Université Fédérale

Toulouse Midi-Pyrénées

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut Supérieur de l'Aéronautique et de l'Espace

Présentée et soutenue par :
Alexandre SCOTTO DI PERROTOLO
le vendredi 26 août 2022

Titre :

Méthodes aléatoires en algèbre linéaire numérique appliquées à l'assimilation de données

---

École doctorale et discipline ou spécialité :
ED MITT : Mathématiques et Applications

Unité de recherche :
Équipe d'accueil ISAE-ONERA MOIS

Directeur(s) de Thèse :

M. Xavier VASSEUR (directeur de thèse)
M. Youssef DIOUANE (co-directeur de thèse)

Jury :
M. Serge GRATTON Professeur INP Toulouse - Président
M. Xavier VASSEUR Ingénieur de recherche ISAE-SUPAERO - Directeur de thèse
M. Youssef DIOUANE Professeur Polytechnique Montréal - Co-directeur de thèse
Mme Stefania BELLAVIA Professeure Université de Florence, Italie - Rapporteure
Mme Melina FREITAG Professeur Institut für Mathematik, Université de Potsdam, Allemagne - Rapporteure
M. Marcin CHRUST Scientifique ECMWF, Reading, Royaume-Uni - Examinateur

Au cheminement hasardeux des idées.

Remerciements

En premier lieu, je voudrais évidemment remercier mes directeurs de thèse, Xavier Vasseur et Youssef Diouane, pour avoir tenu bon au cours de ces quatre années pour le moins mouvementées. Cela vaut tout particulièrement pour les moments de découragement qui se sont inévitablement manifestés, et que vous avez toujours su m'aider à dépasser. Ces moments paraissent bien loin au moment où j'écris ces lignes, et je suis très heureux que vous m'ayez poussé pour mener ce travail jusqu'au bout et de pouvoir être fier du résultat. C'était loin d'être gagné.

Ensuite, je tiens à remercier les membres du jury à commencer par Melina Freitag et Stefania Bellavia, qui ont accepté de rapporter mon manuscrit de thèse, et ce avec grande rigueur malgré des délais contraints. Vos retours ont été grandement appréciés. Merci également à Marcin Chrust, d'avoir accepté d'examiner ma thèse sur ses vacances, ainsi qu'à Serge Gratton, pour avoir assuré la présidence de mon jury avec bienveillance. Enfin, je tiens à remercier Selime Gürol et Michel Salaün, que d'obscures règles administratives ont empêché de figurer officiellement en tant qu'invité du jury, quand bien même leur participation à mon doctorat est clairement de mesure non nulle. À ce propos, mes sincères remerciements à Michel pour avoir été une oreille disponible et attentive à mes nombreux doutes et questionnements.

Place à présent à l'essai libre. Puisqu'être exhaustif semble hors d'atteinte je vais m'abstenir d'en avoir la prétention. Sans grande inspiration, je vais procéder chronologiquement, et commencer par l'équipe de doctorants/postdocs de l'ISAE : Valentin, Erwan, Franco, pour nos déjeuners et nos pauses café toujours bien animées, Guillaume pour nos interminables discussions lors de nos mardis bière/burger, Narjès et Zoé pour l'animation de notre club de déculpabilisation des doctorants anonymes, Maxime pour le massacre régulier de petites balles jaunes en feutrine entre deux conversations sur l'avenir d'un matheux dans la société moderne. Un mot également pour Adrien, Noémie et Marina, rares rescapés de l'ENSICA à ce déraisonnable niveau d'études supérieures (quelle idée), et pour Ilyes, mon co-bureau arrivé en cours de route et qui doit être, en ce moment même, en train de terminer la rédaction de son manuscrit. Une mention aussi à l'équipe de basket de l'AS ISAE-ONERA, Arnaud en tête, pour ces sessions aussi sympathiques qu'intenses qui ont participé à mon équilibre et qui m'ont permis de souffler.

Ensuite, je voudrais remercier le département de Génie Mathématiques et Modélisation de l'INSA, pour m'avoir accueilli pendant cette excellente année d'ATER. Pêle-mêle, merci à Sandrine, Violaine, Olivier, Frédéric, Pierre, Laure, Loïc pour avoir fait de cette première (mais, je l'espère, pas dernière) année d'enseignement une expérience formidable. Fort à parier qu'il soit difficile de faire mieux dans un avenir proche. Merci à Julie, co-bureau exemplaire pour nos bonnes rigolades (évidemment constructives) lors de sessions mouvementées de corrections de copies, et à Clément pour les conversations mathématico-politique de très bon aloi.

Enfin, je voudrais également remercier ma famille, qui a suivi avec un mélange d'enthousiasme et de perplexité cette longue marche, mention spéciale à Julia et Tiffany pour leur pression psychologique constante : "tu finis quand ta thèse alors ?", question à laquelle je n'ai longtemps eu aucune réponse. Une pensée toute particulière et un peu émue à mes deux grands-pères,

Robert Scotto Di Perrotolo et Guy Gschwind, qui sont partis avant que je puisse terminer ce travail, mais qui auraient certainement été fiers. Merci aussi à mes parents pour avoir essayé tant bien que mal de me faire relativiser au cours de ces montagnes russes émotionnelles. Un mot également pour Alexandre, Jérémy, Julien et Rémi, amis de (très) longues dates, toujours fidèles au poste et avec qui je peux constater non sans plaisir que notre maturité collective progresse peu au fil des années. Pour finir, je voudrais remercier Julie, qui partage mon quotidien depuis maintenant onze ans, et qui a dû supporter mon adorable caractère de cochon durant ces quatre années de doctorat.

Abstract

Keywords: data assimilation, eigenpair approximation, limited memory preconditioners, low-rank approximation, nonlinear least-squares, randomized algorithms.

Randomized methods for computing approximate singular value/eigenvalue decompositions have gained a lot of attention during the past decades. These methods have proven to perform well, are computationally efficient, and are well suited for large scale applications. In this regard, recent researches have proposed successful applications of randomized algorithms in data assimilation, where the huge size of the problems is prohibitive for a large number of standard approaches. In this thesis, we propose three interconnected contributions in randomized methods for low-rank approximation, extraction of eigenpairs, and preconditioning within variational data assimilation.

First, we propose a general error analysis of randomized low-rank approximation in Frobenius and spectral norms. This generalization extends the possibilities of analysis to a larger class of randomized methods by allowing general covariance matrices and non-zero mean for the Gaussian sample matrix. Particularization of our bounds to the Randomized Singular Value Decomposition (RSVD) shows that we improve the reference error bounds due to Halko, Martinsson and Tropp (2011).

Then, we develop randomized algorithms to address specific eigenvalue problems that naturally arise in data assimilation. The proposed methods are versatile, and generalize the contributions from Saibaba, Lee and Kitanidis (2016) and Daužickaitė et al. (2021). We then provide a theoretical analysis of the methods, which gives insights regarding the number of subspace iterations, number of random samples, and optimal covariance matrix of the Gaussian sample matrix. Numerical illustrations on a data assimilation test problem confirm the potential of our algorithms.

Finally, we propose a class of randomized spectral limited memory preconditioners for variational data assimilation. We provide such preconditioners for two given Krylov subspace methods: an inverse-free approach in the primal space introduced by Guröl (2013) and a dual space method proposed by Gratton and Tshimanga (2009). The reduced dimension of the dual space makes this latter approach computationally efficient both in terms of cost and storage. Our randomized spectral limited memory preconditioners are based on appropriate expressions identified by Gürol (2013) where we replace expensive computations of exact eigenpairs by approximations obtained with a randomized procedure. Illustrations on a benchmark four-dimensional variational data assimilation problem prove that our randomized preconditioners perform well, opening interesting perspectives.

Résumé

Mots-clefs : approximation de rang faible, approximation spectrale, assimilation de données, méthodes aléatoires, moindres carrés non-linéaires, préconditionnement à mémoire limitée.

Les méthodes aléatoires pour le calcul approché de décomposition aux valeurs singulières/valeurs propres ont suscité beaucoup d'intérêt au cours des dernières décennies. Ces méthodes se sont avérées performantes, efficaces en termes de coût de calcul et particulièrement bien adaptées aux problèmes de grande taille. À cet égard, des recherches récentes ont proposé des applications de ces méthodes en assimilation de données, où la taille des problèmes est prohibitive pour un grand nombre d'approches classiques. Dans cette thèse, nous proposons trois contributions interconnectées aux méthodes aléatoires pour l'approximation de rang faible, l'extraction d'information spectrale et le préconditionnement en assimilation de données variationnelle.

Premièrement, nous proposons une analyse générale de l'erreur d'approximation de rang faible aléatoire en norme de Frobenius et en norme spectrale. Cette généralisation étend les possibilités d'analyse à un plus grand nombre de méthodes aléatoires en autorisant des matrices de covariance générales et un vecteur de moyenne non nulle pour la matrice gaussienne d'échantillonnage. La particularisation de nos bornes à la méthode dite de Randomized Singular Value Decomposition (RSVD) montre que nous améliorons les bornes d'erreur de référence proposées par Halko, Martinsson et Tropp (2011).

Ensuite, nous présentons des algorithmes aléatoires pour la résolution de problèmes aux valeurs propres spécifiques qui apparaissent notamment en assimilation de données. Les méthodes proposées sont polyvalentes et généralisent les contributions de Saibaba, Lee et Kitanidis (2016) et Daužickaité et al. (2021). Nous fournissons ensuite une analyse théorique de nos méthodes qui éclaire sur la sensibilité de l'erreur au nombre d'itérations de sous-espace, au nombre d'échantillons aléatoires et à la matrice de covariance pour la matrice gaussienne d'échantillonnage. Des illustrations numériques sur un problème d'assimilation de données confirment le potentiel de nos algorithmes.

Enfin, nous proposons une classe de préconditionnement à mémoire limitée aléatoire dédiée à l'assimilation de données variationnelle. Nous proposons ces préconditionnements pour deux méthodes de Krylov en particulier: une approche dite inverse-free dans l'espace primal introduite par Guröl (2013) et une méthode d'espace dual proposée par Gratton et Tshimanga (2009). La dimension réduite de l'espace dual rend cette dernière approche plus intéressante à la fois en termes de coût de calcul et de stockage. Les préconditionnements aléatoires proposés sont basés sur des expressions adaptées identifiées par Gürol (2013) pour lesquelles les calculs coûteux d'information spectrale exacte sont remplacés par des approximations obtenues avec une procédure aléatoire. Des illustrations sur un problème d'assimilation de données variationnel quadridimensionnel de référence démontrent le potentiel de nos préconditionnements aléatoires, ouvrant ainsi des perspectives intéressantes.

Contents

CHAPTER 1

Introduction

Context. Data assimilation is a general framework where observations and a priori information are coupled to estimate the underlying state of a complex dynamical system. Initially, the research in data assimilation was driven by practitioners in the field of weather prediction and ocean modelling [23, 40], but it has since been used in many more domains such as, among others, geosciences [21] and mechanical engineering [3]. With the increasing complexity of the models and the growing number of observations [9], solving data assimilation problems has become particularly challenging from a numerical perspective. Practically, there are mainly two different approaches to get a solution: either sequential or variational [5]. The sequential approach corrects the model state estimate whenever the observations are available, but will not be considered in this thesis. Instead, we will focus on the variational approach, where the model fitting problem is converted into an optimization problem. This optimization problem is traditionally solved using descent algorithms, typically a truncated Gauss-Newton method [34, 45]. With this method, the successive descent directions are obtained as the solution of linear systems involving linear operators (matrix-free) of very large size.

Solving a data assimilation problem with the variational approach [5, Chapter 2] thus reduces to solving a sequence of large scale linear systems involving symmetric positive definite operators. This can be efficiently achieved using preconditioned Krylov subspace methods [69, Chapter 2]. For symmetric positive definite linear systems, the Krylov subspace method of choice is the conjugate gradient method [55] whose convergence rate can be improved by the use of a preconditioner [73, 92]. In data assimilation, the problem structure yields a natural preconditioner, which is generally improved using so-called two-level preconditioners [81] such as the limited memory preconditioner [66]. This class of preconditioners integrates eigeninformation to further improve the convergence rate of the conjugate gradient method. Given the large size of data assimilation problems, the eigeninformation is never computed with dedicated eigensolvers, but is rather replaced by Ritz pairs computed from Krylov subspaces [39] associated to the previous linear systems. This technique has proven to perform well in data assimilation [89] but has certain drawbacks. First, the number of approximate eigenvectors is constrained by the number of preconditioned conjugate gradient method iterations. Then, since the eigeninformation is related to a previous system, the obtained preconditioner is not perfectly adapted to the current system, in particular in the beginning of the optimization process. Finally, updating the approximate eigeninformation along the sequence is far from trivial.

In the last decades, randomized methods have gained a lot of importance in the numerical linear algebra community and have been identified as one of the key elements for future advances in numerical linear algebra [19]. The principle of randomization is to construct a sketch [93] of a matrix using random sampling. The sketch acts as a reduced dimension surrogate of the matrix hopefully containing most of the information. The idea is then to compute quantities

of interest on the sketch, at a lower computational cost, to deduce information on the original matrix. In this regard, a relevant sketch provides quantities of interest that are close to the ones of the original matrix/linear operator. To efficiently sketch a matrix, the distribution of the random samples is crucial and generally integrates information related to the relative importance of the columns/rows/coefficients of the matrix [1, 18]. In contexts where the matrix is not available and is replaced by a black-box linear map, sketching amounts to apply the linear map to random sample vectors, usually standard Gaussian vectors. The essential of the arithmetic cost induced by such algorithms thus concentrates in these matrix-vector products which can easily be parallelized. This characteristic makes randomized methods structurally scalable.

Randomized methods using sketching have proven to perform well on a number of fundamental numerical linear algebra problems such as least-squares problems [6, 29], matrix factorizations [63, Section 16] and low-rank approximations [2]. The flagship method for this latter is the randomized singular value decomposition method introduced in [94] and popularized in [53]. In light of the Eckart–Young–Mirsky theorem [30], this algorithm addresses the problem of computing a low rank approximation of a general matrix via the computation of an approximate truncated singular value decomposition. Beyond its performance, the interest for this method was also popularized by the rigorous theoretical analysis of the resulting low rank approximation error proposed in [53, Section 9 and 10]. This analysis gave important theoretical guarantees, and was later completed by results focusing more on the approximate singular vectors/values accuracy [49, 77]. These analysis identified key elements monitoring the randomized singular value decomposition performance such as the number of random samples and the singular value distribution. The key ideas behind this algorithm have then been widely exploited to design randomized methods tackling more sophisticated singular value/eigenvalue problems such as generalized Hermitian eigenvalue problems [80] and generalized singular value decomposition [78].

Altogether, the properties of randomized methods make them particularly adapted to the computationally intensive context of data assimilation. In this regard, recent researches have proposed randomized procedures for solving problems in variational data assimilation. In [17], the authors proposed the randomized incremental optimal technique, which uses randomized low rank approximations to approximately compute the new descent directions. The objective of this approach is to replace the iterative Krylov subspace method by a fully parallel randomized procedure. More recently, the authors in [24] proposed an alternative, where the preconditioned conjugate gradient is maintained and the randomized low rank approximation is rather used to construct a limited memory preconditioner. The idea is to construct the two-level preconditioner using approximate eigenpairs obtained from the randomized method instead of exact eigenpairs. Here, the use of randomized methods within the Krylov subspace method allows to maintain theoretical guarantees on the overall convergence process. The results obtained on a toy data assimilation problem confirmed that such randomized approaches behave similarly to the preconditioner constructed using exact eigenpairs.

Scope and goals. The existing randomized methods for variational data assimilation are essentially limited to formulations involving symmetric positive definite matrices. Practically, this requires the availability of matrix factorizations that may not be affordable in operational contexts. In the absence of such factorizations, alternative formulations have been proposed which implies the use of dedicated Krylov subspace methods. In this thesis, we will focus on two of them. The first one is referred to as the inverse-free preconditioned conjugate gradient [50, Section 3.1], and has been proposed to solve data assimilation problems in the absence of a particular operator inverse. The second one is referred to as the augmented restricted preconditioned conjugate gradient [48] and is based on the dual formulation of the problem. The interest in the dual formulation is that the dual space (i.e. the observation space) has a reduced dimension, yielding significant improvements in the overall arithmetic cost and storage. For these Krylov subspace methods, specific formulations of the limited memory preconditioner have

been proposed [50]. These variants require to compute eigenpairs of specific eigenvalue problems involving operators that are no longer symmetric with respect to the standard inner product.

The main objective of this thesis is to propose and study randomized methods for computing approximate eigenpairs notably adapted to the preconditioning of the inverse-free and the augmented restricted preconditioned conjugate gradient. We propose a rigorous theoretical analysis of the two proposed randomized methods based on a generalization of the randomized singular value decomposition analysis. Then, we propose specific implementations of our methods adapted to the construction of spectral limited memory preconditioners for both Krylov subspace methods. The obtained algorithms are flexible and could for instance be combined with usual deterministic strategies such as Krylov subspace recycling methods. The numerical experiments conducted on a toy test problem showed that the randomized preconditioners have similar performance as the exact spectral limited memory preconditioner, which suggests that such method could become an efficient component in operational variational data assimilation.

Outline. This thesis contains four main chapters.

In Chapter 2, we introduce the background material. We begin with defining and recalling elementary notions in linear algebra. Then, we introduce the conjugate gradient method and the notion of preconditioning. Next, we present the essential material related to randomized numerical linear algebra. We also introduce the variational data assimilation problem in the framework of weighted nonlinear least-squares problems. In the end, we describe the randomized approaches that have been proposed in this context.

Chapter 3 is devoted to a general error analysis of randomized low rank approximation methods. First, we propose a refined deterministic analysis for the low rank approximation error in Frobenius and spectral norms that is then used to derive bounds both in expectation and probability. The novelty is that the proposed stochastic bounds are tighter and hold for general Gaussian matrices. Then, we specialize our bounds to the analysis of the randomized singular value decomposition, which demonstrates that our analysis both generalizes and improves the reference error bounds proposed in [53]. Finally, we propose numerical illustrations on an instructional test problem where our bounds are compared to both the reference bounds and to the empirical error.

In Chapter 4, we develop randomized algorithms to address two related generalized eigenvalue problems in a non-Euclidean inner product. Such specific eigenvalue problems notably arise in the data assimilation formulations of interest in Chapter 5. Our algorithms are based on the randomized subspace iteration, and use the Rayleigh-Ritz method to extract the approximate eigenpairs. We propose two different extraction methods, a direct one, and an inverse one that can be related to the harmonic Rayleigh-Ritz analysis. Based on the general analysis presented in Chapter 3, an average-case analysis of our algorithms is proposed in both weighted spectral and Frobenius norms. A comparison between our bounds and prior bounds is also proposed. Finally, we investigate the performance of our algorithm in terms of eigenpair accuracy on a three-dimensional variational data assimilation test problem.

Then, in Chapter 5, we use the randomized algorithms introduced in Chapter 4 to propose randomized spectral limited memory preconditioners for the inverse-free and augmented restricted preconditioned conjugate gradient methods. Variants are proposed depending on the availability of a first-level preconditioner. Numerical experiments on the Lorentz-95 model demonstrate the potential of the proposed randomized preconditioners. Several sets of parameters are used, considering different number of observations, and in all cases, our randomized preconditioners behave competitively compared to the exact spectral and Ritz limited memory preconditioners.

Finally, we conclude and give some perspectives in Chapter 6.

Notation

$\|\cdot\|_2$	Euclidean norm for vectors, spectral norm for matrices
$\|\cdot\|_F$	Frobenius norm
$\langle\cdot,\cdot\rangle$	Euclidean inner-product
I_n	Identity matrix of order n
A^T	Transpose of the matrix A
A^{-1}	Inverse of the matrix A when defined
$A^\dagger$	Moore-Penrose pseudo-inverse of the matrix A
$\mathcal{R}(S)$	Subspace spanned by the columns of the matrix S
$\lambda_i(A)$	i-th largest eigenvalue of the square matrix A

CHAPTER 2

Background material

Contents

Abstract

In this first chapter, we introduce the fundamental notions and background material that will be helpful throughout the manuscript.

Section 2.1 is intended to recall basic linear algebra notions. Its main objective is to go through standard definitions of norms, orthogonal projectors, orthogonality when the underlying inner product is non-Euclidean.

In Section 2.2, we introduce and derive the conjugate gradient method. We recall the relations between the conjugate gradient method and the Lanczos procedure to derive approximate eigenpairs, and theoretical results related to its convergence. Then, we recall basic elements on the notion of preconditioning, and we present the preconditioned conjugate gradient method algorithm. Finally, we introduce a particular class of preconditioners that are the Limited Memory Preconditioner.

Next, we introduce in Section 2.3 the key notions and algorithms in the randomized numerical linear algebra literature. We present the randomized subspace iteration, aimed at approximating the range of general matrices, and then the Randomized Singular Value Decomposition method. Then, we recall key results on the analysis of these methods. We finish by introducing the Nyström method, which is a randomized method dedicated to symmetric positive definite matrices.

Finally, in Section 2.4, we present the mathematical framework of variational data assimilation, i.e. the weighted nonlinear least-squares problem. We then introduce relevant deterministic and stochastic preconditioning strategies.

2.1 Preliminaries

2.1.1 Norm induced by a non standard inner product

In this thesis, we will make an intensive use of non-Euclidean inner products. Let $\mathsf{W} \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix. The matrix W defines an inner product on $\mathbb{R}^n$ denoted by $\langle\cdot,\cdot\rangle_{\mathsf{W}}$ and defined as

$$\forall x, y \in \mathbb{R}^n, \quad \langle x, y\rangle_{\mathsf{W}} = x^{\mathsf{T}}\mathsf{W}y.$$

The corresponding norm is thus defined as

$$\|x\|_{\mathsf{W}} = \sqrt{\langle x, x\rangle_{\mathsf{W}}} = \sqrt{x^{\mathsf{T}}\mathsf{W}x}.$$

We will refer to this norm as the W-norm.

Remark 2.1. With this definition, the Euclidean norm corresponds to $\|\cdot\|_{I_n}$. Nevertheless, we will avoid this notation and rather maintain the usual notation $\|\cdot\|_2$.

In the following, our objective is to briefly introduce the usual notions related to inner products, such as symmetry or orthogonality when considered with inner products of the form $\langle\cdot,\cdot\rangle_{\mathsf{W}}$.

W-symmetry. A matrix $A \in \mathbb{R}^{n\times n}$ is said to be W-symmetric if it is self-adjoint with respect to the inner product defined by W, that is if it satisfies

$$\forall x, y \in \mathbb{R}^n, \quad \langle Ax, y\rangle_{\mathsf{W}} = \langle x, Ay\rangle_{\mathsf{W}}.$$

In a matrix form, this definition implies that A satisfies $\mathsf{W}A = A^{\mathsf{T}}\mathsf{W}$. Accordingly, a matrix A is W-symmetric if and only if the matrix $\mathsf{W}A$ is symmetric in the usual Euclidean way. The notions of positiveness and definiteness can thus be extended similarly by verifying whether $\mathsf{W}A$ is positive and/or definite.

Orthogonality with respect to W. Another fundamental notion that can be extended is the notion of orthogonality. Two vectors $x, y \in \mathbb{R}^n$ are said to be W-conjugate if they satisfy

$$\langle x, y \rangle_{\mathsf{W}} = 0.$$

From this, we define the notion of conjugation of a set of vectors. A set of vectors $\{x_1, \dots, x_k\}$ of $\mathbb{R}^n$ is said to be W-conjugate if it satisfies

$$i \neq j \implies \langle x_i, x_j \rangle_{\mathsf{W}} = 0 \quad \forall\, 1 \leq i, j \leq k.$$

If we denote $X = [x_1 \ \cdots \ x_k] \in \mathbb{R}^{n \times k}$, then the W-conjugation of $\{x_1, \dots, x_k\}$ can be written in matrix form as

$$X^{\mathsf{T}} \mathsf{W} X = \operatorname{diag}(\alpha_1, \dots, \alpha_k).$$

By definition, it is clear that $\alpha_i = \|x_i\|_{\mathsf{W}}^2$ for all $1 \leq i \leq k$. Then, the counterpart of orthogonal matrices can be obtained by imposing to a set of W-conjugate vectors to also be of unit W-norm. Therefore, a matrix $X \in \mathbb{R}^{n \times k}$ is said to be W-orthogonal if it satisfies

$$X^{\mathsf{T}} \mathsf{W} X = I_k.$$

In this case, the set of columns of X are orthonormal with respect to the inner product induced by W. With this definition, orthogonal matrices in $\mathbb{R}^{n \times n}$ are I_n-orthogonal. Finally, let $X \in \mathbb{R}^{n \times k_1}$ and $Y \in \mathbb{R}^{n \times k_2}$ be two matrices and let us denote $\mathcal{X} = \mathcal{R}(X)$ and $\mathcal{Y} = \mathcal{R}(Y)$ the respective spanned subspaces. Then we note

$$\mathcal{X} \perp_{\mathsf{W}} \mathcal{Y} \iff X^{\mathsf{T}} \mathsf{W} Y = 0_{k_1, k_2},$$

where $0_{k_1, k_2} \in \mathbb{R}^{k_1 \times k_2}$ is the zero matrix.

2.1.2 Orthogonal projector

Orthogonal projectors are a fundamental class of linear maps which naturally arise when solving distance minimization problems. Let $\mathcal{S} \subset \mathbb{R}^n$ denote a k dimensional subspace of $\mathbb{R}^n$. The orthogonal projection of a vector $x \in \mathbb{R}^n$ onto $\mathcal{S}$, denoted by $\pi(\mathcal{S})(x)$, is defined as the unique solution of

$$\pi(\mathcal{S})(x) = \operatorname*{argmin}_{y \in \mathcal{S}} \|x - y\|_2. \tag{2.1}$$

The map $x \mapsto \pi(\mathcal{S})(x)$ is a linear map of rank $k = \dim(\mathcal{S})$ called the orthogonal projector onto $\mathcal{S}$. Let $S \in \mathbb{R}^{n \times k}$ denote a full column rank matrix such that $\mathcal{S} = \mathcal{R}(S)$. Then the matrix form of $\pi(\mathcal{S})$ reads

$$\pi(\mathcal{S}) = S(S^{\mathsf{T}} S)^{-1} S^{\mathsf{T}}. \tag{2.2}$$

Remark 2.2. The full rankness assumption on S can be relaxed, if we replace the inverse of $S^{\mathsf{T}} S$ by its Moore-Penrose pseudo inverse.

By definition, the orthogonal projector $\pi(\mathcal{S})$ is only determined by $\mathcal{S}$. Accordingly, the matrix representation of $\pi(\mathcal{S})$ remains unchanged when performing a change of basis for S. In particular, if the columns of S form an orthonormal basis for $\mathcal{S}$, then the matrix form of $\pi(\mathcal{S})$ simplifies to just SS^{T}.

Remark 2.3. In this thesis, we will manipulate matrices more than linear subspaces. Accordingly, and for convenience, if S is a column vector matrix, we will denote $\pi(S)$ the orthogonal projector onto $\mathcal{R}(S)$.

The definition of the orthogonal projector given in (2.1) can be straightforwardly generalized to non-Euclidean inner products. Let W be a symmetric positive definite matrix. We define the W-orthogonal projector onto $\mathcal{S}$ as the unique solution of the problem

$$\pi_{\mathsf{W}}(\mathcal{S}) = \operatorname*{argmin}_{y \in \mathcal{S}} \|x - y\|_{\mathsf{W}}.$$

In a similar manner, if the columns of the matrix S form a basis if $\mathcal{S}$, then one has the following matrix representation

$$\pi_{\mathsf{W}}(\mathcal{S}) = S(S^{\mathsf{T}}\mathsf{W}S)^{-1}S^{\mathsf{T}}\mathsf{W}. \tag{2.3}$$

Here we note that this expression simplifies if the columns of S form a W-orthogonal basis of $\mathcal{S}$.

2.1.3 Angle between subspaces

The notion of angle between subspaces has been introduced in [25] to quantify the separation between subspaces. It generalizes the notion of geometrical angles between vectors to the case of linear subspaces. The theory has been developed and is now referred to as the Closeness-Separation (CS) decomposition. We refer the interested reader to [71] for an historical review on this topic. Here, we will not need the CS decomposition in itself, and we rather focus on defining the principal angles between subspaces.

The principal angles between subspaces can be defined whenever there exists an inner product. Here, we immediately define them in the generalized setting of non-Euclidean inner product as introduced in [58]. Let $F \in \mathbb{R}^{n\times k}$ and $G \in \mathbb{R}^{n\times p}$ be two full column rank matrices with $p \geq k$ and let $\mathsf{W} \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix. Let us consider the eigenvalue decomposition of the symmetric positive semidefinite matrix

$$F^{\mathsf{T}}\mathsf{W}(I_n - \pi_{\mathsf{W}}(G))F = U\Lambda U^{\mathsf{T}},$$

with $U \in \mathbb{R}^{k\times k}$ an orthogonal matrix and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_k) \in \mathbb{R}^{k\times k}$ a diagonal matrix with $\lambda_1 \geq \cdots \geq \lambda_k \geq 0$. Then by definition one has

$$\sin(\theta_j)^2 = \lambda_j, \quad 1 \leq j \leq k.$$

where $\theta_1, \ldots, \theta_k$ are the principal angles between $\mathcal{R}(F)$ and $\mathcal{R}(G)$. In a similar manner, the eigenvalues μ_j of the symmetric positive semidefinite matrix $F^{\mathsf{T}}\mathsf{W}\pi_{\mathsf{W}}(G)F$ with $0 \leq \mu_1 \leq \cdots \leq \mu_k$ are such that

$$\cos(\theta_j)^2 = \mu_j, \quad 1 \leq j \leq k.$$

The principal angles are therefore accessible by two means, either via their cosine or sine. Of course getting one immediately yields the other since $\sin(\theta_j)^2 + \cos(\theta_j)^2 = 1$. In practice, we favor the sine measures because they quantify the distance between two linear subspaces. Indeed, if $\sin(\theta_1) = \cdots = \sin(\theta_k) = 0$, then it implies that $\mathcal{R}(F) \subset \mathcal{R}(G)$.

An important property satisfied by the principal angles, which might not be obvious at first sight from the above definition, is that they only depend on $\mathcal{R}(F)$ and $\mathcal{R}(G)$ and not on the particular basis. This is a consequence of the fact that we can define them by means of projections only (see [39, Section 2.3]). This operator definition is more abstract and more general but will not be needed here, which is the reason why we decided to focus on the matrix definitions.

2.1.4 Miscellaneous

We introduce several concise results or definitions that do not deserve a full section on their own.

Partial ordering on symmetric matrices

Let $A, B \in \mathbb{R}^{n\times n}$ be two symmetric matrices. We write $A \preccurlyeq B$ if the matrix $B - A$ is positive semidefinite. This relation defines a partial ordering on symmetric matrices referred to as the Loewner order. Here, we will focus on a few properties and we refer the reader to [57, Section 7.7] for a detailed study of this ordering relation.

Let us assume that $A \preccurlyeq B$, then one has

1. $\|A\|_{2,F} \leq \|B\|_{2,F}$ (Monotonicity).
2. $Q^\mathsf{T} A Q \preccurlyeq Q^\mathsf{T} B Q, \quad \forall Q \in \mathbb{R}^{m\times n}$ (Conjugation rule).

Remark 2.4. The first item is actually true for any Schatten norm.

Gaussian matrices

Gaussian matrices will play an important role through this thesis, in particular in Chapter 3 where this distribution is at the core of the theoretical analysis. We will consider Gaussian matrices $Z \in \mathbb{R}^{n\times p}$ whose columns are independent Gaussian vectors. Thus, let $z_1, \ldots, z_p \in \mathbb{R}^n$ be the Gaussian vectors such that $Z = [z_1 \ \cdots \ z_p]$. Then each Gaussian vector is fully characterized by its mean vector and covariance matrix [51, Theorem 4.1]. Therefore, we use the classical notation

$$z_i \sim \mathcal{N}(\widehat{z}_i, \mathbf{Cov}(z_i)), \quad 1 \leq i \leq p,$$

where $\widehat{z}_i \in \mathbb{R}^n$ is the mean vector and $\mathbf{Cov}(z_i) \in \mathbb{R}^{n\times n}$ the covariance matrix of z_i. From elementary properties of Gaussian vectors (e.g. [51, Theorem 3.1]), one can write

$$z_i = \widehat{z}_i + \mathbf{Cov}(z_i)^{\frac{1}{2}} g_i, \quad \text{with} \quad g_i \sim \mathcal{N}(0, I_n), \quad 1 \leq i \leq p.$$

Here $\mathbf{Cov}(z_i)^{\frac{1}{2}}$ refers either to the positive definite square root of $\mathbf{Cov}(z_i)$ if $\mathbf{Cov}(z_i)$ is positive definite or to the unique positive semidefinite square root of $\mathbf{Cov}(z_i)$ if $\mathbf{Cov}(z_i)$ is positive semidefinite [57, Theorem 7.2.6].

If we further assume that the covariance matrices of each Gaussian vector are identical, that is $\mathbf{Cov}(z_1) = \cdots = \mathbf{Cov}(z_p)$, then one can write

$$Z = \widehat{Z} + \mathbf{Cov}(Z)^{\frac{1}{2}} G, \tag{2.4}$$

where $\widehat{Z} = [\widehat{z}_1 \ \cdots \ \widehat{z}_p] \in \mathbb{R}^{n\times p}$, $\mathbf{Cov}(Z) = \mathbf{Cov}(z_1)$ and $G = [g_1 \ \cdots \ g_p] \in \mathbb{R}^{n\times p}$. In this case, Z is entirely determined by $\widehat{Z}$ and $\mathbf{Cov}(Z)$ and we thus write by analogy that

$$Z \sim \mathcal{N}(\widehat{Z}, \mathbf{Cov}(Z)).$$

With this notation, we say that a matrix $G \in \mathbb{R}^{n\times p}$ is a standard Gaussian matrix if $G \sim \mathcal{N}(0, I_n)$.

2.2 The conjugate gradient method

Let us now consider the solution of a linear system of the form

$$Ax = b \tag{2.5}$$

where $A \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix and $b \in \mathbb{R}^n$ the right-hand side. In the context of this thesis, the linear systems that will arise are of very large size, so that A is not explicitly stored as matrix, but is rather a black-box linear transformation that can be

interacted with only via the map $y \mapsto Ay$. Direct methods for the solution of linear systems [26] are thus ineffective, and we must turn to iterative methods. Iterative methods substitute the exact computation of the solution with the computation of a series of approximate solutions hopefully converging towards $x = A^{-1}b$. Although very old (see [76] for a must-see historical review), the study of iterative methods for the solution of linear systems remains an active area of research, and we refer the reader to [69] for an overview of the existing methods.

