathbf{s}_k}{\mathbf{s}_k^T \overline{G} \mathbf{s}_k} = \frac{\mathbf{z}_k^T \overline{G} \mathbf{z}_k}{\mathbf{z}_k^T \mathbf{z}_k} \leq \sup_{\tau \in [0,1]} \|G(\mathbf{x}_k + \tau \mathbf{s}_k)\| \leq M. \tag{5.14}$$

We recall that condition (5.12) *on the Powell's ratio is typically used to prove the global convergence of $BFGS$ method [87] and $\mathcal{L}QN$ methods [40].*

5.3 Conditions for the convergence of the $\mathcal{S}$ecant and $\mathcal{N}$on $\mathcal{S}$ecant Broyden Class-type

The matrices which generate the descent directions in the $\mathcal{S}$ case exhibit explicitly second order information (or, in other words, they satisfy the secant equation). Moreover, in contrast with the limited memory versions of Quasi-Newton methods, they store, in an approximate way, the second order information generated in all the previous steps of the algorithm. In this section we will prove that both $\mathcal{S}$ and $\mathcal{NS}$ versions of Algorithm 1 are convergent if $\tilde{B}_k$ is suitably chosen.
The following result generalizes what proven in [27] for $BFGS$-type $\mathcal{S}$ methods using techniques and ideas from [16, 15] .

Theorem 5.3.1. *If the $\mathcal{S}$ version of Algorithm 2 with $\phi \in [0,1)$ is applied to a function that satisfies Assumption 1 and $\tilde{B}_k$ is chosen such that*

$$tr\,\tilde{B}_k \leq \; tr\,B_k \tag{5.15}$$

$$\det \tilde{B}_k \geq \det B_k \tag{5.16}$$

$$\frac{||B_k\mathbf{s}_k||^2}{(\mathbf{s}_k^T B_k \mathbf{s}_k)^2} \leq \frac{||\tilde{B}_k\mathbf{s}_k||^2}{(\mathbf{s}_k^T \tilde{B}_k \mathbf{s}_k)^2}\,. \tag{5.17}$$

for all k, then

$$\liminf_{k\to\infty} \|\mathbf{g}_k\| = 0 \tag{5.18}$$

for any starting point $\mathbf{x}_0$ and any positive definite matrix B_0.

The main idea to prove Theorem 5.3.1 is to compare the third and fifth term of (5.6). Let us define ψ_k

$$\frac{\|\mathbf{y}_k\|^2}{\mathbf{y}_k^T\mathbf{s}_k}\frac{\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k}{\mathbf{y}_k^T\mathbf{s}_k} - 2\frac{\mathbf{y}_k^T\tilde{B}_k\mathbf{s}_k}{\mathbf{y}_k^T\mathbf{s}_k} = \psi_k\frac{\|\tilde{B}_k\mathbf{s}_k\|^2}{\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k} \tag{5.19}$$

so that (5.6) becomes

$$\mathrm{tr}\,(B_{k+1}) = \; \mathrm{tr}\,(\tilde{B}_k) + \frac{\|\mathbf{y}_k\|^2}{\mathbf{y}_k^T\mathbf{s}_k} - (1-\phi-\psi_k\phi)\frac{\|\tilde{B}_k\mathbf{s}_k\|^2}{\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k}. \tag{5.20}$$

Remark 15. *Let us estimate the first term in* (5.19)*. We have*

$$\begin{aligned}\frac{\|\mathbf{y}_k\|^2}{\mathbf{y}_k^T\mathbf{s}_k}\frac{\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k}{\mathbf{y}_k^T\mathbf{s}_k}\frac{\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k}{\|\tilde{B}_k\mathbf{s}_k\|^2} &\leq M\frac{(\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k)^2}{\mathbf{y}_k^T\mathbf{s}_k\|\tilde{B}_k\mathbf{s}_k\|^2} \\ \leq M\frac{(\mathbf{s}_k^T B_k\mathbf{s}_k)^2}{\mathbf{y}_k^T\mathbf{s}_k\|B_k\mathbf{s}_k\|^2} &= \frac{M(\mathbf{s}_k^T(-\mathbf{g}_k))^2}{\mathbf{y}_k^T\mathbf{s}_k\|-\mathbf{g}_k\|^2} \\ &\leq \frac{M(\mathbf{s}_k^T(-\mathbf{g}_k))}{(1-\beta)\|-\mathbf{g}_k\|^2},\end{aligned} \tag{5.21}$$

where first inequality follows using (5.12)*, the second using* (5.17) *and last inequality follows using* (5.7)*.*

Remark 16. *Let us estimate the second term in* (5.19)*. We have*

$$\begin{aligned}\frac{|\mathbf{y}_k^T\tilde{B}_k\mathbf{s}_k|}{\mathbf{y}_k^T\mathbf{s}_k}\frac{\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k}{\|\tilde{B}_k\mathbf{s}_k\|^2} &\leq \frac{\|\mathbf{y}_k\|\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k}{\mathbf{y}_k^T\mathbf{s}_k\|\tilde{B}_k\mathbf{s}_k\|} \\ &\leq \frac{\sqrt{M}\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k}{\sqrt{\mathbf{y}_k^T\mathbf{s}_k}\|\tilde{B}_k\mathbf{s}_k\|} \\ \leq \frac{\sqrt{M}\mathbf{s}_k^T B_k\mathbf{s}_k}{\sqrt{\mathbf{y}_k^T\mathbf{s}_k}\|B_k\mathbf{s}_k\|} &= \frac{\sqrt{M}(\mathbf{s}_k^T(-\mathbf{g}_k))}{\sqrt{\mathbf{y}_k^T\mathbf{s}_k}\|-\mathbf{g}_k\|} \\ &\leq \frac{\sqrt{M(\mathbf{s}_k^T(-\mathbf{g}_k))}}{\sqrt{1-\beta}\|-\mathbf{g}_k\|},\end{aligned} \tag{5.22}$$

where the first inequality follows from Cauchy-Schwarz inequality, the second from (5.17)*, the third from* (5.12) *and the fourth from* (5.7)*.*

We can now prove Theorem 5.3.1.

Proof. Arguing by contradiction, let us assume $\|\mathbf{g}_k\|$ bounded away from zero, i.e., there exists $\gamma > 0$ such that

$$\|\mathbf{g}_k\| \geq \gamma > 0. \tag{5.23}$$

From Remark 13 we obtain

$$\lim_{k\to\infty} \frac{\mathbf{s}_k^T(-\mathbf{g}_k)}{\|-\mathbf{g}_k\|^2} = 0. \tag{5.24}$$

Now we show that (5.24) leads to a contradiction, thus (5.23) cannot hold. From (5.19), using Remark 15, Remark 16 and (5.24) we obtain

$$\lim_{k\to\infty} \psi_k = 0. \tag{5.25}$$

Using (5.25), since $\phi \in [0, 1)$, we have that there exist an index s and constants $l_1 > 0$, $l_2 > 0$ such that

$$l_2 \geq (1 - \phi - \psi_k\phi) \geq l_1 > 0 \text{ for all } k \geq s. \tag{5.26}$$

Then we can write (for $j \geq s$), using (5.20),

$$\mathrm{tr}B_{j+1} \leq \mathrm{tr}B_s + \sum_{k=s}^{j} \frac{1}{\mathbf{y}_k^T\mathbf{s}_k}\|\mathbf{y}_k\|^2 - \sum_{k=s}^{j} \frac{1}{\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k}\|\tilde{B}_k\mathbf{s}_k\|^2(1-\phi-\psi_k\phi), \tag{5.27}$$

and hence

$$\mathrm{tr}B_{j+1} \leq \mathrm{tr}B_s + M(j+1-s) \leq c_1(j+2-s)$$

where $c_1 = \max\{\,\mathrm{tr}\,B_s, M\}$ (the trace grows at most linearly for all $j \geq s$).

Let us remember that, given n real positive numbers a_i, it holds:

$$\prod_{i=1}^{n} a_i \leq \left(\frac{\sum_{i=1}^{n} a_i}{n}\right)^n \tag{5.28}$$

from which we obtain:

$$\det B_{j+1} = \prod_{i=1}^{n} \nu_i(B_{j+1}) \leq \left(\frac{\sum_{i=1}^{n} \nu_i(B_{j+1})}{n}\right)^n \leq \left(\frac{c_1(j+2-s)}{n}\right)^n. \tag{5.29}$$

Let us note, moreover, that from (5.27), since B_{j+1} is positive definite, we have:

$$\sum_{k=s}^{j} \frac{1}{\mathbf{s}_k^T \tilde{B}_k \mathbf{s}_k} \|\tilde{B}_k \mathbf{s}_k\|^2 (1-\phi-\psi_k\phi) \leq \mathrm{tr}B_s - \mathrm{tr}B_{j+1} + \sum_{k=s}^{j} \frac{1}{\mathbf{y}_k^T \mathbf{s}_k} \|\mathbf{y}_k\|^2$$
$$\leq \mathrm{tr}B_s + \sum_{k=s}^{j} \frac{1}{\mathbf{y}_k^T \mathbf{s}_k} \|\mathbf{y}_k\|^2 \leq c_1(j+2-s) \tag{5.30}$$

and applying once more (5.28) we have:

$$\prod_{k=s}^{j} \frac{1}{\mathbf{s}_k^T \tilde{B}_k \mathbf{s}_k} \|\tilde{B}_k \mathbf{s}_k\|^2 (1-\phi-\psi_k\phi) \leq (2c_1)^{j+1-s}. \tag{5.31}$$

From (5.16) and from direct calculation of the determinant we have:

$$\det B_{j+1} \geq \frac{\mathbf{s}_j^T \mathbf{y}_j}{\mathbf{s}_j^T \tilde{B}_j \mathbf{s}_j} \det \tilde{B}_j \geq \frac{\mathbf{s}_j^T \mathbf{y}_j}{\mathbf{s}_j^T \tilde{B}_j \mathbf{s}_j} \det B_j,$$

from which we obtain:

$$\prod_{k=s}^{j} \frac{\mathbf{s}_k^T \mathbf{y}_k}{\mathbf{s}_k^T \tilde{B}_k \mathbf{s}_k} \leq \frac{\det B_{j+1}}{\det B_s} \ . \tag{5.32}$$

From (5.7) we have

$$(1-\beta)^{j+1-s} \leq \prod_{k=s}^{j} \frac{\mathbf{s}_k^T \mathbf{y}_k}{-\mathbf{g}_k^T \mathbf{s}_k},$$

and hence

$$\begin{aligned}
&(1-\beta)^{j+1-s} \prod_{k=s}^{j} \frac{\|\mathbf{g}_k\|^2}{\mathbf{s}_k^T(-\mathbf{g}_k)} (1-\phi-\psi_k\phi) \\
&\leq \prod_{k=s}^{j} (1-\phi-\psi_k\phi) \frac{\|-\lambda_k \mathbf{g}_k\|^2}{\mathbf{s}_k^T(-\lambda_k \mathbf{g}_k)} \frac{\mathbf{s}_k^T \mathbf{y}_k}{\mathbf{s}_k^T(-\lambda_k \mathbf{g}_k)} \\
&= \prod_{k=s}^{j} (1-\phi-\psi_k\phi) \frac{\|B_k \mathbf{s}_k\|^2}{\mathbf{s}_k^T B_k \mathbf{s}_k} \frac{\mathbf{s}_k^T \mathbf{y}_k}{\mathbf{s}_k^T \ B_k \mathbf{s}_k} \\
&\leq \prod_{k=s}^{j} (1-\phi-\psi_k\phi) \frac{\|\tilde{B}_k \mathbf{s}_k\|^2}{\mathbf{s}_k^T \tilde{B}_k \mathbf{s}_k} \frac{\mathbf{s}_k^T \mathbf{y}_k}{\mathbf{s}_k^T \tilde{B}_k \mathbf{s}_k} \\
&\leq (2c_1)^{j+1-s} \left(\frac{c_1(j+2-s)}{n} \right)^n \frac{1}{\det B_s},
\end{aligned} \tag{5.33}$$

i.e.