An efficient class of iterative methods are Krylov subspace methods. Krylov subspace methods compute the i-th iterate as an approximate solution of (2.5) sought within the affine space $x_0 + \mathcal{K}_i(A, r_0)$ where $x_0 \in \mathbb{R}^n$ is the initial guess, $r_0 = b - Ax_0$ the initial residual and

$$\mathcal{K}_i(A, r_0) = \text{span}\left\{r_0, Ar_0, \ldots, A^{i-1}r_0\right\}, \tag{2.6}$$

the i-th order Krylov subspace. If there is no ambiguity, we will denote the Krylov subspace by $\mathcal{K}_i$ instead of $\mathcal{K}_i(A, r_0)$. Due to the Cayley-Hamilton theorem, it is known that $x - x_0$ is a polynomial in A applied to r_0. Consequently, searching an approximate solution within $\mathcal{K}_i$ is natural since elements y in $\mathcal{K}_i$ are of the form

$$y = \sum_{j=0}^{i-1} \alpha_j A^j r_0 = p(A)r_0,$$

with p a polynomial of degree at most $i-1$ with $p(0) = 1$. Krylov subspaces form an increasing sequence of subspaces, that is

$$\mathcal{K}_i(A, r_0) \subset \mathcal{K}_{i+1}(A, r_0).$$

Also, if $y \in \mathcal{K}_i(A, r_0)$, then $Ay \in \mathcal{K}_{i+1}(A, r_0)$. Both statements can be verified straightforwardly.

2.2.1 Derivation of the method

Krylov subspace methods differ from one another in the criterion used to select the i-th iterate within $x_0 + \mathcal{K}_i$ [69, Section 1.2]. For instance, computing x_i such that $b - Ax_i \perp x_0 + \mathcal{K}_i(A, r_0)$ defines the well-known Generalized Minimum Residual method [69, Section 1.2.5]. When A is symmetric positive definite, the Krylov subspace method of choice is the conjugate gradient method, initially introduced by the authors in [55]. The conjugate gradient method is defined such that the i-th iterate satisfies

$$\begin{aligned} x_i &= \underset{y \in x_0 + \mathcal{K}_i(A, r_0)}{\text{argmin}} \|y - x\|_A \\ &= x_0 + \underset{z \in \mathcal{K}_i(A, r_0)}{\text{argmin}} \|z - (x - x_0)\|_A. \end{aligned} \tag{2.7}$$

Remark 2.5. It can readily be deduced from (2.7) that this method for solving linear systems is closely related to the constrained minimization of the convex quadratic function $y \mapsto \|y - x\|_A^2$. This explains why the term gradient appears in the name of the method [43, Section 11.3.1].

As a norm minimization problem, the solution of (2.7) is obtained as the A-orthogonal projection of $x - x_0$ onto $\mathcal{K}_i$, that is

$$x_i = x_0 + \pi_A(\mathcal{K}_i)(x - x_0). \tag{2.8}$$

Assuming that a basis for $\mathcal{K}_i$ is available, x_i can be computed explicitly using the matrix form of $\pi_A(\mathcal{K}_i)$ as in (2.3). As already mentioned, the matrix form of $\pi_A(\mathcal{K}_i)$ can be simplified if we choose a matrix P_i whose columns are A-conjugate and form a basis for $\mathcal{K}_i$. This would avoid forming and inverting the reduced matrix $P_i^\mathsf{T} A P_i$. Since $\mathcal{K}_i \subset \mathcal{K}_{i+1}$, such an A-conjugate basis can be computed incrementally by successive augmentation, that is,

$$P_{i+1} = \begin{bmatrix} P_i & p_{i+1} \end{bmatrix}, \quad \text{with} \quad p_{i+1} \in \mathcal{K}_{i+1} \perp_A \mathcal{K}_i = \mathcal{R}(P_i).$$

Let us thus assume for the moment that such a basis is available, and let $p_1, \dots, p_i \in \mathbb{R}^n$ denote the basis vectors such that $P_i = [p_1 \ \cdots \ p_i]$. Then one can rewrite (2.8) as

$$\begin{aligned} x_i &= x_0 + P_i (P_i^\mathsf{T} A P_i)^{-1} P_i^\mathsf{T} A (x - x_0) \\ &= x_0 + \sum_{j=1}^{i} \frac{p_j^\mathsf{T} r_0}{p_j^\mathsf{T} A p_j} p_j, \end{aligned} \tag{2.9}$$

where we have used that $A(x - x_0) = r_0$. Assuming that the basis P_i is indeed constructed by successive augmentation, one has $P_i = [P_{i-1} \ p_i]$ which in turn implies that

$$x_i = x_{i-1} + \alpha_i p_i, \quad \text{with} \quad \alpha_i = \frac{p_i^\mathsf{T} r_0}{p_i^\mathsf{T} A p_i}.$$

This results in a simple recurrence relation.

Due to (2.9), $r_i = b - A x_i$ satisfies

$$r_i = r_0 - A P_i (P_i^\mathsf{T} A P_i)^{-1} P_i^\mathsf{T} r_0 = r_0 - \sum_{j=1}^{i} \frac{p_j^\mathsf{T} A r_0}{p_j^\mathsf{T} A p_j} A p_j.$$

By construction, we readily obtain that $P_i^\mathsf{T} r_i = 0$, meaning that the i-th residual satisfies $r_i \perp \mathcal{K}_i$. We also have a simple recurrence relation between the residuals which reads

$$r_i = r_{i-1} - \alpha_i A p_i.$$

By definition of the Krylov subspace (2.6), it is also clear that since $x_i - x_0 \in \mathcal{K}_i$, then $r_i \in \mathcal{K}_{i+1}(A, r_0)$. Therefore, the i-th residual is a natural candidate to augment the Krylov subspace basis from P_i to P_{i+1}. However, to maintain the A-conjugation of P_i, one cannot augment P_i with r_i directly. Instead, we augment P_i with the A-orthogonal projection of r_i onto $\mathcal{K}_{i+1}(A, r_0)^\perp$, that is we compute

$$p_{i+1} = \big(I_n - \pi_A(P_i)\big)\, r_i = r_i - \sum_{j=1}^{i} \frac{p_j^\mathsf{T} A r_i}{p_j^\mathsf{T} A p_j} p_j.$$

Here, since $r_i \in \mathcal{K}_{i+1}(A, r_0)$ and $\pi_A(P_i) r_i \in \mathcal{K}_i(A, r_0)$, we verify that one has indeed $p_{i+1} \in \mathcal{K}_{i+1}(A, r_0)$.

So far, we have obtained a one term recurrence for both the iterates and the residuals. However, it seems that computing the new basis vector p_{i+1} requires a full recurrence over all the previous basis vectors p_j, $j \leq i$. Counter intuitively, this is not the case since all the terms in the sum actually cancel out except for $j = i$. This is a consequence of the two properties, namely: (i) $r_i \perp \mathcal{K}_i$, and (ii) $A p_j \in \mathcal{K}_{j+1}$ since $p_j \in \mathcal{K}_j$. This implies that

$$p_j^\mathsf{T} A r_i = \langle A p_j, r_i \rangle = 0 \quad \forall\, j < i.$$

In the end, we obtain

$$p_{i+1} = r_i - \beta_i p_i, \quad \text{with} \quad \beta_i = \frac{p_i^\mathsf{T} A r_i}{p_i^\mathsf{T} A p_i}.$$

This short term recurrence is a fundamental property of the conjugate gradient method which makes it numerically attractive. It implies that the conjugate gradient method does not require large memory requirements, since a fixed number of vectors must be stored for the algorithm to be performed. This is for instance a major difference with the generalized minimum residual method, where the full basis of the Krylov subspace must be kept, thus increasing the storage at each iteration.

The overall procedure is given in Algorithm 2.1. We point out that the implementation slightly differs from the theoretical derivation because of additional simplifications that can be performed to improve the numerical efficiency, especially for computing α_i and β_i. We refer the reader to [43, Section 11.3] for further details on the derivation of the conjugate gradient method and additional algorithmic considerations.

Algorithm 2.1: Conjugate gradient method

Input: Symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$, right-hand side $b \in \mathbb{R}^n$, initial guess $x_0 \in \mathbb{R}^n$, and tolerance $\varepsilon > 0$.

$r_0 = b - Ax_0$
$\rho_0 = r_0^\mathsf{T} r_0$
$p_0 = r_0$
while *convergence is not reached* **do**
  $q_i = Ap_i$
  $\alpha_i = \rho_i / q_i^\mathsf{T} p_i$
  $x_{i+1} = x_i + \alpha_i p_i$
  $r_{i+1} = r_i - \alpha_i q_i$
  $\rho_{i+1} = r_{i+1}^\mathsf{T} r_{i+1}$
  if $\|r_{i+1}\|_2 \leq \varepsilon \|r_0\|_2$ **then**
    Stop the method.
  end
  $\beta_i = \rho_{i+1}/\rho_i$
  $p_{i+1} = r_{i+1} + \beta_i p_i$
end

Output: Final iterate x_f such that $\|Ax_f - b\|_2 \leq \varepsilon \|Ax_0 - b\|_2$.

2.2.2 Convergence analysis of the conjugate gradient method

The convergence of the conjugate gradient method has been well-studied [90] since the early 50's. In particular, refined analysis have been proposed which related the conjugate gradient method convergence to the convergence of the Ritz values (see Section 2.2.5). In particular, different phases in the convergence have been identified [7, 10]. Nevertheless, we will not enter the details of the refined convergence analysis, and rather focus on two main results that give a general idea of the convergence properties. Both results relate the convergence behavior to the eigenvalue distribution of A.

The first result we want to highlight is related to the clustering of the eigenvalue distribution of A. It has been shown [84, Theorem 38.5] that if $A \in \mathbb{R}^{n\times n}$ has exactly $d \leq n$ distinct eigenvalues, then the conjugate gradient method converges in at most d iterations. Consequently, a clustered spectrum is a very desirable feature to achieve fast convergence.

Then, let us rapidly introduce the probably best known convergence bound. The conjugate gradient method is dedicated to the minimization of the direct error in the A-norm, that is the i-th iterate minimizes $\|y - x\|_A$. Accordingly, it is natural to look at the decrease of this error along the iterations. From (2.8), we obtain that

$$\|x_i - x_0\|_A = \min_{z\in\mathcal{K}_i} \|z - (x - x_0)\|_A.$$

Using the definition of $\mathcal{K}_i$ and recalling that $A(x - x_0) = r_0$ we have

$$\begin{aligned} z - (x - x_0) &= \sum_{j=1}^{i} a_j A^{j-1} r_0 - (x - x_0) \\ &= \sum_{j=1}^{i} a_j A^j (x - x_0) - (x - x_0) \\ &= \left(\sum_{j=1}^{i} a_j A^j - I_n \right) (x - x_0) \\ &= q(A)(x - x_0), \end{aligned}$$

where q is a polynomial of degree i satisfying $q(0) = -1$. Let us denote $\mathcal{P}_i$ the set of such polynomials. Then the minimum over $z \in \mathcal{K}_i$ can be replaced by the minimum over $p \in \mathcal{P}_i$. Then, it can be shown [39, Theorem 2.54] that

$$\|x_i - x_0\|_A = \min_{p \in \mathcal{P}_i} \|p(A)(x - x_0)\|_A \leq \min_{p \in \mathcal{P}_i} \max_{i=1}^{n} \left| p(\lambda_i(A)) \right| \|x - x_0\|_A.$$

The min-max problem can be solved using the appropriately scaled and shifted Chebyshev polynomials yielding the well-known asymptotic convergence bound of the conjugate gradient method

$$\|x_i - x\|_A \leq 2 \left(\frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1} \right)^i \|x_0 - x\|_A,$$

where $\kappa_2(A)$ is the 2-norm condition number of A. This bound tells us that the condition number, and therefore the clustering of the eigenvalue distribution, monitors the asymptotic convergence of the conjugate gradient method. Nevertheless, whenever $\kappa_2(A) \gg 1$, this bound becomes ineffective. In particular, since a moderate number of iterations is generally performed, this asymptotic bound is not of great practical use. However, it suggests that the eigenvalue distribution of A has an important effect, yielding the introduction of preconditioners, which we discuss next.

2.2.3 Preconditioning

The notion of preconditioning is deeply connected to iterative methods [73, 92]. This consists in modifying the linear system such that the iterative method exhibits a faster convergence on the modified linear system. Formally, a preconditioner for the linear system in (2.5) is a symmetric positive definite operator $M \in \mathbb{R}^{n \times n}$, which can be used in several ways:

1. Left preconditioning: $MAx = Mb$.
2. Right preconditioning: $AMy = b$ with $x = My$.
3. Split preconditioning: $L^\mathsf{T} ALz = L^\mathsf{T} b$ with $LL^\mathsf{T} = M$ and $x = Lz$.

In the three different approaches the system matrices are similar and therefore enjoy the same eigenvalue distribution. Applying M as a split preconditioner requires a factorization of M, and allows us to apply the conjugate gradient method directly since the new system matrix remains symmetric positive definite. However, this is no longer the case with either left or right preconditioning, where the system matrix is no longer symmetric. For those approaches, the conjugate gradient method cannot be applied directly and one must derive an alternative variant.

The variant for left preconditioning is the most common one and is referred to as the Preconditioned conjugate gradient. The idea is to remark that although MA is not symmetric, it is M^{-1}-symmetric. It turns out that the theoretical derivation of the preconditioned conjugate gradient method can be performed analogously. The Krylov subspace is now $\mathcal{K}_i(MA, Mr_0)$, and the direct error is still minimized in the A-norm since $M^{-1}MA = A$ which yields the same expression for the iterates. However, the expression for the residual is modified and now reads

$$z_i = Mb - MAx_i = M(b - Ax_i) = Mr_i.$$

This suggests to first compute r_i as in the conjugate gradient method, before computing $z_i = Mr_i$. The modification of the residual modifies the expression of ρ_i which now reads $r_i^\mathsf{T} z_i$. The overall procedure is given in Algorithm 2.2. The change in the inner product implies that the preconditioned residuals z_i are now M^{-1}-conjugate instead.

Algorithm 2.2: Preconditioned conjugate gradient method adapted from [69, Algorithm 1.4].

Input: Symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$ and preconditioner $M \in \mathbb{R}^{n\times n}$, right-hand side $b \in \mathbb{R}^n$, initial guess $x_0 \in \mathbb{R}^n$, tolerance $\varepsilon > 0$.

$r_0 = b - Ax_0$
$z_0 = Mr_0$
$\rho_0 = r_0^\mathsf{T} z_0$
$p_0 = z_0$
while *convergence is not reached* **do**
  $q_i = Ap_i$
  $\alpha_i = \rho_i / q_i^\mathsf{T} p_i$
  $x_{i+1} = x_i + \alpha_i p_i$
  $r_{i+1} = r_i - \alpha_i q_i$
  $z_{i+1} = Mr_{i+1}$
  $\rho_{i+1} = r_{i+1}^\mathsf{T} z_{i+1}$
  if $\|r_{i+1}\|_M \leq \varepsilon \|r_0\|_M$ **then**
    Stop the method.
  end
  $\beta_i = \rho_{i+1} / \rho_i$
  $p_{i+1} = z_{i+1} + \beta_i p_i$
end

Output: Final iterate x_f such that $\|Ax_f - b\|_M \leq \varepsilon \|Ax_0 - b\|_M$.

The variant related to right preconditioning can be obtained similarly, remarking that AM is M-symmetric. This approach is less considered in the literature because it entails an additional application of M per iteration. Consequently we do not present this variant here. However, it turns out that this right preconditioning variant is central in data assimilation as discussed in Chapter 5.

2.2.4 The Limited Memory Preconditioner

The design of efficient preconditioners is a vast problem and generally uses additional knowledge of the underlying problem. Depending on the applications, relevant classes of preconditioners have been developed such as multigrid [88] or domain decomposition methods [83] which are particularly efficient when the system arises from discretized partial differential equations. In this thesis we will not address the problem of multilevel preconditioners. We will assume that

an efficient preconditioner M for the linear system (2.5) is readily available from the application. This is indeed the case in variational data assimilation, as will be discussed in Section 2.4.3.

Let us thus denote M a symmetric positive definite preconditioner for (2.5). In general, the action of M has a known effect on the eigenvalue distribution of MA and typically yields a large cluster of eigenvalues around 1. In this thesis, we consider a class of two-level preconditioners [81] called Limited Memory Preconditioners [47, 66] whose expression reads

$$P = \left[I_n - S(S^\mathsf{T}AS)^{-1}S^\mathsf{T}A\right] M \left[I_n - AS(S^\mathsf{T}AS)^{-1}S^\mathsf{T}\right] + S(S^\mathsf{T}AS)^{-1}S^\mathsf{T}. \tag{2.10}$$

Here, M is referred to as the first-level preconditioner, and $S \in \mathbb{R}^{n\times k}$ is a full column rank matrix. The limited memory preconditioner is used to further improve the preconditioner M and is entirely determined by S. In practice, using P can improve the eigenvalue clustering when the columns of S are approximate eigenvectors of MA associated to the eigenvalues left out by M. Such approximations can be computed either using dedicated methods, or using Ritz vectors (see Section 2.2.5). Indeed, observing that

$$PA = \left[I_n - \pi_A(S)\right] MA \left[I_n - \pi_A(S)\right] + \pi_A(S),$$

it is clear that $PAS = S$, which means that the columns of S are eigenvectors of PA associated to the eigenvalue 1. This implies that PA has a cluster of at least k eigenvalues at 1. Consequently, assuming that MA already has an important cluster of eigenvalues around 1, the LMP will be efficient if S contains eigenvectors associated to the left out eigenvalues. In this case, we expect P to perform significantly better that M. In addition, the limited memory preconditioner has the non-expansion of the spectrum property that has been identified in [46, Theorem 3.4]. This property implies that, apart from the eigenvalues clustered at 1 due to S, the remaining eigenvalues lie within the spectrum of MA. Consequently, the eigenvalue clustering is never worsen whenever 1 was already in the spectrum of MA.

Remark 2.6. Interestingly, the expression (2.10) have emerged from various fields. In [66], it is obtained as a block BFGS update of M, when approximating the inverse of A, while in [44] it appears as the solution of an optimization problem. It has also been proposed in the domain decomposition community as the balancing Neumann-Neumann preconditioner [62].

2.2.5 Eigenvalue approximations from the conjugate gradient method

Krylov subspaces can also be used to compute approximate eigenpairs of A. This is at the core of the Lanczos procedure which was first introduced by Lanczos in [59] to address the solution of the eigenvalue problems for symmetric positive definite matrices. The connections between the Lanczos procedure and the conjugate gradient method lead Lanczos to propose in [60] a method to solve linear systems with symmetric positive definite matrices which turned out to be mathematically equivalent to the conjugate gradient method.

The method consists in iteratively constructing an orthonormal basis of the Krylov subspace. In this process, the symmetric positive definite matrix A is reduced to a tridiagonal form, which is used to obtained approximate eigenpairs of A. In this regard, the Lanczos method is a particular case of the Arnoldi's method [4] when applied to symmetric positive definite matrices. In the general case, the matrix A is rather reduced to an upper Hessenberg form.

Let us present how to recover the Lanczos relation from the preconditioned conjugate gradient method (Algorithm 2.2). Assume that l iterations of the preconditioned conjugate gradient method have been performed and let us define the matrix

$$Z_l = \begin{bmatrix} \frac{z_0}{\sqrt{\rho_0}} & \cdots & \frac{z_l}{\sqrt{\rho_l}} \end{bmatrix} \in \mathbb{R}^{n\times(l+1)},$$

where $z_i \in \mathbb{R}^n$ and $\rho_i \in \mathbb{R}$ are defined in Algorithm 2.2. Then Z_l is M^{-1}-orthogonal and its columns form a basis of $\mathcal{K}_{i+1}$. It can then be shown [42, Section 10.2.5] that

$$MAZ_l = Z_l T_l - \frac{\sqrt{\beta_l}}{\alpha_l} z_{l+1} e_l^\mathsf{T}, \tag{2.11}$$

where e_l is the l-th vector of the canonical basis of $\mathbb{R}^l$, and $T_l \in \mathbb{R}^{(l+1)\times(l+1)}$ a symmetric tridiagonal matrix whose coefficients can be computed using the sequences of α_i, β_i and ρ_i for $i \leq l$ (see [75, Relation 6.103]). From the M^{-1}-orthogonality of Z_l, we obtain in particular that

$$Z_l^\mathsf{T} A Z_l = T_l. \tag{2.12}$$

Let us consider the eigenvalue decomposition $T_l = U\Lambda U^\mathsf{T}$ with $U = [u_1 \ \cdots \ u_{l+1}] \in \mathbb{R}^{(l+1)\times(l+1)}$ orthogonal and $\Lambda = \mathrm{diag}(\mu_1, \ldots, \mu_{l+1})$. Then we define the matrix

$$V_l = Z_l U = \begin{bmatrix} v_0 & \cdots & v_l \end{bmatrix}.$$

The vectors $v_0, \ldots, v_l \in \mathbb{R}^n$ and scalars $\mu_1, \ldots, \mu_{l+1}$ are respectively referred to as the Ritz vectors and Ritz values [65, p.8]. Plugging the definition of the Ritz pairs in (2.11) and (2.12) then yields

$$MAV_l = V_l\Lambda - \frac{\sqrt{\beta_l}}{\alpha_l} z_{l+1} e_l^\mathsf{T} U \quad \text{and} \quad V_l^\mathsf{T} A V_l = \Lambda. \tag{2.13}$$

To compute the Ritz pairs from an application of the preconditioned conjugate gradient method, it is needed to store the sequence of residuals z_i, and the sequences of scalars ρ_i, α_i and β_i. This procedure thus slightly increases the memory requirements of the method, but it allows us to obtain approximate eigeninformation at a reasonable cost.

2.3 Randomized numerical linear algebra

In this section, we introduce the minimal fundamental material related to randomized numerical linear algebra. Therefore, although randomized methods have been developed to address a large class of numerical linear algebra problems, we focus here on introducing the methods addressing the approximation of singular value/eigenvalue decompositions. We refer the reader to [63] for a recent review of the existing methods.

2.3.1 The Randomized Singular Value Decomposition

Let us begin with the Randomized Singular Value Decomposition (RSVD) algorithm introduced in [94] and popularized in [53]. Let $A \in \mathbb{R}^{n\times m}$ be any rectangular matrix satisfying $m \leq n$. The singular value decomposition of A reads

$$A = U\Sigma V^\mathsf{T},$$

with $U \in \mathbb{R}^{n\times n}$, $V \in \mathbb{R}^{m\times m}$ orthogonal matrices and $\Sigma \in \mathbb{R}^{n\times m}$ a diagonal matrix. For a given $k \leq m$, one can partition the SVD of A as follows,

$$A = \begin{bmatrix} U_k & \underline{U}_k \end{bmatrix} \begin{bmatrix} \Sigma_k & \\ & \underline{\Sigma}_k \end{bmatrix} \begin{bmatrix} V_k^\mathsf{T} \\ \underline{V}_k^\mathsf{T} \end{bmatrix}$$

with $U_k \in \mathbb{R}^{n\times k}$, $\underline{U}_k \in \mathbb{R}^{n\times(n-k)}$, $V_k \in \mathbb{R}^{m\times k}$, $\underline{V}_k \in \mathbb{R}^{m\times(m-k)}$, $\Sigma_k \in \mathbb{R}^{k\times k}$ and $\underline{\Sigma}_k \in \mathbb{R}^{(n-k)\times(m-k)}$. We also set $A_k = U_k\Sigma_k V_k^\mathsf{T}$ and $\underline{A}_k = \underline{U}_k\underline{\Sigma}_k\underline{V}_k^\mathsf{T}$ so that $A = A_k + \underline{A}_k$.

The randomized singular value decomposition method has been proposed to address the low rank approximation problem. For a given value $k \leq \text{rank}(A)$, the low rank approximation problem consists in finding a rank k matrix $\widetilde{A}_k$ such that A is close to $\widetilde{A}_k$ in a given norm. In the randomized linear algebra literature, this problem is generally addressed in the spectral or Frobenius norms. The Eckart-Young theorem [30] states that the optimal rank-k approximation of A is given by $A_k = U_k U_k^\mathsf{T} A = \pi(U_k)A$.

The objective of the randomized singular value decomposition is to compute an approximation of U_k using a randomized procedure. Such methods are generally referred to as range-finder methods, since U_k contains the dominant left singular vectors and can thus be interpreted as an approximate range of A. The most elementary procedure to get such an approximation is the subspace iteration scheme. The idea is to consider approximations of U_k of the form $Y = (AA^\mathsf{T})^q A\Omega$ with $\Omega \in \mathbb{R}^{m\times k}$ a random matrix. Here we notice that in [63, Algorithm 9] the randomized subspace iteration rather considers $Y = (AA^\mathsf{T})^q\Omega$ with $\Omega \in \mathbb{R}^{n\times k}$, but there is no fundamental difference in the algorithm structure. This procedure, referred to as the randomized subspace iteration is given in Algorithm 2.3.

Algorithm 2.3: Randomized subspace iteration adapted from [63, Algorithm 9]

Input: Matrix $A \in \mathbb{R}^{n\times n}$, target rank k, number of subspace iterations q.

1 Draw a random matrix $\Omega \in \mathbb{R}^{m\times k}$
2 Perform the thin QR factorization $A\Omega = QR$ and set $Z = Q$
3 **for** $i = 1, \ldots, q$ **do**
4 Compute $Z = (AA^\mathsf{T})Z$
5 Perform the thin QR factorization $Z = QR$ and set $Z = Q$
6 **end**

Output: Orthogonal matrix $Z \in \mathbb{R}^{n\times k}$.

Algorithm 2.3 produces an approximation Z of U_k, from which the low rank approximation $\pi(Z)A = ZZ^\mathsf{T}A$ for A can be computed. The method to construct the low rank approximation of A given an approximate range is called the randomized singular value decomposition and is presented in Algorithm 2.4. More sophisticated approaches than the randomized subspace iteration method have been proposed to approximate the range of A, based on the use of Krylov subspaces and referred to as the Krylov range finder [63, Algorithm 10]. However, the improvements brought by these methods appear when several applications of A are affordable. This is out of the scope in this thesis, since applying A in the targeted applications in data assimilation is very expensive.

Practically, several points must be clarified. The random matrix Ω is generally drawn from a standard Gaussian distribution. However, as highlighted in [70], the performance of the algorithm is fairly insensitive to the choice of the distribution. In this regard, alternative distributions can be preferred to improve the computational efficiency (see [63, Section 9]). Then, Algorithms 2.3 and 2.4 use the matrix A only via matrix multiplications to a block of p vectors. This routine is highly optimized when A is a matrix, but can easily be parallelized also when A is a black-box linear operator. Finally, we observe that the matrix Z in Algorithms 2.3 tends to be highly ill-conditioned since the columns will mostly be aligned in the direction of the dominant eigenmodes of AA^T. Consequently, the orthogonalization step must be performed using a numerically stable algorithm [43, Chapter 5]. In case of very large scale problems, it is also worth mentioning the existence of communication-avoiding parallel methods for performing QR factorization such as [27, 82, 96].

Algorithm 2.4: Randomized Singular Value Decomposition [63, Algorithm 8]

Input: Matrix $A \in \mathbb{R}^{n\times m}$, orthogonal matrix $Z \in \mathbb{R}^{n\times p}$ obtained for instance using Algorithm 2.3, target rank $k \leq p$.

1 Compute $C = Z^{\mathsf{T}}A \in \mathbb{R}^{p\times m}$

2 Perform the SVD $C = \widetilde{U}\Sigma V^{\mathsf{T}}$ with $\widetilde{U} \in \mathbb{R}^{p\times p}$ and $V \in \mathbb{R}^{m\times m}$ two orthogonal matrices and $\Sigma \in \mathbb{R}^{p\times m}$ a diagonal matrix containing the singular values in decreasing order

3 Remove the last $p-k$ columns of $\widetilde{U}$ and of Σ

4 Remove the last $m-k$ rows of V^{T} and of Σ

5 Compute $U = Z\widetilde{U} \in \mathbb{R}^{n\times k}$

Output: Orthogonal matrices $U \in \mathbb{R}^{n\times k}$, $V \in \mathbb{R}^{m\times k}$ and diagonal matrix $\Sigma \in \mathbb{R}^{k\times k}$ such that $A \approx U\Sigma V^{\mathsf{T}}$.

2.3.2 Theoretical analysis of the Randomized Singular Value Decomposition

The theoretical analysis of the randomized singular value decomposition can be conducted in two different ways. The first approach consists in looking at the corresponding low rank approximation error, that is

$$\|A - ZZ^{\mathsf{T}}A\|_{2,F} = \|[I_n - \pi(Z)]A\|_{2,F}. \tag{2.14}$$

In [53, Theorem 10.5 and 10.6], Halko, Martinsson and Tropp have proposed popular error bounds in expectation and in probability for the low-rank approximation error (2.14). The bounds in probability are derived from the bounds in expectation, so we focus on the bounds in expectation. Let $Z = A\Omega$ with $\Omega \in \mathbb{R}^{m\times p}$ a standard Gaussian matrix, then Theorem 10.5 [53] states that for all $k \leq p-2$ one has

$$\mathbb{E}\left[\, \|[I_n - \pi(Z)]A\|_F \,\right] \leq \left(1 + \frac{k}{p-k-1}\right)^{\frac{1}{2}} \|\Sigma_k\|_F.$$

Similarly, let $Z = (AA^{\mathsf{T}})^q A\Omega$ with $\Omega \in \mathbb{R}^{m\times p}$ a standard Gaussian matrix, then for all $k \leq p-2$ one has from [53, Corollary 10.10]

$$\mathbb{E}\left[\, \|[I_n - \pi(Z)]A\|_2 \,\right] \leq \left[\left(1 + \sqrt{\frac{k}{p-k-1}}\right)^{\frac{1}{2}} \|\Sigma_k^{2q+1}\|_2 + \frac{e\sqrt{p}}{p-k}\|\Sigma_k^{2q+1}\|_F\right]^{\frac{1}{2q+1}},$$

where $e = \exp(1)$. The second approach consists in directly measuring the sine of the principal canonical angles $\theta_1, \ldots, \theta_k$ between $\mathcal{R}(Z)$ and $\mathcal{R}(U_k)$. This has notably been done in [49] and [77]. In [77, Theorem 6], the following result has been obtained

$$\mathbb{E}\left[\sin(\theta_j)\right] \leq \frac{\gamma_j^{2q+1} c_e}{\sqrt{1 + \gamma_j^{4q+2} c_e^2}}, \quad 1 \leq j \leq k$$

with $\gamma_j = \sigma_{k+1}/\sigma_j$ and the constant c_e defined as

$$c_e = \sqrt{\frac{k}{p-k-1}} + \frac{e\sqrt{p(n-k)}}{p-k}.$$

The main message behind those bounds is that the accuracy of the randomized singular value decomposition, regardless of the quantity of interest, is monitored by: (i) the number of random samples p; (ii) the number of random subspace iterations q; (iii) the singular value distribution. These considerations lead to the heuristic that randomized methods for the low rank approximation problems perform well on matrices whose singular values decay rapidly.

2.3.3 The Nyström method

In the particular case of symmetric positive definite matrices $A \in \mathbb{R}^{n\times n}$, the notion of singular vectors and eigenvectors merges. Indeed, in this case the eigenvectors of A are equal to both its the left and right singular vectors. In this particular case, the matrix U_k has both orthonormal and A-conjugate columns. Consequently, we observe that

$$\pi(U_k) = \pi_A(U_k) = \pi_A(U_k)^{\mathsf{T}}.$$

This suggests that in this particular case, a rank k approximation of A can rather be obtained considering either $\pi_A(Z)A$ or $\pi_A(Z)^{\mathsf{T}}A$. It turns out that only the latter is symmetric, which is desirable since A is. Consequently, for a given approximation Z of U_k, the Nyström method consists in constructing rather the low rank approximation of the form

$$A_{\text{Nys.}} = \pi_A(Z)^{\mathsf{T}}A = AZ(Z^{\mathsf{T}}AZ)^{-1}Z^{\mathsf{T}}A. \tag{2.15}$$

Interestingly, although A appears several times in (2.15), $A_{\text{Nys.}}$ can be formed with only one application of A to Z, and is therefore no more expensive than the randomized singular value decomposition. Also, $A_{\text{Nys.}}$ can be well defined even when A is semidefinite. In this case, one must ensure that $\mathcal{R}(Z) \cap \mathcal{N}(A) = \{0\}$.

Error bounds for $\|A - A_{\text{Nys.}}\|_{2,F}$ can be obtained by observing that

$$A_{\text{Nys.}} = A^{\frac{1}{2}}\pi(A^{\frac{1}{2}}Z)A^{\frac{1}{2}}.$$

Then, one has

$$A - A_{\text{Nys.}} = A^{\frac{1}{2}}\left[I_n - \pi(A^{\frac{1}{2}}Z)\right]A^{\frac{1}{2}},$$

which in turn yields

$$\begin{aligned}\|A - A_{\text{Nys.}}\|_{2,F} &= \|A^{\frac{1}{2}}\left[I_n - \pi(A^{\frac{1}{2}}Z)\right]A^{\frac{1}{2}}\|_{2,F} \\ &= \|\left[I_n - \pi(A^{\frac{1}{2}}Z)\right]A^{\frac{1}{2}}\|_{2,F}^2.\end{aligned}$$

From this identity, it is possible to derive an average-case analysis for the Nyström method using the average-case analysis of the randomized singular value decomposition. The obtained error bounds are stated in [86, Theorem 4.1] and in the particular case of real matrices one has

$$\mathbb{E}\left[\,\|A - A_{\text{Nys.}}\|_*\,\right] \leq \left(1 + \frac{k}{p-k-1}\right)\|\Sigma_k\|_*,$$

and

$$\mathbb{E}\left[\,\|A - A_{\text{Nys.}}\|_2\,\right] \leq \|\Sigma_k\|_2 + \frac{k}{p-k-1}\|\Sigma_k\|_*,$$

where $\|\cdot\|_*$ is the nuclear norm [43, Section 2.3.2]. The Nyström method for symmetric positive definite matrices is given in Algorithm 2.5. We point out here that a variant for symmetric positive semidefinite matrices has been proposed (see [63, Algorithm 16]). This variant provides a numerically stable algorithm.

Algorithm 2.5: Nyström method [53, Algorithm 5.5]

Input: Symmetric positive definite matrix $A \in \mathbb{R}^{n\times m}$, orthogonal matrix $Z \in \mathbb{R}^{n\times p}$ obtained for instance using Algorithm 2.3, target rank $k \leq p$.