$$\prod_{k=s}^{j}(1-\phi-\psi_k\phi)\frac{\|\mathbf{g}_k\|^2}{\mathbf{s}_k^T(-\mathbf{g}_k)} \leq c_2^{j+1-s} \tag{5.34}$$

for a suitable constant c_2, which is a contradiction since we supposed

$$\lim_{k\to\infty}\frac{\mathbf{s}_k^T(-\mathbf{g}_k)}{\|-\mathbf{g}_k\|^2}=0.$$

We have hence proved that (5.18) holds. □

Observe that (5.17) is in particular satisfied when $\tilde{B}_k$ is such that

$$\tilde{B}_k\mathbf{s}_k = B_k\mathbf{s}_k. \tag{5.35}$$

In the next Sections 5.4 and 5.5 we will investigate some further consequences of condition (5.35) and we will prove that it can be imposed by choosing $\tilde{B}_k$ as the projection of B_k on algebras of matrices diagonalized by a fixed number of orthogonal Householder transforms.
The following result generalizes what proven in [40] for BFGS-type $\mathcal{NS}$ methods.

Theorem 5.3.2. *If the $\mathcal{NS}$ version of Algorithm 2 with $\phi \in [0,1)$ is applied to a function that satisfies Assumption 1 and $\tilde{B}_k$ is chosen such that*

$$tr\,\tilde{B}_k \leq\ tr\,B_k \tag{5.36}$$

$$\det\tilde{B}_k \geq \det B_k \tag{5.37}$$

for all k, then

$$\liminf_{k\to\infty}\|\mathbf{g}_k\|=0 \tag{5.38}$$

for any starting point $\mathbf{x}_0$ and any positive definite matrix B_0.

Proof. Proceed as in the proof of Theorem 5.3.1 noting that the hypothesis (5.17) on $\tilde{B}_k$ is no longer necessary to obtain Remark 15 (see (5.21)), Remark 16 (see (5.22)) and (5.33), since in $\mathcal{NS}$ methods $\tilde{B}_k\mathbf{s}_k$ turns out to be equal to $-\lambda_k\mathbf{g}_k$. □

In Figure 5.1 we illustrate in a pictorial way the restricted Broyden Class-type $\mathcal{S}$ecant and $\mathcal{N}$on $\mathcal{S}$ecant methods satisfying the conditions $\mathrm{tr}\,\tilde{B}_k \leq \mathrm{tr}\,B_k$, $\det\tilde{B}_k \geq \det B_k$ and $f \in C^2$, which appear basic in proving convergence results for both classes of methods. At the moment only a subset of the pictured $\mathcal{S}$ecant methods are certainly convergent, those satisfying the surplus condition (5.17). Let us observe that in [27] we investigated *BFGS*-type methods where $\sigma_k\tilde{B}_k\mathbf{s}_k = B_k\mathbf{s}_k$ for some $\sigma_k > 0$, and thus verifying condition (5.17). In the following we will focus on Broyden Class-type methods such that $\tilde{B}_k\mathbf{s}_k = B_k\mathbf{s}_k$, which form a subset of the intersection between convergent $\mathcal{S}$ and $\mathcal{NS}$, with the aim to define new efficient *BFGS*-type algorithms.

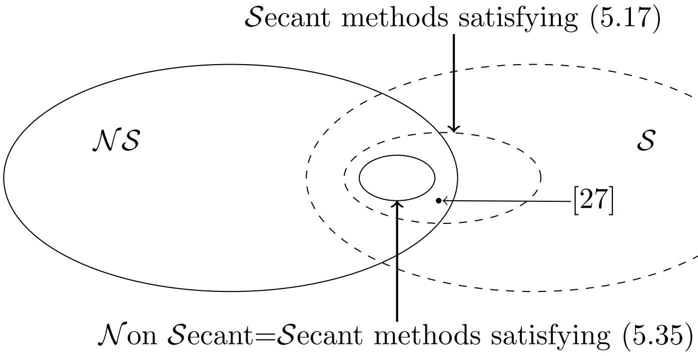

$\mathcal{S}$ecant methods satisfying (5.17)
$\mathcal{NS}$
$\mathcal{S}$
[27]
$\mathcal{N}$on $\mathcal{S}$ecant=$\mathcal{S}$ecant methods satisfying (5.35)

Figure 5.1: Restricted Broyden Class-type methods satisfying the conditions on trace, determinant.

5.4 Self correcting properties implied by convergence conditions

In this section, assuming $\phi = 0$ in Algorithm 2, we will study how (5.35) reverberate on self correcting properties of the algorithm.
There are experimental evidences (in the case the matrix $\tilde{B}_k$ is chosen in some fixed matrix algebra $\mathcal{L}$), that the $\mathcal{S}$ version of Algorithm 2 perform better if compared with the $\mathcal{NS}$ one (see [7] and [17]). In this section we will try to motivate theoretically this experimental observation by comparing $\operatorname{tr} B_{k+1}$ and $\det\{B_{k+1}\}$ produced by classic *BFGS* and Algorithm 2 when $\phi = 0$. Observe moreover, that in [27] some preliminary experimental experiences have shown that even if condition (5.35) is imposed in an approximate way (i.e $\tilde{B}_k\mathbf{s}_k \approx B_k\mathbf{s}_k$) performances of Algorithm 2 are competitive with those of $\mathcal{H}QN$, which, in turn, has been proved to be competitive with *L-BFGS* on some neural networks problem (see [40, 7]).

Finally let us stress the fact that, even if "the Quasi-Newton updating is inherently an overwriting process rather than an averaging process" (see [14]), the following analysis will show how algorithms proposed in this chapter exhibit an interaction between averaging and overwriting phases more similar to *BFGS* than to *L-BFGS* (remember that the curvature information constructed by *BFGS* are good enough to endow the algorithm with a superlinear rate of convergence, see [82]).

Performing one step of the "classic" *BFGS*, one has

$$\begin{aligned} B_{k+1}^{BFGS} &= \Phi^{BFGS}(B_k, \mathbf{s}_k, \mathbf{y}_k) \\ \operatorname{tr} B_{k+1}^{BFGS} &= \operatorname{tr} B_k - \frac{\|B_k\mathbf{s}_k\|^2}{\mathbf{s}_k^T B_k \mathbf{s}_k} + \frac{\|\mathbf{y}_k\|^2}{\mathbf{y}_k^T \mathbf{s}_k} \end{aligned} \tag{5.39}$$

$$\det\big(B_{k+1}^{BFGS}\big) = \det(B_k)\frac{\mathbf{y}_k^T\mathbf{s}_k}{\mathbf{s}_k^T B_k\mathbf{s}_k} = \det(B_k)\frac{\mathbf{s}_k^T(\overline{G}\mathbf{s}_k)}{\mathbf{s}_k^T B_k\mathbf{s}_k} \tag{5.40}$$

from which it is clear that $BFGS$ (and all updates in the restricted Broyden class) "have a strong self correcting property with respect to the determinant" (see [16] and Remark 11 for (5.40)). In particular curvatures of the model are inflated or deflated (and hence corrected) accordingly to the ratio

$$\frac{\mathbf{s}_k^T(\overline{G}\mathbf{s}_k)}{\mathbf{s}_k^T B_k\mathbf{s}_k},$$

allowing the algorithm to compare the computed model with the true Hessian. In fact, the above ratio is used to correct the spectrum of the operator defining the descent direction at next step.

On the contrary, by performing one step of Algorithm 2 we obtain

$$\begin{aligned} B_{k+1} &= \Phi(\tilde{B}_k, \mathbf{s}_k, \mathbf{y}_k) \\ \operatorname{tr} B_{k+1} &= \operatorname{tr}\tilde{B}_k - \frac{\|\tilde{B}_k\mathbf{s}_k\|^2}{\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k} + \frac{\|\mathbf{y}_k\|^2}{\mathbf{y}_k^T\mathbf{s}_k} \\ \det(B_{k+1}) &= \det(\tilde{B}_k)\frac{\mathbf{y}_k^T\mathbf{s}_k}{\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k} = \det(\tilde{B}_k)\frac{\mathbf{s}_k^T(\overline{G}\mathbf{s}_k)}{\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k} \end{aligned} \tag{5.41}$$

from which it is clear that if $\tilde{B}_k\mathbf{s}_k$ is not suitably chosen, then the ratio

$$\frac{\mathbf{s}_k^T(\overline{G}\mathbf{s}_k)}{\mathbf{s}_k^T\tilde{B}_k\mathbf{s}_k},$$

could not exhibit a reasonable behavior, making the algorithm not able to self-correct bad estimated curvatures and hence loosing efficiency. In the hypothesis (5.35), we have

$$\begin{aligned} B_{k+1} &= \Phi(\tilde{B}_k, \mathbf{s}_k, \mathbf{y}_k) \\ \operatorname{tr} B_{k+1} &= \operatorname{tr}\tilde{B}_k - \frac{\|B_k\mathbf{s}_k\|^2}{\mathbf{s}_k^T B_k\mathbf{s}_k} + \frac{\|\mathbf{y}_k\|^2}{\mathbf{y}_k^T\mathbf{s}_k} \\ \det(B_{k+1}) &= \det(\tilde{B}_k)\frac{\mathbf{y}_k^T\mathbf{s}_k}{\mathbf{s}_k^T B_k\mathbf{s}_k} = \det(\tilde{B}_k)\frac{\mathbf{s}_k^T(\overline{G}\mathbf{s}_k)}{\mathbf{s}_k^T B_k\mathbf{s}_k} \end{aligned} \tag{5.42}$$

which is then a reasonable choice even under the self-correcting properties point of view. Observe that if we choose $\tilde{B}_k = \mathcal{L}_{B_k}$, the error we introduce contributes to inappropriately inflate the curvatures of the model because by Theorem 1.4.9, even if $\operatorname{tr}\tilde{B}_k = \operatorname{tr} B_k$, we have $\det(\tilde{B}_k) \geq \det(B_k)$ (see [74] and references therein for more information regarding the inappropriate inflations problems affecting $BFGS$). Recall that by the same Theorem 1.4.9, $\det(\tilde{B}_k) = \det(B_k)$ iff U diagonalizes B_k. Thus, in order to reduce the inappropriate inflation of the curvatures of the model, U should be chosen, in principle, besides of low complexity, as close as possible to a matrix which diagonalizes B_k. So, the problem concerning the possibility to exploit $\tilde{B}_k$ in order to improve such self correcting properties as much as possible remains open.