1 Compute $Y = AZ \in \mathbb{R}^{n\times p}$
2 Compute $C = Z^{\mathsf{T}}Y \in \mathbb{R}^{p\times p}$
3 Perform the Cholesky factorization $C = LL^{\mathsf{T}}$
4 Compute $B = YL^{-1} \in \mathbb{R}^{n\times p}$
5 Perform the thin SVD $B = U\Sigma V^{\mathsf{T}}$ with $\widetilde{U} \in \mathbb{R}^{n\times p}$ and $V \in \mathbb{R}^{p\times p}$ two orthogonal matrices and $\Sigma \in \mathbb{R}^{p\times p}$ a diagonal matrix containing the singular values of B in decreasing order
6 Remove the last $p-k$ columns of U and Σ
7 Remove the last $p-k$ rows of Σ
8 Set $\Lambda = \Sigma^2$

Output: Orthogonal matrix $U \in \mathbb{R}^{n\times k}$ and diagonal matrix $\Lambda \in \mathbb{R}^{k\times k}$ such that $A \approx U\Lambda U^{\mathsf{T}}$.

2.4 The weighted nonlinear least-squares problem

Nonlinear least-squares problems naturally arise when observations of a dynamical system are used to estimate its true under-lying state. This is a common problem in data fitting and optimal control. The idea is to find the parameters of a model that best fit the observations made on the physical system described by the model.

2.4.1 Presentation

Let us assume that a given dynamical system is described by a set $x_1, \ldots, x_n$ of parameters, that we gather in the state vector $x = [x_1, \ldots, x_n] \in \mathbb{R}^n$. Let us denote $y_1, \ldots, y_m$ a set of observations of this dynamical system, obtained for the different time steps $t_1, \ldots, t_m$. We consider that a prediction model $\mathcal{H} : \mathbb{R}^n \mapsto \mathbb{R}^m$ is available, allowing to map the state vector x to the observations, that is

$$y = \mathcal{H}(x, t).$$

The prediction model is generally nonlinear. The model fitting problem then consists in simultaneously minimizing each individual error

$$d_i(x) = y_i - \mathcal{H}(x, t_i),$$

for all $1 \leq i \leq m$. In this regard, we define $d(x) = [d_1(x), \ldots, d_m(x)] \in \mathbb{R}^m$, the residual vector gathering all the residuals. Fitting the observations to the model parameters reduces to solving the unconstrained optimization problem

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{2}\|d(x)\|_2^2.$$

In concrete applications, the observations can be noisy, yielding an uncertainty on the residual vector $d(x)$. To model the noise, we consider a symmetric positive definite matrix $\Gamma_o \in \mathbb{R}^{m\times m}$ playing the role of the observation error covariance matrix. Consequently, the optimization problem should be transformed into

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{2}\|d(x)\|_{\Gamma_o^{-1}}^2,$$

Here we focus on the under-determined setting, that is the number of observations m is smaller than the number of parameters n. In turn, there is no longer uniqueness of the solution. To tackle the non-uniqueness, it is common to add a regularization term, typically a Tikhonov regularization. Let $x_c \in \mathbb{R}^n$ be the center vector. This vector is problem-dependent and generally contains a priori information on the true state parameters. Again, this a priori can be noisy, which motivates the introduction of the center vector covariance matrix $\Gamma_b \in \mathbb{R}^{n\times n}$, assumed to be symmetric positive definite. The resulting regularized optimization problem takes the form

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{2}\|d(x)\|^2_{\Gamma_o^{-1}} + \frac{1}{2}\|x - x_c\|^2_{\Gamma_b^{-1}}. \tag{2.16}$$

Remark 2.7. Here, we have deviated from the notation generally adopted in variational data assimilation, where the observation error covariance matrix is usually denoted by R instead of Γ_o and the center vector error covariance matrix B instead of Γ_b. This is a deliberate choice to avoid conflicts in notation, especially in Chapter 4.

2.4.2 The Gauss-Newton method

Solving (2.16) can be done with various approaches, and we refer the interested reader to [68, Section 10] for a detailed overview. Since the targeted applications are in operational data assimilation, we are going to focus on the (truncated) Gauss-Newton method which has proven to be efficient for such problems [45]. It is an iterative method where the increment s_j at the j-th iteration is obtained from the minimization of a convex quadratic subproblem. Let us denote x_j the current iterate of the Gauss-Newton. A linearization around x_j yields

$$d(x_j + s) \approx d_j - H_j s,$$

with $d_j = d(x_j) \in \mathbb{R}^m$ and $H_j \in \mathbb{R}^{m\times n}$ denotes the Jacobian of $\mathcal{H}$ at x_j. The obtained subproblem to be solved is thus

$$\min_{x\in\mathbb{R}^n} q_j(s) = \frac{1}{2}\|H_j s - d_j\|^2_{\Gamma_o^{-1}} + \frac{1}{2}\|s + x_j - x_c\|^2_{\Gamma_b^{-1}}, \tag{2.17}$$

whose solution $s_j \in \mathbb{R}^n$ is computed from the normal equations

$$A_j s_j = b_j. \tag{2.18}$$

where $A_j = \Gamma_b^{-1} + H_j^\mathsf{T}\Gamma_o^{-1}H_j$ and $b_j = \Gamma_b^{-1}(x_c - x_j) + H_j^\mathsf{T}\Gamma_o^{-1}d_j$. The next iterate is then obtained via $x_{j+1} = x_j + s_j$. Applying the Gauss-Newton then yields to the solution of a sequence of symmetric positive definite linear systems, whose dimensions can be extremely large in operational applications where the dynamical system is complex.

2.4.3 Solving the linearized subproblem with the preconditioned conjugate gradient method

The solution of (2.18) is generally obtained using iterative methods since the concrete problems are of very large scale. Nevertheless, alternative approaches have been proposed such as low-rank approximation-based methods [35]. Since the system matrix is symmetric positive definite, the solution is usually obtained using the preconditioned conjugate gradient method (Algorithm 2.2). To prevent any ambiguity, the iterations of the Gauss-Newton method will be referred to as the *outer-loop* iterations, while the conjugate gradient method iterations will be referred to as the *inner-loop* iterations. In practical applications, few iterations of the preconditioned conjugate gradient method are performed because first, the problems are so large that the computational resources consumption rapidly becomes important and second, the solution of (2.18) need not

be computed precisely for the Gauss-Newton method to converge. Consequently, an efficient preconditioner for A_j should result in fast convergence of the preconditioned conjugate gradient method in the first iterations.

The structure of (2.18) suggests that a natural candidate for a preconditioner is Γ_b. Indeed, $\Gamma_b A_j = I_n + \Gamma_b H_j^\mathsf{T} \Gamma_o^{-1} H_j$ is a rank m update of the identity, meaning that the eigenvalue distribution of $\Gamma_b A_j$ enjoys a nice cluster of $n - m$ eigenvalues at 1, the remaining eigenvalues being larger than one. This is particularly appealing whenever $m \ll n$, which is frequent in data assimilation applications where the number of model parameters largely exceeds the number of observations.

Preconditioning with deterministic limited memory preconditioners

Given the particular eigenvalue distribution of $\Gamma_b A_j$, using a limited memory preconditioner on top of Γ_b seems natural to try capturing the large eigenvalues left out by the action of Γ_b. Also, such preconditioners are well suited for sequences of linear systems since it allows to update the preconditioner along the sequence. However, alternatives based on updating (incomplete) factorizations have been proposed [11, 12]. Particular forms of the limited memory preconditioner have been derived depending on the content of S.

Spectral LMP. Let us define $S = [s_1 \ \cdots \ s_k] \in \mathbb{R}^{n \times k}$ and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ with $s_1, \ldots, s_k \in \mathbb{R}^n$ eigenvectors of $\Gamma_b A_j$ associated to the eigenvalues $\lambda_1, \ldots, \lambda_k$. Plugging S into the expression of the LMP (2.10) yields the so-called spectral LMP [47, Section 4.2.1] whose expression reads

$$P_{\mathrm{sp}} = \Gamma_b + V(\Lambda^{-1} - I_k)V^\mathsf{T}. \tag{2.19}$$

Ritz LMP. In practice, the computation of exact eigenpairs is generally out of reach. However, we have already evoked in Section 2.2.5 that approximate eigenpairs can be obtained almost for free. Let $V \in \mathbb{R}^{n \times k}$ denote the matrix whose columns are the Ritz vectors and $\Lambda \in \mathbb{R}^{k \times k}$ the diagonal matrix containing the corresponding Ritz values. Plugging the relations satisfied by V and Λ given in (2.13) in the expression of the LMP yields an alternative form for the LMP, referred to as the Ritz LMP [47, Section 4.2.2] whose expression reads

$$P_{\mathrm{Ritz}} = \Gamma_b + V(\Lambda^{-1} - I_k)V^\mathsf{T} + z_{k+1}\omega^\mathsf{T}V + V\omega z_{k+1}^\mathsf{T} - V\omega\omega^\mathsf{T}V, \tag{2.20}$$

where

$$\omega = \frac{\sqrt{\beta_k}}{\alpha_k}\Lambda^{-1}Ue_k \in \mathbb{R}^k.$$

The main disadvantage of this variant is that, thus constructed, P_{Ritz} is defined using approximate eigenpairs of A_j, but is intended to be used as a preconditioner for the next linear system involving A_{j+1}. In practice, this is not problematic when the systems are slowly varying, that is $A_{j+1} \approx A_j$, and this strategy has proven to perform well and is commonly used in operational data assimilation codes [89].

2.4.4 Solving the linearized subproblem with randomized methods

Let us now present the randomized methods that have been proposed in the literature to address variational data assimilation problems. The first one, called the Ritzit method and proposed in [24] consists in forming a spectral LMP as in (2.19) replacing true eigenpairs by approximations computed using a randomized procedure. The second one is the Randomized Incremental Optimal Technique introduced in [17]. Here, low rank approximations using the Nyström method are used to directly approximate the solution of (2.18).

The Ritzit method

The construction of limited memory preconditioners based on randomized methods have been proposed recently in [24]. The idea is to construct P_{sp} as in (2.19) with approximate eigenpairs instead of exact eigenpairs, that are computed using randomized methods. Doing so, all the properties of the LMP are theoretically lost. However, if the approximate eigenpairs are accurate enough, one may expect the resulting preconditioner to behave similarly as the exact spectral limited memory preconditioner. This fact has been observed in [89] and theoretically quantified in [47, Theorem 4.5] when the approximate eigenpairs are Ritz pairs. With this approach, the resulting LMP is constructed with eigeninformation of the current linear system, and is therefore no longer dependent on the fact that $A_{j+1} \approx A_j$. Also, the number of approximate eigenvectors to compute is defined by the user, while for the Ritz LMP it is dependent on the number of preconditioned conjugate gradient method iterations performed.

The different randomized approaches considered in [24] all rely on randomized methods for symmetric positive definite matrices. Consequently, this imposes to have access to the symmetric preconditioned matrix $\Gamma_b^{1/2} A_j \Gamma_b^{1/2}$, and thus implies that a square root $\Gamma_b^{1/2}$ of Γ_b is available. This is a reasonable assumption since such factorizations can indeed be available in certain data assimilation problems.

The first approach consists in using the Nyström method (Algorithm 2.5). The matrix Z in Algorithm 2.5 is obtained by performing an orthogonalization of the matrix $\Gamma_b^{1/2} A_j \Gamma_b^{1/2} \Omega$, with $\Omega \in \mathbb{R}^{n \times p}$ a standard Gaussian matrix. Constructing Z requires one application of $\Gamma_b^{1/2} A_j \Gamma_b^{1/2}$, and performing the Nyström method requires another application of $\Gamma_b^{1/2} A_j \Gamma_b^{1/2}$. If we denote $U \in \mathbb{R}^{n \times k}$ and $\Lambda \in \mathbb{R}^{k \times k}$ the output of Algorithm 2.5 obtained with this configuration, then the authors consider the following preconditioner:

$$P_{\text{Nys.}} = I_n + U(\Lambda^{-1} - I_k)U^{\mathsf{T}}.$$

In addition to the Nyström method, the authors in [24] have also proposed another randomized approach called Ritzit, given in Algorithm 2.6. The Ritzit method is rather based on the approximation of eigenpairs of the symmetric positive definite matrix A^2 instead of A. This approach appeared to perform well on a four-dimensional variational data assimilation problem, and has the advantage of requiring only one application of $\Gamma_b^{1/2} A_j \Gamma_b^{1/2}$.

Algorithm 2.6: Ritzit method from [24, Algorithm 5]

Input: Symmetric positive definite matrix $A \in \mathbb{R}^{n \times n}$, number of random samples p, target rank $k \leq p$.

1 Draw a standard Gaussian matrix $G \in \mathbb{R}^{n \times p}$
2 Perform the thin QR factorization $G = Q_G R_G$
3 Compute $Y = AQ_G \in \mathbb{R}^{n \times p}$
4 Perform the thin QR factorization $Y = Q_Y R_Y$
5 Form the matrix $K = R_G R_G^{\mathsf{T}} \in \mathbb{R}^{p \times p}$
6 Perform the eigenvalue decomposition $K = W\Lambda^2 W^{\mathsf{T}}$ with $W \in \mathbb{R}^{p \times p}$ orthogonal and $\Lambda \in \mathbb{R}^{p \times p}$ diagonal containing the eigenvalues in decreasing order
7 Remove the last $p - k$ columns of W and of Λ
8 Remove the last $p - k$ rows of Λ
9 Form $U = Q_Y W \in \mathbb{R}^{n \times k}$

Output: Orthogonal matrix $U \in \mathbb{R}^{n \times k}$ and diagonal matrix $\Lambda \in \mathbb{R}^{k \times k}$ such that $A \approx U\Lambda U^{\mathsf{T}}$.

The Randomized Incremental Optimal Technique method

Another randomized approach for solving the sequence of linear systems (2.18) has been proposed in [17]. The Randomized Incremental Optimal Technique (RIOT) provides a method where the iterations of the conjugate gradient method are replaced with a fully parallel randomized approach. The method is dedicated to solve symmetric positive linear systems. Let us assume that M_j is a first-level preconditioner for the j-th linear system (2.18) used as a split preconditioner, then the algorithm addresses the solution of

$$M_j^{\frac{1}{2}} A_j M_j^{\frac{1}{2}} y = M_j^{\frac{1}{2}} b_j,$$

with $x_j = M_j^{\frac{1}{2}} y$. Thus, one can apply the Nyström method to obtain an approximate eigenvalue decomposition as

$$M_j^{\frac{1}{2}} A_j M_j^{\frac{1}{2}} \approx U \Lambda U^{\mathsf{T}},$$

where $U \in \mathbb{R}^{n \times k}$ is orthogonal and $\Lambda \in \mathbb{R}^{k \times k}$ is diagonal. This allows us to compute an approximate solution of (2.18) as

$$x \approx U \Lambda^{-1} U^{\mathsf{T}} M_j^{\frac{1}{2}} b_j.$$

Another way to obtain an approximate solution relies on the spectral limited memory preconditioner as in (2.19). This preconditioner can indeed be interpreted as an approximate inverse of A, that is $P_{\text{sp}} \approx A^{-1}$. Hence, another approximate solution can be computed as

$$x \approx \left(I_n + U \left[\Lambda^{-1} - I_k \right] U^{\mathsf{T}} \right) M_j^{\frac{1}{2}} b_j.$$

The authors in [17] have proposed a criterion to select between the two methods depending on the obtained approximate eigenvalues. The overall method is given in Algorithm 2.7.

In practice, Γ_b is used to precondition the first linear system, that is $M_1 = \Gamma_b$. Then the preconditioners for the next linear systems are computed as

$$M_j = M_{j-1} + U \left(I_k + \Lambda \right)^{-1} U^{\mathsf{T}}, \quad \forall\ j > 1.$$

2.5 Conclusions

In this chapter, we have presented the essential material required for the rest of the manuscript. In particular, we have recalled the main algorithms of randomized numerical linear algebra that are at the core of the randomized methods proposed recently in variational data assimilation. These methods rely on the proven ability of randomized methods to efficiently capture dominant eigenmodes and use the approximate eigeninformation either to precondition the conjugate gradient method (Ritzit method), or to construct low rank approximations to directly get approximate descent directions (RIOT method). However, both methods are essentially limited to the cases where the problem can be formulated in terms of symmetric positive definite operators, which may not be always possible.

In this thesis, we explore aspects that are at the interface of randomized numerical linear algebra and preconditioning. Our objective is to present randomized methods to construct limited memory preconditioners adapted to formulations of variational data assimilation where the operators are no longer symmetric positive definite with respect to the Euclidean inner product. Although they ended up being fairly general, the developments proposed in this thesis were initially motivated by needs and constraints of variational data assimilation. Consequently, we will mostly propose numerical illustrations of the proposed algorithms in such a setting. Similarly, the algorithmic discussions will be systematically oriented towards operational aspects and computational complexity.

Algorithm 2.7: Cycle of Randomized Incremental Optimal Technique

Input: Factorization $\Gamma_b = L_b L_b^\mathsf{T}$, boolean `precond` and `rotation`, number of random samples k and number of Gauss-Newton steps J

Set $x_0 = x_c$, $v_0 = 0$ and $P'_0 = I_n$.
for $j = 1, \dots, J$ **do**
  Integrate and store trajectory $\mathcal{H}(x_{j-1})$
  Compute and store the innovation vector $d_{j-1} = \mathcal{H}(x_{j-1}) - y \in \mathbb{R}^m$
  Define $\widehat{A}_0 = L_b^\mathsf{T} H_{j-1}^\mathsf{T} \Gamma_o^{-1} H_{j-1} L_b \in \mathbb{R}^{n \times n}$ with H_{j-1} the linearization of $\mathcal{H}$ around x_{j-1}
  if *precond* **then**
    if $k \geq 2$ **then**
      $P'_{j-1} = I_n + \widetilde{V}_{j-1} \left(\widetilde{\Lambda}_{j-1}^{-\frac{1}{2}} - I_k \right) \widetilde{V}_{j-1}^\mathsf{T}$
      $P'^{-1}_{j-1} = I_n + \widetilde{V}_{j-1} \left(\widetilde{\Lambda}_{j-1}^{\frac{1}{2}} - I \right) \widetilde{V}_{j-1}^\mathsf{T}$
    $P_k = P'_0 \dots P'_{j-1}$
    $P_k^{-1} = P'^{-1}_{j-1} \dots P'^{-1}_0$
    $\widehat{A}_k = P_k^\mathsf{T} P_k - I_n + P_k^\mathsf{T} \widehat{A}_0 P_k$
  else
    $\widehat{A}_k = \widehat{A}_0$
    $P_k = P'_0$
  end
  Draw $\Omega_k \in \mathbb{R}^{n \times m} \sim \mathcal{N}(0, I_n)$
  for $i = 1, \dots, m$ **do**
    $\omega_i = \Omega_k(:, i)$
    if $k \geq 2$ *and* *precond* *and* *rotation* **then**
      Compute the SVD $[P_1 \widetilde{V}_1, \dots, P_k \widetilde{V}_k] = U_k D Z_k^\mathsf{T}$
      $\omega_i = P_k^{-1} (I_n - U_k U_k^\mathsf{T}) \omega_i$
    end
  end
  for $i = 0, \dots, m$ **do**
    if $i = 0$ **then**
      Compute the gradient $b_j = P_k^\mathsf{T} L_b^\mathsf{T} H_{j-1}^\mathsf{T} \Gamma_o^{-1} d_{j-1}$
    else
      $y_i = \widehat{A}_k \omega_i$
    end
  end
  Form $Y = [y_1, \dots, y_m]$
  Perform the thin QR factorization $Y = QR$
  Solve for K in $K Q^\mathsf{T} \Omega = Q^\mathsf{T} Y$
  Perform the eigenvalue decomposition $K = Z \Lambda Z^\mathsf{T}$ with $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_k)$
  Compute $\widetilde{V}_k = QZ$
  $\widetilde{\Lambda}_k = \Lambda + I_k$
  if $\lambda_k > 1$ **then**
    $s_j = -\left(\sum_{i=1}^m \frac{1}{1+\lambda_i} u_i u_i^\mathsf{T} \right) (b_j + v_{j-1})$
  else
    $s_j = -\left(I_n - \sum_{i=1}^m \frac{\lambda_i}{1+\lambda_i} u_i u_i^\mathsf{T} \right) (b + v_{j-1})$
  end
  $v_j = v_{j-1} + s_j$
  $x_j = x_{j-1} + L_b P_k s_j$
end

CHAPTER 3

A general error analysis for randomized low-rank approximation methods

Contents

This chapter is based on a preprint available at https://arxiv.org/abs/2206.08793v2. Complements have been added and are spotted by a red line in the left margin as:

Example of complementary material.

Abstract

In this chapter, we are interested in generalizing the theoretical analysis of the randomized low rank approximation error. We begin with deriving deterministic error bounds in spectral and Frobenius norms, from which we obtain stochastic bounds both in expectation and in probability. The obtained error bounds generalize the existing bounds to Gaussian matrices with non-zero mean term and general covariance matrix. An application of our general analysis to the Randomized Singular Value Decomposition shows that our bounds improve the reference error bounds from Halko, Martinsson and Tropp (2011). Finally, we propose a numerical illustration on an instructional test problem aimed at demonstrating the tightness of the proposed error bounds. This generalization allows us to extend the theoretical analysis to a larger class of randomized methods. In particular, it will be at the core of the average-case analysis of the algorithms introduced in the next chapter.

3.1 Introduction

Low-rank approximation of large-scale matrices is a key ingredient in numerous applications in data analysis and scientific computing. These applications include principal component analysis [74], data compression [61] and approximation algorithms for partial differential and integral equations [52], to name a few. Namely, let $A \in \mathbb{R}^{n\times m}$ be a rectangular matrix satisfying $n \geq m$ and $Z \in \mathbb{R}^{n\times p}$ any full column rank random matrix (with $p \leq \operatorname{rank}(A) \leq m$). If $\pi(Z)$ denotes the orthogonal projection onto the range of Z and $I_n \in \mathbb{R}^{n\times n}$ the identity matrix of order n, we aim at analyzing the general quantity of interest

$$\|[I_n - \pi(Z)]A\|_{2,F}, \tag{3.1}$$

where $\|.\|_{2,F}$ represents a shortcut for either the spectral norm ($\|.\|_2$) or the Frobenius norm ($\|.\|_F$), respectively.

In our analysis, we will consider the standard Singular Value Decomposition (SVD) of A, i.e., $A = U\Sigma V^\mathsf{T}$, where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{m\times m}$ are orthogonal matrices containing the left and right singular vectors of A respectively and $\Sigma \in \mathbb{R}^{n\times m}$ is a diagonal matrix containing the singular values of A denoted as $\sigma_1, \ldots, \sigma_m$ (sorted in decreasing order, i.e. $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m \geq 0$). For a target rank $k \in \{1, \ldots, \operatorname{rank}(A)\}$, which is assumed to be much smaller than n, a central decomposition appearing in the error analysis is then given by

$$A = \begin{bmatrix} U_k & \underline{U}_k \end{bmatrix} \begin{bmatrix} \Sigma_k & \\ & \underline{\Sigma}_k \end{bmatrix} \begin{bmatrix} V_k^\mathsf{T} \\ \underline{V}_k^\mathsf{T} \end{bmatrix}, \tag{3.2}$$

with $U_k \in \mathbb{R}^{n\times k}$, $\underline{U}_k \in \mathbb{R}^{n\times(n-k)}$, $V_k \in \mathbb{R}^{m\times k}$, $\underline{V}_k \in \mathbb{R}^{m\times(m-k)}$, $\Sigma_k \in \mathbb{R}^{k\times k}$ and $\underline{\Sigma}_k \in \mathbb{R}^{(n-k)\times(m-k)}$. We also set $A_k = U_k\Sigma_k V_k^\mathsf{T}$ and $\underline{A}_k = \underline{U}_k\underline{\Sigma}_k\underline{V}_k^\mathsf{T}$ so that $A = A_k + \underline{A}_k$.

The Eckart-Young theorem [30] states that the optimal rank-k approximation of A is given by $A_k = \pi(U_k)A$. In practice, for large-scale applications, computing U_k can be computationally challenging or too expensive. In this context, randomized algorithms for approximating U_k have become increasingly popular [53, 63] in the past few years. They have been proved to be easy

to implement, computationally efficient and numerically robust. The general idea of randomized subspace iteration methods is to use random sampling to identify a subspace that approximates the range of a given matrix. In this chapter, we propose a general error analysis related to the low-rank approximation to a given real matrix in the spectral and Frobenius norms.

3.1.1 Related research

The error analysis of randomized algorithms for low-rank approximation to a given matrix has been extensively considered in the literature; see, e.g., the survey papers [53, 61, 63, 93] for a general overview. This theoretical analysis takes into account the distribution of the random matrix to derive either error bounds in expectation or tail bounds of the error distribution. The case of standard Gaussian matrices is usually considered, even though alternative theoretical results exist, based on either random column selection matrices [18] or Subsampled Random Fourier Transform matrices [53]. Halko, Martinsson and Tropp have developed a reference error analysis in expectation [53, Theorems 10.5 and 10.6] in the Frobenius and spectral norms for $Z = A\Omega$, Ω being a standard Gaussian matrix. Later, Gu has refined these error bounds [49, Theorem 5.7]. More recently, Saibaba [77] has proposed a complementary average case error analysis in terms of principal angles between appropriate subspaces that is available in any unitarily invariant norm.

3.1.2 Contributions

In certain situations, a priori knowledge of Z (or of the corresponding projection subspace) may be available. This naturally arises when considering, e.g., the solution of large-scale nonlinear systems of equations, requiring the solution of a sequence of linear systems of equations. Approximate spectral information based on Ritz or Harmonic Ritz vectors can be easily retrieved to form such a subspace. Exploiting this a priori knowledge is thus of primary interest to design fast, robust and efficient low-rank approximation algorithms. Hence we aim at developing a general theoretical error analysis of randomized algorithms for the low-rank approximation, assuming the existence of a non-trivial mean and of a general covariance matrix for Z.

In this chapter, for a given target rank, we propose to analyze theoretically the low-rank approximation to a given matrix in both the spectral and Frobenius norms, when Z is drawn from a non-standard Gaussian distribution. Namely, we will first derive in Theorems 3.4 and 3.8 deterministic error bounds that hold with some minimal assumptions. Second, we will derive error bounds in expectation in the non-standard Gaussian case (with a non-trivial mean value and a general covariance matrix) in Theorems 3.18, 3.20 and 3.21, respectively. Our analysis simultaneously generalizes and improves the error bounds for spectral and Frobenius norms proposed in [53]. We specialize our error bounds to the Randomized Singular Value Decomposition (RSVD) in Corollaries 3.33, 3.35 and 3.36, respectively and provide numerical experiments on a synthetic test case that illustrate the tightness of the obtained error bounds.

Section 3.2 introduces specific results useful later in our analysis. Section 3.3 details our error analysis related to the low-rank approximation of the matrix A. First, in Section 3.3.1, we derive deterministic error bounds that hold with some minimal assumptions. Then, in Section 3.3.2, we derive error bounds in expectation, with Z drawn following a non-standard Gaussian distribution. In Section 3.4, we specialize our error bounds to the Randomized Singular Value Decomposition. In Section 3.5, we provide detailed numerical illustrations, including a broad comparison with reference error bounds. Conclusions are finally drawn in Section 3.6.

3.2 Preliminaries

We first introduce notations used throughout the chapter and remind specific results.

Sherman-Morrison formula

Let $M \in \mathbb{R}^{m\times n}$ and $N \in \mathbb{R}^{n\times m}$ such that $I_n + NM$ is nonsingular. Then, $I_m + MN$ is also nonsingular and one has [43, Section 2.1.4]

$$(I_m + MN)^{-1} = I_m - M(I_n + NM)^{-1}N. \tag{3.3}$$

Strong submultiplicativity of Frobenius and spectral norms

The strong submultiplicativity of the spectral and Frobenius norms [56, Relation(B.7)] reads

$$\forall M \in \mathbb{R}^{n\times p}, \forall N \in \mathbb{R}^{p\times q}, \forall Q \in \mathbb{R}^{q\times m}, \|MNQ\|_{2,F} \leq \|M\|_2 \, \|N\|_{2,F} \, \|Q\|_2. \tag{3.4}$$

3.3 Error bounds for the low-rank approximation of a matrix

Our main objective is to derive error bounds related to an approximation of rank k to A using the orthogonal projection $\pi(Z)$ where $Z \in \mathbb{R}^{n\times p}$ and $k \in \{1, \ldots, p\}$, $p \leq \operatorname{rank}(A)$. First, in Section 3.3.1, we consider the general case with the minimal assumption that Z is full column rank. In this case, we are able to derive deterministic error bounds using a systematic approach. Second, in Section 3.3.2, we focus on error bounds that are tractable now from a stochastic point of view, where we assume that the matrix Z corresponds to a general Gaussian matrix. Namely, $Z \in \mathbb{R}^{n\times p}$ will be drawn as a Gaussian full column rank matrix of mean $\widehat{Z} \in \mathbb{R}^{n\times p}$ and covariance matrix $\mathbf{Cov}(Z)$, that is, $Z \sim \mathcal{N}(\widehat{Z}, \mathbf{Cov}(Z))$. In this case, we develop an error analysis in expectation with respect to the random variable Z.

3.3.1 Deterministic analysis

Without loss of generality, we aim at deriving error bounds for the following quantity

$$\|[I_n - \pi(Z)]A\|_{2,F}^2 - \|[I_n - \pi(Z)]\underline{A}_k\|_{2,F}^2, \tag{3.5}$$

where $Z \in \mathbb{R}^{n\times p}$ is a full column rank matrix (with $p \leq \operatorname{rank}(A)$). Since $\|[I_n - \pi(Z)]\underline{A}_k\|_{2,F}^2 \leq \|\underline{A}_k\|_{2,F}^2$, we note that (3.5) is an upper bound of

$$\|[I_n - \pi(Z)]A\|_{2,F}^2 - \|\underline{A}_k\|_{2,F}^2, \tag{3.6}$$

a quantity which is frequently considered in the analysis of low-rank approximation methods, see, e.g., [53]. In this sense, our error bounds will naturally cover existing bounds from the literature. The next lemma will be helpful to derive the deterministic error bound for (3.5).

Lemma 3.1. *Let $A \in \mathbb{R}^{n\times m}$ such that $n \geq m$ and $Z \in \mathbb{R}^{n\times p}$ a full column rank matrix with $p \leq \operatorname{rank}(A)$. For a given $k \in \{1, \ldots, p\}$, set $\Omega_k = U_k^{\mathsf{T}} Z \in \mathbb{R}^{k\times p}$ and $\underline{\Omega}_k = \underline{U}_k^{\mathsf{T}} Z \in \mathbb{R}^{(n-k)\times p}$ such that Ω_k is full row rank (i.e., $\operatorname{rank}(\Omega_k) = k$). Then, one has*

$$U_k^{\mathsf{T}}[I_n - \pi(Z)]U_k \quad \preccurlyeq \quad S_k^{\mathsf{T}} S_k \quad \preccurlyeq \quad T_k^{\mathsf{T}} T_k, \tag{3.7}$$

with $T_k = \underline{\Omega}_k \Omega_k^{\dagger} \in \mathbb{R}^{(n-k)\times k}$ and $S_k = (I_{n-k} + T_k T_k^{\mathsf{T}})^{-\frac{1}{2}} T_k \in \mathbb{R}^{(n-k)\times k}$.

Proof. By assumption, Ω_k has full row rank and therefore has a right multiplicative inverse $\Omega_k^{\dagger}$. Hence $\bar{Z}_k = Z\Omega_k^{\dagger}$ satisfies the two relations

$$U_k^{\mathsf{T}} \bar{Z}_k = I_k \quad \text{and} \quad \underline{U}_k^{\mathsf{T}} \bar{Z}_k = T_k.$$

Moreover, we have $\mathcal{R}(\bar{Z}_k) \subset \mathcal{R}(Z)$. Hence, by applying [53, Proposition 8.5], one gets $I_n - \pi(Z) \preccurlyeq I_n - \pi(\bar{Z}_k)$. By using the conjugation rule ((1)) and the identity $U_k U_k^\mathsf{T} + \underline{U}_k \underline{U}_k^\mathsf{T} = I_n$, we obtain

$$\begin{aligned} U_k^\mathsf{T}[I_n - \pi(Z)]U_k \preccurlyeq U_k^\mathsf{T}[I_n - \pi(\bar{Z}_k)]U_k &= U_k^\mathsf{T}\left(I_n - \bar{Z}_k(\bar{Z}_k^\mathsf{T}\bar{Z}_k)^{-1}\bar{Z}_k^\mathsf{T}\right)U_k, \\ &= I_k - U_k^\mathsf{T}\bar{Z}_k(\bar{Z}_k^\mathsf{T}\bar{Z}_k)^{-1}\bar{Z}_k^\mathsf{T}U_k, \\ &= I_k - (\bar{Z}_k^\mathsf{T}\bar{Z}_k)^{-1}, \\ &= I_k - \left(\bar{Z}_k^\mathsf{T}(U_kU_k^\mathsf{T} + \underline{U}_k\underline{U}_k^\mathsf{T})\bar{Z}_k\right)^{-1}, \\ &= I_k - \left(I_k + T_k^\mathsf{T}T_k\right)^{-1}, \\ &= T_k^\mathsf{T}\left(I_{n-k} + T_kT_k^\mathsf{T}\right)^{-1}T_k, \end{aligned}$$

where the last equality is obtained using the Sherman-Morrison formula. Then we observe that $T_k^\mathsf{T}\left(I_{n-k} + T_kT_k^\mathsf{T}\right)^{-1}T_k = S_k^\mathsf{T}S_k$ and $\left(I_{n-k} + T_kT_k^\mathsf{T}\right)^{-1} \preccurlyeq I_{n-k}$. We conclude the proof by applying the conjugation rule to deduce $S_k^\mathsf{T}S_k \preccurlyeq T_k^\mathsf{T}T_k$. □

Remark 3.2. In the particular case of $Z = A\Omega$ ($\Omega \in \mathbb{R}^{n\times p}$), we have $\Omega_k = \Sigma_k(V_k^\mathsf{T}\Omega)$, $\underline{\Omega}_k = \underline{\Sigma}_k(\underline{V}_k^\mathsf{T}\Omega)$ and $T_k = \underline{\Sigma}_k(\underline{V}_k^\mathsf{T}\Omega)(V_k^\mathsf{T}\Omega)^\dagger\Sigma_k^{-1}$. With the notation of [53], this gives $\Omega_k := \Sigma_1\Omega_1$, $\underline{\Omega}_k := \Sigma_2\Omega_2$ and $T_k := \Sigma_2\Omega_2\Omega_1^\dagger\Sigma_1^{-1}$, respectively. We therefore point out that our notation is related but not directly equivalent to the one employed in [53].

Remark 3.3. We note that the positive singular values of T_k represent the *tangent* of the canonical angles between $\mathcal{R}(Z\Omega_k^\dagger)$ and $\mathcal{R}(U_k)$ [97, Theorem 3.1 and Remark 3.1], while the positive singular values of S_k represent the *sine* of the canonical angles between the same subspaces. Hence we stress this information in the notation and refer the reader to [77] for a theoretical analysis of low-rank approximation methods in terms of subspace angles.

> Let $\theta_1, \ldots, \theta_k$ denote the principal canonical angles between $\mathcal{R}(Z)$ and $\mathcal{R}(U_k)$, and $\bar{\theta}_1, \ldots, \bar{\theta}_k$ the ones between $\mathcal{R}(Z\Omega_k^\dagger)$ and $\mathcal{R}(U_k)$ (see Section 2.1.3). The statement of Lemma 3.1 can be rephrased geometrically as
>
> $$\sin(\theta_i)^2 \leq \sin(\bar{\theta}_i)^2 = \frac{\tan(\bar{\theta}_i)^2}{1+\tan(\bar{\theta}_i)^2} \leq \tan(\bar{\theta}_i)^2, \quad 1 \leq i \leq k.$$
>
> Less formally, the forthcoming analysis is based on how close $\mathcal{R}(U_k)$ and $\mathcal{R}(Z\Omega_k^\dagger)$ are, while the truly computed error rather concerns how close $\mathcal{R}(Z)$ is from $\mathcal{R}(U_k)$. Since the latter is of dimension p, while $\mathcal{R}(Z\Omega_k^\dagger)$ is of dimension k, it is clear that the looseness of the derived bounds will increase with p.

The next theorem introduces unified deterministic error bounds for the quantity of interest (3.5).

Theorem 3.4. *(Deterministic error bounds in Frobenius and spectral norms) Let $A \in \mathbb{R}^{n\times m}$ such that $n \geq m$ and $Z \in \mathbb{R}^{n\times p}$ a full column rank matrix with $p \leq \operatorname{rank}(A)$. For a given $k \in \{1, \ldots, p\}$, set $\Omega_k = U_k^\mathsf{T}Z \in \mathbb{R}^{k\times p}$ and $\underline{\Omega}_k = \underline{U}_k^\mathsf{T}Z \in \mathbb{R}^{(n-k)\times p}$ such that Ω_k is full row rank (i.e., $\operatorname{rank}(\Omega_k) = k$). Then, one has*

$$\|[I_n - \pi(Z)]A\|_{2,F}^2 - \|[I_n - \pi(Z)]\underline{A}_k\|_{2,F}^2 \leq \|[I_n - \pi(Z)]A_k\|_{2,F}^2 \leq \min\left\{\|S_k\|_{2,F}^2\|\Sigma_k\|_2^2,\ \|T_k\Sigma_k\|_{2,F}^2\right\}, \tag{3.8}$$

where $T_k = \underline{\Omega}_k\Omega_k^\dagger \in \mathbb{R}^{(n-k)\times k}$ and $S_k = (I_{n-k} + T_kT_k^\mathsf{T})^{-\frac{1}{2}}T_k \in \mathbb{R}^{(n-k)\times k}$.

Proof. Using the identity $AA^{\mathsf{T}} = A_k A_k^{\mathsf{T}} + \underline{A}_k \underline{A}_k^{\mathsf{T}}$, we obtain for the spectral norm case

$$\begin{aligned}\|[I_n - \pi(Z)]A\|_2^2 &= \|[I_n - \pi(Z)]AA^{\mathsf{T}}[I_n - \pi(Z)]\|_2, \\ &= \|[I_n - \pi(Z)][A_k A_k^{\mathsf{T}} + \underline{A}_k \underline{A}_k^{\mathsf{T}}][I_n - \pi(Z)]\|_2.\end{aligned}$$

Hence we obtain

$$\|[I_n - \pi(Z)]A\|_2^2 \leq \|[I_n - \pi(Z)]A_k\|_2^2 + \|[I_n - \pi(Z)]\underline{A}_k\|_2^2.$$

Thus, by using the unitarily invariance of the spectral norm, we get

$$\|[I_n - \pi(Z)]A\|_2^2 - \|[I_n - \pi(Z)]\underline{A}_k\|_2^2 \leq \|[I_n - \pi(Z)]U_k\Sigma_k\|_2^2. \tag{3.9}$$

Moreover, using Lemma 3.1 and the conjugation rule, we get

$$\Sigma_k^{\mathsf{T}} U_k^{\mathsf{T}}[I_n - \pi(Z)]U_k\Sigma_k \preccurlyeq \Sigma_k^{\mathsf{T}} S_k^{\mathsf{T}} S_k \Sigma_k \preccurlyeq \Sigma_k T_k^{\mathsf{T}} T_k \Sigma_k. \tag{3.10}$$

Hence, since $[I_n - \pi(Z)]^2 = I_n - \pi(Z)$, we obtain

$$\|[I_n - \pi(Z)]U_k\Sigma_k\|_2^2 = \|\Sigma_k^{\mathsf{T}} U_k^{\mathsf{T}}[I_n - \pi(Z)]U_k\Sigma_k\|_2 \leq \|\Sigma_k^{\mathsf{T}} S_k^{\mathsf{T}} S_k \Sigma_k\|_2 \leq \|\Sigma_k^{\mathsf{T}} T_k^{\mathsf{T}} T_k \Sigma_k\|_2,$$

or equivalently,

$$\|[I_n - \pi(Z)]U_k\Sigma_k\|_2^2 \leq \|S_k\Sigma_k\|_2^2 \leq \|T_k\Sigma_k\|_2^2.$$

Using the strong submultiplicativity of the spectral norm, the latter inequality implies

$$\|[I_n - \pi(Z)]U_k\Sigma_k\|_2^2 \leq \min\left\{\|S_k\|_2^2\|\Sigma_k\|_2^2,\ \|T_k\Sigma_k\|_2^2\right\}. \tag{3.11}$$

Combining (3.9) and (3.11) completes the proof for the spectral norm case.
For the Frobenius norm, similar arguments are used. In fact, we have

$$\begin{aligned}\|[I_n - \pi(Z)]A\|_F^2 &= \operatorname{tr}\left([I_n - \pi(Z)]AA^{\mathsf{T}}[I_n - \pi(Z)]\right), \\ &= \operatorname{tr}\left([I_n - \pi(Z)][A_k A_k^{\mathsf{T}} + \underline{A}_k \underline{A}_k^{\mathsf{T}}][I_n - \pi(Z)]\right), \\ &= \operatorname{tr}\left([I_n - \pi(Z)]A_k A_k^{\mathsf{T}}[I_n - \pi(Z)]\right) + \operatorname{tr}\left([I_n - \pi(Z)]\underline{A}_k \underline{A}_k^{\mathsf{T}}[I_n - \pi(Z)]\right), \\ &= \|[I_n - \pi(Z)]A_k\|_F^2 + \|[I_n - \pi(Z)]\underline{A}_k\|_F^2.\end{aligned}$$

Thus,

$$\|[I_n - \pi(Z)]A\|_F^2 - \|[I_n - \pi(Z)]\underline{A}_k\|_F^2 = \|[I_n - \pi(Z)]U_k\Sigma_k\|_F^2. \tag{3.12}$$

Using (3.10), we obtain

$$\|[I_n - \pi(Z)]U_k\Sigma_k\|_F^2 = \operatorname{tr}\left(\Sigma_k^{\mathsf{T}} U_k^{\mathsf{T}}[I_n - \pi(Z)]U_k\Sigma_k\right) \leq \operatorname{tr}\left(\Sigma_k^{\mathsf{T}} S_k^{\mathsf{T}} S_k \Sigma_k\right) \leq \operatorname{tr}\left(\Sigma_k^{\mathsf{T}} T_k^{\mathsf{T}} T_k \Sigma_k\right).$$

Hence we obtain

$$\|[I_n - \pi(Z)]U_k\Sigma_k\|_F^2 \leq \min\left\{\|S_k\|_F^2\|\Sigma_k\|_2^2,\ \|T_k\Sigma_k\|_F^2\right\}. \tag{3.13}$$

Combining relations (3.12) and (3.13) completes the proof for the Frobenius norm case. □

Remark 3.5. We have shown in Theorem 3.4 that

$$\|[I_n - \pi(Z)]A\|_{2,F}^2 - \|[I_n - \pi(Z)]\underline{A}_k\|_{2,F}^2 \leq \|S_k\Sigma_k\|_{2,F}^2 \leq \|T_k\Sigma_k\|_{2,F}^2. \tag{3.14}$$

Hence, $\|S_k\Sigma_k\|_{2,F}^2$ definitively represents a sharper upper bound in the deterministic case. Nevertheless we have decided to use the submultiplicativity to bound $\|S_k\Sigma_k\|_{2,F}^2$, since only this formulation authorizes a possible treatment in the stochastic setting as detailed in Section 3.3.2.

Remark 3.6. In the particular case of $Z = A\Omega$ ($\Omega \in \mathbb{R}^{n\times p}$), with the notation of [53], we have $\|T_k\Sigma_k\|_{2,F} := \|\Sigma_2\Omega_2\Omega_1^\dagger\|_{2,F}$, a quantity which precisely appears in the reference upper bound [53, Theorem 9.1]. We note that our error bound (3.8) is tighter whenever $\|S_k\|_{2,F}^2\|\Sigma_k\|_2^2 < \|T_k\Sigma_k\|_{2,F}^2$. Hence Theorem 3.4 either recovers or improves the reference error bound [53, Theorem 9.1] in this setting.

Remark 3.7. Using $\|[I_n - \pi(Z)]\underline{A}_k\|_{2,F}^2 \leq \|\underline{A}_k\|_{2,F}^2$ and $\sqrt{a^2+b^2} \leq a+b$ for any real positive scalars a and b, we note that Theorem 3.4 implies [28, Theorem 3.3] when stated in the spectral and Frobenius norms. As [28, Theorem 3.3], Theorem 3.4 can be extended to the case of any Schatten-p class of norms.

If we target to bound $\|[I_n-\pi(Z)]A\|_2^2 - \|\underline{A}_k\|_2^2$ instead of $\|[I_n-\pi(Z)]A\|_2^2 - \|[I_n-\pi(Z)]\underline{A}_k\|_2^2$, we are able to improve the result given in Theorem 3.4. This is detailed next.

Theorem 3.8. *(Improved deterministic error bound in spectral norm) Let $A \in \mathbb{R}^{n\times m}$ such that $n \geq m$ and $Z \in \mathbb{R}^{n\times p}$ a full column rank matrix with $p \leq \operatorname{rank}(A)$. For a given $k \in \{1,\dots,p\}$, set $\Omega_k = U_k^\mathsf{T} Z \in \mathbb{R}^{k\times p}$ and $\underline{\Omega}_k = \underline{U}_k^\mathsf{T} Z \in \mathbb{R}^{(n-k)\times p}$ such that Ω_k is full row rank (i.e., $\operatorname{rank}(\Omega_k) = k$). Then, one has*

$$\|[I_n - \pi(Z)]A\|_2^2 - \|\underline{A}_k\|_2^2 \leq \min\left\{\|S_k\|_2^2\|\widehat{\Sigma}_k\|_2^2,\ \|T_k\widehat{\Sigma}_k\|_2^2\right\}, \tag{3.15}$$

where $\widehat{\Sigma}_k = \left(\Sigma_k^2 - \sigma_{k+1}^2 I_k\right)^{\frac{1}{2}} \in \mathbb{R}^{k\times k}$, $T_k = \underline{\Omega}_k\Omega_k^\dagger \in \mathbb{R}^{(n-k)\times k}$ and $S_k = (I_{n-k} + T_kT_k^\mathsf{T})^{-\frac{1}{2}}T_k \in \mathbb{R}^{(n-k)\times k}$.

Proof. First, we note that $\|\underline{A}_k\|_2^2 = \sigma_{k+1}^2$. Then, by definition of σ_{k+1} and $\underline{\Sigma}_k$, one has $\underline{\Sigma}_k\underline{\Sigma}_k^\mathsf{T} \preccurlyeq \sigma_{k+1}^2 I_{n-k}$. Thus, we obtain

$$\begin{aligned} AA^\mathsf{T} &\preccurlyeq U_k\Sigma_k^2U_k^\mathsf{T} + \sigma_{k+1}^2\underline{U}_k\underline{U}_k^\mathsf{T}, \\ &\preccurlyeq U_k\Sigma_k^2U_k^\mathsf{T} + \sigma_{k+1}^2(I_n - U_kU_k^\mathsf{T}), \\ &\preccurlyeq U_k\widehat{\Sigma}_k^2U_k^\mathsf{T} + \sigma_{k+1}^2 I_n. \end{aligned}$$

Hence,

$$\begin{aligned} \|[I_n - \pi(Z)]A\|_2^2 &= \|[I_n - \pi(Z)]AA^\mathsf{T}[I_n - \pi(Z)]\|_2, \\ &\leq \sigma_{k+1}^2\|I_n - \pi(Z)\|_2 + \|[I_n - \pi(Z)]U_k\widehat{\Sigma}_k^2U_k^\mathsf{T}[I_n - \pi(Z)]\|_2, \\ &\leq \|\underline{A}_k\|_2^2 + \|[I_n - \pi(Z)]U_k\widehat{\Sigma}_k\|_2^2. \end{aligned}$$

Then, using (3.11) with $\widehat{\Sigma}_k$ instead of Σ_k, the rest of the proof follows straightforwardly. □

By definition of $\widehat{\Sigma}_k$, one has $\widehat{\Sigma}_k \preccurlyeq \Sigma_k$. Thus we deduce that if we target to bound $\|[I_n - \pi(Z)]A\|_2^2 - \|\underline{A}_k\|_2^2$, then the error bound (3.15) is tighter than (3.8). To the best of our knowledge, Theorem 3.8 is new and provides an improved error bound in the spectral norm for a fairly general choice of Z.

Finally, we emphasize that both Theorems 3.4 and 3.8 will play a key role for deriving new improved error bounds in expectation for the low-rank approximation to a given matrix. This is detailed next.

3.3.2 Analysis in expectation

We now provide error bounds in expectation and consider the case where $Z \in \mathbb{R}^{n\times p}$ is drawn as a Gaussian matrix of mean $\widehat{Z} \in \mathbb{R}^{n\times p}$ and covariance matrix $\mathbf{Cov}(Z)$, that is, $Z \sim \mathcal{N}(\widehat{Z}, \mathbf{Cov}(Z))$, with $2 < p \leq \min\left\{\operatorname{rank}(A), \operatorname{rank}(\mathbf{Cov}(Z))\right\}$. For $k \in \{1,\dots,p-2\}$, we define $\Omega_k \in \mathbb{R}^{k\times p}$

and $\underline{\Omega}_k \in \mathbb{R}^{(n-k)\times p}$ as $\Omega_k = U_k^{\mathsf{T}}(Z - \widehat{Z})$ and $\underline{\Omega}_k = \underline{U}_k^{\mathsf{T}}(Z - \widehat{Z})$, respectively. The condition $p \leq \operatorname{rank}(\mathbf{Cov}(Z))$ ensures that $Z - \widehat{Z}$ has full column rank with probability one [31]. We assume that Ω_k has full row rank in this section.

This general approach offers several advantages but rises additional technical difficulties. In particular, Ω_k and $\underline{\Omega}_k$ are now stochastically dependent on the distribution law of Z. Thus, before stating our main results in Section 3.3.2, we provide in Section 3.3.2 preparatory lemmas that extend existing probabilistic results to the non-standard Gaussian case. In this section, we consider the following block partitioning of the projected covariance matrix $U^{\mathsf{T}}\,\mathbf{Cov}(Z)\,U$

$$U^{\mathsf{T}}\,\mathbf{Cov}(Z)\,U = \begin{bmatrix} U_k^{\mathsf{T}} \\ \underline{U}_k^{\mathsf{T}} \end{bmatrix} \mathbf{Cov}(Z) \begin{bmatrix} U_k & \underline{U}_k \end{bmatrix} = \begin{bmatrix} \mathbf{Cov}_k(Z) & \mathbf{Cov}_{\perp,k}(Z)^{\mathsf{T}} \\ \mathbf{Cov}_{\perp,k}(Z) & \underline{\mathbf{Cov}}_k(Z) \end{bmatrix},$$

with

$$\begin{aligned} \mathbf{Cov}_k(Z) &= U_k^{\mathsf{T}}\,\mathbf{Cov}(Z)U_k \in \mathbb{R}^{k\times k}, \\ \mathbf{Cov}_{\perp,k}(Z) &= \underline{U}_k^{\mathsf{T}}\,\mathbf{Cov}(Z)U_k \in \mathbb{R}^{(n-k)\times k}, \\ \underline{\mathbf{Cov}}_k(Z) &= \underline{U}_k^{\mathsf{T}}\,\mathbf{Cov}(Z)\underline{U}_k \in \mathbb{R}^{(n-k)\times(n-k)}. \end{aligned} \tag{3.16}$$

Preparatory lemmas

Given that $\Omega_k = U_k^{\mathsf{T}}(Z - \widehat{Z})$ and $\underline{\Omega}_k = \underline{U}_k^{\mathsf{T}}(Z - \widehat{Z})$, then by using elementary properties of Gaussian vectors, one gets that $\Omega_k \sim \mathcal{N}(0, \mathbf{Cov}_k(Z))$ and $\underline{\Omega}_k \sim \mathcal{N}(0, \underline{\mathbf{Cov}}_k(Z))$. We note that, although Ω_k and $\underline{\Omega}_k$ are centered Gaussian matrices, the conditional law of $\underline{\Omega}_k$ with respect to Ω_k follows a Gaussian distribution that is not necessarily centered [67, Theorem 1.2.11]. We therefore adapt this result to our setting in the next lemma.

Lemma 3.9. *Let $A \in \mathbb{R}^{n\times m}$ such that $n \geq m$ and $Z \in \mathbb{R}^{n\times p}$ a Gaussian full column rank matrix of mean $\widehat{Z} \in \mathbb{R}^{n\times p}$ and covariance matrix $\mathbf{Cov}(Z)$, that is, $Z \sim \mathcal{N}(\widehat{Z}, \mathbf{Cov}(Z))$ satisfying $2 < p \leq \min\{\operatorname{rank}(A), \operatorname{rank}(\mathbf{Cov}(Z))\}$. For a given $k \in \{1, \ldots, p-2\}$, set $\Omega_k = U_k^{\mathsf{T}}(Z - \widehat{Z}) \in \mathbb{R}^{k\times p}$, $\underline{\Omega}_k = \underline{U}_k^{\mathsf{T}}(Z - \widehat{Z}) \in \mathbb{R}^{(n-k)\times p}$ such that Ω_k is full row rank (i.e. $\operatorname{rank}(\Omega_k) = k$). If the projected covariance matrix $\mathbf{Cov}_k(Z)$ is nonsingular, then the random matrix $\underline{\Omega}_k$ conditioned by Ω_k follows a Gaussian distribution of mean*

$$\mathbb{E}\left[\,\underline{\Omega}_k \mid \Omega_k\,\right] = \mathbf{Cov}_{\perp,k}(Z)[\mathbf{Cov}_k(Z)]^{-1}\Omega_k, \tag{3.17}$$

and of covariance matrix given by

$$\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right) = \underline{\mathbf{Cov}}_k(Z) - \mathbf{Cov}_{\perp,k}(Z)\left[\mathbf{Cov}_k(Z)\right]^{-1}\mathbf{Cov}_{\perp,k}(Z)^{\mathsf{T}}, \tag{3.18}$$

where $\mathbf{Cov}_k(Z) = U_k^{\mathsf{T}}\,\mathbf{Cov}(Z)U_k$, $\mathbf{Cov}_{\perp,k}(Z) = \underline{U}_k^{\mathsf{T}}\,\mathbf{Cov}(Z)U_k$ and $\underline{\mathbf{Cov}}_k(Z) = \underline{U}_k^{\mathsf{T}}\,\mathbf{Cov}(Z)\underline{U}_k$.

Remark 3.10. When $\mathbf{Cov}(Z) = AA^{\mathsf{T}}$, we note that $\mathbf{Cov}_{\perp,k}(Z) = 0$, which yields $\mathbb{E}\left[\,\underline{\Omega}_k \mid \Omega_k\,\right] = 0$ and $\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right) = \underline{\mathbf{Cov}}_k(Z)$.

The next two lemmas aim at proposing key results about Gaussian matrices in the non-standard case. In particular, we extend [53, Propositions 10.1 and 10.2] to the case of a non-standard Gaussian matrix with general covariance matrix and potentially nonzero mean term.

Lemma 3.11. *Let $N \in \mathbb{R}^{p\times p}$ be a given matrix and $M \in \mathbb{R}^{k\times p}$ be a Gaussian matrix such that $M \sim \mathcal{N}(\widehat{M}, \mathbf{Cov}(M))$ with mean $\widehat{M} \in \mathbb{R}^{k\times p}$ and covariance matrix $\mathbf{Cov}(M) \in \mathbb{R}^{k\times k}$. Then*

$$\mathbb{E}\left[\,\|MN\|_2\,\right] \leq \|\widehat{M}N\|_2 + \|\,\mathbf{Cov}(M)^{\frac{1}{2}}\|_2\|N\|_F + \|\,\mathbf{Cov}(M)^{\frac{1}{2}}\|_F\|N\|_2, \tag{3.19}$$

and

$$\mathbb{E}\left[\,\|MN\|_F^2\,\right] = \|\widehat{M}N\|_F^2 + \|\,\mathbf{Cov}(M)^{\frac{1}{2}}\|_F^2\|N\|_F^2. \tag{3.20}$$

Proof. Using the definition of a non-standard Gaussian matrix (2.4), we have $M = \widehat{M} + \mathbf{Cov}(M)^{\frac{1}{2}}G$ with $G \in \mathbb{R}^{k\times p}$ a standard Gaussian matrix (i.e. $G \sim \mathcal{N}(0, I_k)$). Thus, by applying the triangle inequality in the spectral norm, we get

$$\|MN\|_2 = \|(\widehat{M} + \mathbf{Cov}(M)^{\frac{1}{2}}G)N\|_2 \leq \|\widehat{M}N\|_2 + \|\,\mathbf{Cov}(M)^{\frac{1}{2}}GN\|_2.$$

Then, by applying [53, Proposition 10.1] to obtain an upper bound of $\mathbb{E}\left[\,\|\,\mathbf{Cov}(M)^{\frac{1}{2}}GN\|_2\,\right]$, we deduce (3.19). In the Frobenius norm, we have

$$\|MN\|_F^2 = \|\widehat{M}N+\mathbf{Cov}(M)^{\frac{1}{2}}GN\|_F^2 = \|\widehat{M}N\|_F^2+\|\,\mathbf{Cov}(M)^{\frac{1}{2}}GN\|_F^2+2\,\mathrm{tr}(N^{\mathsf{T}}\widehat{M}^{\mathsf{T}}\,\mathbf{Cov}(M)^{\frac{1}{2}}GN).$$

Taking the expectation and using [53, Proposition 10.1], we get

$$\mathbb{E}\left[\,\|MN\|_F^2\,\right] = \|\widehat{M}N\|_F^2 + \|\,\mathbf{Cov}(M)^{\frac{1}{2}}\|_F^2\|N\|_F^2 + 2\,\mathbb{E}\left[\,\mathrm{tr}(N^{\mathsf{T}}\widehat{M}^{\mathsf{T}}\,\mathbf{Cov}(M)^{\frac{1}{2}}GN)\,\right].$$

By using the linearity of the expectation and the fact that $\mathbb{E}\,[\,G\,] = 0$, we remark that

$$\mathbb{E}\left[\,\mathrm{tr}(N^{\mathsf{T}}\widehat{M}^{\mathsf{T}}\,\mathbf{Cov}(M)^{\frac{1}{2}}GN)\,\right] = \mathrm{tr}\left(N^{\mathsf{T}}\widehat{M}^{\mathsf{T}}\,\mathbf{Cov}(M)^{\frac{1}{2}}\mathbb{E}\,[\,G\,]\,N\right) = 0.$$

This concludes the proof. □

In the next lemma, we show how to bound (in expectation) the quantity $\|M^{\dagger}N\|_2$ or $\|M^{\dagger}N\|_F^2$, where $N \in \mathbb{R}^{k\times k}$ is a given fixed matrix and $M \in \mathbb{R}^{k\times p}$ follows a centered Gaussian distribution.

Lemma 3.12. *For a fixed integer $k > 1$, let $N \in \mathbb{R}^{k\times k}$ be a given matrix and $M \in \mathbb{R}^{k\times p}$ (with $p > k+1$) be a centered Gaussian matrix $M \sim \mathcal{N}(0, \mathbf{Cov}(M))$. If the covariance matrix $\mathbf{Cov}(M)$ is nonsingular, then*

$$\mathbb{E}\left[\,\|M^{\dagger}N\|_F^2\,\right] = \frac{\|(N^{\mathsf{T}}[\mathbf{Cov}(M)]^{-1}N)^{\frac{1}{2}}\|_F}{p-k-1} \quad \textit{and} \quad \mathbb{E}\left[\,\|M^{\dagger}N\|_2\,\right] \leq \frac{e\sqrt{p}}{p-k}\sqrt{\|N^{\mathsf{T}}[\mathbf{Cov}(M)]^{-1}N\|_2},$$

where e denotes exponential of 1, i.e., $e = \exp(1)$.

Proof. Since M is a random matrix with p independent columns, where each column is a multivariate Gaussian distribution with zero mean and covariance matrix $\mathbf{Cov}(M)$, the matrix MM^{T} follows a Wishart distribution of the form $\mathcal{W}_k(p, \mathbf{Cov}(M))$[67, Definition 3.1.3]. One has also $\|M^{\dagger}N\|_F^2 = \mathrm{tr}(N^{\mathsf{T}}[(M^{\dagger})^{\mathsf{T}}M^{\dagger}]N) = \mathrm{tr}(N^{\mathsf{T}}[MM^{\mathsf{T}}]^{-1}N)$, where the second equality holds with probability one since $p > k+1$. In fact, if $\mathbf{Cov}(M)$ is nonsingular, the matrix MM^{T} is nonsingular almost surely, see [67, Theorem 3.1.4]. In this case, according to [67, Theorem 3.2.12] for $p > k+1$, one has, for any matrix $N \in \mathbb{R}^{k\times k}$,

$$\mathbb{E}\left[\,N^{\mathsf{T}}[MM^{\mathsf{T}}]^{-1}N\,\right] = N^{\mathsf{T}}\mathbb{E}\left[\,[MM^{\mathsf{T}}]^{-1}\,\right]N = \frac{N^{\mathsf{T}}\,\mathbf{Cov}(M)^{-1}N}{p-k-1}.$$

Hence,

$$\mathbb{E}\left[\,\mathrm{tr}(N^{\mathsf{T}}[MM^{\mathsf{T}}]^{-1}N)\,\right] = \frac{\mathrm{tr}(N^{\mathsf{T}}\,\mathbf{Cov}(M)^{-1}N)}{p-k-1} = \frac{\|(N^{\mathsf{T}}\,\mathbf{Cov}(M)^{-1}N)^{\frac{1}{2}}\|_F}{p-k-1},$$

which concludes the proof for the Frobenius norm. In the spectral norm, using (2.4), we have $M = \mathbf{Cov}(M)^{\frac{1}{2}}G$ with $G \in \mathbb{R}^{k\times p}$ following a standard Gaussian distribution. If $\mathbf{Cov}(M)$ is nonsingular, then one has $M^{\dagger} = G^{\dagger}\,\mathbf{Cov}(M)^{-\frac{1}{2}}$ and thus, for any matrix $N \in \mathbb{R}^{k\times k}$, we have

$$\|M^{\dagger}N\|_2 = \|G^{\dagger}\,\mathbf{Cov}(M)^{-\frac{1}{2}}N\|_2 \leq \|G^{\dagger}\|_2\|\,\mathbf{Cov}(M)^{-\frac{1}{2}}N\|_2 = \|G^{\dagger}\|_2\sqrt{\|N^{\mathsf{T}}\,\mathbf{Cov}(M)^{-1}N\|_2}.$$

Then, we take the expectation and apply [53, Proposition 10.2] to bound the quantity $\mathbb{E}\left[\,\|G^{\dagger}\|_2\,\right]$, which concludes the proof. □

We now introduce a final lemma based on Lemmas 3.9, 3.11 and 3.12, which is central in the derivation of our error bounds in expectation.

Lemma 3.13. *Let $A \in \mathbb{R}^{n\times m}$ such that $n \geq m$ and $Z \in \mathbb{R}^{n\times p}$ a Gaussian matrix of mean $\widehat{Z} \in \mathbb{R}^{n\times p}$ and covariance matrix $\mathbf{Cov}(Z)$, that is, $Z \sim \mathcal{N}(\widehat{Z}, \mathbf{Cov}(Z))$, satisfying $2 < p \leq \min\{\mathrm{rank}(A), \mathrm{rank}(\mathbf{Cov}(Z))\}$. For a given $k \in \{1,\dots,p-2\}$, set $\Omega_k = U_k^{\mathsf{T}}(Z-\widehat{Z}) \in \mathbb{R}^{k\times p}$, $\underline{\Omega}_k = \underline{U}_k^{\mathsf{T}}(Z-\widehat{Z}) \in \mathbb{R}^{(n-k)\times p}$ such that Ω_k is full row rank (i.e. $\mathrm{rank}(\Omega_k) = k$) and $T_k = \underline{\Omega}_k\Omega_k^{\dagger}$.*

If the covariance matrix $\mathbf{Cov}_k(Z)$ is nonsingular, then, for any matrix $N \in \mathbb{R}^{k\times k}$, one has

$$\mathbb{E}\left[\,\|T_kN\|_2\,\right] \leq \mathsf{c}_2^{tot}(\mathbf{Cov}(Z),N) := \mathsf{c}_2^{dep}(\mathbf{Cov}(Z),N) + \mathsf{c}_2(\mathbf{Cov}(Z),N), \tag{3.21}$$

and

$$\mathbb{E}\left[\,\|T_kN\|_F^2\,\right] = \mathsf{c}_F^{tot}(\mathbf{Cov}(Z),N) := \mathsf{c}_F^{dep}(\mathbf{Cov}(Z),N)^2 + \mathsf{c}_F(\mathbf{Cov}(Z),N)^2. \tag{3.22}$$

The (positive) constants are defined as

$$\mathsf{c}_{2,F}^{dep}(\mathbf{Cov}(Z),N) = \|\,\mathbf{Cov}_{\perp,k}(Z)[\mathbf{Cov}_k(Z)]^{-1}N\|_{2,F}, \tag{3.23}$$

$$\begin{aligned}\mathsf{c}_2(\mathbf{Cov}(Z),N) = &\frac{\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_2\|(N^{\mathsf{T}}[\mathbf{Cov}_k(Z)]^{-1}N)^{\frac{1}{2}}\|_F}{\sqrt{p-k-1}}\\ &+\frac{e\sqrt{p}}{p-k}\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_F\|(N^{\mathsf{T}}[\mathbf{Cov}_k(Z)]^{-1}N)^{\frac{1}{2}}\|_2,\end{aligned} \tag{3.24}$$

and

$$\mathsf{c}_F(\mathbf{Cov}(Z),N) = \frac{\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_F\|(N^{\mathsf{T}}[\mathbf{Cov}_k(Z)]^{-1}N)^{\frac{1}{2}}\|_F}{\sqrt{p-k-1}}, \tag{3.25}$$

where $\mathbf{Cov}_k(Z) = U_k^{\mathsf{T}}\,\mathbf{Cov}(Z)U_k$, $\mathbf{Cov}_{\perp,k}(Z) = \underline{U}_k^{\mathsf{T}}\,\mathbf{Cov}(Z)U_k$, $\underline{\mathbf{Cov}}_k(Z) = \underline{U}_k^{\mathsf{T}}\,\mathbf{Cov}(Z)\underline{U}_k$ and $\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right) = \underline{\mathbf{Cov}}_k(Z) - \mathbf{Cov}_{\perp,k}(Z)\left[\mathbf{Cov}_k(Z)\right]^{-1}\mathbf{Cov}_{\perp,k}(Z)^{\mathsf{T}}$.

Proof. In the spectral norm, Lemma 3.11 gives, for any matrix $N \in \mathbb{R}^{k\times k}$,

$$\begin{aligned}\mathbb{E}\left[\,\|\underline{\Omega}_k\Omega_k^{\dagger}N\|_2 \mid \Omega_k\,\right] \leq &\|\mathbb{E}\left[\,\underline{\Omega}_k \mid \Omega_k\,\right]\Omega_k^{\dagger}N\|_2 + \|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_2\|\Omega_k^{\dagger}N\|_F\\ &+\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_F\|\Omega_k^{\dagger}N\|_2.\end{aligned} \tag{3.26}$$

From Lemma 3.9, one has for any matrix $N \in \mathbb{R}^{k\times k}$,

$$\mathbb{E}\left[\,\underline{\Omega}_k \mid \Omega_k\,\right]\Omega_k^{\dagger}N = \mathbf{Cov}_{\perp,k}(Z)[\mathbf{Cov}_k(Z)]^{-1}\Omega_k\Omega_k^{\dagger}N = \mathbf{Cov}_{\perp,k}(Z)[\mathbf{Cov}_k(Z)]^{-1}N,$$

which is a deterministic constant. Hence, taking the total expectation of (3.26) leads to

$$\begin{aligned}\mathbb{E}\left[\,\|T_kN\|_2\,\right] = \mathbb{E}\left[\,\|\underline{\Omega}_k\Omega_k^{\dagger}N\|_2\,\right] &= \mathbb{E}\left[\,\mathbb{E}\left[\,\|\Omega_{\perp}\Omega_k^{\dagger}N\|_2 \mid \Omega_k\,\right]\right],\\ &\leq \|\,\mathbf{Cov}_{\perp,k}(Z)[\mathbf{Cov}_k(Z)]^{-1}N\|_2 + \|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_2\mathbb{E}\left[\,\|\Omega_k^{\dagger}N\|_F\,\right]\\ &\quad+\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_F\mathbb{E}\left[\,\|\Omega_k^{\dagger}N\|_2\,\right].\end{aligned}$$

Finally, by using Lemma 3.12, one has

$$\mathbb{E}\left[\,\|\Omega_k^{\dagger}N\|_F\,\right] \leq \mathbb{E}\left[\,\|\Omega_k^{\dagger}N\|_F^2\,\right]^{\frac{1}{2}} = \frac{\|(N^{\mathsf{T}}[\mathbf{Cov}_k(Z)]^{-1}N)^{\frac{1}{2}}\|_F^{\frac{1}{2}}}{\sqrt{p-k-1}}$$

and

$$\mathbb{E}\left[\,\|\Omega_k^\dagger N\|_2\,\right] \leq \frac{e\sqrt{p}}{p-k}\|(N^\mathsf{T}[\mathbf{Cov}_k(Z)]^{-1}N)^{\frac{1}{2}}\|_2.$$

This concludes the proof in the spectral norm case. The proof related to the Frobenius norm can be derived similarly. □

Remark 3.14. We note that both $\mathsf{c}_2^{\text{dep}}(\mathbf{Cov}(Z),N)$ and $\mathsf{c}_F^{\text{dep}}(\mathbf{Cov}(Z),N)$ do depend on $\mathbf{Cov}_{\perp,k}(Z)$, which is the term related to the statistical dependence between Ω_k and $\underline{\Omega}_k$. Those two terms cancel out whenever Ω_k and $\underline{\Omega}_k$ are independent. By contrast, $\mathsf{c}_2(\mathbf{Cov}(Z),N)$ and $\mathsf{c}_F(\mathbf{Cov}(Z),N)$ will not cancel out if Ω_k and $\underline{\Omega}_k$ are independent, but do approach zero as the number of samples p increases.