5.5 How to ensure $\mathcal{S}$ecant convergence conditions by low complexity matrices

In this section we will show that it is always possible to satisfy hypothesis of Theorem 5.3.1 by a low complexity matrix $\tilde{B}_k$. In particular, a matrix $\tilde{B}_k$ satisfying (5.15), (5.16) and (5.35) will be explicitly constructed.
As noticed in Theorem 1.4.9, spectral conditions

$$\mathrm{tr}\,(\tilde{B}_k) \leq \mathrm{tr}\,(B_k),$$

$$\det(\tilde{B}_k) \geq \det(B_k),$$

on the approximation are always fulfilled when we choose

$$\tilde{B}_k = \mathcal{L}_{B_k} \text{ for some } \mathcal{L} = \mathrm{sd}\,U.$$

Nevertheless, the condition

$$\mathcal{L}_{B_k}\mathbf{s}_k = B_k\mathbf{s}_k. \tag{5.43}$$

is not satisfied for a generic matrix algebra $\mathcal{L}$ and we have to face the following Problem 3, named Totally Non Linear Problem in [27]:
Problem 3. *Given a spd matrix $A \in \mathbb{R}^{n\times n}$ and a vector $\mathbf{s} \in \mathbb{R}^n$, find a low complexity orthogonal matrix U such that defining $\mathcal{L} = sdU$ it holds*

$$\mathcal{L}_A\mathbf{s} = A\mathbf{s}. \tag{5.44}$$

Observe that Problem 3 has been solved in [26] in the particular case when $\mathbf{s}$ is an eigenvector of A with the aim to speed-up the Pagerank computation by the preconditioned Euler-Richardson method. The following Lemma 5.5.1 completely characterizes solution of Problem 3 in this case.

Lemma 5.5.1. *If $\mathbf{s}$ is such that $A\mathbf{s} = \gamma\mathbf{s}$, for any unitary matrix L such that $\mathbf{s}/\|\mathbf{s}\|$ is among its columns, defining $\mathcal{L} = sdL$, it holds*

$$\mathcal{L}_A\mathbf{s} = A\mathbf{s}.$$

In particular L can be chosen as an orthogonal Householder matrix.

Proof. Consider an orthogonal L such that $L\mathbf{e}_k = \mathbf{s}/\|\mathbf{s}\|$ for some fixed $k \in \{1, \dots, n\}$. From 1. in Theorem 1.4.9 we have $\mathcal{L}_A = Ld(\mathbf{z}_A)L^T$ being $\mathbf{z}_A = [\dots, (L^TAL)_{ii}, \dots]^T$ and hence

$$\mathcal{L}_A\mathbf{s} = (\mathbf{z}_A)_k\mathbf{s} = \frac{\mathbf{s}^TA\mathbf{s}}{\|\mathbf{s}\|^2}\mathbf{s} = \gamma\mathbf{s} = A\mathbf{s}. \tag{5.45}$$

For the second part see Lemma 2.2.5 in Chapter 2. □

By the following Theorem 5.5.3 we solve Problem 3 in the general case and at the same time we shed light on algorithmic details necessary for the construction of the solution. We note that in Chapter 2 ([28]) it has been proved a more general result where the projection $\mathcal{L}_A$ retains the action of A on a set of vectors instead on a single one. Nevertheless, we repeat here the proof in this particular case since it will be useful later in connection to the optimization algorithms we will introduce and develop.

Let us begin recalling the well-known Arnoldi algorithm for finding an orthogonal basis of the Krylov subspace

$$\mathcal{K}_m(A, \mathbf{v}) :=< \mathbf{v}, A\mathbf{v}, \dots, A^{m-1}\mathbf{v} > .$$

In what follows we will assume $\dim \mathcal{K}_m(A, \mathbf{v}) = m$.

```
Data: $A$, $\mathbf{v}_1 := \mathbf{v}/\|\mathbf{v}\|_2$;
while $j \leq m$ do
    Compute $\mathbf{w} := A\mathbf{v}_j$ ;
    while $i \leq j$ do
        Compute $h_{i,j} = (\mathbf{w}, \mathbf{v}_i)$ ;
        Compute $\mathbf{w} := \mathbf{w} - h_{i,j}\mathbf{v}_i$ ;
    end
    Compute $h_{j+1,j} := \|\mathbf{w}\|_2$ and $\mathbf{v}_{j+1} := \mathbf{w}/h_{j+1,j}$ ;
end
```

Algorithm 3: Arnoldi Algorithm

The above algorithm produces an orthonormal basis $V_m = [\mathbf{v}_1, \dots, \mathbf{v}_m]$ of the Krylov subspace $K_m(A, \mathbf{v})$ such that

$$AV_m = V_mH_m + h_{m+1,m}\mathbf{v}_{m+1}\mathbf{e}_m^T,$$

where the matrix H_m denotes the $m \times m$ upper Hessenberg matrix consisting of the coefficients $h_{i,j}$ computed by the algorithm. From the above observations we get

$$V_m^TAV_m = H_m. \tag{5.46}$$

Moreover, the following lemma holds:

Lemma 5.5.2 ([90])**.** *Let A be any matrix and V_m, H_m the results of m steps of the Arnoldi or Lanczos method applied to A. Then for any polynomial p_j of degree $j \leq m-1$ the following equality holds:*

$$p_j(A)\mathbf{v}_1 = V_m p_j(H_m)\mathbf{e}_1. \tag{5.47}$$

Theorem 5.5.3. *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. For every fixed integer m and $1 \leq m \leq n$ and for any $\mathbf{s} \in \mathbb{R}^n$ there exists an orthogonal matrix $L \in \mathbb{R}^{n\times n}$ such that if $\mathcal{L} = sd\,L$ and $\mathcal{L}_A$ is the best approximation of A in $\mathcal{L}$, then*

$$p_j(\mathcal{L}_A)\mathbf{s} = p_j(A)\mathbf{s} \tag{5.48}$$

for any polynomial p_j of degree $j \leq m-1$. Moreover, the thesis is satisfied also by any other orthogonal matrix having, among its columns, m particular columns of L (see (5.51)*).*

Proof. Consider the matrices V_m and H_m constructed from Arnoldi Algorithm applied to $\mathcal{K}_m(A,\mathbf{s})$ (observe that the first column of V_m is $\mathbf{v}_1 := \mathbf{s}/\|\mathbf{s}\|$). From Lemma 5.5.2 with $j=1$ we have

$$A\mathbf{v}_1 = V_m H_m V_m^T \mathbf{v}_1.$$

From (5.46) we can write

$$A\mathbf{v}_1 = V_m Q Q^T V_m^T A V_m Q Q^T V_m^T \mathbf{v}_1 \tag{5.49}$$

for any orthogonal matrix Q. In particular, being $V_m^T A V_m$ symmetric, we can choose in (5.49) Q as the orthogonal matrix which diagonalizes $V_m^T A V_m$, i.e.

$$A\mathbf{v}_1 = V_m Q \begin{bmatrix} x_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \dots & 0 & x_m \end{bmatrix} Q^T V_m^T \mathbf{v}_1, \tag{5.50}$$

where $x_i = \mathbf{e}_i^T Q^T V_m^T A V_m Q \mathbf{e}_i$ for $i = 1,\dots,m$. Consider now the matrix

$$L = [V_m Q\mathbf{e}_1|\dots|V_m Q\mathbf{e}_m|\mathbf{q}_{m+1}|\dots|\mathbf{q}_n] \tag{5.51}$$

where $\{\mathbf{q}_{m+1},\dots,\mathbf{q}_n\}$ is an orthonormal basis for

$$< V_m Q\mathbf{e}_1,\dots,V_m Q\mathbf{e}_m >^{\perp} = < V_m\mathbf{e}_1,\dots,V_m\mathbf{e}_m >^{\perp}, \tag{5.52}$$

set $\mathcal{L} = \mathrm{sd}\,L$ and consider $\mathcal{L}_A$ the best approximation of A in $\mathcal{L}$.
In order to prove that $\mathcal{L}_A$ satisfies (5.48) it is sufficient to prove that

$$\mathcal{L}_A^j \mathbf{v}_1 = A^j \mathbf{v}_1 \text{ for } 0 \leq j \leq m-1. \tag{5.53}$$

Of course, (5.53) is true for $j = 0$. The equality $\mathcal{L}_A\mathbf{v}_1 = A\mathbf{v}_1$ follows observing that using the first formula in Theorem 1.4.9 we have

$$\begin{aligned}\mathcal{L}_A\mathbf{v}_1 &= (\sum_i^n (L^TAL)_{ii}L\mathbf{e}_i(L\mathbf{e}_i)^T)\mathbf{v}_1 \\ &= (\sum_i^m x_i(V_mQ\mathbf{e}_i)(V_mQ\mathbf{e}_i)^T)\mathbf{v}_1 = A\mathbf{v}_1\end{aligned} \tag{5.54}$$

where in the second equality we take into account that $\mathbf{q}_i^T\mathbf{v}_1 = \mathbf{0}^T$ for $i \in \{m+1,\dots,n\}$ (see (5.52)) and (5.51).

Suppose now (5.53) true for all indexes $j \in [1,k]$, $k \leq m-2$, and let us prove it for $j = k+1$. From inductive hypothesis and Lemma 5.5.2 we have

$$\mathcal{L}_A^{k+1}\mathbf{v}_1 = \mathcal{L}_A\mathcal{L}_A^k\mathbf{v}_1 = \mathcal{L}_AA^k\mathbf{v}_1 = \mathcal{L}_AV_mH_m^k\mathbf{e}_1.$$

From direct computation using (5.52) and the definition of Q, we have $\mathcal{L}_AV_m = V_mH_m$ and thus

$$\mathcal{L}_AV_mH_m^k\mathbf{e}_1 = V_mH_m^{k+1}\mathbf{e}_1 = A^{k+1}\mathbf{v}_1,$$

where the last equality follows using again Lemma 5.5.2. Hence (5.53) holds also for $j \in [1,k+1]$.

□

Corollary 5.5.4. *Solutions U of Problem 3 are obtained by using Theorem 5.5.3 for $m = 2$, $j = 1$. Observe that just two of the columns of such orthogonal matrices U are uniquely determined (they are suitable linear combinations of the vectors* $\mathbf{s}$ *and* $A\mathbf{s}$*), and hence one of such U can be chosen as the product of two Householder matrices that can be determined by performing two matrix-vector products involving A plus $O(n)$ FLOPs.*

Proof. See the proof of Theorem 5.5.3 and Lemma 2.2.5 in Chapter 2. □

5.5.1 Convergent $\mathcal{L}^{(k)}QN$ scheme

In order to impose (5.43) for each k, an adaptive choice of the space $\mathcal{L} = \mathrm{sd}\,U$ is necessary. Any method obtained in this way will be called $\mathcal{L}^{(k)}QN$ extending the notation $\mathcal{L}QN$ introduced in [40] to denote the $BFGS$-type methods with $\tilde{B}_k = \mathcal{L}_{B_k}$ being $\mathcal{L}$ fixed. As a result of what discussed in Section 5.3 and in the first part of this section we report here the following Algorithm 4 which can be considered a refinement and an extension of the scheme proposed in [27]:

```
Data: $\mathbf{x}_0 \in \mathbb{R}^n, B_0 = I$ spd, toll, $\mathbf{d}_0 = -\mathbf{g}_0$, $k = 0$;
while $\mathbf{g}_k \neq 0$ do
    $\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{d}_k$ ;            /* $\lambda_k$ verifies (5.4), (5.5) */
    $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$;
    $\mathbf{y}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$;
    /* Definition of $\mathcal{L}^{(k)} :=$ sd $U_k$ s.t.  $\mathcal{L}^{(k)}_{B_k}\mathbf{s}_k = B_k\mathbf{s}_k$:   */
    if $\|B_k\mathbf{s}_k - \frac{\mathbf{s}_k B_k \mathbf{s}_k}{\|\mathbf{s}_k\|^2}\mathbf{s}_k\| <$ toll then
        apply Lemma 5.5.1;
    else
        apply Lemma 5.5.4;
    end
    $B_{k+1} = \Phi(\mathcal{L}^{(k)}_{B_k}, \mathbf{s}_k, \mathbf{y}_k, \phi)$ ;
    Compute $\mathbf{d}_{k+1} = -B_{k+1}^{-1}\mathbf{g}_{k+1}$;
    Set $k := k + 1$ ;
end
```

Algorithm 4: A convergent $\mathcal{L}^{(k)}$QN

5.5.2 Computational Complexity

The computational complexity of Algorithm 4 is $O(n)$ in space and time. This follows observing that

$$\mathcal{L}^{(k+1)}_{B_{k+1}} = \mathcal{L}^{(k+1)}_{\mathcal{L}^{(k)}_{B_k}} - \mathcal{L}^{(k+1)}_{\frac{\mathcal{L}^{(k)}_{B_k}\mathbf{s}_k\mathbf{s}_k^T\mathcal{L}^{(k)}_{B_k}}{\mathbf{s}_k^T\mathcal{L}^{(k)}_{B_k}\mathbf{s}_k}} + \mathcal{L}^{(k+1)}_{\frac{\mathbf{y}_k\mathbf{y}_k^T}{\mathbf{y}_k^T\mathbf{s}_k}} + (\phi\ \mathbf{s}_k^T\mathcal{L}^{(k)}_{B_k}\mathbf{s}_k)\mathcal{L}^{(k+1)}_{\mathbf{v}_k\mathbf{v}_k^T},$$

i.e.,

$$\begin{aligned}\lambda(\mathcal{L}^{(k+1)}_{B_{k+1}}) &= d(U_{k+1}^T B_{k+1} U_{k+1}) \\ &= d(U_{k+1}^T \mathcal{L}^{(k)}_{B_k} U_{k+1} - U_{k+1}^T \frac{\mathcal{L}^{(k)}_{B_k}\mathbf{s}_k\mathbf{s}_k^T\mathcal{L}^{(k)}_{B_k}}{\mathbf{s}_k^T\mathcal{L}^{(k)}_{B_k}\mathbf{s}_k} U_{k+1}) + \\ &+ d(U_{k+1}^T \frac{\mathbf{y}_k\mathbf{y}_k^T}{\mathbf{y}_k^T\mathbf{s}_k} U_{k+1} + (\phi\ \mathbf{s}_k^T\mathcal{L}^{(k)}_{B_k}\mathbf{s}_k) U_{k+1}^T\mathbf{v}_k\mathbf{v}_k^T U_{k+1}).\end{aligned} \tag{5.55}$$

Notice that the above equality is an extension of an eigenvalues updating formula obtained in [40] where $\mathcal{L}^{(k)} \equiv \mathcal{L}$ for all k. Since U_k and U_{k+1} can be chosen as the product of one or two Householder matrices (see Lemma 5.5.1, Lemma 5.5.4), the right hand side of the above expression can be computed in $O(n)$ FLOPs using Proposition 3.2 in [28]. In particular, using Remark 2 in [28], the worst case computational cost per step can be estimated in $16n + O(1)$.

5.6 The quadratic finite termination property

In literature it has been studied which Quasi-Newton methods terminate in a finite number of steps when applied to quadratic functions (quadratic finite termination), see [70, 80] and references therein. In this section, extending the analogous result obtained in [70] for *L-BFGS*, we will introduce conditions on $\tilde{B}_k$ (see (5.57)) which endow the $\mathcal{S}$ *BFGS*-type methods with the quadratic finite termination property. Before continuing let us observe that the notation of this section is not consistent with that of the previous Section 5.5. We prefer to adopt this notation here in order to be consistent with the existing literature.

Let us consider a spd matrix A and the problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \text{ where } f(\mathbf{x}) := \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b}. \tag{5.56}$$

In order to solve Problem (5.56) consider the following Algorithm 5 (which is the $\mathcal{S}$ version of Algorithm 2 where we use the exact line search and where we set $H_k = B_k^{-1}$, $\tilde{H}_k = \tilde{B}_k^{-1}$ and $\phi = 0$):

```
Data: $\mathbf{x}_0 \in \mathbb{R}^n$, $\mathbf{g}_0 = A\mathbf{x}_0 - \mathbf{b}$, $\tilde{H}_0 = H_0$ spd, $\mathbf{d}_0 = -H_0\mathbf{g}_0$, k=0;
while $\mathbf{g}_k \neq 0$ do
    $\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k\mathbf{d}_k$ ;            /* $\lambda_k := \arg\min_\lambda f(\mathbf{x}_k + \lambda\mathbf{d}_k)$ */
    $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$;
    $\mathbf{g}_{k+1} = A\mathbf{x}_{k+1} - \mathbf{b}$;
    $\mathbf{y}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$;
    $\rho_k = 1/\mathbf{s}_k^T\mathbf{y}_k$ ;
    Define $\tilde{H}_k$ spd ;
    $H_{k+1} = (I - \rho_k\mathbf{s}_k\mathbf{y}_k^T)\tilde{H}_k(I - \rho_k\mathbf{y}_k\mathbf{s}_k^T) + \rho_k\mathbf{s}_k\mathbf{s}_k^T$;
    Set $\mathbf{d}_{k+1} = -H_{k+1}\mathbf{g}_{k+1}$;
    Set $k := k+1$ ;
end
```

Algorithm 5: *BFGS*-type for quadratic problems

Theorem 5.6.1. *Let us consider Algorithm 5. If*

$$\tilde{H}_k\mathbf{g}_{k+1} = \beta_k H_0\mathbf{g}_{k+1} \text{ for some } \beta_k \neq 0, \tag{5.57}$$

then we have:

$$\mathbf{g}_{k+1}^T\mathbf{s}_j = 0 \text{ for all } j = 0,\dots,k; \tag{5.58}$$

$$\mathbf{s}_{k+1}^T A\mathbf{s}_j = 0 \text{ for all } j = 0,\dots,k; \tag{5.59}$$

$$span\{\mathbf{s}_0,\dots,\mathbf{s}_{k+1}\} = span\{H_0\mathbf{g}_0,\dots,H_0\mathbf{g}_{k+1}\}; \tag{5.60}$$

Proof. By induction. The case $k = 0$ can be easily verified. Let us suppose the thesis true for $k = 0,\dots,\hat{k}-1$ and prove it for $k = \hat{k}$. Let us prove (5.58): $\mathbf{g}_{\hat{k}+1}^T\mathbf{s}_{\hat{k}} = 0$ since we are using exact line search; if $j < \hat{k}$ we have

$$\mathbf{g}_{\hat{k}+1}^T\mathbf{s}_j = \mathbf{g}_{\hat{k}}^T\mathbf{s}_j + \mathbf{y}_{\hat{k}}^T\mathbf{s}_j = \mathbf{g}_{\hat{k}}^T\mathbf{s}_j + \mathbf{s}_{\hat{k}}^T A\mathbf{s}_j = 0 \tag{5.61}$$

by induction hypothesis. To prove (5.59) observe that for $j < \hat{k}$

$$\begin{aligned} &\mathbf{s}^T_{\hat{k}+1} A\mathbf{s}_j = -\lambda_{\hat{k}+1}\mathbf{g}^T_{\hat{k}+1} H_{\hat{k}+1}\mathbf{y}_j = \\ &-\lambda_{\hat{k}+1}\mathbf{g}^T_{\hat{k}+1}((I - \rho_{\hat{k}}\mathbf{s}_{\hat{k}}\mathbf{y}^T_{\hat{k}})\tilde{H}_{\hat{k}}(I - \rho_{\hat{k}}\mathbf{y}_{\hat{k}}\mathbf{s}^T_{\hat{k}}) + \rho_{\hat{k}}\mathbf{s}_{\hat{k}}\mathbf{s}^T_{\hat{k}})\mathbf{y}_j = \\ &-\lambda_{\hat{k}+1}\mathbf{g}^T_{\hat{k}+1}\tilde{H}_{\hat{k}}\mathbf{y}_j = -\beta_{\hat{k}}\lambda_{\hat{k}+1}\mathbf{g}^T_{\hat{k}+1} H_0\mathbf{y}_j = 0 \end{aligned} \tag{5.62}$$

where the third equality follows observing that $\mathbf{g}^T_{\hat{k}+1}\mathbf{s}_{\hat{k}} = 0$ and that $\mathbf{s}^T_{\hat{k}}\mathbf{y}_j = 0$ for $j < \hat{k}$ by induction hypothesis; the fourth equality follows by (5.57); the last equality follows observing that, since $\mathbf{g}^T_{\hat{k}+1}\mathbf{s}_{\hat{i}} = 0$ for all $j = 0, \dots, \hat{k}$ and $\text{span}\{\mathbf{s}_0, \dots, \mathbf{s}_{\hat{k}}\} = \text{span}\{H_0\mathbf{g}_0, \dots, H_0\mathbf{g}_{\hat{k}}\}$ by induction hypothesis, it holds that

$$\mathbf{g}^T_{\hat{k}+1} H_0\mathbf{g}_j = 0 \text{ for all } j = 0, \dots, \hat{k}. \tag{5.63}$$

Now let us consider the case $j = \hat{k}$. Since $\mathbf{s}_{\hat{k}+1} = -\lambda_{\hat{k}+1} H_{\hat{k}+1}\mathbf{g}_{\hat{k}+1}$, by direct computation using the expression of $H_{\hat{k}+1}$, it can be verified that $\mathbf{s}^T_{\hat{k}+1} A\mathbf{s}_{\hat{k}} = \mathbf{s}^T_{\hat{k}+1}\mathbf{y}_{\hat{k}} = 0$. Let us prove now (5.60): we have

$$\begin{aligned} &\mathbf{s}_{\hat{k}+1} = -\lambda_{\hat{k}+1} H_{\hat{k}+1}\mathbf{g}_{\hat{k}+1} = -\lambda_{\hat{k}+1}\tilde{H}_{\hat{k}}\mathbf{g}_{\hat{k}+1} + \lambda_{\hat{k}+1}\rho_{\hat{k}}\mathbf{y}^T_{\hat{k}}\tilde{H}_{\hat{k}}\mathbf{g}_{\hat{k}+1}\mathbf{s}_{\hat{k}} = \\ &-\beta_{\hat{k}}\lambda_{\hat{k}+1} H_0\mathbf{g}_{\hat{k}+1} + \lambda_{\hat{k}+1}\rho_{\hat{k}}\mathbf{y}^T_{\hat{k}}\tilde{H}_{\hat{k}}\mathbf{g}_{\hat{k}+1}\mathbf{s}_{\hat{k}} \end{aligned} \tag{5.64}$$

and hence

$$\text{span}\{H_0\mathbf{g}_0, \dots, H_0\mathbf{g}_{\hat{k}+1}\} = \text{span}\{\mathbf{s}_0, \dots, \mathbf{s}_{\hat{k}+1}\}$$

since $\text{span}\{H_0\mathbf{g}_0, \dots, H_0\mathbf{g}_{\hat{k}}\} = \text{span}\{\mathbf{s}_0, \dots, \mathbf{s}_{\hat{k}}\}$ and $\{\mathbf{s}_0, \dots, \mathbf{s}_{\hat{k}+1}\}$ are linearly independent since they are A-conjugate. □