Remark 3.15. Following Theorem 1 in [77], Lemma 3.13 could serve to generalize the analysis of the singular vector accuracy.

Error bounds in expectation

We are now able to provide a first key result. Given $Z \sim \mathcal{N}(\widehat{Z},\mathbf{Cov}(Z))$, we aim at bounding (in expectation) the quantity $\|S_k\|_{2,F}$, where $S_k = (I_{n-k}+T_kT_k^\mathsf{T})^{-\frac{1}{2}}T_k$ and $T_k = \underline{\Omega}_k\Omega_k^\dagger$. We recall that the positive singular values of S_k represent the sine of the canonical angles between $\mathcal{R}\left(Z\Omega_k^\dagger\right)$ and $\mathcal{R}\left(U_k\right)$. This will make our proposed error bounds accurate in the presence of large canonical angles, compared to [53] where the analysis is based on the tangent of the canonical angles (via the matrix T_k). We highlight this result in the next proposition.

Proposition 3.16. *Let $A \in \mathbb{R}^{n\times m}$ such that $n \geq m$ and $Z \in \mathbb{R}^{n\times p}$ a Gaussian matrix of mean $\widehat{Z} \in \mathbb{R}^{n\times p}$ and covariance matrix $\mathbf{Cov}(Z)$, that is, $Z \sim \mathcal{N}(\widehat{Z},\mathbf{Cov}(Z))$, satisfying $2 < p \leq \min\left\{\text{rank}(A), \text{rank}(\mathbf{Cov}(Z))\right\}$. For a given $k \in \{1,\ldots,p-2\}$, set $\Omega_k = U_k^\mathsf{T}(Z-\widehat{Z}) \in \mathbb{R}^{k\times p}$, $\underline{\Omega}_k = \underline{U}_k^\mathsf{T}(Z-\widehat{Z}) \in \mathbb{R}^{(n-k)\times p}$ such that Ω_k is full row rank (i.e. $\text{rank}(\Omega_k)=k$), $T_k = \underline{\Omega}_k\Omega_k^\dagger \in \mathbb{R}^{(n-k)\times k}$ and $S_k = (I_{n-k}+T_kT_k^\mathsf{T})^{-\frac{1}{2}}T_k \in \mathbb{R}^{(n-k)\times k}$. Let $\varphi : x \mapsto x/\sqrt{1+x^2}$ for $x \geq 0$. We have*

$$\mathbb{E}\left[\,\|S_k\|_2\,\right] \leq \varphi\left(\mathsf{c}_2^{tot}(\mathbf{Cov}(Z),I_k)\right) \quad \text{and} \quad \mathbb{E}\left[\,\|S_k\|_F\,\right] \leq \sqrt{k}\varphi\left(\frac{1}{\sqrt{k}}\sqrt{\mathsf{c}_F^{tot}(\mathbf{Cov}(Z),I_k)}\right),$$

where the positive constants $\mathsf{c}_2^{tot}(\mathbf{Cov}(Z),I_k)$ and $\mathsf{c}_F^{tot}(\mathbf{Cov}(Z),I_k)$ are given in Lemma 3.13 (with $N = I_k$).

Proof. We define $\sigma_1(T_k) \geq \cdots \geq \sigma_k(T_k)$ (resp. $\sigma_1(S_k) \geq \cdots \geq \sigma_k(S_k)$) as the positive singular values of T_k (resp. S_k). Since φ is an increasing map, one has[1]

$$\sigma_i(S_k) = \frac{\sigma_i(T_k)}{\sqrt{1+\sigma_i(T_k)^2}} = \varphi\left(\sigma_i(T_k)\right), \quad i \in \{1,\ldots,k\}. \tag{3.27}$$

Then, we obtain

$$\|S_k\|_2 = \sigma_1(S_k) = \varphi\left(\sigma_1(T_k)\right) = \varphi(\|T_k\|_2). \tag{3.28}$$

Taking the expectation then leads to

$$\mathbb{E}\left[\,\|S_k\|_2\,\right] \leq \varphi\left(\mathbb{E}\left[\,\|T_k\|_2\,\right]\right) \leq \varphi\left(\mathsf{c}_2^{\text{tot}}(\mathbf{Cov}(Z),I_k)\right),$$

[1] Since the positive singular values of T_k represent the *tangent* of the canonical angles between $\mathcal{R}(Z\Omega_k^\dagger)$ and $\mathcal{R}(U_k)$, relation (3.27) shows that the positive singular values of S_k are the sine of the canonical angles between the same subspaces.

where the first inequality relies on both the concavity of φ and Jensen's inequality, while the second one uses (3.21) of Lemma 3.13 and the fact that φ is an increasing map. In the Frobenius norm, one has

$$\|S_k\|_F^2 = \mathrm{tr}(S_k^{\mathsf{T}} S_k) = \sum_{i=1}^{k} \sigma_i(S_k)^2 = \sum_{i=1}^{k} \frac{\sigma_i(T_k)^2}{1+\sigma_i(T_k)^2}.$$

We note that the scalar map $\psi : x \to \dfrac{x}{1+x}$ is concave (for all $x \geq 0$). Thus, the Jensen's inequality yields

$$\frac{1}{k}\|S_k\|_F^2 = \frac{1}{k}\sum_{i=1}^{k} \psi\left(\sigma_i(T_k)^2\right) \leq \psi\left(\frac{1}{k}\sum_{i=1}^{k}\sigma_i(T_k)^2\right) = \left[\varphi\left(\frac{1}{\sqrt{k}}\|T_k\|_F\right)\right]^2. \tag{3.29}$$

Then, by taking the expectation and exploiting both the concavity of ψ and (3.22) of Lemma 3.13, we finally obtain the result. □

By exploiting key projection perturbation results from [28], we are now able to extend our analysis to the general setting, where Z is drawn from a Gaussian distribution of mean $\widehat{Z} \in \mathbb{R}^{n\times p}$ and of covariance matrix $\mathbf{Cov}(Z)$. This is the second key result of our stochastic analysis proposed next.

Proposition 3.17. *Let $A \in \mathbb{R}^{n\times m}$ such that $n \geq m$ and $Z \in \mathbb{R}^{n\times p}$ a Gaussian matrix of mean $\widehat{Z} \in \mathbb{R}^{n\times p}$ and covariance matrix $\mathbf{Cov}(Z)$, that is, $Z \sim \mathcal{N}(\widehat{Z}, \mathbf{Cov}(Z))$, satisfying $2 < p \leq \min\{\mathrm{rank}(A), \mathrm{rank}(\mathbf{Cov}(Z))\}$. Let $\pi(Z)$ and $\pi(Z-\widehat{Z})$ denote the orthogonal projections onto the vector spaces spanned by the columns of Z and $Z-\widehat{Z}$, respectively.*

Then, for any $k \in \{1,\ldots,p-2\}$, one has

$$\mathbb{E}\left[\,\|[I_n-\pi(Z)]A\|_{2,F} - \|[I_n-\pi(Z)]\underline{A}_k\|_{2,F}\,\right] \leq \frac{e\sqrt{r}}{r-p}\frac{\|\widehat{Z}\|_2}{\sqrt{\lambda_r}}\|A_k\|_{2,F} + \mathbb{E}\left[\,\|[I_n-\pi(Z-\widehat{Z})]A_k\|_{2,F}\,\right],$$

where r denotes the rank of $\mathbf{Cov}(Z)$ and λ_r the smallest nonzero eigenvalue of $\mathbf{Cov}(Z)$.

Proof. First, we begin with the trivial observation that

$$\|[I_n-\pi(Z)]A\|_{2,F} - \|[I_n-\pi(Z)]\underline{A}_k\|_{2,F} \leq \|[I_n-\pi(Z)]A_k\|_{2,F}.$$

Let us define $W = Z-\widehat{Z}$, which is a centered Gaussian matrix with covariance matrix $\mathbf{Cov}(W) = \mathbf{Cov}(Z)$. It yields

$$\begin{aligned}
\|[I_n-\pi(Z)]A_k\|_{2,F} &= \|[I_n-\pi(W)+\pi(W)-\pi(Z)]A_k\|_{2,F},\\
&= \|[I_n-\pi(W)]A_k + [\pi(W)-\pi(Z)]A_k\|_{2,F},\\
&\leq \|[I_n-\pi(W)]A_k\|_{2,F} + \|[\pi(W)-\pi(Z)]A_k\|_{2,F},\\
&\leq \|[I_n-\pi(W)]A_k\|_{2,F} + \|[\pi(W)-\pi(Z)]\|_2\|A_k\|_{2,F},\\
&= \|[I_n-\pi(W)]A_k\|_{2,F} + \|[I_n-\pi(Z)]\pi(W)\|_2\|A_k\|_{2,F},
\end{aligned}$$

where the last inequality follows from [28, Lemma 1.5]. Then, adapting the proof of [28, Theorem 2.2], we observe that

$$\begin{aligned}
[I_n-\pi(Z)]\pi(W) &= [I_n-\pi(Z)]WW^{\dagger},\\
&= [I_n-\pi(Z)](W-Z+Z)W^{\dagger},\\
&= [I_n-\pi(Z)](W-Z)W^{\dagger} + [I_n-\pi(Z)]ZW^{\dagger},\\
&= [I_n-\pi(Z)](W-Z)W^{\dagger}.
\end{aligned}$$

Therefore, taking the spectral norm and using the submultiplicativity yields

$$\|[I_n - \pi(Z)]\pi(W)\|_2 \leq \|W - Z\|_2 \|W^\dagger\|_2 = \|\widehat{Z}\|_2 \|W^\dagger\|_2.$$

Overall, we have

$$\|[I_n - \pi(Z)]A_k\|_{2,F} \leq \|[I_n - \pi(W)]A_k\|_{2,F} + \|\widehat{Z}\|_2 \|W^\dagger\|_2 \|A_k\|_{2,F}.$$

Since $W \in \mathbb{R}^{n\times p}$ is full column rank, one has $W^\dagger = (W^\mathsf{T}W)^{-1}W^\mathsf{T}$ and due to the unitarily invariance of the spectral norm,

$$\|W^\dagger\|_2 = \|(W^\mathsf{T}W)^{-\frac{1}{2}}\|_2 = \sqrt{\|(W^\mathsf{T}W)^{-1}\|_2}.$$

Given $r = \operatorname{rank}(\mathbf{Cov}\, Z)$, the covariance matrix $\mathbf{Cov}(Z)$ has the following thin eigendecomposition, $\mathbf{Cov}(Z) = Q_r \Lambda_r Q_r^\mathsf{T}$ where $Q_r \in \mathbb{R}^{n\times r}$ is an orthogonal matrix containing the first λ_r eigenvectors of $\mathbf{Cov}(Z)$ and $\Lambda_r = \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$ contains the corresponding nonzero eigenvalues sorted in decreasing order, i.e. $\lambda_1 \geq \cdots \geq \lambda_r > 0$. Then, using (2.4), one has $W = \mathbf{Cov}(Z)^{\frac{1}{2}}G$ with $G \in \mathbb{R}^{n\times p}$ such that $G \sim \mathcal{N}(0, I_n)$. Thus,

$$W^\mathsf{T}W = G^\mathsf{T}\,\mathbf{Cov}(Z)G = G^\mathsf{T}Q_r\Lambda_r Q_r^\mathsf{T}G = Y_r^\mathsf{T}\Lambda_r Y_r \succcurlyeq \lambda_r Y_r^\mathsf{T}Y_r,$$

where $Y_r = Q_r^\mathsf{T}G \sim \mathcal{N}(0, I_r)$. Then, $(W^\mathsf{T}W)^{-1} \preccurlyeq \lambda_r^{-1}(Y_r^\mathsf{T}Y_r)^{-1}$ leading to

$$\|W^\dagger\|_2 \leq \frac{1}{\sqrt{\lambda_r}}\sqrt{\|(Y_r^\mathsf{T}Y_r)^{-1}\|_2} = \frac{1}{\sqrt{\lambda_r}}\|Y_r^\dagger\|_2 = \frac{1}{\sqrt{\lambda_r}}\|(Y_r^\mathsf{T})^\dagger\|_2,$$

with $Y_r^\mathsf{T} \in \mathbb{R}^{p\times r}$ (with $p \leq r$ by assumption) and $Y_r^\mathsf{T} \sim \mathcal{N}(0, I_p)$. We then take the expectation and apply Lemma 3.12 (with $N = I_r$) to deduce an upper bound of $\mathbb{E}\left[\,\|(Y_r^\mathsf{T})^\dagger\|_2\,\right]$, which completes the proof. □

We are now able to state the main theorems concerning the error bounds in expectation. Namely, given $Z \sim \mathcal{N}(\widehat{Z}, \mathbf{Cov}(Z))$, we target to bound the following quantity

$$\mathbb{E}\left[\,\|[I_n - \pi(Z)]A\|_{2,F} - \|[I_n - \pi(Z)]\underline{A}_k\|_{2,F}\,\right].$$

The proof of the next two theorems can be straightforwardly obtained by using the inequality $\sqrt{a^2+b^2} \leq a+b$ for any real positive scalars a and b, and then by taking the expectation of (3.8) of Theorem 3.4. The right-hand side terms are bounded using Lemma 3.13 (with $N = \Sigma_k$) and Propositions 3.16 and 3.17, respectively. We first state our error bound in the Frobenius norm case.

Theorem 3.18. *(Average analysis error bound in Frobenius norm) Let $A \in \mathbb{R}^{n\times m}$ such that $n \geq m$ and $Z \in \mathbb{R}^{n\times p}$ a Gaussian matrix of mean $\widehat{Z} \in \mathbb{R}^{n\times p}$ and covariance matrix $\mathbf{Cov}(Z)$, that is, $Z \sim \mathcal{N}(\widehat{Z}, \mathbf{Cov}(Z))$, satisfying $2 < p \leq \min\{\operatorname{rank}(A), \operatorname{rank}(\mathbf{Cov}(Z))\}$. Let $\pi(Z)$ denote the orthogonal projection onto the vector space spanned by the columns of Z. For a given $k \in \{1, \ldots, p-2\}$, set $\Omega_k = U_k^\mathsf{T}(Z - \widehat{Z}) \in \mathbb{R}^{k\times p}$, $\underline{\Omega}_k = \underline{U}_k^\mathsf{T}(Z - \widehat{Z}) \in \mathbb{R}^{(n-k)\times p}$ such that Ω_k is full row rank (i.e. $\operatorname{rank}(\Omega_k) = k$). Let $\varphi : x \mapsto x/\sqrt{1+x^2}$ for $x \geq 0$.*

If the covariance matrix $\mathbf{Cov}_k(Z)$ is nonsingular, then one has

$$\begin{aligned}\mathbb{E}\left[\,\|[I_n - \pi(Z)]A\|_F - \|[I_n - \pi(Z)]\underline{A}_k\|_F\,\right] \leq \\ \frac{e\sqrt{r}}{r-p}\frac{\|\widehat{Z}\|_2}{\sqrt{\lambda_r}}\|A_k\|_F + \min\left\{\sqrt{a_k},\ \sqrt{k}\,\varphi\left(\frac{1}{\sqrt{k}}\sqrt{b_k}\right)\|\Sigma_k\|_2\right\},\end{aligned} \tag{3.30}$$

where r denotes the rank of $\mathbf{Cov}(Z)$, λ_r the smallest nonzero eigenvalue of $\mathbf{Cov}(Z)$ and

$$a_k = \|\,\mathbf{Cov}_{\perp,k}(Z)[\mathbf{Cov}_k(Z)]^{-1}\Sigma_k\|_F^2 + \frac{\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_F^2\|(\Sigma_k^\mathsf{T}[\mathbf{Cov}_k(Z)]^{-1}\Sigma_k)^{\frac{1}{2}}\|_F^2}{p-k-1},$$

$$b_k = \|\,\mathbf{Cov}_{\perp,k}(Z)[\mathbf{Cov}_k(Z)]^{-1}\|_F^2 + \frac{\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_F^2\|[\mathbf{Cov}_k(Z)]^{-\frac{1}{2}}\|_F^2}{p-k-1},$$

with $\mathbf{Cov}_k(Z) = U_k^\mathsf{T}\,\mathbf{Cov}(Z)U_k$, $\mathbf{Cov}_{\perp,k}(Z) = \underline{U}_k^\mathsf{T}\,\mathbf{Cov}(Z)U_k$, $\underline{\mathbf{Cov}}_k(Z) = \underline{U}_k^\mathsf{T}\,\mathbf{Cov}(Z)\underline{U}_k$ and $\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right) = \underline{\mathbf{Cov}}_k(Z) - \mathbf{Cov}_{\perp,k}(Z)\left[\mathbf{Cov}_k(Z)\right]^{-1}\mathbf{Cov}_{\perp,k}(Z)^\mathsf{T}$.

Remark 3.19. We note that in the case of $\widehat{Z} = 0$, we are able to improve the tightness of our error bound in the Frobenius norm. More precisely, we have

$$\mathbb{E}\left[\,\|[I_n - \pi(Z)]A\|_F^2 - \|[I_n - \pi(Z)]\underline{A}_k\|_F^2\,\right] \leq \min\left\{a_k,\ k\,\varphi\left(\frac{1}{\sqrt{k}}\,\sqrt{b_k}\right)^2\|\Sigma_k\|_2^2\right\},$$

where a_k and b_k are defined in Theorem 3.18.

Similarly, we state our error bound in the spectral norm next.

Theorem 3.20. *(Average analysis error bound in spectral norm) Let $A \in \mathbb{R}^{n\times m}$ such that $n \geq m$ and $Z \in \mathbb{R}^{n\times p}$ a Gaussian matrix of mean $\widehat{Z} \in \mathbb{R}^{n\times p}$ and covariance matrix $\mathbf{Cov}(Z)$, that is, $Z \sim \mathcal{N}(\widehat{Z}, \mathbf{Cov}(Z))$, satisfying $2 < p \leq \min\left\{\mathrm{rank}(A), \mathrm{rank}(\mathbf{Cov}(Z))\right\}$. Let $\pi(Z)$ denote the orthogonal projection onto the vector space spanned by the columns of Z. For a given $k \in \{1,\ldots,p-2\}$, set $\Omega_k = U_k^\mathsf{T}(Z - \widehat{Z}) \in \mathbb{R}^{k\times p}$, $\underline{\Omega}_k = \underline{U}_k^\mathsf{T}(Z - \widehat{Z}) \in \mathbb{R}^{(n-k)\times p}$ such that Ω_k is full row rank (i.e. $\mathrm{rank}(\Omega_k) = k$). Let $\varphi : x \mapsto x/\sqrt{1+x^2}$ for $x \geq 0$.*

If the covariance matrix $\mathbf{Cov}_k(Z)$ is nonsingular, then one has

$$\mathbb{E}\left[\,\|[I_n - \pi(Z)]A\|_2 - \|[I_n - \pi(Z)]\underline{A}_k\|_2\,\right] \leq \frac{e\sqrt{r}}{r-p}\frac{\|\widehat{Z}\|_2}{\sqrt{\lambda_r}}\|A_k\|_2 + \min\left\{c_k,\ \varphi(d_k)\|\Sigma_k\|_2\right\}, \tag{3.31}$$

where r denotes the rank of $\mathbf{Cov}(Z)$, λ_r the smallest nonzero eigenvalue of $\mathbf{Cov}(Z)$ and

$$\begin{aligned} c_k &= \|\,\mathbf{Cov}_{\perp,k}(Z)[\mathbf{Cov}_k(Z)]^{-1}\Sigma_k\|_2 + \frac{\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_2\|(\Sigma_k^\mathsf{T}[\mathbf{Cov}_k(Z)]^{-1}\Sigma_k)^{\frac{1}{2}}\|_F}{\sqrt{p-k-1}} \\ &\quad + \frac{e\sqrt{p}}{p-k}\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_F\|(\Sigma_k^\mathsf{T}[\mathbf{Cov}_k(Z)]^{-1}\Sigma_k)^{\frac{1}{2}}\|_2, \\ d_k &= \|\,\mathbf{Cov}_{\perp,k}(Z)[\mathbf{Cov}_k(Z)]^{-1}\|_2 + \frac{\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_2\|[\mathbf{Cov}_k(Z)]^{-\frac{1}{2}}\|_F}{\sqrt{p-k-1}} \\ &\quad + \frac{e\sqrt{p}}{p-k}\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_F\|[\mathbf{Cov}_k(Z)]^{-\frac{1}{2}}\|_2, \end{aligned}$$

with $\mathbf{Cov}_k(Z) = U_k^\mathsf{T}\,\mathbf{Cov}(Z)U_k$, $\mathbf{Cov}_{\perp,k}(Z) = \underline{U}_k^\mathsf{T}\,\mathbf{Cov}(Z)U_k$, $\underline{\mathbf{Cov}}_k(Z) = \underline{U}_k^\mathsf{T}\,\mathbf{Cov}(Z)\underline{U}_k$ and $\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right) = \underline{\mathbf{Cov}}_k(Z) - \mathbf{Cov}_{\perp,k}(Z)\left[\mathbf{Cov}_k(Z)\right]^{-1}\mathbf{Cov}_{\perp,k}(Z)^\mathsf{T}$.

As in Section 3.3.1, if we target to bound the quantity of interest $\mathbb{E}\left[\,\|[I_n - \pi(Z)]A\|_2\,\right] - \|\underline{A}_k\|_2$ instead of $\mathbb{E}\left[\,\|[I_n - \pi(Z)]A\|_2 - \|[I_n - \pi(Z)]\underline{A}_k\|_2\,\right]$, it is then possible to significantly improve the error bound given in Theorem 3.20. The proof is similar to the proof of Theorem 3.20, exception made that we take the expectation of the result given in Theorem 3.8 and that the right-hand side terms are now bounded using Lemma 3.13 with the choice $N = \widehat{\Sigma}_k$. We give this result in the next theorem.

Theorem 3.21. *(Average analysis error bound in spectral norm, improved bound) Let $A \in \mathbb{R}^{n\times m}$ such that $n \geq m$ and $Z \in \mathbb{R}^{n\times p}$ a Gaussian matrix of mean $\widehat{Z} \in \mathbb{R}^{n\times p}$ and covariance matrix $\mathbf{Cov}(Z)$, that is, $Z \sim \mathcal{N}(\widehat{Z}, \mathbf{Cov}(Z))$, satisfying $2 < p \leq \min\{\mathrm{rank}(A), \mathrm{rank}(\mathbf{Cov}(Z))\}$. Let $\pi(Z)$ denote the orthogonal projection onto the vector space spanned by the columns of Z. For a given $k \in \{1, \ldots, p-2\}$, set $\Omega_k = U_k^{\mathsf{T}}(Z - \widehat{Z}) \in \mathbb{R}^{k\times p}$, $\underline{\Omega}_k = \underline{U}_k^{\mathsf{T}}(Z - \widehat{Z}) \in \mathbb{R}^{(n-k)\times p}$ such that Ω_k is full row rank (i.e. $\mathrm{rank}(\Omega_k) = k$). Let $\varphi : x \mapsto x/\sqrt{1+x^2}$ for $x \geq 0$.*

If the covariance matrix $\mathbf{Cov}_k(Z)$ is nonsingular, then, one has

$$\mathbb{E}\left[\, \|[I_n - \pi(Z)]A\|_2 \,\right] - \|\underline{A}_k\|_2 \leq \frac{e\sqrt{r}}{r-p}\frac{\|\widehat{Z}\|_2}{\sqrt{\lambda_r}}\|A_k\|_2 + \min\left\{\widehat{c}_k,\ \varphi(\widehat{d}_k)\|\widehat{\Sigma}_k\|_2\right\}, \tag{3.32}$$

where $\widehat{\Sigma}_k = \left(\Sigma_k^2 - \sigma_{k+1}^2 I_k\right)^{\frac{1}{2}}$, r denotes the rank of $\mathbf{Cov}(Z)$, λ_r the smallest nonzero eigenvalue of $\mathbf{Cov}(Z)$ and

$$\begin{aligned}
\widehat{c}_k &= \|\,\mathbf{Cov}_{\perp,k}(Z)[\mathbf{Cov}_k(Z)]^{-1}\widehat{\Sigma}_k\|_2 + \frac{\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_2\|(\widehat{\Sigma}_k^{\mathsf{T}}[\mathbf{Cov}_k(Z)]^{-1}\widehat{\Sigma}_k)^{\frac{1}{2}}\|_F}{\sqrt{p-k-1}} \\
&\quad + \frac{e\sqrt{p}}{p-k}\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_F\|(\widehat{\Sigma}_k^{\mathsf{T}}[\mathbf{Cov}_k(Z)]^{-1}\widehat{\Sigma}_k)^{\frac{1}{2}}\|_2, \\
\widehat{d}_k &= \|\,\mathbf{Cov}_{\perp,k}(Z)[\mathbf{Cov}_k(Z)]^{-1}\|_2 + \frac{\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_2\|[\mathbf{Cov}_k(Z)]^{-\frac{1}{2}}\|_F}{\sqrt{p-k-1}} \\
&\quad + \frac{e\sqrt{p}}{p-k}\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_F\|[\mathbf{Cov}_k(Z)]^{-\frac{1}{2}}\|_2,
\end{aligned}$$

with $\mathbf{Cov}_k(Z) = U_k^{\mathsf{T}}\,\mathbf{Cov}(Z)U_k$, $\mathbf{Cov}_{\perp,k}(Z) = \underline{U}_k^{\mathsf{T}}\,\mathbf{Cov}(Z)U_k$ and $\underline{\mathbf{Cov}}_k(Z) = \underline{U}_k^{\mathsf{T}}\,\mathbf{Cov}(Z)\underline{U}_k$ and $\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right) = \underline{\mathbf{Cov}}_k(Z) - \mathbf{Cov}_{\perp,k}(Z)\left[\mathbf{Cov}_k(Z)\right]^{-1}\mathbf{Cov}_{\perp,k}(Z)^{\mathsf{T}}$.

Remark 3.22. Since $\|\widehat{\Sigma}_k\|_2 \leq \|\Sigma_k\|_2$ and due to the partial ordering property, we deduce that $\widehat{c}_k \leq c_k$ and remark that $\widehat{d}_k = d_k$. Hence if one targets to bound $\mathbb{E}\left[\, \|[I_n - \pi(Z)]A\|_2 \,\right] - \|\underline{A}_k\|_2$, then the error bound (3.32) is tighter than (3.31). Hence Theorem 3.21 provides a new error bound in the spectral norm in a fairly general setting.

We provide in Section 3.5 numerical illustrations to highlight the potential of the error bounds.

3.3.3 Analysis in probability

To complement the theoretical analysis given in Section 3.3.2, we provide probabilistic error bounds in both the Frobenius and the spectral norms. Then we consider the Randomized Singular Value Decomposition and detail in Section 3.4 the new constants involved in the probabilistic error bounds. Preparatory technical lemmas are first given.

Preparatory lemmas

The first two lemmas generalize Propositions 10.3 and 10.4 of [53] to the non-standard Gaussian case. We first provide a result related to concentration inequalities for functions of a non-standard Gaussian matrix.

Lemma 3.23 (Concentration for functions of a non-standard Gaussian matrix)**.** *Let h be a real function on matrices in $\mathbb{R}^{k\times p}$ satisfying the following Lipschitz condition for a given positive constant L*

$$\left|h(X) - h(Y)\right| \leq L\ \|X - Y\|_F, \quad \forall\ X, Y \in \mathbb{R}^{k\times p}.$$

Let $M \in \mathbb{R}^{k\times p}$ be a Gaussian matrix such that $M \sim \mathcal{N}(\widehat{M}, \mathbf{Cov}(M))$ with mean $\widehat{M} \in \mathbb{R}^{k\times p}$ and covariance matrix $\mathbf{Cov}(M) \in \mathbb{R}^{k\times k}$. Then for all $t \geq 0$ one has

$$\mathbb{P}\left\{h(M) \geq \mathbb{E}\left[h(M)\right] + L\sqrt{\|\mathbf{Cov}(M)\|_2}\cdot t\right\} \leq e^{-t^2/2}. \tag{3.33}$$

Proof. By the definition of a non-standard Gaussian matrix (2.4), we have $M = \widehat{M} + \mathbf{Cov}(M)^{\frac{1}{2}}G$ with $G \in \mathbb{R}^{k\times p}$ a standard Gaussian matrix (i.e. $G \sim \mathcal{N}(0, I_k)$). We define the real function g on matrices in $\mathbb{R}^{k\times p}$ such that $g : X \mapsto h(\widehat{M} + \mathbf{Cov}(M)^{\frac{1}{2}}X)$, implying that $g(G)$ and $h(M)$ have the same distribution. Since h is L-Lipschitz by assumption, we have

$$\left|g(X) - g(Y)\right| \leq L\,\|\mathbf{Cov}(M)^{\frac{1}{2}}(X-Y)\|_F \leq L\,\|\mathbf{Cov}(M)^{\frac{1}{2}}\|_2\|X-Y\|_F,\ \forall\ X, Y \in \mathbb{R}^{k\times p}.$$

This shows that g is a $L\,\|\mathbf{Cov}(M)^{\frac{1}{2}}\|_2$-Lipschitz function. Applying [53, Proposition 10.3] to g then gives ((3.33)). □

Secondly, we state a result related to large deviation bounds for the norm of a pseudo-inverted non-standard Gaussian matrix.

Lemma 3.24 (Norm bounds for a pseudoinverted non-standard Gaussian matrix)**.** *Let $N \in \mathbb{R}^{k\times k}$ be a given matrix and let $M \in \mathbb{R}^{k\times p}$ with $p \geq k+4$ be a Gaussian matrix such that $M \sim \mathcal{N}(0, \mathbf{Cov}(M))$ with covariance matrix $\mathbf{Cov}(M) \in \mathbb{R}^{k\times k}$. Then if $\mathbf{Cov}(M)$ is invertible, the spectral (resp. Frobenius) norm of a pseudoinverted non-standard Gaussian matrix satisfies for all $t \geq 1$,*

$$\mathbb{P}\left\{\|M^\dagger N\|_2 \leq \sqrt{\|N^\mathsf{T}\,\mathbf{Cov}(M)^{-1}N\|_2}\,\frac{e\sqrt{p}}{p-k+1}t\right\} \geq 1 - t^{-(p-k+1)}, \tag{3.34}$$

and

$$\mathbb{P}\left\{\|M^\dagger N\|_F \leq \sqrt{\|N^\mathsf{T}\,\mathbf{Cov}(M)^{-1}N\|_2}\,\frac{\sqrt{3k}}{\sqrt{p-k+1}}t\right\} \geq 1 - t^{-(p-k)}, \tag{3.35}$$

respectively.

Proof. We write $M = \mathbf{Cov}(M)^{\frac{1}{2}}G$ with $G \in \mathbb{R}^{n\times k}$ a standard Gaussian matrix and use the invertibility of $\mathbf{Cov}(M)$ to obtain

$$\|M^\dagger N\|_{2,F} \leq \|G^\dagger\|_{2,F}\|\,\mathbf{Cov}(M)^{-\frac{1}{2}}N\|_2 = \|G^\dagger\|_{2,F}\sqrt{\|N^\mathsf{T}\,\mathbf{Cov}(M)^{-1}N\|_2}.$$

Applying [53, Proposition 10.4] to $\|G^\dagger\|_{2,F}$ allows us to deduce ((3.34)) and ((3.35)), respectively. □

Then we are now able to provide a probabilistic counterpart of Lemma 3.13.

Lemma 3.25. *Let $A \in \mathbb{R}^{n\times m}$ such that $n \geq m$ and $Z \in \mathbb{R}^{n\times p}$ a Gaussian matrix of mean $\widehat{Z} \in \mathbb{R}^{n\times p}$ and covariance matrix $\mathbf{Cov}(Z)$, that is, $Z \sim \mathcal{N}(\widehat{Z}, \mathbf{Cov}(Z))$, satisfying $4 < p \leq \min\left\{\mathrm{rank}(A), \mathrm{rank}(\mathbf{Cov}(Z))\right\}$. For a given $k \in \{1, \ldots, p-4\}$, set $\Omega_k = U_k^\mathsf{T}(Z - \widehat{Z}) \in \mathbb{R}^{k\times p}$, $\underline{\Omega}_k = \underline{U}_k^\mathsf{T}(Z - \widehat{Z}) \in \mathbb{R}^{(n-k)\times p}$ such that Ω_k is full row rank (i.e. $\mathrm{rank}(\Omega_k) = k$) and $T_k = \underline{\Omega}_k\Omega_k^\dagger$.*

If the covariance matrix $\mathbf{Cov}_k(Z)$ *is nonsingular, then, for any matrix* $N \in \mathbb{R}^{k\times k}$ *and all* $u, t \geq 1$*, it holds with probability of failure at most* $e^{-u^2/2} + t^{-(p-k+1)}$

$$\|T_k N\|_2 \leq \mathsf{c}_2^{tot}(\mathbf{Cov}(Z), N) + \mathsf{c}_2^{prob}(\mathbf{Cov}(Z), N) \cdot tu \tag{3.36}$$

with

$$\mathsf{c}_2^{prob}(\mathbf{Cov}(Z), N) = \|\underline{\mathbf{Cov}}_k(Z)^{\frac{1}{2}}\|_2 \|(N^{\mathsf{T}}[\mathbf{Cov}_k(Z)]^{-1}N)^{\frac{1}{2}}\|_2 \frac{e\sqrt{p}}{p-k+1}. \tag{3.37}$$

Similarly, for all $u, t \geq 1$*, it holds with probability of failure at most* $e^{-u^2/2} + t^{-(p-k)}$

$$\|T_k N\|_F \leq \sqrt{\mathsf{c}_F^{tot}(\mathbf{Cov}(Z), N)} + \mathsf{c}_F^{prob}(\mathbf{Cov}(Z), N) \cdot tu, \tag{3.38}$$

with

$$\mathsf{c}_F^{prob}(\mathbf{Cov}(Z), N) = \|\underline{\mathbf{Cov}}_k(Z)^{\frac{1}{2}}\|_2 \|(N^{\mathsf{T}}[\mathbf{Cov}_k(Z)]^{-1}N)^{\frac{1}{2}}\|_F \frac{\sqrt{3k}}{\sqrt{p-k+1}}, \tag{3.39}$$

where the positive constants $\mathsf{c}_2^{tot}(\mathbf{Cov}(Z), N)$ *and* $\mathsf{c}_F^{tot}(\mathbf{Cov}(Z), N)$ *are defined in Lemma 3.13, respectively.*

Proof. We first consider the case of the spectral norm. We define the following event for $t \geq 1$

$$E_t = \left\{ \Omega_k \mid \|\Omega_k^\dagger N\|_2 \leq \|(N^{\mathsf{T}}[\mathbf{Cov}_k(Z)]^{-1}N)^{\frac{1}{2}}\|_2 \frac{e\sqrt{p}}{p-k+1} \cdot t \right\}.$$

Using relation (3.34) of Lemma 3.24, we obtain $\mathbb{P}(E_t) \geq 1 - t^{-(p-k+1)}$. We define the function h on matrices in $\mathbb{R}^{(n-k)\times p}$ such that $h : Y \mapsto \|Y\Omega_k^\dagger N\|_2$. In order to use Lemma 3.23, we must compute the Lipschitz constant related to h and $\mathbb{E}\left[h(\underline{\Omega}_k)\right]$. Using the reverse triangle inequality, we get

$$\left|h(Y_1) - h(Y_2)\right| \leq \|(Y_1 - Y_2)\Omega_k^\dagger N\|_2 \leq \|\Omega_k^\dagger N\|_2 \|Y_1 - Y_2\|_2 \leq \|\Omega_k^\dagger N\|_2 \|Y_1 - Y_2\|_F, \ \forall Y_1, Y_2 \in \mathbb{R}^{(n-k)\times p}.$$

Therefore, h is at least a $\|\Omega_k^\dagger N\|_2$-Lipschitz function. The application of Lemma 3.13 gives

$$\mathbb{E}\left[h(\underline{\Omega}_k)\right] \leq \mathsf{c}_2^{\mathrm{tot}}(\mathbf{Cov}(Z), N).$$

Since $\underline{\Omega}_k \sim \mathcal{N}(0, \underline{\mathbf{Cov}}_k(Z))$, applying Lemma 3.23 gives for all $u \geq 1$

$$\mathbb{P}\left\{ \|T_k N\|_2 \geq \mathsf{c}_2^{\mathrm{tot}}(\mathbf{Cov}(Z), N) + \|\underline{\mathbf{Cov}}_k(Z)^{\frac{1}{2}}\|_2 \|\Omega_k^\dagger N\|_2 \cdot u \right\} \leq e^{-u^2/2}.$$

The law of total probability then reads

$$\mathbb{P}\left\{ \|T_k N\|_2 \geq \mathsf{c}_2^{\mathrm{tot}}(\mathbf{Cov}(Z), N) + \|\underline{\mathbf{Cov}}_k(Z)^{\frac{1}{2}}\|_2 \|\Omega_k^\dagger N\|_2 \cdot u \mid E_t \right\} \mathbb{P}\{E_t\} \leq e^{-u^2/2}.$$

We then define $\mathsf{c}_2^{\mathrm{prob}}(\mathbf{Cov}(Z), N)$ as in ((3.37)).Under the event E_t, we have

$$\|\Omega_k^\dagger N\|_2 \leq \|(N^{\mathsf{T}}[\mathbf{Cov}_k(Z)]^{-1}N)^{\frac{1}{2}}\|_2 \frac{e\sqrt{p}}{p-k+1} \cdot t,$$

which leads to

$$\mathbb{P}\left\{\|T_k N\|_2 \geq \mathsf{c}_2^{\mathrm{tot}}(\mathbf{Cov}(Z), N) + \mathsf{c}_2^{\mathrm{prob}}(\mathbf{Cov}(Z), N) \cdot tu \mid E_t\right\} \mathbb{P}\{E_t\} \leq e^{-u^2/2}.$$

We use the law of total probability to remove the conditioning and finally obtain

$$\begin{aligned}\mathbb{P}\left\{\|T_k N\|_2 \geq \mathsf{c}_2^{\mathrm{tot}}(\mathbf{Cov}(Z), N) + \mathsf{c}_2^{\mathrm{prob}}(\mathbf{Cov}(Z), N) \cdot tu\right\} &\leq e^{-u^2/2} + \mathbb{P}\{E_t^c\}, \\ &\leq e^{-u^2/2} + t^{-(p-k+1)}.\end{aligned}$$

The proof for the Frobenius norm follows similar arguments using the event for $t \geq 1$

$$E_t = \left\{\Omega_k \mid \|\Omega_k^\dagger N\|_F \leq \|(N^\mathsf{T}[\mathbf{Cov}_k(Z)]^{-1}N)^{\frac{1}{2}}\|_2 \frac{\sqrt{3k}}{\sqrt{p-k+1}} \cdot t\right\},$$

and the following function h defined on matrices in $\mathbb{R}^{(n-k)\times p}$ such that $Y \mapsto \|Y\Omega_k^\dagger N\|_F$. Using successively Hölder's inequality and Lemma 3.13 then gives

$$\mathbb{E}\left[h(\underline{\Omega}_k)\right] \leq \mathbb{E}\left[h(\underline{\Omega}_k)^2\right]^{\frac{1}{2}} = \sqrt{\mathsf{c}_F^{\mathrm{tot}}(\mathbf{Cov}(Z), N)}. \tag{3.40}$$

□

We end this section by stating a final proposition which later allows us to derive the error bounds in probability in the next section.

Proposition 3.26. *Let $A \in \mathbb{R}^{n\times m}$ such that $n \geq m$ and $Z \in \mathbb{R}^{n\times p}$ a Gaussian matrix of mean $\widehat{Z} \in \mathbb{R}^{n\times p}$ and covariance matrix $\mathbf{Cov}(Z)$, that is, $Z \sim \mathcal{N}(\widehat{Z}, \mathbf{Cov}(Z))$, satisfying $4 < p \leq \min\{\mathrm{rank}(A), \mathrm{rank}(\mathbf{Cov}(Z))\}$. For a given $k \in \{1, \ldots, p-4\}$, set $\Omega_k = U_k^\mathsf{T}(Z - \widehat{Z}) \in \mathbb{R}^{k\times p}$, $\underline{\Omega}_k = \underline{U}_k^\mathsf{T}(Z - \widehat{Z}) \in \mathbb{R}^{(n-k)\times p}$ such that Ω_k is full row rank (i.e. $\mathrm{rank}(\Omega_k) = k$) and $T_k = \underline{\Omega}_k\Omega_k^\dagger$ and $S_k = (I_{n-k} + T_k T_k^\mathsf{T})^{-\frac{1}{2}} T_k$. Let $\varphi : x \mapsto x/\sqrt{1+x^2}$ for $x \geq 0$.*

If the covariance matrix $\mathbf{Cov}_k(Z)$ is nonsingular, then for any $t, u \geq 1$ it holds with probability of failure at most $e^{-u^2/2} + t^{-(p-k+1)}$

$$\|S_k\|_2 \leq \varphi\left(\mathsf{c}_2^{tot}(\mathbf{Cov}(Z), I_k) + \mathsf{c}_2^{prob}(\mathbf{Cov}(Z), I_k) \cdot tu\right), \tag{3.41}$$

and with a probability of failure at most $e^{-u^2/2} + t^{-(p-k)}$

$$\|S_k\|_F \leq \sqrt{k}\ \varphi\left(\frac{1}{\sqrt{k}}\ \left[\sqrt{\mathsf{c}_F^{tot}(\mathbf{Cov}(Z), I_k)} + \mathsf{c}_F^{prob}(\mathbf{Cov}(Z), I_k) \cdot tu\right]\right). \tag{3.42}$$

The positive constants c_2^{tot} and c_F^{tot} are given in Lemma 3.13 (with $N = I_k$), while c_2^{prob} and c_F^{prob} are given in Lemma 3.25 (with $N = I_k$), respectively.

Proof. We first consider the case of the spectral norm. Due to Lemma 3.25 (relation ((3.36)) with $N = I_k$), the inequality

$$\|T_k\|_2 \leq \mathsf{c}_2^{\mathrm{tot}}(\mathbf{Cov}(Z), I_k) + \mathsf{c}_2^{\mathrm{prob}}(\mathbf{Cov}(Z), I_k) \cdot tu$$

holds with probability of failure at most $e^{-u^2/2} + t^{-(p-k+1)}$. Then, we use the relation ((3.28)) and the fact that φ is an increasing and bijective map to obtain ((3.41)). In the Frobenius case, we deduce the following inequality from relation ((3.29))

$$\|S_k\|_F \leq \sqrt{k}\ \varphi\left(\frac{1}{\sqrt{k}}\|T_k\|_F\right).$$

The proof then follows similar arguments as in the spectral norm case, by first bounding $\|T_k\|_F$ using Lemma 3.25 (relation ((3.38)) with $N = I_k$). □

Error bounds in probability

We now provide the corresponding probabilistic error bounds related to Theorems 3.18, 3.20 and 3.21, respectively. The probabilistic error bounds can be easily derived from Lemma 3.25 and Proposition 3.26. Probabilistic error bounds for the approximation error in the Frobenius norm are provided in Theorem 3.27, which extends Theorem 3.18.

Theorem 3.27. *Probabilitic error bound in Frobenius norm Let $A \in \mathbb{R}^{n\times m}$ such that $n \geq m$ and $Z \in \mathbb{R}^{n\times p}$ a Gaussian matrix of mean $\widehat{Z} \in \mathbb{R}^{n\times p}$ and covariance matrix $\mathbf{Cov}(Z)$, that is, $Z \sim \mathcal{N}(\widehat{Z}, \mathbf{Cov}(Z))$, satisfying $4 < p \leq \min\left\{\mathrm{rank}(A), \mathrm{rank}(\mathbf{Cov}(Z))\right\}$. For a given $k \in \{1, \ldots, p-4\}$, set $\Omega_k = U_k^{\mathsf{T}}(Z - \widehat{Z}) \in \mathbb{R}^{k\times p}$, $\underline{\Omega}_k = \underline{U}_k^{\mathsf{T}}(Z - \widehat{Z}) \in \mathbb{R}^{(n-k)\times p}$ such that Ω_k is full row rank (i.e. $\mathrm{rank}(\Omega_k) = k$) and $T_k = \underline{\Omega}_k \Omega_k^{\dagger}$.*

If the covariance matrix $\mathbf{Cov}_k(Z)$ is nonsingular, then, for all $u, t \geq 1$, it holds with probability of failure at most $e^{-u^2/2} + t^{-(p-k)}$

$$\|[I_n - \pi(Z)]A\|_F - \|[I_n - \pi(Z)]\underline{A}_k\|_F \quad \leq \quad \min\left\{\sqrt{a_k} + \alpha_k\ tu,\ \sqrt{k}\ \varphi\left(\frac{1}{\sqrt{k}}\ [\sqrt{b_k} + \beta_k\ tu]\right)\|\Sigma_k\|_2\right\},$$

where

$$a_k = \|\,\mathbf{Cov}_{\perp,k}(Z)[\mathbf{Cov}_k(Z)]^{-1}\Sigma_k\|_F^2 + \frac{\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_F^2 \|(\Sigma_k^{\mathsf{T}}[\mathbf{Cov}_k(Z)]^{-1}\Sigma_k)^{\frac{1}{2}}\|_F^2}{p-k-1},$$

$$b_k = \|\,\mathbf{Cov}_{\perp,k}(Z)[\mathbf{Cov}_k(Z)]^{-1}\|_F^2 + \frac{\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_F^2 \|[\mathbf{Cov}_k(Z)]^{-\frac{1}{2}}\|_F^2}{p-k-1},$$

$$\alpha_k = \|\,\underline{\mathbf{Cov}}_k(Z)^{\frac{1}{2}}\|_2 \|(\Sigma_k^{\mathsf{T}}[\mathbf{Cov}_k(Z)]^{-1}\Sigma_k)^{\frac{1}{2}}\|_F \frac{\sqrt{3k}}{\sqrt{p-k+1}},$$

$$\beta_k = \|\,\underline{\mathbf{Cov}}_k(Z)^{\frac{1}{2}}\|_2 \|[\mathbf{Cov}_k(Z)]^{-\frac{1}{2}}\|_F \frac{\sqrt{3k}}{\sqrt{p-k+1}}.$$

with $\mathbf{Cov}_k(Z) = U_k^{\mathsf{T}}\,\mathbf{Cov}(Z)U_k$, and $\mathbf{Cov}_{\perp,k}(Z) = \underline{U}_k^{\mathsf{T}}\,\mathbf{Cov}(Z)U_k$, $\underline{\mathbf{Cov}}_k(Z) = \underline{U}_k^{\mathsf{T}}\,\mathbf{Cov}(Z)\underline{U}_k$, $\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right) = \underline{\mathbf{Cov}}_k(Z) - \mathbf{Cov}_{\perp,k}(Z)\left[\mathbf{Cov}_k(Z)\right]^{-1}\mathbf{Cov}_{\perp,k}(Z)^{\mathsf{T}}$.

Probabilistic error bounds for the approximation error in the spectral norm are provided in Theorems 3.28 and 3.29, respectively. They are extensions of Theorems 3.20 and 3.21, respectively.

Theorem 3.28. *Probabilitic error bound in spectral norm Let $A \in \mathbb{R}^{n\times m}$ such that $n \geq m$ and $Z \in \mathbb{R}^{n\times p}$ a Gaussian matrix of mean $\widehat{Z} \in \mathbb{R}^{n\times p}$ and covariance matrix $\mathbf{Cov}(Z)$, that*

is, $Z \sim \mathcal{N}(\widehat{Z}, \mathbf{Cov}(Z))$*, satisfying* $4 < p \leq \min\left\{\mathrm{rank}(A),\, \mathrm{rank}(\mathbf{Cov}(Z))\right\}$*. For a given* $k \in \{1, \ldots, p-4\}$*, set* $\Omega_k = U_k^{\mathsf{T}}(Z - \widehat{Z}) \in \mathbb{R}^{k\times p}$*,* $\underline{\Omega}_k = \underline{U}_k^{\mathsf{T}}(Z - \widehat{Z}) \in \mathbb{R}^{(n-k)\times p}$ *such that* Ω_k *is full row rank (i.e.* $\mathrm{rank}(\Omega_k) = k$*) and* $T_k = \underline{\Omega}_k \Omega_k^{\dagger}$*.*

If the covariance matrix $\mathbf{Cov}_k(Z)$ *is nonsingular, then, for all* $u, t \geq 1$*, it holds with probability of failure at most* $e^{-u^2/2} + t^{-(p-k+1)}$

$$\|[I_n - \pi(Z)]A\|_2 - \|[I_n - \pi(Z)]\underline{A}_k\|_2 \leq \min\left\{c_k + \theta_k\, tu, \varphi\left(d_k + \delta_k\, tu\right)\|\Sigma_k\|_2\right\},$$

where

$$\begin{aligned}
c_k &= \|\,\mathbf{Cov}_{\perp,k}(Z)[\mathbf{Cov}_k(Z)]^{-1}\Sigma_k\|_2 + \frac{\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_2 \|(\Sigma_k^{\mathsf{T}}[\mathbf{Cov}_k(Z)]^{-1}\Sigma_k)^{\frac{1}{2}}\|_F}{\sqrt{p-k-1}} \\
&\quad + \frac{e\sqrt{p}}{p-k}\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_F \|(\Sigma_k^{\mathsf{T}}[\mathbf{Cov}_k(Z)]^{-1}\Sigma_k)^{\frac{1}{2}}\|_2, \\
d_k &= \|\,\mathbf{Cov}_{\perp,k}(Z)[\mathbf{Cov}_k(Z)]^{-1}\|_2 + \frac{\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_2 \|[\mathbf{Cov}_k(Z)]^{-\frac{1}{2}}\|_F}{\sqrt{p-k-1}} \\
&\quad + \frac{e\sqrt{p}}{p-k}\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_F \|[\mathbf{Cov}_k(Z)]^{-\frac{1}{2}}\|_2, \\
\theta_k &= \|\,\underline{\mathbf{Cov}}_k(Z)^{\frac{1}{2}}\|_2 \|(\Sigma_k^{\mathsf{T}}[\mathbf{Cov}_k(Z)]^{-1}\Sigma_k)^{\frac{1}{2}}\|_2 \frac{e\sqrt{p}}{p-k+1}, \\
\delta_k &= \|\,\underline{\mathbf{Cov}}_k(Z)^{\frac{1}{2}}\|_2 \|[\mathbf{Cov}_k(Z)]^{-1})^{\frac{1}{2}}\|_2 \frac{e\sqrt{p}}{p-k+1},
\end{aligned}$$

with $\mathbf{Cov}_k(Z) = U_k^{\mathsf{T}}\,\mathbf{Cov}(Z)U_k$*, and* $\mathbf{Cov}_{\perp,k}(Z) = \underline{U}_k^{\mathsf{T}}\,\mathbf{Cov}(Z)U_k$*,* $\underline{\mathbf{Cov}}_k(Z) = \underline{U}_k^{\mathsf{T}}\,\mathbf{Cov}(Z)\underline{U}_k$*,* $\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right) = \underline{\mathbf{Cov}}_k(Z) - \mathbf{Cov}_{\perp,k}(Z)\left[\mathbf{Cov}_k(Z)\right]^{-1}\mathbf{Cov}_{\perp,k}(Z)^{\mathsf{T}}$*.*

Theorem 3.29. *Probabilitic error bound in spectral norm, improved bound Let* $A \in \mathbb{R}^{n\times m}$ *such that* $n \geq m$ *and* $Z \in \mathbb{R}^{n\times p}$ *a Gaussian matrix of mean* $\widehat{Z} \in \mathbb{R}^{n\times p}$ *and covariance matrix* $\mathbf{Cov}(Z)$*, that is,* $Z \sim \mathcal{N}(\widehat{Z}, \mathbf{Cov}(Z))$*, satisfying* $4 < p \leq \min\left\{\mathrm{rank}(A),\, \mathrm{rank}(\mathbf{Cov}(Z))\right\}$*. For a given* $k \in \{1, \ldots, p-4\}$*, set* $\Omega_k = U_k^{\mathsf{T}}(Z - \widehat{Z}) \in \mathbb{R}^{k\times p}$*,* $\underline{\Omega}_k = \underline{U}_k^{\mathsf{T}}(Z - \widehat{Z}) \in \mathbb{R}^{(n-k)\times p}$ *such that* Ω_k *is full row rank (i.e.* $\mathrm{rank}(\Omega_k) = k$*) and* $T_k = \underline{\Omega}_k \Omega_k^{\dagger}$*.*

If the covariance matrix $\mathbf{Cov}_k(Z)$ *is nonsingular, then, for all* $u, t \geq 1$*, it holds with probability of failure at most* $e^{-u^2/2} + t^{-(p-k+1)}$

$$\|[I_n - \pi(Z)]A\|_2 - \|\underline{A}_k\|_2 \leq \min\left\{\widehat{c}_k + \widehat{\theta}_k\; tu,\; \varphi\left(\widehat{d}_k + \widehat{\delta}_k\; tu\right)\|\widehat{\Sigma}_k\|_2\right\},$$

where $\widehat{\Sigma}_k = \left(\Sigma_k^2 - \sigma_{k+1}^2 I_k\right)^{\frac{1}{2}}$ *and*

$$\widehat{c}_k = \|\,\mathbf{Cov}_{\perp,k}(Z)[\mathbf{Cov}_k(Z)]^{-1}\widehat{\Sigma}_k\|_2 + \frac{\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_2\|(\widehat{\Sigma}_k^{\mathsf{T}}[\mathbf{Cov}_k(Z)]^{-1}\widehat{\Sigma}_k)^{\frac{1}{2}}\|_F}{\sqrt{p-k-1}}$$
$$+ \frac{e\sqrt{p}}{p-k}\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_F\|(\widehat{\Sigma}_k^{\mathsf{T}}[\mathbf{Cov}_k(Z)]^{-1}\widehat{\Sigma}_k)^{\frac{1}{2}}\|_2,$$

$$d_k = \|\,\mathbf{Cov}_{\perp,k}(Z)[\mathbf{Cov}_k(Z)]^{-1}\|_2 + \frac{\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_2\|[\mathbf{Cov}_k(Z)]^{-\frac{1}{2}}\|_F}{\sqrt{p-k-1}}$$
$$+ \frac{e\sqrt{p}}{p-k}\|\,\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_F\|[\mathbf{Cov}_k(Z)]^{-\frac{1}{2}}\|_2,$$

$$\widehat{\theta}_k = \|\,\underline{\mathbf{Cov}}_k(Z)^{\frac{1}{2}}\|_2\|(\widehat{\Sigma}_k^{\mathsf{T}}[\mathbf{Cov}_k(Z)]^{-1}\widehat{\Sigma}_k)^{\frac{1}{2}}\|_2\frac{e\sqrt{p}}{p-k+1},$$

$$\delta_k = \|\,\underline{\mathbf{Cov}}_k(Z)^{\frac{1}{2}}\|_2\|[\mathbf{Cov}_k(Z)]^{-1}\|_2\frac{e\sqrt{p}}{p-k+1}.$$

with $\mathbf{Cov}_k(Z) = U_k^{\mathsf{T}}\,\mathbf{Cov}(Z)U_k$, *and* $\mathbf{Cov}_{\perp,k}(Z) = \underline{U}_k^{\mathsf{T}}\,\mathbf{Cov}(Z)U_k$, $\underline{\mathbf{Cov}}_k(Z) = \underline{U}_k^{\mathsf{T}}\,\mathbf{Cov}(Z)\underline{U}_k$, $\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right) = \underline{\mathbf{Cov}}_k(Z) - \mathbf{Cov}_{\perp,k}(Z)\left[\mathbf{Cov}_k(Z)\right]^{-1}\mathbf{Cov}_{\perp,k}(Z)^{\mathsf{T}}$.

Remark 3.30. Unlike the bounds for the expected low-rank approximation error, the derivation of the probability bounds required to step aside from the proofs in [53]. Consequently, we cannot consider the results from Theorems 3.27, 3.28 and 3.29 as generalizations of Theorems 10.7 and 10.8 in [53].

3.4 Application to the Randomized Singular Value Decomposition

As an illustration, we consider the Randomized Singular Value Decomposition (RSVD), a popular algorithm for obtaining a low-rank approximation to a given matrix [53, 63]. In this method, $Z \in \mathbb{R}^{n\times p}$ is constructed to approximate the k dominant left singular vectors of A. Given $q \in \mathbb{N}$ and $A_q = (AA^{\mathsf{T}})^q A$, the Randomized Singular Value Decomposition considers $Z = A_q G$. In the single-pass setting ($q = 0$), error bounds have been notably provided in [53, 54] for the spectral norm. For $q \in \mathbb{N}^\star$, we provide error bounds which, to the best of our knowledge, are new. As detailed next, the relation $Z \sim \mathcal{N}(0, A_q A_q^{\mathsf{T}})$ leads to significant simplifications in the expressions of the error bounds given in Theorems 3.18, 3.20 and 3.21, respectively. Deriving these error bounds implies to first express the projected covariance matrices defined in (3.16). This is the main purpose of the next lemma.

Lemma 3.31. *Let* $A \in \mathbb{R}^{n\times m}$ *such that* $n \geq m$ *and* $Z \in \mathbb{R}^{n\times p}$ *with* $2 < p \leq \operatorname{rank}(A)$ *such that* $Z = A_q G$ *with* $A_q = (AA^{\mathsf{T}})^q A \in \mathbb{R}^{n\times m}$ *($q \in \mathbb{N}$) and* $G \in \mathbb{R}^{m\times p}$ *a standard Gaussian matrix* ($G \sim \mathcal{N}(0, I_m)$). *For a given integer* $k \in \{1, \ldots, p\}$, *set* $\Omega_k = U_k^{\mathsf{T}} Z$ *and* $\underline{\Omega}_k = \underline{U}_k^{\mathsf{T}} Z$. *The projected covariance matrices are then given by*

$$\begin{aligned} \mathbf{Cov}_k(Z) &= \Sigma_k^{4q+2}, \\ \mathbf{Cov}_{\perp,k}(Z) &= 0, \\ \underline{\mathbf{Cov}}_k(Z) &= (\underline{\Sigma}_k\underline{\Sigma}_k^{\mathsf{T}})^{2q+1}. \end{aligned} \tag{3.43}$$

Then the random matrix $\underline{\Omega}_k$ *conditioned by* Ω_k *follows a Gaussian distribution of mean*

$$\mathbb{E}\left[\,\underline{\Omega}_k \mid \Omega_k\,\right] = 0, \tag{3.44}$$

and of covariance matrix

$$\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right) = (\underline{\Sigma}_k \underline{\Sigma}_k^{\mathsf{T}})^{2q+1}. \tag{3.45}$$

Proof. A straightforward calculation gives $\Omega_k = \Sigma_k^{2q+1} V_k^{\mathsf{T}} G$. Hence Ω_k has full row rank with probability one [31]. Since $Z \sim \mathcal{N}(0, A_q A_q^{\mathsf{T}})$, $\mathbf{Cov}(Z) = A_q A_q^{\mathsf{T}}$ can be expressed as

$$\mathbf{Cov}(Z) = U(\Sigma\Sigma^{\mathsf{T}})^{2q+1} U^{\mathsf{T}}.$$

Thus $\mathbf{Cov}(Z)$ and A have the same rank and we deduce

$$\begin{aligned}
\mathbf{Cov}_k(Z) &= \Sigma_k^{4q+2}, \\
\mathbf{Cov}_{\perp,k}(Z) &= 0, \\
\underline{\mathbf{Cov}}_k(Z) &= (\underline{\Sigma}_k \underline{\Sigma}_k^{\mathsf{T}})^{2q+1}.
\end{aligned}$$

We apply relations ((3.17)) and ((3.18)) of Lemma 3.9 to deduce

$$\begin{aligned}
\mathbb{E}\left[\,\underline{\Omega}_k \mid \Omega_k\,\right] &= 0, \\
\mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right) &= (\underline{\Sigma}_k \underline{\Sigma}_k^{\mathsf{T}})^{2q+1},
\end{aligned}$$

which concludes the proof. □

Remark 3.32. From Lemma 3.31, we deduce that the projected covariance matrix $\mathbf{Cov}_k(Z)$ is nonsingular, allowing us to apply Theorems 3.18, 3.20 and 3.21, respectively. Since $\mathbf{Cov}_{\perp,k}(Z) = 0$, we also note that $\underline{\Omega}_k$ and Ω_k become statistically independent.

With the help of relations (3.43), (3.44) and (3.45), we specialize the main theorems proposed in Section 3.3.2 to the setting of the Randomized Singular Value Decomposition. To enhance the readability of the constants arising in the error bounds, we introduce the singular value ratios related to the matrix A as

$$\gamma_i = \frac{\sigma_{k+1}}{\sigma_i}, \quad i = 1, \ldots, \operatorname{rank}(A). \tag{3.46}$$

3.4.1 Error bounds in Frobenius norm

We first consider the case of the Frobenius norm. A direct application of Theorem 3.18, 3.27 and Lemma 3.31 leads to the following corollaries.

Corollary 3.33. *(Average analysis error bound in Frobenius norm, RSVD) Let* $A \in \mathbb{R}^{n\times m}$ *such that* $n \geq m$ *and* $Z \in \mathbb{R}^{n\times p}$ *with* $2 < p \leq \operatorname{rank}(A)$ *such that* $Z = A_q G$ *with* $A_q = (AA^{\mathsf{T}})^q A \in \mathbb{R}^{n\times m}$ $(q \in \mathbb{N})$ *and* $G \in \mathbb{R}^{m\times p}$ *a standard Gaussian matrix* $(G \sim \mathcal{N}(0, I_m))$*. Let* $\pi(Z)$ *denote the orthogonal projection onto the vector space spanned by the columns of* Z*. For a given* $k \in \{1, \ldots, p-2\}$*, set* $\Omega_k = U_k^{\mathsf{T}} Z \in \mathbb{R}^{k\times p}$ *such that* Ω_k *is full row rank (i.e.* $\operatorname{rank}(\Omega_k) = k$*). Let* $\varphi : x \mapsto x/\sqrt{1+x^2}$ *for* $x \geq 0$*.*

Then, one has

$$\mathbb{E}\left[\,\|[I_n - \pi(Z)]A\|_F - \|[I_n - \pi(Z)]\underline{A}_k\|_F\,\right] \leq \min\left\{\sqrt{a_k},\ \sqrt{k}\,\varphi\left(\frac{1}{\sqrt{k}}\,\sqrt{b_k}\right)\sigma_1\right\}, \tag{3.47}$$

where

$$a_k = \frac{\sigma_{k+1}^2}{p-k-1}\left(\sum_{i=k+1}^{\operatorname{rank}(A)}\frac{1}{\gamma_i^{4q+2}}\right)\left(\sum_{i=1}^{k}\gamma_i^{4q}\right),$$
$$b_k = \frac{1}{p-k-1}\left(\sum_{i=k+1}^{\operatorname{rank}(A)}\frac{1}{\gamma_i^{4q+2}}\right)\left(\sum_{i=1}^{k}\gamma_i^{4q+2}\right),$$

with the singular value ratios γ_i given in (3.46).

Assume further that $p \geq k+4$, then for any $t, u \geq 1$, it holds with probability of failure at most $e^{-u^2/2} + t^{-(p-k)}$

$$\|[I_n - \pi(Z)]A\|_F - \|[I_n - \pi(Z)]\underline{A}_k\|_F \leq \min\left\{\sqrt{a_k} + \alpha_k\, tu,\ \sqrt{k}\,\varphi\left(\frac{\sqrt{b_k} + \beta_k\, tu}{\sqrt{k}}\right)\sigma_1\right\},$$

where

$$\alpha_k = \sigma_{k+1}\left(\sum_{i=1}^{k}\gamma_i^{4q}\right)\frac{\sqrt{3k}}{\sqrt{p-k+1}} \quad \text{and} \quad \beta_k = \left(\sum_{i=1}^{k}\gamma_i^{4q+2}\right)\frac{\sqrt{3k}}{\sqrt{p-k+1}}.$$

Remark 3.34. In [15], Boullé and Townsend have considered the low-rank approximation of partial differential operators in infinite dimension using a zero mean and a general covariance matrix for the random matrix variable. They have recently improved their error bound to the finite-dimensional case using the Frobenius norm in [14]. It can be shown that the error bound (3.47) provided in Corollary 3.33 is always tighter. An illustration is given in Section 3.5.2.

3.4.2 Error bounds in spectral norm

We next consider the case of the spectral norm. A straightforward application of Theorem 3.20 and Lemma 3.31 first leads to the following corollary.

Corollary 3.35. *(Average analysis error bound in spectral norm, RSVD) Let $A \in \mathbb{R}^{n\times m}$ such that $n \geq m$ and $Z \in \mathbb{R}^{n\times p}$ with $2 < p \leq \operatorname{rank}(A)$ such that $Z = A_q G$ with $A_q = (AA^\mathsf{T})^q A \in \mathbb{R}^{n\times m}$ ($q \in \mathbb{N}$) and $G \in \mathbb{R}^{m\times p}$ a standard Gaussian matrix ($G \sim \mathcal{N}(0, I_m)$). Let $\pi(Z)$ denote the orthogonal projection onto the vector space spanned by the columns of Z. For a given $k \in \{1, \ldots, p-2\}$, set $\Omega_k = U_k^\mathsf{T} Z \in \mathbb{R}^{k\times p}$, such that Ω_k is full row rank (i.e. $\operatorname{rank}(\Omega_k) = k$). Let $\varphi : x \mapsto x/\sqrt{1+x^2}$ for $x \geq 0$.*

Then, one has

$$\mathbb{E}\left[\,\|[I_n - \pi(Z)]A\|_2 - \|[I_n - \pi(Z)]\underline{A}_k\|_2\,\right] \leq \min\left\{c_k,\ \varphi(d_k)\sigma_1\right\}, \tag{3.48}$$

where

$$c_k = \frac{\sigma_{k+1}}{\sqrt{p-k-1}}\left(\sum_{i=1}^{k}\gamma_i^{4q}\right)^{\frac{1}{2}} + \sigma_k\left(\sum_{i=k+1}^{\operatorname{rank}(A)}\left(\frac{\sigma_i}{\sigma_k}\right)^{4q+2}\right)^{\frac{1}{2}}\frac{e\sqrt{p}}{p-k},$$
$$d_k = \frac{1}{\sqrt{p-k-1}}\left(\sum_{i=1}^{k}\gamma_i^{4q+2}\right)^{\frac{1}{2}} + \left(\sum_{i=k+1}^{\operatorname{rank}(A)}\left(\frac{\sigma_i}{\sigma_k}\right)^{4q+2}\right)^{\frac{1}{2}}\frac{e\sqrt{p}}{p-k},$$

with the singular value ratios γ_i given in (3.46).

Assume further that $p \geq k+4$, then for any $t, u \geq 1$, it holds with probability of failure at most $e^{-u^2/2} + t^{-(p-k+1)}$

$$\|[I_n - \pi(Z)]A\|_2 - \|[I_n - \pi(Z)]\underline{A}_k\|_2 \leq \min\left\{c_k + \theta_k\, tu,\ \varphi\left(d_k + \delta_k\, tu\right)\sigma_1\right\}, \tag{3.49}$$

where

$$\theta_k = \sigma_{k+1}\,\gamma_k^{2q}\,\frac{e\sqrt{p}}{p-k+1} \quad \text{and} \quad \delta_k = \gamma_k^{2q+1}\,\frac{e\sqrt{p}}{p-k+1}.$$

We finally state the improved error bound related to Theorem 3.21 and Lemma 3.31 in the next corollary.

Corollary 3.36. *(Average analysis error bound in spectral norm, improved bound, RSVD) Let $A \in \mathbb{R}^{n\times m}$ such that $n \geq m$ and $Z \in \mathbb{R}^{n\times p}$ with $2 < p \leq \operatorname{rank}(A)$ such that $Z = A_q G$ with $A_q = (AA^{\mathsf{T}})^q A \in \mathbb{R}^{n\times m}$ $(q \in \mathbb{N})$ and $G \in \mathbb{R}^{m\times p}$ a standard Gaussian matrix $(G \sim \mathcal{N}(0, I_m))$. Let $\pi(Z)$ denote the orthogonal projection onto the vector space spanned by the columns of Z. For a given $k \in \{1, \ldots, p-2\}$, set $\Omega_k = U_k^{\mathsf{T}} Z \in \mathbb{R}^{k\times p}$, such that Ω_k is full row rank (i.e. $\operatorname{rank}(\Omega_k) = k$). Let $\varphi : x \mapsto x/\sqrt{1+x^2}$ for $x \geq 0$.*

Then, one has

$$\mathbb{E}\left[\,\|[I_n - \pi(Z)]A\|_2 - \sigma_{k+1}\,\right] \leq \min\left\{\widehat{c}_k,\ \varphi(\widehat{d}_k)\sqrt{\sigma_1^2 - \sigma_{k+1}^2}\right\}, \tag{3.50}$$

where

$$\widehat{c}_k = \frac{\sigma_{k+1}}{\sqrt{p-k-1}}\left(\sum_{i=1}^{k}\gamma_i^{4q}(1-\gamma_i^2)\right)^{\frac{1}{2}} + \sqrt{1-\gamma_\ell^2}\,\sigma_\ell\left(\sum_{i=k+1}^{\operatorname{rank}(A)}\left(\frac{\sigma_i}{\sigma_\ell}\right)^{4q+2}\right)^{\frac{1}{2}}\frac{e\sqrt{p}}{p-k},$$

$$\widehat{d}_k = \frac{1}{\sqrt{p-k-1}}\left(\sum_{i=1}^{k}\gamma_i^{4q+2}\right)^{\frac{1}{2}} + \left(\sum_{i=k+1}^{\operatorname{rank}(A)}\left(\frac{\sigma_i}{\sigma_k}\right)^{4q+2}\right)^{\frac{1}{2}}\frac{e\sqrt{p}}{p-k},$$

with

$$\ell = \begin{cases} 1, & (q=0 \quad \text{or} \quad \text{if } \sigma_{k+1}\sqrt{1+1/(2q)} \geq \sigma_1, \quad (q \in \mathbb{N}^\star)), \\ \arg\max\limits_i \left(\dfrac{\sqrt{1-\gamma_i^2}}{\sigma_i^{2q}}, \dfrac{\sqrt{1-\gamma_{i+1}^2}}{\sigma_{i+1}^{2q}}\right), & \text{with } \sigma_{k+1}\sqrt{1+1/(2q)} \in [\sigma_i, \sigma_{i+1}], \quad (q \in \mathbb{N}^\star), \\ k, & \text{if } \sigma_{k+1}\sqrt{1+1/(2q)} \leq \sigma_k, \quad (q \in \mathbb{N}^\star), \end{cases} \tag{3.51}$$

with the singular value ratios γ_i given in (3.46).