Corollary 5.6.2. *If the spd matrix $\tilde{H}_k$ satisfy hypothesis of Theorem 5.6.1, then Algorithm 5 generates the same iterates as the Conjugate Gradient method preconditioned with H_0 and hence it converges in at most n steps.*

Proof. Analogous to the proof of Corollary 2.3 in [70], observing that under hypothesis of Theorem 5.6.1 conditions (5.58), (5.59) and (5.60) hold for Algorithm 5. □

Interestingly enough, using the above corollary it can be shown that the iterates of Algorithm 5 coincide with those from *BFGS* and *L-BFGS* since they all coincide with the Preconditioned Conjugate Gradient (see [80, 70]). We can now prove that the convergence condition (5.35) and the quadratic termination condition (5.57) can be verified simultaneously if $\tilde{H}_k = \mathcal{L}^{-1}_{B_k}$ provided that H_0 in (5.57) is a multiple of the identity.

Lemma 5.6.3. *For any pair of vectors* $\mathbf{s}_k$, $\mathbf{g}_{k+1}$, *and any spd matrix* B_k, *there exists a low complexity orthogonal matrix* $L^{(k)}$ *and hence a matrix algebra* $\mathcal{L}^{(k)} = sd\, L^{(k)}$ *such that*

$$\begin{aligned} \mathcal{L}^{(k)}_{B_k}\mathbf{s}_k &= B_k\mathbf{s}_k, \\ \mathcal{L}^{(k)}_{B_k}\mathbf{g}_{k+1} &= \alpha_k\mathbf{g}_{k+1} \textit{ for some } \alpha_k \neq 0. \end{aligned} \tag{5.65}$$

Thus, it is well defined a $\mathcal{L}^{(k)}QN$ *version of Algorithm 5, with* $\tilde{H}_k = \mathcal{L}^{-1}_{B_k}$, *convergent in at most* n *steps, provided that* H_0 *is a multiple of the identity.*

Proof. For the sake of simplicity we use in the following the symbols L and $\mathcal{L}$ in place of $L^{(k)}$ and $\mathcal{L}^{(k)}$. Case $B_k\mathbf{s}_k = \gamma\mathbf{s}_k$.
From Theorem 5.6.1 we have $\mathbf{g}^T_{k+1}\mathbf{s}_k = 0$. Any orthogonal matrix L which has among its columns $\mathbf{s}_k/\|\mathbf{s}_k\|$ and $\mathbf{g}_{k+1}/\|\mathbf{g}_{k+1}\|$ is such that, defining $\mathcal{L} = \mathrm{sd}\, L$, $\mathcal{L}_{B_k}$ satisfies conditions in (5.65) (the columns of L are eigenvectors of any matrix in $\mathcal{L}$). One of such orthogonal matrix L can be constructed as the product of two Householder matrices (see Lemma 2.2.5 and see [28] for more details).
Case $B_k\mathbf{s}_k \neq \gamma\mathbf{s}_k$.
Any matrix L in (5.51) satisfies $\mathcal{L}_{B_k}\mathbf{s}_k = B_k\mathbf{s}_k$ if $\mathcal{L} = \mathrm{sd}\, L$; it is then enough to consider a matrix L in (5.51) where $\mathbf{g}_{k+1}/\|\mathbf{g}_{k+1}\|$ is chosen to be one of the vectors $\mathbf{q}_i$ (see the proof of Theorem 5.5.3 with $m = 2$ and $\mathbf{s}_k$, B_k in the roles of $\mathbf{s}$ and A, respectively). Observe that this can be done since, from Theorem 5.6.1, $\mathbf{g}^T_{k+1}\mathbf{s}_k = 0 = \mathbf{g}^T_{k+1}\mathbf{g}_k$ (see (5.63)) and since the first two columns of L are suitable linear combinations of $\mathbf{s}_k$ and $B_k\mathbf{s}_k = -\lambda_k\mathbf{g}_k$. A matrix L with the required properties can be constructed as the product of three Householder matrices (see Lemma 2.2.5 and see [28] for more details).
□

5.7 A convergent $\mathcal{L}^{(k)}$QN method with quadratic termination property

The $\mathcal{L}^{(k)}QN$ scheme that we consider in this section, combines the results obtained in Section 5.3 for the $\mathcal{S}$ecant scheme with $\phi = 0$ and in Section 5.6 for quadratic termination, setting in both $\tilde{B}_k = \mathcal{L}^{(k)}_{B_k}$. In particular it combines the convergence result stated in Theorem 5.3.1 for general non linear problems with the quadratic termination result obtained in Theorem 5.6.1. The main motivation for this choice can be traced in the key observation that in this way the resulting method coincides, as already pointed out in Section 5.6, with *BFGS* and *L-BFGS* when applied on quadratic problems using exact line search.

5.7.1 The proposed method

```
   Data: $\mathbf{x}_0 \in \mathbb{R}^n, B_0 = I$ spd, $\mathbf{d}_0 = -\mathbf{g}_0$, $k = 0$;
1  while $\mathbf{g}_k \neq 0$ do
2      $\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{d}_k$ ;            /* $\lambda_k$ verifies (5.4), (5.5) */
3      $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$;
4      $\mathbf{y}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$;
5      if $\|B_k\mathbf{s}_k - \frac{\mathbf{s}_k B_k \mathbf{s}_k}{\|\mathbf{s}_k\|^2}\mathbf{s}_k\| < toll$ then
6          Define $\overline{\mathbf{g}}_{k+1}$ as the projection of $\mathbf{g}_{k+1}$ on $< \mathbf{s}_k >^{\perp}$ ;
7      else
8          Define $\overline{\mathbf{g}}_{k+1}$ as the projection of $\mathbf{g}_{k+1}$ on $< \mathbf{s}_k, B_k\mathbf{s}_k >^{\perp}$ ;
9      end
10     Define $\mathcal{L}^{(k)} :=$ sd $U_k$ s.t. $\mathcal{L}^{(k)}_{B_k}\mathbf{s}_k = B_k\mathbf{s}_k$ and
       $\mathcal{L}^{(k)}_{B_k}\overline{\mathbf{g}}_{k+1} = \alpha_k\overline{\mathbf{g}}_{k+1}$;
11     $B_{k+1} = \mathcal{L}^{(k)}_{B_k} - \frac{\mathcal{L}^{(k)}_{B_k}\mathbf{s}_k\mathbf{s}_k^T\mathcal{L}^{(k)}_{B_k}}{\mathbf{s}_k^T\mathcal{L}^{(k)}_{B_k}\mathbf{s}_k} + \frac{\mathbf{y}_k\mathbf{y}_k^T}{\mathbf{y}_k^T\mathbf{s}_k}$;
12     Compute $\mathbf{d}_{k+1} = -B_{k+1}^{-1}\mathbf{g}_{k+1}$;
13     Set $k := k + 1$ ;
14 end
```

Algorithm 6: A convergent $\mathcal{L}^{(k)}$QN method with quadratic termination property verified if exact line search is used.

Remark 17. *Observe that the existence of a matrix $\mathcal{L}^{(k)}_{B_k}$ satisfying conditions at line 10 of Algorithm 6 can be proved using techniques analogous to those used in the proof of Lemma 5.6.3. In particular the required orthogonal matrices U_k can be constructed as the product of two or three orthogonal Householder matrices depending if condition at line 5 of Algorithm 6 is satisfied or not.*

5.7.2 Computational Complexity

The computational complexity of Algorithm 6 is $O(n)$ in space and time. This follows using equation (5.55) with $\phi = 0$. In fact we have

$$\lambda(\mathcal{L}^{(k+1)}_{B_{k+1}}) = d(U_{k+1}^T B_{k+1} U_{k+1}) =$$

$$d(U_{k+1}^T \mathcal{L}^{(k)}_{B_k} U_{k+1} - U_{k+1}^T \frac{\mathcal{L}^{(k)}_{B_k}\mathbf{s}_k\mathbf{s}_k^T\mathcal{L}^{(k)}_{B_k}}{\mathbf{s}_k^T\mathcal{L}^{(k)}_{B_k}\mathbf{s}_k} U_{k+1} + U_{k+1}^T \frac{\mathbf{y}_k\mathbf{y}_k^T}{\mathbf{y}_k^T\mathbf{s}_k} U_{k+1}).$$

Since U_k and U_{k+1} can be chosen as the product of two or three Householder matrices (see Lemma 5.6.3 and Remark 17), the right hand side of the above expression can be computed in $O(n)$ FLOPs using Proposition 3.2 in [28]. In particular, using Remark 2 in [28], the worst case computational cost per step can be estimated in $36n + O(1)$.

5.7.3 Numerical Results

We have used performance profiles (see [46]) in order to investigate the numerical behavior of Algorithm 4 with $\phi = 0$ (refinement of the method introduced in [27]) when compared with Algorithm 6, and Algorithm 6 when compared with $\mathcal{D}QN$ [17], $\mathcal{H}QN$ [7, 40] and *L-BFGS* with $M = 30$ [75]. The latter method, that has been implemented by the Poblano toolbox [47], has a computational complexity per step analogous to Algorithm 6; however it requires more memory space to be implemented. We have tested the algorithms on a set of medium/large scale problems from CUTEst [56] (see Table 5.1), using the line-search routine provided by Poblano, i.e., the Moré-Thuente cubic interpolation line search (which implements the Strong-Wolfe conditions) enforcing the reproducibility of our results. In order to make a fair comparison we have used for all the algorithms the same stopping criteria as those from Poblano. We have used the following parameters where the names of the variables are the same as those from Poblano:

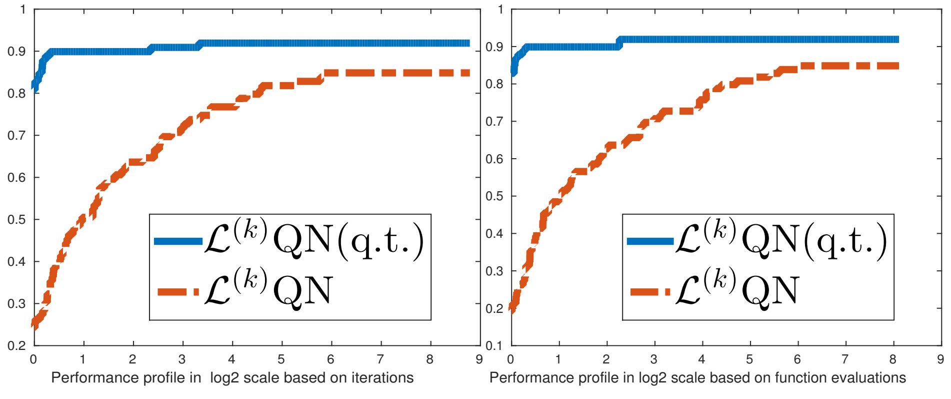
```
LineSearch_xtol=1e-15;
LineSearch_ftol=1e-4;
LineSearch_gtol=0.9;
LineSearch_stpmin=1e-15;
LineSearch_stpmax=1e15;
LineSearch_maxfev=20;

StopTol=1e-6;
MaxIters=10000;
MaxFuncEvals=50000;
RelFuncTol=1e-20;
```

LineSearch_ftol=α in (5.4) and LineSearch_gtol=β in (5.5).
Let us point out that, as in Poblano, the successful termination is achieved when $\|g_k\|_2/n \leq \mathit{StopTol}$ being n the dimension of the problem. In the following Figures 5.2 and 5.3 the caption '$\mathcal{L}^{(k)}$QN' will indicate Algorithm 4 and '$\mathcal{L}^{(k)}$QN(q.t.)' (quadratic termination) will indicate Algorithm 6.

5.7.4 Conclusions

In this chapter we have proposed novel optimization schemes $\mathcal{L}^{(k)}$QN obtained generalizing the updates in the restricted Broyden class by means of projections of the Hessian approximations B_k on adaptive low complexity matrix algebras, and, in particular, we have studied in detail a new *BFGS*-type method. Even if it is known that finite quadratic termination is not relevant for general Quasi-Newton methods [70], the numerical results presented in Section 5.7.3 confirm that exploiting the adaptivity of the spaces where to choose the approximations $\tilde{B}_k$ of B_k also in order to endow the



(a)

(b)

Figure 5.2: Performance profiles for Algorithm 4 when compared with Algorithm 6 on a set of 99 problems from CUTEst [56]. LineSearch_ftol=1e-4; LineSearch_gtol=0.9;

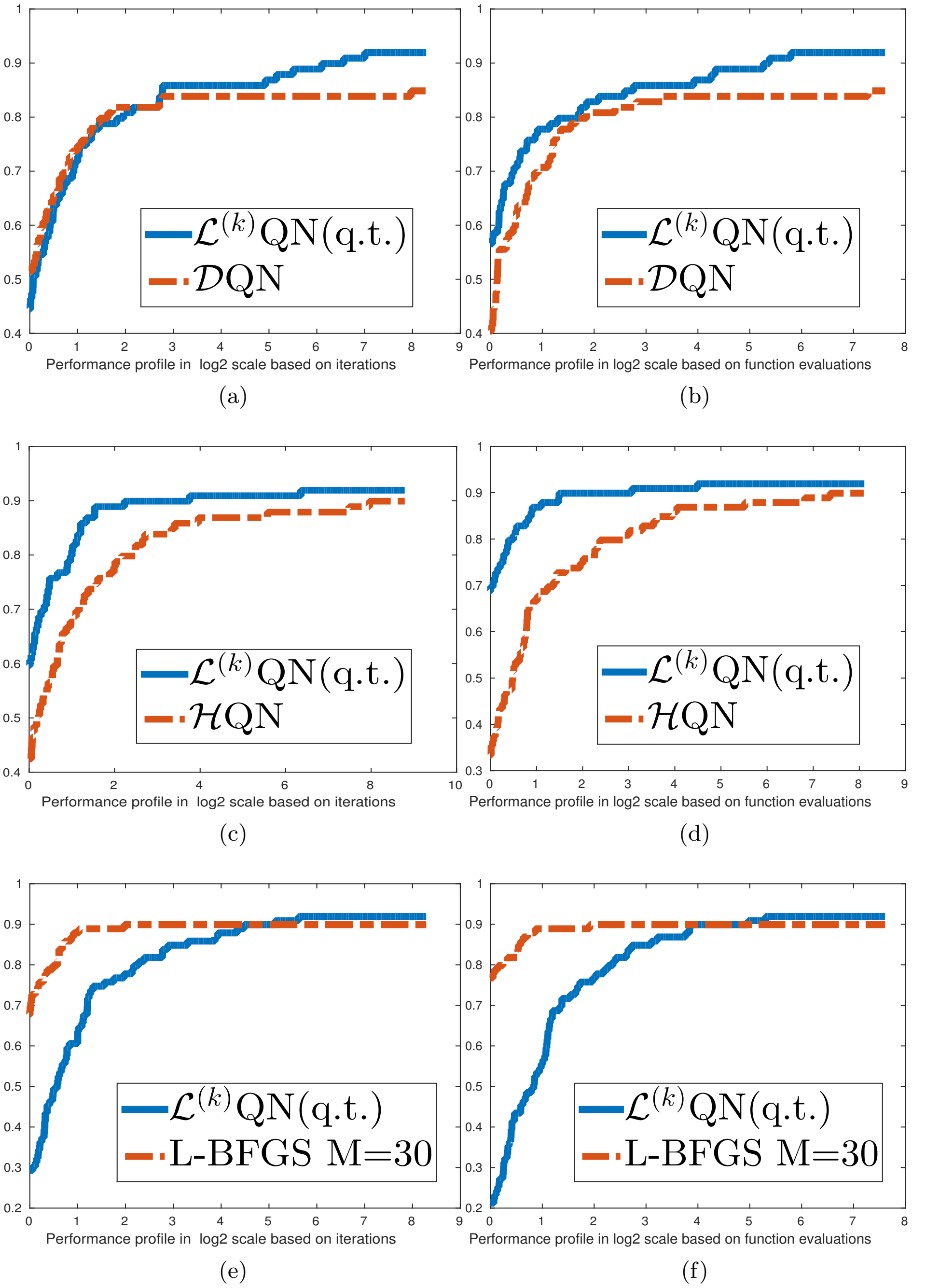
$BFGS$-type method with this property, considerably improves the performances of the basic $\mathcal{L}^{(k)}$QN scheme in Algorithm 4 (see Figure 5.2), which is a convergent refinement of the methods considered in [27]. As numerical results confirm, the introduction of such adaptive choice of the matrix algebras $\mathcal{L}^{(k)}$, permits to show, moreover, that existing fixed algebras $\mathcal{L}QN$ methods, $\mathcal{H}QN$ and $\mathcal{D}QN$, can be overcome (see Figure 5.3). Concerning the results obtained comparing our $\mathcal{L}^{(k)}QN$ method with L-$BFGS$, even if the comparison on this set of problems is unfavorable for the method we propose with respect to the probability of win (see Figure 5.3), it is important to note that $\mathcal{D}QN$ and $\mathcal{H}QN$ have been proved to be competitive with L-$BFGS$ on some real world problems (see [7, 17, 48]). For the above reasons further investigation urges in order to understand if the method we propose could be a valid competitor of L-$BFGS$ on those problems where large values of the parameter M must be chosen in order to guarantee satisfactory performances (see also [64]). However, let us point out that the increasing performances of our $\mathcal{L}^{(k)}QN$ schemes on general problems with respect to [40, 27] (see Figure 5.2 and 5.3) are encouraging because one guesses that certainly exist and deserve investigation more valuable criteria for the choice at each step of the algebra $\mathcal{L}^{(k)}$ where to project the matrix B_k; thus $\mathcal{L}^{(k)}QN$ could really become competitive with $L - BFGS$.

It is clear that $\mathcal{L}^{(k)}QN$ methods should be also compared with the class of nonlinear conjugate gradient methods. Moreover, it would be important to understand if the matrices generated by means of our Quasi Newton-type updates could be useful as preconditioners for nonlinear conjugate gradient methods as in [18]. Of course, further investigation should be devoted, in future, in order to understand if the Broyden Class-version of Algorithm

6 can produce better performances for $\phi \in (0, 1)$. Last but not least, it could be interesting to understand if the results presented in this chapter can be extended to the modified $BFGS$ method for non-convex functions as in [73].

Table 5.1: Problem Set

Prob	**Dim.**	**N.Z.**	**Prob**	**Dim.**	**N.Z.**
1] ARGLINA	200	20100	50] FLETCHBV	10000	19999
2] ARGLINB	50	1275	51] FLETCHCR	1000	1999
3] ARGLINC	100	4851	52] FMINSRF2	15625	77377
4] ARGTRIGLS	200	20100	53] FMINSURF	15625	122078125
5] ARWHEAD	5000	9999	54] FREUROTH	5000	9999
6] BA-L1LS	57	438	55] GENHUMPS	5000	9999
7] BDQRTIC	1000	4990	56] GENROSE	500	999
8] BOX	10000	39994	57] HILBERTB	50	1275
9] BOXPOWER	10000	29997	58] HYDC20LS	99	1125
10] BROWNAL	1000	500500	59] INDEFM	10000	29997
11] BROYDN3DLS	10000	29997	60] INTEQNELS	502	125252
12] BROYDN7D	5000	17497	61] JIMACK	3549	118824
13] BROYDNBDLS	10000	69979	62] LIARWHD	10000	19999
14] BRYBND	10000	69979	63] MANCINO	100	5050
15] CHAINWOO	10000	19999	64] MODBEALE	20000	39999
16] COSINE	10000	19999	65] MOREBV	1000	2997
17] CRAGGLVY	5000	9999	66] MSQRTALS	4900	12007450
18] CURLY10	1000	10945	67] MSQRTBLS	4900	12007450
19] CURLY20	1000	20790	68] NCB20	5010	99821
20] CURLY30	1000	30535	69] NCB20B	5000	99810
21] DECONVU	63	1111	70] NONCVXU2	10000	39987
22] DIXMAANA	3000	8999	71] NONCVXUN	100	386
23] DIXMAANB	3000	8999	72] NONDIA	10000	19997
24] DIXMAANC	3000	8999	73] NONDQUAR	10000	29997
25] DIXMAAND	3000	8999	74] NONMSQRT	4900	173950
26] DIXMAANE	3000	8999	75] OSCIPATH	500	999
27] DIXMAANF	3000	8999	76] PENALTY1	1000	500500
28] DIXMAANG	3000	8999	77] PENALTY2	100	5050
29] DIXMAANH	3000	8999	78] POWELLSG	10000	20000
30] DIXMAANI	3000	8999	79] POWER	10000	50005000
31] DIXMAANJ	3000	8999	80] QUARTC	10000	10000
32] DIXMAANK	3000	8999	81] SBRYND	1000	6979
33] DIXMAANL	3000	8999	82] SCHMVETT	10000	29997
34] DIXMAANM	3000	8999	83] SCOSINE	10000	19999
35] DIXMAANN	3000	8999	84] SENSORS	1000	500500
36] DIXMAANO	3000	8999	85] SINQUAD	100	199
37] DIXMAANP	3000	8999	86] SPARSINE	100	1232
38] DIXON3DQ	10000	19998	87] SPARSQUR	10000	159494
39] DQDRTIC	5000	5000	88] SPMSRTLS	10000	43326
40] DQRTIC	5000	5000	89] SROSENBR	10000	15000
41] EDENSCH	2000	3999	90] SSBRYBND	5000	34979
42] EIGENALS	110	6105	91] TESTQUAD	1000	1000
43] EIGENBLS	2550	3252525	92] TOINTGSS	10000	29997
44] EIGENCLS	2652	3517878	93] TOINTPSP	50	165
45] ENGVAL1	5000	9999	94] TOINTQOR	50	165
46] EXTROSNB	1000	1999	95] TQUARTIC	5000	9999
47] FLETCBV2	100	199	96] TRIDIA	5000	9999
48] FLETCBV3	100	199	97] VARDIM	100	5050
49] FLETBV3M	10000	19999	98] VAREIGVL	5000	12502500
			99] WOODS	10000	17500