Assume further that $p \geq k+4$, then for any $t, u \geq 1$, it holds with probability of failure at most $e^{-u^2/2} + t^{-(p-k+1)}$

$$\|[I_n - \pi(Z)]A\|_2 - \|\underline{A}_k\|_2 \leq \min\left\{\widehat{c}_k + \widehat{\theta}_k\, tu,\ \varphi\left(\widehat{d}_k + \widehat{\delta}_k\, tu\right)\|\widehat{\Sigma}_k\|_2\right\}, \tag{3.52}$$

where

$$\widehat{\theta}_k = \sigma_{k+1}\,\gamma_\ell^{2q}\sqrt{1-\gamma_\ell^2}\,\frac{e\sqrt{p}}{p-k+1} \quad \text{and} \quad \widehat{\delta}_k = \gamma_k^{2q+1}\,\frac{e\sqrt{p}}{p-k+1}.$$

Proof. To deduce the expression of $\widehat{c}_k$, we need to obtain $\|(\widehat{\Sigma}_k^{\mathsf{T}}[\mathbf{Cov}_k(Z)]^{-1}\widehat{\Sigma}_k)^{\frac{1}{2}}\|_2$. Since

$$\widehat{\Sigma}_k^{\mathsf{T}}[\mathbf{Cov}_k(Z)]^{-1}\widehat{\Sigma}_k = \operatorname{diag}\left(\frac{\sigma_i^2 - \sigma_{k+1}^2}{\sigma_i^{4q+2}}\right), \quad i = 1, \ldots, k,$$

we introduce the map $\psi : x \mapsto (x - \sigma_{k+1}^2)/x^{2q+1}$ for $x \geq \sigma_{k+1} > 0$ for $q \in \mathbb{N}^\star$. A simple calculation shows that the extremum of ψ is reached for $x = \sigma_{k+1}^2(1 + 1/(2q))$. Hence we deduce

$$\|(\widehat{\Sigma}_k^{\mathsf{T}}[\mathbf{Cov}_k(Z)]^{-1}\widehat{\Sigma}_k)^{\frac{1}{2}}\|_2 := \frac{\sqrt{1-\gamma_\ell^2}}{\sigma_\ell^{2q}},$$

with ℓ defined in (3.51). The other quantities arising in $\widehat{c}_k$ can be obtained straightforwardly. □

To the best of our knowledge, we note that the error bounds provided in Corollaries 3.33, 3.35 and 3.36 are new. In Section 3.5.3, we provide numerical illustrations to show the relevance of the bounds in expectation the context of the Randomized Singular Value Decomposition.

3.5 Numerical illustrations

In this section, we aim at illustrating the numerical behaviour of the error bounds introduced in Sections 3.3 and 3.4, respectively. We first show the tightness of the error bounds compared to empirical experimental errors in Section 3.5.1. Then, we propose a detailed comparison with the state-of-the-art error bounds [53] in Sections 3.5.2 and 3.5.3, respectively. In the following, we consider a square matrix A (with $n = m = 1000$) obtained as follows. First, we select two orthogonal matrices $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{n\times n}$. Each matrix is obtained independently by drawing a random standard Gaussian matrix and taking its QR factorization. Then, given

$$\Sigma = \operatorname{diag}(\underbrace{1, \ldots, 1}_{10 \text{ times}}, \ 2^{-\frac{1}{2}}, \ 3^{-\frac{1}{2}}, \ \ldots, \ (n-9)^{-\frac{1}{2}}) \in \mathbb{R}^{n\times n},$$

we conduct all the numerical experiments with $A = U\Sigma V^{\mathsf{T}}$. This test case is directly inspired from [78, 87]. We consider two different target ranks k ($k = 5$ and $k = 15$, respectively) to allow a variety of results and comments. We believe that performing such an analysis is instructional since, in practice, we are not aware of the ideal value for the target rank k. The data and the scripts to reproduce all the numerical results and the resulting figures are publicly available at https://github.com/a-scotto/R-SVD-Analysis.

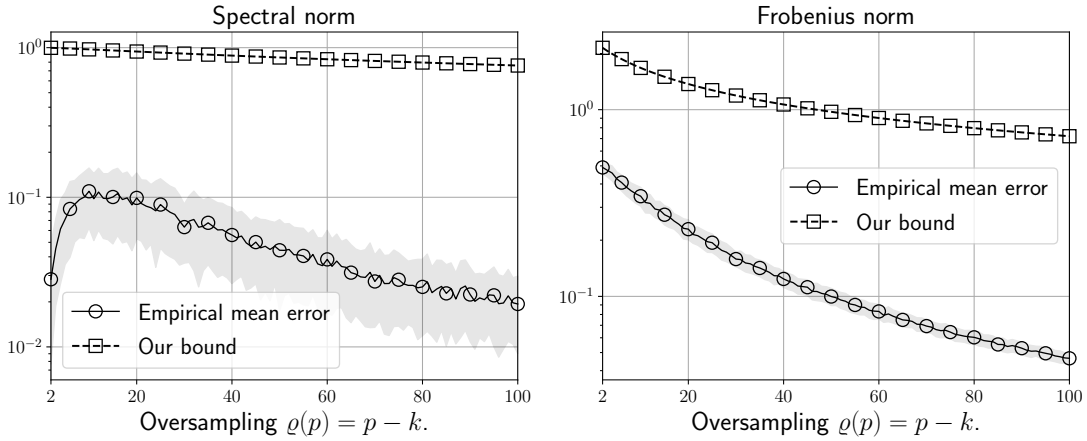
3.5.1 Error bounds in expectation versus the empirical error

We first focus on the tightness of the error bounds in expectation, given in Theorem 3.18 for the Frobenius norm and in Theorem 3.20 for the spectral norm, respectively. To quantify their tightness, for a given target rank k and for a fixed value of the oversampling parameter $\varrho(p) = p - k$, we compare our bounds with the empirical mean error over 100 samples, i.e.,

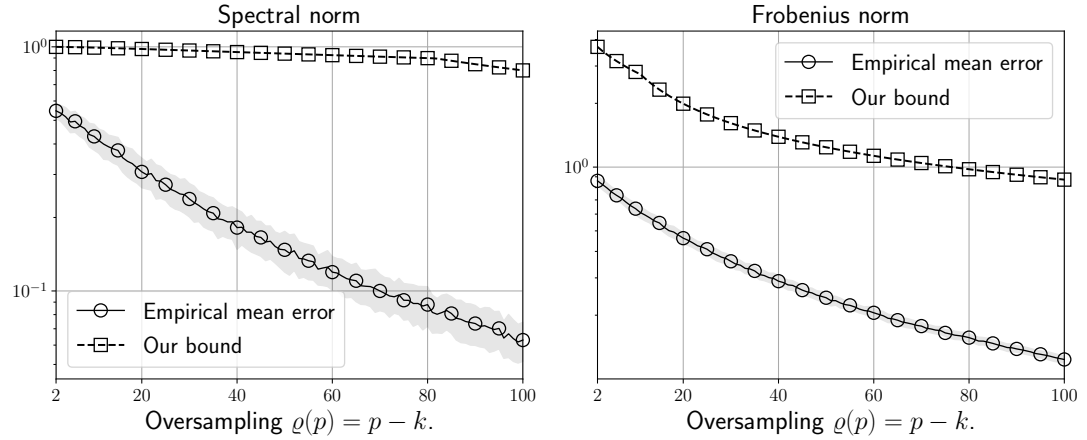
$$\frac{1}{100}\sum_{i=1}^{100}\left(\|[I_n - \pi(AG_i)]A\|_{2,F} - \|[I_n - \pi(AG_i)]\underline{A}_k\|_{2,F}\right),$$

where $G_i \in \mathbb{R}^{n\times p}$ ($1 \leq i \leq 100$) are 100 independent standard Gaussian matrices ($G_i \sim \mathcal{N}(0, I_n)$). We note that, in this setting, one has $\widehat{Z} = 0$ and $\mathbf{Cov}(Z) = AA^{\mathsf{T}}$.

With respect to the target rank k, Figure 3.1 shows the empirical error as well as our error bounds (in both norms) with respect to the oversampling parameter $\varrho(p)$ for both target ranks

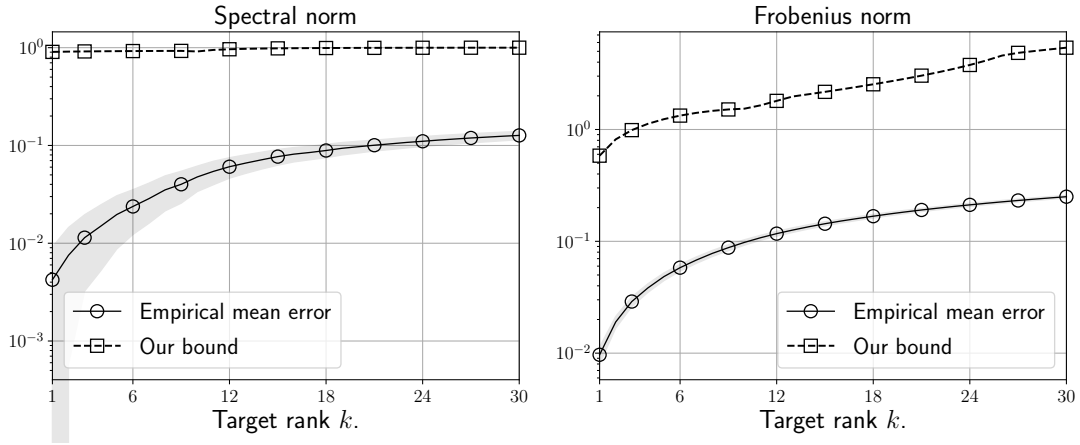


(a) Results for $k = 5$.

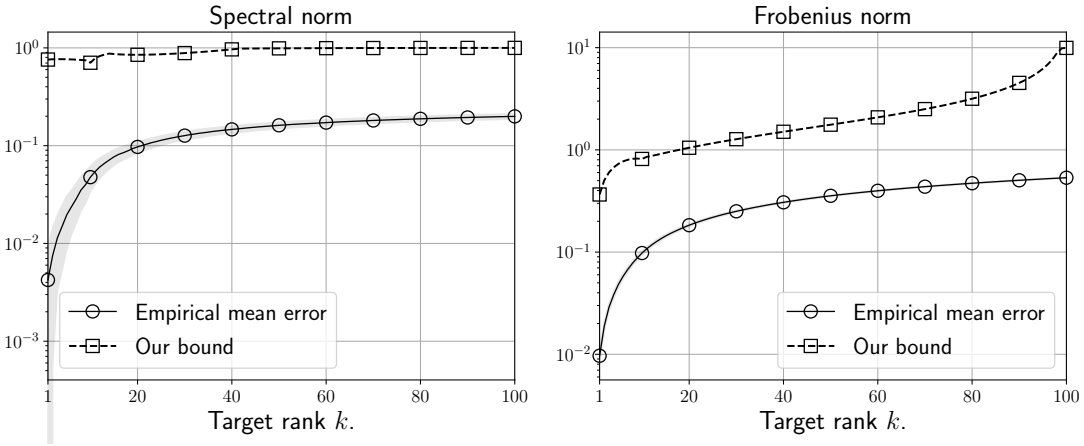


(b) Results for $k = 15$.

Figure 3.1: Empirical mean error and our error bounds in expectation with respect to the oversampling parameter $\varrho(p) = p - k$. Case of $k = 5$ (top), $k = 15$ (bottom), spectral norm (left) and Frobenius norm (right). Statistics on errors are also shown: empirical mean (circle mark) $\pm$ one standard deviation (grey area).



(a) Results for $p = 32$.



(b) Results for $p = 102$.

Figure 3.2: Empirical mean error and our error bounds in expectation with respect to the target rank k. Cases of $p = 32$ (top) and of $p = 102$ (bottom), spectral norm (left) and Frobenius norm (right). Statistics on errors are also shown: empirical mean (circle mark) $\pm$ one standard deviation (grey area).

$k = 5$ and $k = 15$, respectively. We remark that, as the number of samples increases, our bounds predict a smaller error in both norms. We also note that the error bounds in expectation in the Frobenius norm are relatively accurate compared to the spectral case. In fact, the decrease rate seems to be slower in the spectral norm compared to the Frobenius norm. This is indeed expected since our error bounds in the spectral norm have been obtained in a looser way compared to the Frobenius norm.

With respect to the oversampling parameter $\varrho(p)$, Figure 3.2 shows the empirical error and our error bounds with respect to the target rank k for two fixed values of the sample parameter ($p = 32$ and $p = 102$, respectively). As expected, the error bounds are found to be more accurate for a lower target rank. In the spectral norm, for small target ranks, the gap between the empirical mean error and our error bound does not seem to be as tight as in the Frobenius norm. We also note that by increasing the value of p, our bound reaches a smaller value for the same target rank k. For instance, for a target rank of order 20, our bound is of order 10^0 for $p = 102$, while it exceeds 5×10^0 in the case of $p = 32$.

3.5.2 Error bounds in expectation versus the state-of-the-art

We now compare our error bounds with respect to the reference error bounds [53, Theorems 10.5 and 10.6]. We stress that various reference error bounds exist in the literature but they can be mostly considered as adaptations[2] of the error bounds proposed by Halko, Martinsson and Tropp [53] in different settings, see e.g., [14]. For simplicity reasons, we later refer to the Halko, Martinsson and Tropp error bounds [53] by "`HMT bound`". We note that the "`HMT bound`" corresponds to an upper bound for the following expected error quantity

$$\mathbb{E}\left[\, \|[I_n - \pi(Z)]A\|_{2,F} - \|\underline{A}_k\|_{2,F} \,\right], \tag{3.53}$$

where $Z = AG$ with G drawn following a standard Gaussian distribution ($G \sim \mathcal{N}(0, I_n)$).

In this case, since $\|[I_n - \pi(Z)]\underline{A}_k\|_{2,F} \leq \|\underline{A}_k\|_{2,F}$, we deduce that the error bounds given in Theorems 3.18 and 3.20 can be also considered as upper bounds of the error quantity (3.53) with $\mathbf{Cov}(Z) = AA^{\mathsf{T}}$. In the spectral norm case, when we consider the error quantity (3.53), we have been able to derive an improved error bound in Theorem 3.21. We refer to such a bound as "`Improved bound`" in the numerical experiments.

Figure 3.3 shows a comparison of the three error bounds for the error quantity (3.53) using two different target ranks for k in both norms. In the spectral norm case, for $k = 5$, "`Our improved bound`" outperforms by far the other error bounds (i.e., "`HMT bound`" and "`Our bound`"). The "`HMT bound`" is in particular found to be very loose for small values of the oversampling parameter $\varrho(p)$. When the target rank k is getting larger (i.e., $k = 15$), "`Our bound`" and "`Our improved bound`" behave similarly (with a slight advantage for "`Our improved bound`"). Also, for $\varrho(p) \approx 80$, we observe a break in the curves related to both bounds. This is mainly due to a change in the minimum in our bounds. The "`HMT bound`" is very loose for small value of p, but as far as the sample parameter gets larger, the bound gets close to our two error bounds.

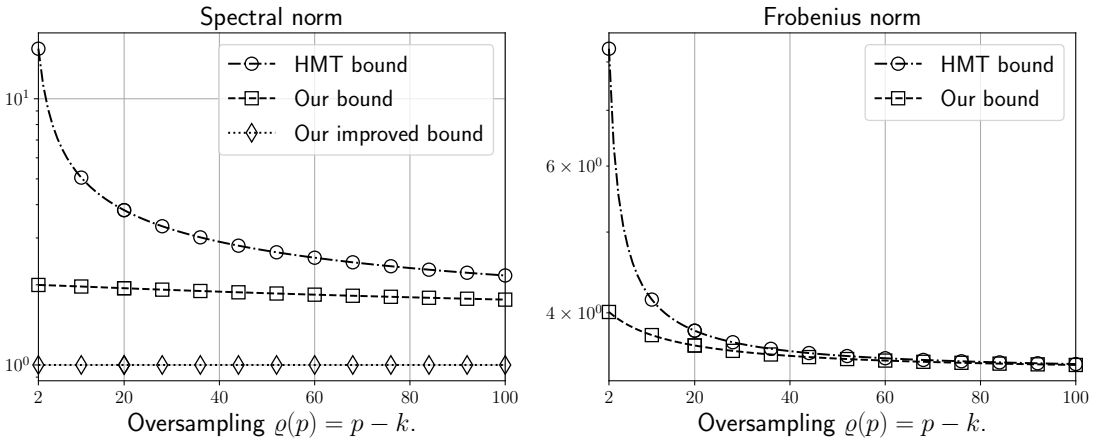
In the Frobenius norm case, as shown in Figure 3.3, we again notice the clear benefit of "`Our bound`" which, compared to "`HMT bound`", is leading to a moderate overestimation of the error, in particular for small values of the oversampling parameter. The asymptotic behaviour of the error bounds is consistent with the fact that, for a large value of p, "`Our bound`" reduces to the "`HMT bound`" in this norm. For further comparison, we have also plotted the behaviour of the error bounds proposed by Boullé and Townsend ("`BT bound`") [14, Proposition 6] using $K = I_n$ in their general setting.

3.5.3 Error bounds for the Randomized Singular Value Decomposition

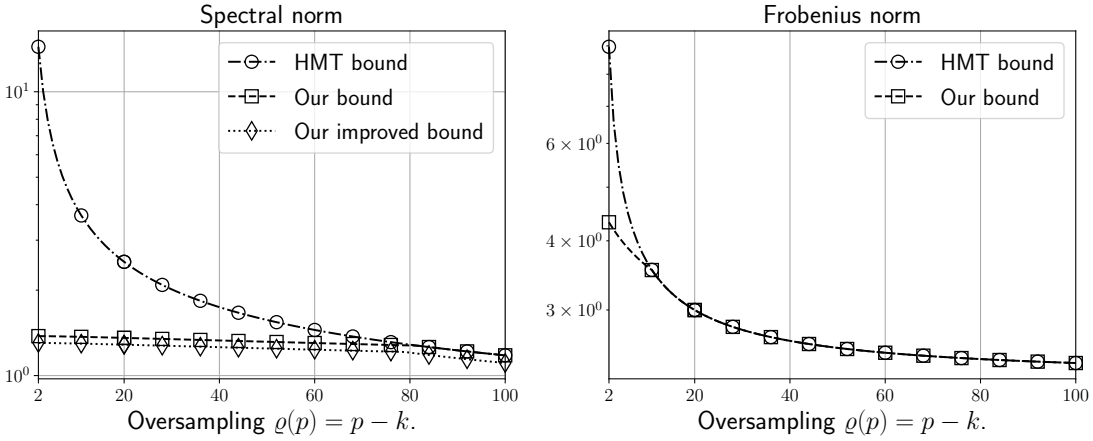
Finally, we illustrate the relevance of our error bounds in the context of the Randomized Singular Value Decomposition. We consider the quantity of interest ((3.53)) now with $Z = A_q G$, with $A_q = (AA^{\mathsf{T}})^q A$ for a given $q \in \mathbb{N}$ and G drawn following a standard Gaussian distribution ($G \sim \mathcal{N}(0, I_n)$). We later consider the single-pass case (i.e., $q = 0$) and the power scheme for two different values of the iteration (i.e., $q = 1$ and $q = 2$).

We denote the error bound [53, Corollary 10.10] by "`HMT bound`" in this section. To the best of our knowledge, in the context of the power iteration scheme for the Randomized Singular Value Decomposition, the "`HMT bound`" has been derived only in the spectral norm. We note that we have been able to provide an error analysis for both the Frobenius and the spectral norms. Hence, we later denote the error bound given in Corollary 3.33 by "`Our bound`". In the spectral norm case, only "`Our improved bound`" given in Corollary 3.36 is considered in the following, since it is found to outperform "`Our bound`".

[2] A noticeable exception is the more involved analysis proposed in [49], which investigates the quality of the singular value approximation and of the randomized low-rank approximation to a given matrix (see Theorem 5.7 therein, notably). To unify our presentation, we have therefore decided to restrict the comparison to the reference bounds proposed in [53].

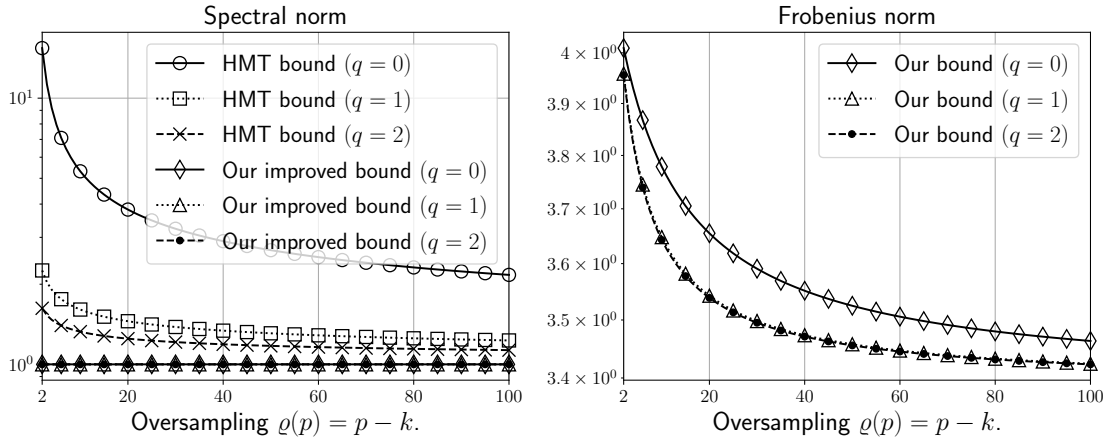


(a) Results for $k = 5$.

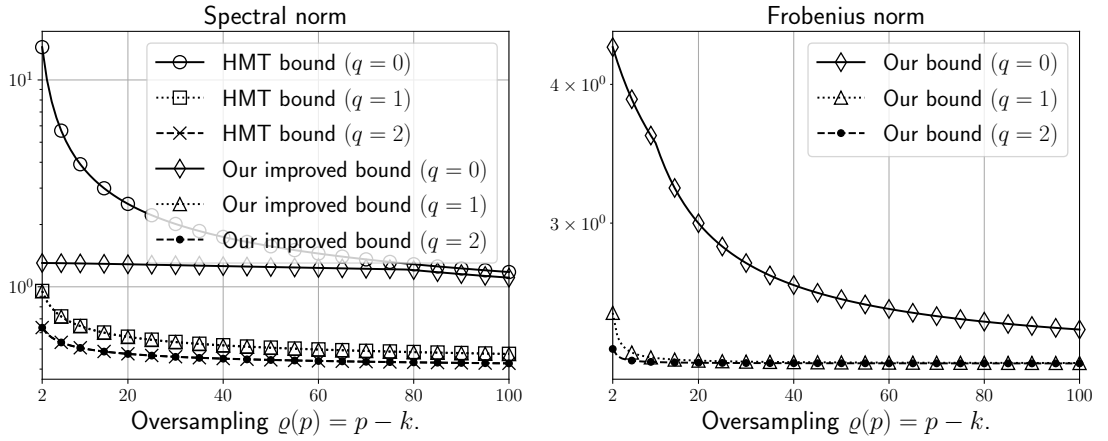


(b) Results for $k = 15$.

Figure 3.3: Comparison of different bounds for the error quantity (3.53) using two different target ranks $k = 5$ (top) and $k = 15$ (bottom), spectral norm (left) and Frobenius norm (right).



(a) Results for $k = 5$.



(b) Results for $k = 15$.

Figure 3.4: Randomized Singular Value Decomposition: comparison of different bounds for the error quantity (3.53) for two different target ranks $k = 5$ (top) and $k = 15$ (bottom), spectral norm (left) and Frobenius norm (right).

Figure 3.4 shows the comparison between the different bounds related to the error quantity (3.53), in the context of the Randomized Singular Value Decomposition. First, as expected, we clearly identify that increasing the value of q in the power iteration scheme does lead to a strong improvement in the tightness of all the error bounds. In fact, while the error bounds are above 10^0 when $q = 0$, they become much smaller when $q \geq 1$. In the spectral norm case, the advantage of "`Our improved bound`" over the "`HMT bound`" is clear. We also note that for a large value of the target rank (i.e., $k = 15$), the gap between the two error bounds gets smaller when $q \geq 1$. In the Frobenius norm case, we observe that the power iteration scheme is indeed very profitable, in particular for the large rank case, i.e., $k = 15$. A value of $q = 1$ seems to be sufficient to get optimal error bounds, since the convergence towards the optimal value $\|\underline{A}_k\|_F$ is almost immediate. In terms of computational cost, our numerical experiments suggest that performing only $q = 1$ iteration is a very satisfactory trade-off in this test case.

3.6 Conclusions and perspectives

We have analyzed theoretically the low-rank approximation to a given matrix in both the spectral and Frobenius norms. First, we have derived in Theorems 3.4 and 3.8 deterministic error bounds that hold with some minimal assumptions. Second, we have derived error bounds in expectation and in probability in the non-standard Gaussian case, assuming a non-trivial mean and a general covariance matrix for the random matrix variable (Theorems 3.18, 3.20 and 3.21). This analysis generalizes and improves the error bounds proposed in [53]. Then, we have applied our analysis to the Randomized Singular Value Decomposition and have deduced the related error bounds in expectation (Corollaries 3.33, 3.35 and 3.36). Numerical experiments on a synthetic test case have shown the tightness of the new error bounds.

In a near future, we plan to apply the contributions from this chapter to generalize the analysis of the randomized subspace iteration method (Alg. 2.3) proposed in [77]. Such a complementary analysis would provide an analysis of the accuracy in terms of singular vectors and singular values for a larger class of randomized subspace iteration methods. Plus, it would be beneficial in applications where the accuracy in terms of singular vectors and singular values.

The randomized methods for low rank approximation have recently been considered in the infinite dimensional case [15]. In this setting, the Frobenius norm is replaced by an Hilbert-Schmidt norm and Gaussian matrices by Gaussian processes. This generalization required to extend the appropriate technical lemmas in [53] to the infinite dimensional case. Consequently, investigating whether the preparatory lemmas presented in Section 3.3.2 can also be extended to the infinite dimensional setting could yield the generalization of the proposed analysis to any Hilbert space.

Finally, in this chapter, we have compared our bounds in expectation with the state-of-the-art error bounds. Consequently, understanding how the proposed error bounds in probability given in Corollaries 3.33, 3.35 and 3.36 behaves compared to the reference bounds in [53, Theorem 10.7 and 10.8] seems also important. In particular, the error bounds in probability are no longer generalizations of the ones in [53], since the proof of Lemma 3.25 differs from [53, Proposition 10.3]. Consequently, at this point, it is not clear which bound is tighter, and under which conditions.

CHAPTER 4

Randomized methods for the generalized symmetric eigenvalue problem in a non-Euclidean inner product

Contents

Abstract

In this chapter, we propose and study randomized methods to address two related generalized eigenvalue problems (GEP) involving matrices with non-Euclidean symmetry. Our interest in such specific eigenvalue problems is notably motivated by later applications in preconditioning within variational data assimilation (see Chapter 5).

Our algorithms compute an approximate dominant eigenspace using randomized subspace iteration. Then, getting the approximate eigenpairs is handled using the Rayleigh-Ritz method. For each GEP we propose two variants depending on whether the Rayleigh-Ritz method is applied to the GEP directly, or to an equivalent formulation based on the inverse. In phase with the theoretical derivation, we propose a numerically robust implementation for each algorithm. We detail the related computational costs and memory requirements, and outline the relations with existing methods, especially from [80] and [24].

Based on the general analysis developed in Chapter 3, we propose an average-case error analysis of the methods in both weighted Frobenius and spectral norms. The obtained bounds give insights regarding the number of subspace iterations, the number of random samples, and the distribution of the Gaussian sample matrix. Our analysis in spectral norm surpasses the state-of-the-art error bounds in [80, Theorem 1], while the analysis in Frobenius norm, to the best of our knowledge, is new.

Finally, we investigate the performance of the proposed algorithms in terms of approximate eigenpair accuracy on a variational data assimilation problem.

4.1 Introduction

Let $\Upsilon \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix defining an inner product on $\mathbb{R}^n$. In this chapter, we are interested in solving the generalized eigenvalue problem (GEP)

$$A\,v = \lambda\,B\,v, \tag{4.1}$$

where $A, B \in \mathbb{R}^{n\times n}$ are Υ-symmetric matrices, B is invertible and $v \in \mathbb{R}^n$ is non zero. The pair $\{A, B\}$ in (4.1) is called a *matrix pencil*. The study of this class of eigenvalue problems is motivated by their applications in preconditioning presented in Chapter 5. Similarly, we are also interested in addressing the related eigenvalue problem

$$AB^{-1}u = \lambda\,u, \tag{4.2}$$

where it can readily be seen that $u = Bv$. In the preconditioning terminology, if B^{-1} is interpreted as a preconditioner for A, then solving (4.1) (resp. (4.2)) would be interpreted as finding eigenpairs of the left (resp. right) preconditioned matrix. From now on, we will refer to (4.1) as the GEP in initial form, and to (4.2) as the GEP with basis transformation.

The eigenvalue problem (4.1) can be transformed into a generalized Hermitian eigenvalue problem by left-multiplying (4.1) by Υ,

$$\Upsilon A\,v = \lambda\,\Upsilon B\,v. \tag{4.3}$$

Here, ΥA and ΥB are indeed symmetric due to the Υ-symmetry of A and B. Thus, we obtain that the eigenvalues are real, and that the eigenvectors are ΥB-orthogonal [72, Theorem 15.3.3]. Similarly, noting that AB^{-1} is ΥB^{-1}-symmetric, left-multiplying (4.2) by ΥB^{-1} yields the equivalent generalized Hermitian eigenvalue problem

$$\Upsilon B^{-1}AB^{-1}\,v = \lambda\,\Upsilon B^{-1}\,v. \tag{4.4}$$

with analogous consequences.

Here, we focus on the case ΥB (or equivalently ΥB^{-1}) is symmetric positive definite, which is less restrictive than it might look at first sight. Indeed, if ΥB is not symmetric positive definite, then there exist real scalars α, β such that $\alpha\Upsilon A + \beta\Upsilon B$ is symmetric positive definite, and we thus rather study the matrix pencil $\{A, \alpha A + \beta B\}$, where we notice that $\alpha A + \beta B$ remains Υ-symmetric. The eigenvectors are identical, and if μ denotes an eigenvalue of the modified pencil, then the eigenvalues of the initial pencil can be recovered via $\lambda = \beta\mu/(1-\alpha\mu)$, given that $\alpha\mu \neq 1$.

Another important assumption that we make is that solving linear systems involving B can be performed accurately. Again, this might seem restrictive, but if B^{-1} is interpreted as a preconditioner for A, then B^{-1} is generally directly available as an operator. In this case, accurately solving linear systems involving B reduces to simply apply B^{-1}. When B^{-1} is not available, then applications of B^{-1} can be replaced by an iterative procedure. This setting is out of the scope of this thesis, although results in the deterministic and non-symmetric case [36] suggest that it could still be viable.

4.1.1 Related research

There are several deterministic methods for solving a generalized Hermitian eigenvalue problem given as (4.3). A first possibility is to use direct methods, which requires to perform the Cholesky factorization of ΥB [43, Theorem 4.2.7] as $\Upsilon B = LL^{\mathsf{T}}$. (4.3) can then be transformed into the following standard Hermitian eigenvalue problem

$$L^{-1}\Upsilon A L^{-\mathsf{T}}\,\tilde{v} = \lambda\,\tilde{v}, \tag{4.5}$$

where $\tilde{v} = L^{\mathsf{T}}v$. Then, solving (4.5) can be performed using standard methods such as Lanczos of subspace iteration method (see [8, Chapter 4]). However, such factorization for ΥB is out of reach in our context for two main reasons. First, we target large-scale applications where the worst-case $O(n^3)$ algorithmic complexity of the Cholesky factorization becomes prohibitively expensive. Second, ΥB is a product of matrices that is neither formed nor stored explicitly, making direct methods unusable. On the other hand, iterative methods (see [8, Chapter 5]) can be directly applied on (4.3). However, they require to accurately solve linear systems involving ΥB, which might neither be affordable nor possible.

Randomized algorithms have been proposed to compute approximate eigen/singular information. These methods have proven to be robust, and particularly efficient when the eigen/singular values are rapidly decaying. They are also *matrix-free*, in that they do not need to access the matrix coefficients, but only matrix-vector products with the involved matrices to block of vectors.

A first valid option to address would thus be to use randomized SVD (Algorithm 2.4), to compute approximate dominant singular vectors/values of $B^{-1}A$. However, one can show that if $w \in \mathbb{R}^n$ is a singular vector of $B^{-1}A$, then its eigenvectors can be recovered via $v = (\Upsilon B)^{-1/2}w$. Thus, it would require a factorization of ΥB, which is excluded. This motivated the authors in [80] to propose three dedicated randomized methods to approximately solve (4.3) in the particular case $\Upsilon = I_n$. Their algorithms directly provide B-orthonormal approximate eigenvectors, and an average-case analysis in weighted spectral norm has been provided (see Theorem 1) for one of the methods. Nevertheless, since their algorithms require applications of A, B and B^{-1}, the solution of (4.3) would especially require to apply $(\Upsilon B)^{-1}$, which might not always be possible. Finally, randomized methods for computing an approximate truncated generalized singular value decomposition (GSVD) in the sense of [91, Definition 3] have been proposed in [78, Algorithm 3]. Nevertheless, the method proposed in [78] leads to very expensive algorithms, because the power iterations are fundamentally different for rectangular and square matrices.

4.1.2 Contributions

In this chapter, we derive randomized methods for computing approximate dominant eigenvectors and eigenvalues of (4.1) and (4.2) under the assumption that ΥB is symmetric positive definite.