(a) (b) (c) (d) (e) (f)

Figure 5.3: Performance profiles for Algorithm 6 when compared with $\mathcal{D}$QN, $\mathcal{H}$QN and *L-BFGS* with $M = 30$ on a set of 99 problems from CUTEst [56]. LineSearch_ftol=1e-4; LineSearch_gtol=0.9;

Chapter 6

Conclusions and Future Works

In this thesis we have shown how suitable projections onto low complexity matrix algebras can produce evident computational benefits in gaining the efficiency of iterative methods when used in order to solve different problems. Nevertheless, we strongly believe that further research should be carried on, not only in order to better clarify some aspects of the techniques and the results contained in the previous chapters, but mainly in order to broaden their field of applicability. In particular, in addition to the possible investigations pointed out at the end of the previous chapters, a further possible line of research can be identified in the introduction of ad-hoc low complexity matrix algebras in order to preconditioning more general linear systems than those introduced in Chapter 4. To this end, consider $B \in \mathbb{C}^{n\times n}$ and denote by $\mathbb{U}_n$ the set of unitary matrices of dimension n. With the aim of defining preconditioners with better *clustering* capabilities (see [101]) than those obtained as simple projections onto matrix algebras, called optimal preconditioners, in [102] the *superoptimal* preconditioner has been introduced, i.e., obtained as

$$\min_{X\in\mathcal{L} \text{ invertible}} \|I - X^{-1}B\|_F. \tag{6.1}$$

As pointed out in [23], if B is Hermitian, then the solution of the above problem can be expressed as $X = \mathcal{L}_{B^2}\mathcal{L}_B^{-1}$. Even if the superoptimal preconditioner provides a cluster of the eigenvalues of Hermitian positive definite Toeplitz matrices – as pointed out in [24] –, in general it holds that (see [25]):

$$\lambda_k((\mathcal{L}_{B^2}\mathcal{L}_B^{-1})^{-1}B) \leq \lambda_k(\mathcal{L}_B^{-1}B) \quad k = 1, \ldots, n. \tag{6.2}$$

The inequality in (6.2) represents a non desirable behavior for the superoptimal operator, when used as preconditioner, since it does not guarantee an improvement in the condition number of the preconditioned matrix with respect to the optimal operator.

Different approaches can be devised in order to produce *good* preconditioners using projections onto low complexity matrix algebras. For example, the fact that point 9. in Theorem 1.4.9 holds for every space $\mathcal{L} = \mathrm{sd}\, U$ suggests that to construct a suitable preconditioner for a given Hermitian positive definite matrix B, one could consider the following problem:

$$\min_{U \in \mathbb{U}_n,\, U \text{ low complexity}} K(\mathcal{L}_B^{-1} B), \tag{6.3}$$

where, of course, if the minimization is performed in $\mathbb{U}_n$ without any further constraints, the minimum is realized by the unitary matrix which diagonalizes B and $K(\mathcal{L}_B^{-1} B) = 1$. The same approach could be considered substituting the K-condition number with the 2-norm condition number. Finally, observe that using results in [101, 85], another possible approach to define a meaningful preconditioner, could be to search for the solution of the problem

$$\min_{U \in \mathbb{U}_n,\, U \text{ low complexity}} rank(\mathcal{L}_B - B). \tag{6.4}$$

Of course, in all the above reasonings, one could investigate the use of low complexity matrix spaces more general than $\mathrm{sd}\, U$ spaces or, using analogous techniques to those introduced in [29], one can consider, instead of $\mathcal{L}_B$, suitable combination of projections of the powers of B.

References

[1] P. Berkhin, A survey on pagerank computing. Internet Mathematics 2(1): 73–120, 2005.

[2] D. Bertaccini, C. Di Fiore, P. Zellini. Complessità e iterazione. Percorsi, matrici e algoritmi veloci nel calcolo numerico. Bollati Boringhieri 2013.

[3] D. Bertaccini, F. Durastante, Iterative Methods Preconditioning for Large Sparse Linear Systems with Applications. CRC Press 2017.

[4] R. Bhatia, Matrix analysis. Graduate Texts in Mathematics, Springer Verlag, 1997.

[5] D. Bini, P. Favati, On a matrix algebra related to the discrete Hartley transform. SIAM Journal on Matrix Analysis and Applications 14(2): 500–507, 1993.

[6] A. Bortoletti, C. Di Fiore, On a set of matrix algebras related to discrete Hartley-type transforms. Linear algebra and its applications 366: 65–85, 2003.

[7] A. Bortoletti, C. Di Fiore, S. Fanelli, P. Zellini, A new class of quasi-newtonian methods for optimal learning in mlp-networks. IEEE Transactions on Neural Networks 14(2): 263–273, 2003.

[8] E. Bozzo, C. Di Fiore, On the use of certain matrix algebras associated with discrete trigonometric transforms in matrix displacement decomposition. SIAM Journal on Matrix Analysis and Applications 16(1): 312-326, 1995.

[9] C. Brezinski, M. Redivo-Zaglia, Rational extrapolation for the PageRank vector. Mathematics of Computation 77(263): 1585–1598, 2008.

[10] C. Brezinski, M. Redivo-Zaglia, S. Serra-Capizzano, Extrapolation methods for PageRank computations. Comptes Rendus Mathematique 340(5): 393–397, 2005.

[11] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems 30(1): 107–117, 1998.

[12] A.Z. Broder, R. Lempel, F. Maghoul, J. Pedersen, Efficient PageRank approximation via graph aggregation. Information Retrieval 9(2): 123–138, 2006.

[13] C. G. Broyden, J. E. Dennis, J. J. Morè, On the local superlinear convergence of quasi-newton methods. IMA Journal of Applied Mathematics 12, 1973.

[14] R.H. Byrd, S.L. Hansen, J. Nocedal, Y. Singer, A stochastic quasi-newton method for large-scale optimization. SIAM Journal on Optimization 26(2): 1008–1031, 2016.

[15] R.H. Byrd, J. Nocedal, A tool for the analysis of quasi-newton methods with application to unconstrained minimization. SIAM Journal on Numerical Analysis 26(3): 727–739 ,1989.

[16] R. H. Byrd, J. Nocedal, Y.-X. Yuan, Global convergence of a class of quasi-newton methods on convex problems. SIAM Journal on Numerical Analysis, 24(5): 1171–1190, 1987.

[17] J.F. Cai, R.H. Chan, C. Di Fiore, Minimization of a detail-preserving regularization functional for impulse noise removal. Journal of Mathematical Imaging and Vision 29(1), 2007.

[18] A. Caliciotti, G. Fasano, M. Roma, Novel preconditioners based on quasi–newton updates for nonlinear conjugate gradient methods. Optimization Letters 11(4): 835–853, 2017.

[19] P. Cannarsa, T. D'Aprile, Introduction to measure theory and functional analysis. Springer 2015.

[20] T.F. Chan, An optimal circulant preconditioner for Toeplitz systems. SIAM journal on scientific and statistical computing 9(4): 766–71, 1988.

[21] R.H. Chan, Circulant preconditioners for Hermitian Toeplitz systems. SIAM Journal on Matrix Analysis and Applications 10(4): 542–550, 1989.

[22] R.H. Chan, X.Q. Jin, An introduction to iterative Toeplitz solvers. Society for industrial and applied mathematics, 2007.

[23] R.H. Chan, X.Q. Jin, M.C. Yeung, The circulant operator in the Banach algebra of matrices. Linear algebra and its applications 149: 41–53, 1991.

[24] R. H. Chan, X. Q. Jin, M. C. Yeung, The spectra of super-optimal circulant preconditioned Toeplitz systems. SIAM journal on numerical analysis 28(3): 871–879, 1991.

[25] C.M. Cheng, X.Q. Jin, S.W. Vong, W. Wang, A note on spectra of optimal and superoptimal preconditioned matrices. Linear algebra and its applications 422, 2007.

[26] S. Cipolla, C. Di Fiore, F. Tudisco, Euler-richardson method preconditioned by weakly stochastic matrix algebras: a potential contribution to pagerank computation. Electronic Journal of Linear Algebra 32: 254–272, 2017.

[27] S. Cipolla, C. Di Fiore, F. Tudisco, P. Zellini, Adaptive matrix algebras in unconstrained minimization. Linear Algebra its Applications 471: 544 – 568, 2015.

[28] S. Cipolla, C. Di Fiore, P. Zellini, Low complexity matrix projections preserving actions on vectors. Submitted for publication, 2017.

[29] S. Cipolla, C. Di Fiore, P. Zellini, Regularizing properties of a class of matrices including the Optimal and the Superoptimal preconditioners. Submitted for publication, 2017.

[30] S. Cipolla, C. Di Fiore, P. Zellini, Updating broyden class-type descent directions by householder adaptive transforms. Submitted for publication, 2017.

[31] T.A. Davis, Y. Hu, The University of Florida sparse matrix collection. ACM Transactions on Mathematical Software (TOMS) 38(1): 1–25, 2011.

[32] G.M. Del Corso, A. Gulli, F. Romani, Fast PageRank computation via a sparse linear system. Internet Mathematics 2(3): 251–273, 2005.

[33] J. E. Dennis, J. J. Moré, A characterization of superlinear convergence and its application to quasi-newton methods. Mathematics of Computation 28(126): 549–560,1974.

[34] F. Di Benedetto, Gram matrices of fast algebras have a rank structure. SIAM Journal on Matrix Analysis and Applications 31(2): 526–545, 2009.

[35] F. Di Benedetto, C. Estatico, S. Serra-Capizzano, Superoptimal preconditioned conjugate gradient iteration for image deblurring. SIAM Journal on Scientific Computing 26(3): 1012–35, 2005.

[36] F. Di Benedetto, S. Serra-Capizzano, A note on the superoptimal matrix algebra operators. Linear and Multilinear Algebra 50(4): 343–72, 2002.

[37] F. Di Benedetto, S. Serra-Capizzano, A unifying approach to abstract matrix algebra preconditioning. Numerische Mathematik, 82: 57–90, 1999.

[38] F. Di Benedetto, S. Serra-Capizzano, Optimal multilevel matrix algebra operators. Linear and Multilinear Algebra 48(1): 35–66, 2000.

[39] C. Di Fiore, Structured matrices in unconstrained minimization methods. In Contemporary mathematics, American Mathematical Society, 205–219, 2001.

[40] C. Di Fiore, S. Fanelli, F. Lepore, P. Zellini, Matrix algebras in Quasi-Newton methods for unconstrained minimization. Numerische Mathematik 94: 479–500, 2003.

[41] C. Di Fiore, S. Fanelli, P. Zellini, Low-complexity minimization algorithms. Numerical Linear Algebra with Applications 12(8), 755–768, 2005.

[42] C. Di Fiore, S. Fanelli, P. Zellini, Low complexity secant quasi-Newton minimization algorithms for nonconvex functions. Journal of Computational and Applied Mathematics 210(1): 167–74, 2007.

[43] C. Di Fiore, F. Lepore, P. Zellini, Hartley-type algebras in displacement and optimization strategies. Linear algebra and its applications 366: 215–232, 2003.

[44] C. Di Fiore, P. Zellini, Matrix algebras in optimal preconditioning. Linear Algebra and its Applications 335: 1 – 54, 2001.

[45] C. Di Fiore, P. Zellini, Matrix decompositions using displacement rank and classes of commutative matrix algebras. Linear algebra and its applications 229: 49–99, 1995.

[46] E.D. Dolan, J.J. Moré, Benchmarking optimization software with performance profiles. Mathematical programming 91(2): 201–213, 2002.

[47] D.M. Dunlavy, T.G. Kolda, E. Acar, Poblano v1. 0: A matlab toolbox for gradient-based optimization. Sandia National Laboratories, Albuquerque, NM and Livermore, CA, Tech. Rep. SAND2010-1422, 2010.

[48] A. Ebrahimi, G. Loghmani, B-spline curve fitting by diagonal approximation BFGS methods. Iranian Journal of Science and Technology, Transactions A: Science 1–12,2018.

[49] C. Estatico, A class of filtering superoptimal preconditioners for highly ill-conditioned linear systems. BIT Numerical Mathematics 42(4): 753–78, 2002.

[50] D. Fasino, F. Tudisco, An algebraic analysis of the graph modularity. SIAM Journal on Matrix Analysis and Applications 35(3): 997–1018, 2014.

[51] D. Fasino, F. Tudisco, Generalized modularity matrices. Linear Algebra and its Applications 502: 327–345, 2016.

[52] R. Fezzani, L. Grigori, F. Nataf, K. Wang, Block filtering decomposition, Numerical Linear Algebra with Applications 21(6): 703–721, 2014.

[53] D. Gleich, L. Zhukov, P. Berkhin, Fast parallel PageRank: A linear system approach. Yahoo! Research Technical Report 13–22, 2004.

[54] G.H. Golub, C. Greif, An Arnoldi-type algorithm for computing page rank. BIT Numerical Mathematics 46(4): 759–771, 2006.

[55] G.H. Golub, C.F. Van Loan, Matrix computations. JHU Press, 2012.

[56] N.I. Gould, D. Orban, P.L. Toint, Cutest: a constrained and unconstrained testing environment with safe threads for mathematical optimization. Computational Optimization and Applications 60(3): 545–557, 2015.

[57] S. A. Goreinov, I. V. Oseledets, D. V. Savostyanov, E. E. Tyrtyshnikov, N. L. Zamarashkin. How to find a good submatrix. Matrix Methods: Theory, Algorithms, Applications, V. Olshevsky and E. Tyrtyshnikov, eds., World Scientific, 247–256, 2010

[58] S. A. Goreinov, E. E. Tyrtyshnikov, N. L. Zamarashkin, A theory of pseudoskeleton approximations. Linear Algebra and its Applications 261, 1–21, 1997.

[59] S. A. Goreinov, N. L. Zamarashkin, E. E. Tyrtyshnikov, Pseudoskeleton approximations by matrices of maximal volume. Mathematical Notes 62, 1997.

[60] C. Gu, L. Wang, On the multi-splitting iteration method for computing PageRank. Journal of Applied Mathematics and Computing 42(1-2): 479–490, 2013.

[61] M. H. Gutknecht, Block Krylov space methods for linear systems with multiple right-hand sides: an introduction. In Modern Mathematical Models, Methods and Algorithms for Real World Systems (A.H. Siddiqi, I.S. Duff, and O. Christensen, eds.) Anamaya Publishers, 2007.

[62] M. Hanke, J.G. Nagy, R.J. Plemmons, Preconditioned iterative regularization. Numerical linear algebra 141–163 1993.

[63] R.A. Horn and C.R. Johnson, Matrix analysis. Cambridge university press, 2012.

[64] L. Jiang, R.H. Byrd, E. Eskow, R.B. Schnabel, A preconditioned L-BFGS algorithm with application to molecular energy minimization. Tech. rep., Colorado University at Boulder Dept. of Computer Science, 2004.

[65] X.-Q. Jin, Y.-M. Wei, A survey and some extensions of T. Chan's preconditioner. Linear Algebra and its Applications 428(2): 403–412, 2008.

[66] T. Kailath, A. H. Sayed, Fast reliable algorithms for matrices with structure. Siam, 1999.

[67] S. Kamvar, T. Haveliwala, C. Manning, G. Golub, Adaptive methods for the computation of PageRank. Linear Algebra and its Applications 386: 51–65, 2004.

[68] S. Kamvar, T. Haveliwala, C. Manning, G. Golub, Exploiting the block structure of the web for computing pagerank. Stanford University Technical Report, 2003.

[69] S.D. Kamvar, T.H. Haveliwala, C.D. Manning, G.H. Golub, Extrapolation methods for accelerating PageRank computations. In Proceedings of the 12th international conference on World Wide Web, ACM 261–270, 2003.

[70] T.G. Kolda, D.P. O'leary, L. Nazareth, BFGS with update skipping and varying memory. SIAM Journal on Optimization 8(4): 1060–1083, 1998.

[71] D. Kressner, R. Luce, Fast computation of the matrix exponential for a Toeplitz matrix. arXiv preprint arXiv:1607.01733, 2016.

[72] A.N. Langville, C.D. Meyer, Google's PageRank and beyond: The science of search engine rankings. Princeton University Press, 2011.

[73] D.H. Li, M. Fukushima, A modified BFGS method and its global convergence in nonconvex minimization. Journal of Computational and Applied Mathematics 129(1): 15–35, 2001.

[74] C. Liu, , S.A. Vander Wiel, Statistical quasi-newton: A new look at least change. SIAM Journal on Optimization 18(4): 1266–1285,2007.

[75] D.C. Liu, J. Nocedal, On the limited memory BFGS method for large scale optimization. Mathematical Programming 45: 1989.

[76] L. Lopez, V. Simoncini, Preserving geometric properties of the exponential matrix by block krylov subspace methods. BIT Numerical Mathematics 46(4): 813–830, 2006.

[77] I. Marek, D.B. Szyld, Iterative and semi-iterative methods for computing stationary probability vectors of Markov operators. Mathematics of Computation 61(204): 719–731, 1993.

[78] J. S. Milne, Algebraic geometry (v6.01), 2015. Available at www.jmilne.org/math/.

[79] G. Ming, X. S. Li, P. S. Vassilevski, Direction-preserving and schur-monotonic semiseparable approximations of symmetric positive definite matrices. SIAM Journal on Matrix Analysis and Applications 31, 2010.

[80] L. Nazareth, A relationship between the BFGS and conjugate gradient algorithms and its implications for new algorithms- SIAM Journal on Numerical Analysis 16(5): 794–800, 1979.

[81] M. E. Newman, A measure of betweenness centrality based on random walks. Social networks 27(1): 39–54, 2005.

[82] J. Nocedal, S. J. Wright, Numerical Optimization. Springer, 2nd ed., 2006.

[83] J.D. Noh, H. Rieger, Random walks on complex networks. Physical review letters 92(11), 2004.

[84] M.A. Olshanskii, E.E. Tyrtyshnikov, Iterative methods for linear systems: theory and applications. Society for Industrial and Applied Mathematics, 2014.

[85] I. Oseledets, E.E. Tyrtyshnikov, A unifying approach to the construction of circulant preconditioners. Linear algebra and its applications 418: 435-449, 2006.

[86] P. Pons, M. Latapy, Computing communities in large networks using random walks. In International Symposium on Computer and Information Sciences. Springer Berlin Heidelberg 284–293, 2005.

[87] M.J.D. Powell, Some global convergence properties of a variable metric algorithm for minimization without exact line searches, Nonlinear Programming, SIAM-AMS Proc. 9: 53–72, 1976.

[88] X. Qi, E. Fuller, Q. Wu, Y. Wu, C.Q. Zhang, Laplacian centrality: A new centrality measure for weighted networks. Information Sciences 194: 240–253, 2012.

[89] U.G. Rothblum, C.P. Tan, Upper bounds on the maximum modulus of subdominant eigenvalues of nonnegative matrices. Linear Algebra and its Applications 66: 45–86, 1985.

[90] Y. Saad, Analysis of some Krylov subspace approximations to the matrix exponential operator, SIAM Journal of Numerical Analysis, 29(1): 209–228, 1992.

[91] Y. Saad, Iterative methods for sparse linear systems. Society for Industrial and Applied Mathematics, 2003.

[92] Y. Saad, Numerical methods for large eigenvalue problems. SIAM, 2011.

[93] M. Sadkane, Block-Arnoldi and Davidson methods for unsymmetric large eigenvalue problems. Numerische Mathematik 64(1): 195–211, 1993.

[94] V. Sanchez, P. Garcia, A.M. Peinado, J.C Segura, A.J. Rubio, Diagonalizing properties of the discrete cosine transforms. IEEE transactions on Signal Processing 43(11): 2631–2641, 1995.

[95] V. Sanchez, A.M. Peinado, J.C Segura, P. Garcia, A.J. Rubio, Generating matrices for the discrete sine transforms. IEEE transactions on Signal Processing 44(10): 2644–2646, 1995.

[96] S. Serra-Capizzano, Superlinear PCG methods for symmetric Toeplitz systems, Mathematics of Computation of the American Mathematical Society 68: 793–803, 1999.

[97] T. Tao, Topics in random matrix theory. American Mathematical Society, 2012.

[98] F. Tudisco, A note on certain ergodicity coeflcients. Special Matrices 3(1): 175–185, 2015.

[99] F. Tudisco, C. Di Fiore, A preconditioning approach to the pagerank computation problem. Linear Algebra and its Applications 435(9): 2222-2246, 2011.

[100] F. Tudisco, C. Di Fiore, E.E Tyrtyshnikov, Optimal rank matrix algebras preconditioners. Linear Algebra and its Applications 438(1): 405–427, 2013.

[101] E. E. Tyrtyshnikov, A unifying approach to some old and new theorems on distribution and clustering. Linear Algebra and its Applications 232: 1–43, 1996.

[102] E. E. Tyrtyshnikov, Optimal and superoptimal circulant preconditioners. SIAM Journal on Matrix Analysis ans Application 13(2): 459–473, 1992.

[103] E.E. Tyrtyshnikov, Lectures of the Rome-Moscow school of Matrix Methods and Applied Linear Algebra, 2016.

[104] R.S. Varga, Matrix iterative analysis. Springer Science & Business Media, 2009.

[105] C. Wagner, Tangential frequency filtering decompositions for unsymmetric matrices. Numerische Mathematik 78(1): 143–163, 1997.

[106] C. Wagner, G. Wittum, Adaptive filtering. Numerische Mathematik 78(2): 305–328, 1997.

[107] R.K. Yarlagadda, J.E. Hershey, Hadamard matrix analysis and synthesis: with applications to communications and signal/image processing. Springer Science & Business Media, 2012.