The existing methods in the literature all require either solving linear systems involving ΥB or factorizations of ΥB, and *our intention in this chapter is to show that algorithms can be derived without these requirements.*

Our algorithms rely on the randomized subspace iteration (Algorithm 2.3) to compute an approximate dominant eigenspace. Then, we propose two different extraction methods to obtain the approximate eigenpairs based on the Rayleigh-Ritz method. Thus, Algorithms 4.1 and 4.2 address the solution of (4.1) and Algorithms 4.3 and 4.4 the solution of (4.2).

The proposed algorithms are versatile and allow us to recover several existing methods such as the Nyström (Algorithm 2.5) and Ritzit (Algorithm 2.6) methods. In particular, Algorithms 4.1 and 4.2 are generalizations of [80, Algorithms 6, 7 and 8]. However, our implementations avoid applications of B and are thus more efficient computationally.

Theoretically, we propose an average-case error analysis of the algorithms in both weighted Frobenius and spectral norms. The obtained bounds directly stem from Theorems 3.18 and 3.20 proposed in Section 3.3.2. Theorems 4.9, 4.10 and 4.11 provide an analysis of Algorithms 4.1 and 4.2, and Theorems 4.13, 4.14 and 4.15 an analysis of Algorithms 4.3 and 4.4. To the best of our knowledge, the bounds obtained in weighted Frobenius norm are new. The bounds in weighted spectral norm are not, but they generalize and improve over the ones proposed in [80, Theorem 1].

4.2 Preliminaries

Let us begin with introducing additional material that will be helpful throughout this chapter.

Schur complement

Let $M \in \mathbb{R}^{n\times n}$ be a symmetric matrix and let $1 \leq k \leq n$. Assume M has the following block partitioning

$$M = \begin{bmatrix} M_{1,1} & M_{1,2} \\ M_{1,2}^{\mathsf{T}} & M_{2,2} \end{bmatrix},$$

where $M_{1,1} \in \mathbb{R}^{k\times k}$, $M_{1,2} \in \mathbb{R}^{k\times(n-k)}$ and $M_{2,2} \in \mathbb{R}^{(n-k)\times(n-k)}$. If $M_{1,1}$ is nonsingular, then the Schur complement of the block $M_{1,1}$ of the matrix M is defined as

$$M_{2,2}/M_{1,1} = M_{2,2} - M_{1,2}^{\mathsf{T}} M_{1,1}^{-1} M_{1,2}.$$

Remark 4.1. Here the notation is **not standard**, but it makes the quantity arising in the error bounds much lighter, hence this choice.

4.3 Derivation of the algorithms

Let $\Upsilon \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix and let $A, B \in \mathbb{R}^{n\times n}$ be two Υ-symmetric matrices. In this section, given Assumptions 1 and 2, we propose randomized methods to solve

$$B^{-1}Av = \lambda v \quad \text{and} \quad AB^{-1}u = \lambda u,$$

where $v, u \in \mathbb{R}^n$ are non zero. We briefly recall the derivation of the Rayleigh-Ritz method in Section 4.3.1. Then, we derive our algorithms for both the GEP and the GEP with basis transformation.

4.3.1 The Rayleigh-Ritz method

The Rayleigh-Ritz method [39, Section 2.5] is a natural framework to derive approximations of eigenvalue problems. Let us briefly introduce the procedure.

Let $M \in \mathbb{R}^{n\times n}$ be any given matrix, and $\mathcal{S}$ be a linear subspace in which the approximate eigenvectors of M are sought. This subspace is thereby usually referred to as the *search space*. For any vector $x \in \mathcal{S}$ and scalar $\mu \in \mathbb{C}$, we define the residual vector associated to the approximate eigenpair (x, μ) of M as

$$r(x, \mu) = Mx - \mu x.$$

To derive the approximation, we impose a Galerkin condition on the residual, that is an orthogonality condition with respect to $\mathcal{S} \subset \mathbb{R}^n$. Performing the Rayleigh-Ritz method consists in finding $x \in \mathcal{S}$ and $\mu \in \mathbb{R}$ such that

$$r(x, \mu) \perp \mathcal{S}. \tag{4.6}$$

Remark 4.2. In our context, the orthogonality will be considered with respect to a non-Euclidean inner-product.

Let S be a matrix whose columns form a basis of $\mathcal{S}$. For any $x \in \mathcal{S}$, there exists some vector y of appropriate dimension such that $x = Sy$. Accordingly, (4.6) can be written in a matrix form as

$$S^\mathsf{T} M S\, y = \mu\, S^\mathsf{T} S\, y. \tag{4.7}$$

If (y, μ) is a solution of (4.7), then (Sy, μ) is called a Ritz pair of M associated to $\mathcal{S}$ [72, Section 11.3].

4.3.2 Algorithms for the generalized eigenvalue problem in initial form

We focus on the GEP in its initial formulation, that is as in (4.1). Let us rewrite it as

$$B^{-1}Av = \lambda v. \tag{4.8}$$

To derive the Ritz approximation, we must first construct the search space. Several methods are available in the randomized linear algebra literature to construct relevant search spaces (see [63, Section 11] for a review). Here, and as mentioned in the introduction, we consider the randomized subspace iteration method. We therefore consider search spaces of the form $\mathcal{S} = \mathcal{R}(V_q)$ with $V_q = (B^{-1}A)^q\Omega$, with $\Omega \in \mathbb{R}^{n\times p}$ a full column rank random matrix satisfying $1 \le p \le \operatorname{rank}(A)$, and $q \in \mathbb{N}$.

Remark 4.3. Although one of the simplest approach, the random subspace iteration is very well suited when the emphasis is made on approximating dominant eigenvectors, which is precisely what will be needed in Chapter 5. However, it is worth noticing that by this choice, we limit de facto the range of applications of our algorithms.

Direct approach

For this first approach, we directly apply the Rayleigh-Ritz method on (4.8) with $S = V_q$. Since $B^{-1}A$ is ΥB-symmetric, the orthogonality in the Galerkin condition is considered with respect to the ΥB inner product. Altogether, we obtain

$$V_q^\mathsf{T}\Upsilon A V_q\, y = \mu\, V_q^\mathsf{T}\Upsilon B V_q\, y.$$

Let us now assume that $q \ge 1$. In this case, one has by definition of V_q that $BV_q = AV_{q-1}$. Therefore, we can rewrite the projected eigenvalue problem as

$$V_q^\mathsf{T}\Upsilon A V_q\, y = \mu\, V_q^\mathsf{T}\Upsilon A V_{q-1}\, y. \tag{4.9}$$

Here, we observe that forming (4.9) no longer requires B. Assuming $q \geq 1$ only excludes the case where we consider a search space of the form Ω with Ω being a random matrix. Since very poor performance can be expected from such a search space, assuming $q \geq 1$ is reasonable.

Remark 4.4. Avoiding the use of B can be important in applications where this matrix is not explicitly available such as in certain variational data assimilation settings [50, Section 3.1].

Finding Ritz pairs (μ, y) satisfying (4.9) is thus equivalent to solving a reduced generalized symmetric eigenvalue problem. Let us denote $T = V_q^\mathsf{T} \Upsilon A V_q$ and $\Phi = V_q^\mathsf{T} \Upsilon A V_{q-1}$. By assumption, ΥB is symmetric positive definite and V_q is full column rank. Consequently, Φ is symmetric positive definite, and there exist a Φ-orthonormal matrix $W \in \mathbb{R}^{p \times p}$ and a diagonal matrix $\Delta \in \mathbb{R}^{p \times p}$ such that

$$TW = \Phi W \Delta, \quad \text{and} \quad W^\mathsf{T} \Phi W = I_p.$$

The approximate eigenvectors $\widetilde{V}$ and approximate eigenvalues $\widetilde{\Lambda}$ of (4.8) are then obtained via $\widetilde{V} = V_q W$ and $\widetilde{\Lambda} = \Delta$.

Remark 4.5. We notice that $W^\mathsf{T} V_q^\mathsf{T} \Upsilon B V_q W = W^\mathsf{T} \Phi W = I_p$, meaning that the obtained approximate eigenvectors are ΥB-orthonormal. Here, we remark that their ΥB-orthonormality is entirely determined by the Φ-orthonormality of W. This implies that the reduced eigenvalue problem must be solved accurately to ensure a satisfactory ΥB-orthonormality of the resulting approximate eigenvectors.

The procedure is summarized in Algorithm 4.1. Algorithmic considerations are discussed in Section 4.3.5.

Algorithm 4.1: Direct approach for the GEP in initial form (4.1).

Input: Matrices $A, B \in \mathbb{R}^{n \times n}$ that are Υ-symmetric, number of random samples $1 \leq p \leq \operatorname{rank}(A)$, number of approximate eigenpairs $1 \leq k \leq p$ to provide, number of subspace iterations $q \geq 1$.

1 Draw a random matrix $\Omega \in \mathbb{R}^{n \times p}$, and set $V = \Omega$
2 **for** $j = 1, \ \ldots, \ q$ **do**
3 Compute $X = AV \in \mathbb{R}^{n \times p}$
4 Perform the thin QR factorization $B^{-1}X = QR$ and set $V = Q$
5 **end**
6 Compute $Z = \Upsilon V \in \mathbb{R}^{n \times p}$ and form $\Phi = R^{-\mathsf{T}} X^\mathsf{T} Z \in \mathbb{R}^{p \times p}$
7 Compute $X = AV \in \mathbb{R}^{n \times p}$ and form $T = X^\mathsf{T} Z \in \mathbb{R}^{p \times p}$
8 Solve the generalized Hermitian eigenvalue problem $TW = \Phi W \Delta$ with $W \in \mathbb{R}^{p \times p}$ a Φ-orthogonal matrix and $\Delta \in \mathbb{R}^{p \times p}$ a diagonal matrix with the eigenvalues sorted in **decreasing** order
9 Remove the last $p - k$ columns of W and Δ
10 Remove the last $p - k$ rows of Δ
11 Set $\widetilde{V} = VW \in \mathbb{R}^{n \times k}$ and $\widetilde{\Lambda} = \Delta \in \mathbb{R}^{k \times k}$.

Output: Matrices $\widetilde{V} \in \mathbb{R}^{n \times k}$ and $\widetilde{\Lambda} \in \mathbb{R}^{k \times k}$ such that $B^{-1}A\widetilde{V} \approx \widetilde{V}\widetilde{\Lambda}$ with $\widetilde{V}^\mathsf{T} \Upsilon B \widetilde{V} = I_k$ and $\widetilde{\Lambda}$ diagonal.

Inverse approach

The particular form of V_q allows us to address (4.8) from another perspective. Let us temporarily assume that A is nonsingular. Hence, we can rather consider the standard eigenvalue problem

$$A^{-1}Bv = \theta v, \tag{4.10}$$

where $\theta = 1/\lambda$. Since $A^{-1}B$ is ΥB symmetric, the Rayleigh-Ritz method applied to (4.10) with V_q yields

$$V_q^{\mathsf{T}}\Upsilon BA^{-1}BV_q\, y = \theta\, V_q^{\mathsf{T}}\Upsilon BV_q\, y.$$

The inverse of A is not available in practice. However, assuming again that $q \geq 1$, we have $A^{-1}BV_q = V_{q-1}$ along with $BV_q = AV_{q-1}$. In addition with the Υ-symmetry of A and B, one can rewrite the projected eigenvalue problem as

$$V_{q-1}^{\mathsf{T}}\Upsilon AV_{q-1}\, y = \theta\, V_q^{\mathsf{T}}\Upsilon AV_{q-1}\, y. \tag{4.11}$$

This resulting reduced eigenvalue problem no longer depends on either A^{-1} or B. Thus, although we assumed A nonsingular to derive (4.11), this projected eigenvalue problem can be formed and solved even in the case A is singular. The procedure is summarized in Algorithm 4.2. Although different, this approach is connected to the harmonic Rayleigh-Ritz method (see the discussion in Section 4.3.4).

Algorithm 4.2: Inverse approach for the GEP in initial form (4.1).

Input: Matrices $A, B \in \mathbb{R}^{n\times n}$ that are Υ-symmetric, number of random samples $1 \leq p \leq \operatorname{rank}(A)$, number of approximate eigenpairs $1 \leq k \leq p$ to provide, number of subspace iterations $q \geq 1$.

1 Draw a random matrix $\Omega \in \mathbb{R}^{n\times p}$, and set $V = \Omega$
2 Perform the thin QR factorization of $AV = QR$ and set $X = Q$
3 **for** $j = 1, \ \ldots, \ q-1$ **do**
4 Compute $V = B^{-1}X$
5 Perform the thin QR factorization of $AV = QR$ and set $X = Q$
6 **end**
7 Compute $Z = \Upsilon X \in \mathbb{R}^{n\times p}$ and form $T = R^{-\mathsf{T}}V^{\mathsf{T}}Z \in \mathbb{R}^{p\times p}$
8 Compute $V = B^{-1}X \in \mathbb{R}^{n\times p}$ and form $\Phi = V^{\mathsf{T}}Z \in \mathbb{R}^{p\times p}$
9 Solve the generalized Hermitian eigenvalue problem $TW = \Phi W\Theta$ with $W \in \mathbb{R}^{p\times p}$ a Φ-orthogonal matrix and $\Theta \in \mathbb{R}^{p\times p}$ a diagonal matrix with the eigenvalues sorted in **increasing** order
10 Remove the last $p-k$ columns of W and Θ
11 Remove the last $p-k$ rows of Θ
12 Set $\widetilde{V} = VW \in \mathbb{R}^{n\times k}$ and $\widetilde{\Lambda} = \Theta^{-1} \in \mathbb{R}^{k\times k}$.

Output: Matrices $\widetilde{V} \in \mathbb{R}^{n\times k}$ and $\widetilde{\Lambda} \in \mathbb{R}^{k\times k}$ such that $B^{-1}A\widetilde{V} \approx \widetilde{V}\widetilde{\Lambda}$ with $\widetilde{V}^{\mathsf{T}}\Upsilon B\widetilde{V} = I_k$ and $\widetilde{\Lambda}$ diagonal.

4.3.3 Algorithms for the generalized eigenvalue problem with basis transformation

Let us now propose two randomized algorithms for the eigenvalue problem with basis transformation, that is,

$$AB^{-1}u = \lambda u. \tag{4.12}$$

Again, deriving the Ritz approximation requires to determine the appropriate search space. With the basis transformation, the randomized subspace iteration considers matrices of the form $U_q = (AB^{-1})^q\Omega$, where $\Omega \in \mathbb{R}^{n\times p}$ is a full column rank random matrix satisfying $1 \leq p \leq \operatorname{rank}(A)$, and $q \in \mathbb{N}$.

Direct approach

To apply the Rayleigh-Ritz method, we recall that AB^{-1} is ΥB^{-1}-symmetric, which defines the inner product with respect to which the Galerkin condition is imposed. Altogether, we obtain

$$U_q^\mathsf{T}\Upsilon B^{-1}AB^{-1}U_q\, y = \mu\, U_q^\mathsf{T}\Upsilon B^{-1}U_q\, y. \tag{4.13}$$

Here, we remark that B does not appear in (4.13), without any further assumption on q. In particular, this means that the method can be derived without requiring B also for $q = 0$. However, as already mentioned, we expect poor performance of such a choice, since $U_0 = \Omega$ is very unlikely to contain eigeninformation.

Finding Ritz pairs (μ, y) satisfying (4.13) is then equivalent to solve a reduced generalized symmetric eigenvalue problem. If we denote $T = U_q^\mathsf{T}\Upsilon B^{-1}AB^{-1}U_q = U_q^\mathsf{T}\Upsilon B^{-1}U_{q+1}$ and $\Phi = U_q^\mathsf{T}\Upsilon B^{-1}U_q$, then Φ is symmetric positive definite by assumption on U_q and ΥB^{-1}. Consequently, there exist a Φ-orthonormal matrix $W \in \mathbb{R}^{p\times p}$ and a diagonal matrix $\Delta \in \mathbb{R}^{p\times p}$ such that

$$TW = \Phi W\Delta.$$

The approximate eigenvectors $\widetilde{U}$ and approximate eigenvalues $\widetilde{\Lambda}$ of (4.12) are then obtained via $\widetilde{U} = U_qW$ and $\widetilde{\Lambda} = \Delta$. As in Section 4.3.2, the approximate eigenvectors $\widetilde{U}$ are indeed ΥB^{-1}-orthonormal, and they inherit this property from the Φ-orthonormality of W.

The overall procedure is summarized in Algorithm 4.3.

Algorithm 4.3: Direct approach for the GEP with basis transformation (4.2).

Input: Matrices $A, B \in \mathbb{R}^{n\times n}$ that are Υ-symmetric, number of random samples $1 \leq p \leq \operatorname{rank}(A)$, number of approximate eigenpairs $1 \leq k \leq p$ to provide, number of subspace iterations $q \geq 0$.

1 Draw a random matrix $\Omega \in \mathbb{R}^{n\times p}$, and set $U = \Omega$
2 Perform the thin QR factorization of $B^{-1}U = QR$ and set $X = Q$
3 **for** $j = 1, \ \ldots, \ q$ **do**
4 Compute $U = AX$
5 Perform the thin QR factorization of $B^{-1}U = QR$ and set $X = Q$
6 **end**
7 Compute $Z = \Upsilon X \in \mathbb{R}^{n\times p}$ and form $\Phi = R^{-\mathsf{T}}U^\mathsf{T}Z \in \mathbb{R}^{p\times p}$
8 Compute $X = AX \in \mathbb{R}^{n\times p}$ and form $T = X^\mathsf{T}Z \in \mathbb{R}^{p\times p}$
9 Solve the generalized Hermitian eigenvalue problem $TW = \Phi W\Delta$ with $W \in \mathbb{R}^{p\times p}$ a Φ-orthogonal matrix and $\Delta \in \mathbb{R}^{p\times p}$ a diagonal matrix with the eigenvalues sorted in **decreasing** order
10 Remove the last $p - k$ columns of W and Δ
11 Remove the last $p - k$ rows of Δ
12 Set $\widetilde{U} = UW \in \mathbb{R}^{n\times k}$ and $\widetilde{\Lambda} = \Delta \in \mathbb{R}^{k\times k}$

Output: Matrices $\widetilde{U} \in \mathbb{R}^{n\times k}$ and $\widetilde{\Lambda} \in \mathbb{R}^{k\times k}$ such that $AB^{-1}\widetilde{U} \approx \widetilde{U}\widetilde{\Lambda}$ with $\widetilde{U}^\mathsf{T}\Upsilon B^{-1}\widetilde{U} = I_k$ and $\widetilde{\Lambda}$ diagonal.

Inverse approach

As in Section 4.3.2, let us assume temporarily that A is nonsingular, and let us rather study the inverse eigenvalue problem

$$BA^{-1}u = \theta u, \tag{4.14}$$

where $\theta = 1/\lambda$. Deriving the Ritz approximation of (4.14) then yields

$$U_q^{\mathsf{T}} \Upsilon A^{-1} U_q \, y = \theta \, U_q^{\mathsf{T}} \Upsilon B^{-1} U_q \, y.$$

Again, since A^{-1} is not available in practice, we assuming $q \geq 1$ so that $A^{-1}U_q = B^{-1}U_{q-1}$. In addition with Υ-symmetry of A and B one can rewrite the projected eigenvalue problem as

$$U_q^{\mathsf{T}} \Upsilon B^{-1} U_{q-1} \, y = \theta \, U_q^{\mathsf{T}} \Upsilon B^{-1} U_q \, y. \tag{4.15}$$

Here, A^{-1} no longer appear, meaning that (4.15) can be formed and solved even when A^{-1} is not available. The procedure is summarized in Algorithm 4.4.

Algorithm 4.4: Inverse approach for GEP with basis transformation (4.2).

Input: Matrices $A, B \in \mathbb{R}^{n\times n}$ that are Υ-symmetric, number of random samples $1 \leq p \leq \operatorname{rank}(A)$, number of approximate eigenpairs $1 \leq k \leq p$ to provide, number of subspace iterations $q \geq 0$.

1 Draw a random matrix $\Omega \in \mathbb{R}^{n\times p}$, and set $U = \Omega$
2 **for** $j = 1, \ \ldots, \ q$ **do**
3 Compute $X = B^{-1}U \in \mathbb{R}^{n\times p}$
4 Perform the thin QR factorization $AX = QR$ and set $U = Q$
5 **end**
6 Compute $Z = \Upsilon U \in \mathbb{R}^{n\times p}$ and form $T = R^{-\mathsf{T}} X^{\mathsf{T}} Z \in \mathbb{R}^{p\times p}$
7 Compute $X = B^{-1}U \in \mathbb{R}^{n\times p}$ and form $\Phi = X^{\mathsf{T}} Z$
8 Solve the generalized Hermitian eigenvalue problem $TW = \Phi W \Theta$ with $W \in \mathbb{R}^{p\times p}$ a Φ-orthogonal matrix and $\Theta \in \mathbb{R}^{p\times p}$ a diagonal matrix with the eigenvalues sorted in **increasing** order
9 Remove the last $p - k$ columns of W and Θ; remove last $p - k$ rows of Θ
10 Set $\widetilde{U} = UW \in \mathbb{R}^{n\times k}$ and $\widetilde{\Lambda} = \Theta^{-1} \in \mathbb{R}^{k\times k}$.

Output: Matrices $\widetilde{U} \in \mathbb{R}^{n\times k}$ and $\widetilde{\Lambda} \in \mathbb{R}^{k\times k}$ such that $AB^{-1}\widetilde{U} \approx \widetilde{U}\widetilde{\Lambda}$ with $\widetilde{U}^{\mathsf{T}} \Upsilon B^{-1} \widetilde{U} = I_k$ and $\widetilde{\Lambda}$ diagonal.

4.3.4 Relation between the inverse approaches and the harmonic Rayleigh-Ritz method

Let us relate the proposed inverse approaches to the harmonic Rayleigh-Ritz method [39, Section 2.5]. Let us first begin with the inverse approach for the initial GEP. In this context, a pair (x, μ) with $x \in \mathcal{S}$ and $\mu \in \mathbb{R}$ is an harmonic Ritz pair of $B^{-1}A$ associated to $\mathcal{S}$ if it satisfies

$$B^{-1}Ax - \mu x \perp_{\Upsilon B} B^{-1}A\mathcal{S}. \tag{4.16}$$

In the proposed algorithms, one has $\mathcal{S} = \mathcal{R}(V_q)$, and the matrix form of (4.16) then reads

$$V_q^{\mathsf{T}} A^{\mathsf{T}} B^{-\mathsf{T}} \Upsilon A V_q \, y = \mu \, V_q^{\mathsf{T}} A^{\mathsf{T}} B^{-\mathsf{T}} \Upsilon B V_q \, y.$$

Using the Υ-symmetry of A and B along with the relation $B^{-1}AV_q = V_{q+1}$ we obtain

$$V_{q+1}^{\mathsf{T}}\Upsilon AV_q\, y = \mu\, V_q^{\mathsf{T}}\Upsilon AV_q\, y.$$

Permuting left and right-hand sides, and defining $\theta = 1/\mu$ ($\mu \neq 0$) it finally yields

$$V_q^{\mathsf{T}}\Upsilon AV_q\, y = \theta\, V_{q+1}^{\mathsf{T}}\Upsilon AV_q\, y. \tag{4.17}$$

This is of the same form as (4.11), except for the index q. Therefore, for a given q, applying the inverse approach with V_{q+1} and the harmonic Rayleigh-Ritz approach with V_q will result in forming and solving the same reduced eigenvalue problem. However, we expect V_{q+1} to contain more accurate dominant eigeninformation than V_q because of the additional subspace iteration. Thus the output dominant eigenvectors using the harmonic Rayleigh-Ritz approach should be of poorer quality. Alternatively, if both methods are applied with the same search space V_q, then forming the reduced eigenvalue problem will be more costly with the harmonic Rayleigh-Ritz method than with the inverse approach. This is the main motivation why we preferred the inverse approach over the harmonic approach.

Similarly, for the inverse approach on the GEP with basis transformation, we say that a pair (x, μ) with $x \in \mathcal{S}$ and $\mu \in \mathbb{R}$ is an harmonic Ritz pair of AB^{-1} associated to $\mathcal{S}$ if it satisfies

$$AB^{-1}x - \mu x \perp_{\Upsilon B^{-1}} AB^{-1}\mathcal{S}. \tag{4.18}$$

Using that $\mathcal{S} = \mathcal{R}(U_q)$, the matrix form of (4.18) is

$$U_q^{\mathsf{T}}B^{-\mathsf{T}}A^{\mathsf{T}}\Upsilon B^{-1}AB^{-1}U_q\, y = \theta\, U_q^{\mathsf{T}}B^{-\mathsf{T}}A^{\mathsf{T}}\Upsilon B^{-1}U_q\, y. \tag{4.19}$$

Noticing that $AB^{-1}U_q = U_{q+1}$, permuting left and right-hand sides, and defining $\theta = 1/\mu$ ($\mu \neq 0$) we obtain

$$U_{q+1}\Upsilon B^{-1}U_{q+1}\, y = \theta\, U_{q+1}\Upsilon B^{-1}U_q\, y. \tag{4.20}$$

Again, this is similar to (4.15) except for the index q. Consequently, an analogous argument leads us to prefer the inverse approach over the harmonic Ritz method.

Nevertheless, let us insist on the fact that our inverse approaches are viable only because of the particular form of the search space we consider, while the harmonic Rayleigh-Ritz method is well defined for any search space.

4.3.5 Algorithmic considerations

For each approach, the implementation guideline is to form the appropriate reduced generalized Hermitian eigenvalue problem as efficiently as possible, using the different simplifications occurring when $q \geq 1$ highlighted in Sections 4.3.2 and 4.3.3. Consequently, the distinction between the construction of the search space, and the Rayleigh-Ritz method itself are not clearly separated. This aspect differs from the traditional derivation of randomized algorithms, where the construction phase of the search space (or range finder) is clearly separated from the approximation extraction phase. However, it is precisely from this interlacing that we obtain a cheaper implementation than the algorithms in [80] (see Section 4.3.6). The truncation performed at the end of each algorithm is optional, although it can improve the accuracy of the result whenever $p > k$.

We have mentioned in Remark 4.5 that the solution of the projected eigenvalue problem (step 8 in Algorithms 4.1 and 4.4 or step 9 in Algorithms 4.2 and 4.3) must be performed accurately because the orthogonality of the resulting eigenvectors is entirely determined by the Φ-orthogonality of the matrix $W \in \mathbb{R}^{p\times p}$. In the applications we target, the number of samples p will be moderate, meaning that solving the matrix pencil $\{T,\, \Phi\}$ can be done using direct methods

such as the QZ process [43, Section 7.7.7]. Thus, the thin QR factorizations are performed to ensure that the conditioning of the matrix pencil remains moderate. As an example, it can be verified that in Algorithm 4.1, one has

$$T = Q^\mathsf{T}\Upsilon AQ \quad \text{and} \quad \Phi = Q^\mathsf{T}\Upsilon BQ$$

with $Q \in \mathbb{R}^{n\times p}$ an orthogonal matrix. Therefore, the condition number of T and Φ is not worse than the one of ΥA and ΥB respectively, and the reduced eigenvalue problem is thus never more ill-conditioned that the initial system. In this regard, an accurate orthonormality of the columns of Q is not necessary, because the condition number of ΥA and ΥB is expected to be significantly larger than the one of $Q^\mathsf{T}Q$. If we denote by $Q \in \mathbb{R}^{n\times p}$ and $R \in \mathbb{R}^{p\times p}$ the matrices obtained from the QR factorization of $X \in \mathbb{R}^{n\times p}$, then the error $\|Q^\mathsf{T}Q - I_p\|$ is not critical in our algorithm. However, since we use the factors R and Q separately, it is crucial that the error $\|X - QR\|$ is as small as possible. This is critical to guarantee the symmetry of T or Φ (depending on the algorithm). In practice, it is also important that the QR factorization can be performed efficiently when n is large, which implies to use parallel algorithms as in [82, 96].

Then, we note that Algorithms 4.2 and 4.4 might break down if Θ is singular, which would mean that the reduced matrix H is singular too. However, this is of no practical concern for two reasons. First, since $\Omega \in \mathbb{R}^{n\times p}$ is such that $p \leq \operatorname{rank}(A)$, then $\mathcal{R}(\Omega)$ lies in the image of A almost surely. Consequently, H is almost surely nonsingular. Nevertheless, if it occurs, one can replace Θ^{-1} by its Moore-Penrose pseudo-inverse. In that case, the approximate eigenvectors associated with 0 will correspond to elements in the null space of A.

Table 4.1 summarizes the computational costs of all the algorithms. The first columns detail the number of applications of each operator to a block of p vectors. The formation of the reduced operators, and the truncation account for $4np^2$ and $2npk \leq 2np^2$ respectively. For the thin QR factorization, we consider for instance the Modified Gram-Schmidt algorithm whose cost is $2np^2$ when applied to a $n \times p$ matrix [43, Algorithm 5.2.6]. In total, the arithmetic operations account for no more than $4np^2 + 2np^2 + 2qnp^2 = 2np^2(q+3)$. In terms of memory requirements, all the algorithms require two blocks of p vectors, that is $2np$ coefficients.

All the methods have a different consumption of applications of A and B^{-1}. Hence, Algorithm 4.2 and 4.4 should be preferred in a context where applying these operators is known to be expensive.

Algorithm	A	B^{-1}	Υ	Other flops	Storage
4.1	$q+1$	q	1	$2np^2(q+3)$	$2np$
4.2	q	q	1	$2np^2(q+3)$	$2np$
4.3	$q+1$	$q+1$	1	$2np^2(q+3)$	$2np$
4.4	q	$q+1$	1	$2np^2(q+3)$	$2np$

Table 4.1: Computational costs and memory requirements of the algorithms. Columns 2 to 4 account for the number of applications of each operator to a block of p vectors. Column 5 accounts for the number of floating point operations induced by the arithmetic operations and column 6 the memory requirements.

4.3.6 Relations with prior algorithms

The proposed methods are actually general enough to allow us to recover a number of existing algorithms.

Algorithms from Saibaba, Lee and Katinidis [80]. It can be shown that Algorithms 4.1 and 4.2 are generalizations of Algorithms 6, 7 and 8 proposed in [80]. Let us denote those algorithms by `SLK_6`, `SLK_7` and `SLK_8` respectively. The scope of [80] was to propose randomized methods to address the generalized Hermitian eigenvalue problem involving the matrix pencil $\{A, B\}$ with $A \in \mathbb{R}^{n\times n}$ symmetric and $B \in \mathbb{R}^{n\times n}$ symmetric positive definite. There are two equivalent possibilities to address the same eigenvalue problem using Algorithms 4.1 and 4.2. Either we apply them to $\{A, B\}$ directly with $\Upsilon = I_n$, or we apply them to $\{B^{-1}A, I_n\}$ with $\Upsilon = B$. In [80], the derivation of the methods is rather based on the latter. However, we observe that the former does not require applications of B. Consequently, Algorithms 4.1 and 4.2 provide, in exact arithmetic, equivalent approximations as `SLK_6`, `SLK_7` and `SLK_8` at a lower computational cost. Numerical illustrations of this fact are given in Section 4.5.2. Table 4.2 gives the choice of parameters for Algorithms 4.1 and 4.2 to recover mathematically equivalent approximations as `SLK_6`, `SLK_7` and `SLK_8`.

	$q = 1$	$q = 2$
Algorithm 4.1	`SLK_6`	-
Algorithm 4.2	`SLK_7`	`SLK_8`

Table 4.2: Equivalence in exact arithmetic between the algorithms derived in Section 4.3.2 (with $\Upsilon = I_n$) and `SLK_6`, `SLK_7` and `SLK_8` (respectively Algorithms 6, 7 and 8 in [80]). The equivalences are verified when the algorithms are applied to the same pencil $\{A, B\}$ with identical values for p and k.

Algorithms from Saibaba and Katinidis [79]. In [79], the authors were interested in solving the generalized Hermitian eigenvalue problem involving the matrix pencil $\{A, B\}$ where A and B arise from the solution of a Bayesian inverse problem in geostatistics. In this regard, they address a problem formally similar to variational data assimilation. Algorithm 1 in [79] can be recovered in two different ways with our algorithms. The first option is to apply Algorithm 4.1 with $\Upsilon = I_n$ and $q = 1$ and to draw $\Omega \in \mathbb{R}^{n\times p}$ as a Gaussian matrix with covariance matrix $\mathbf{Cov}(\Omega) = B^{-2}$, i.e. $\Omega = B^{-1}G$ with $G \sim \mathcal{N}(0, I_n)$. Alternatively, we can apply Algorithm 4.3 with $\Upsilon = I_n$ and $q = 1$, and then postmultiply the approximate eigenvectors $\widetilde{U}$ by B^{-1}. Although the implementations differ, we notice that [79, Algorithm 1], unlike `SLK_6`, `SLK_7` and `SLK_8`, does not require applications of B and is thus similar on this aspect to our algorithms.

Nyström method [53]. Notably, the proposed algorithms also allow us to recover the Nyström method (Algorithm 2.5) for standard symmetric positive definite A. Assuming $\Upsilon = B = I_n$, and A symmetric positive definite, then the algorithms for the GEP in initial form and with basis transformation become identical. We indeed verify that either Algorithm 4.2 or 4.4 with $q = 1$ provides approximations identical to the Nyström method. This allows us to interpret the Nyström method in the frame of approximate eigenvalue problems.

Ritzit method from Daužickaitė et al [24]. Given A symmetric positive definite and $B = I_n$, we can also recover the Ritzit method, introduced in [24, Algorithm 8]. This method applies the Rayleigh-Ritz method to the squared matrix A^2, with a subspace of the form AG with $G \in \mathbb{R}^{n\times p}$ a standard Gaussian matrix. The Ritzit method can be recovered by applying for instance Algorithm 4.1 with A^2 with $B = \Upsilon = I_n$ and $q = 0$ and drawing the random matrix $\Omega \in \mathbb{R}^{n\times p}$ according to a Gaussian distribution with covariance matrix A^2. This distribution for Ω is indeed equivalent to consider $\Omega = AG$ with $G \in \mathbb{R}^{n\times p}$ a standard Gaussian matrix. The main reason why we have to modify the distribution is that the Ritzit method considers

a search subspace with a different structure than ours. However, since the distribution of Ω is fairly general, this can be bypassed.

4.3.7 Exploiting an additional matrix structure

In certain applications, the matrices A and B are related to each other. In the context of weighted nonlinear least-squares problems, the matrices A and B for instance satisfy

$$A = B + C,$$

with $A, B \in \mathbb{R}^{n\times n}$ being symmetric positive definite and $C \in \mathbb{R}^{n\times n}$ being symmetric positive semi-definite. In general, the rank m of C is much smaller than n. This particular structure allows us to significantly improve the approximations that can be obtained with the proposed algorithms. Indeed, we observe that in this case

$$B^{-1}A = I_n + B^{-1}C \quad \text{and} \quad AB^{-1} = I_n + CB^{-1}.$$

Consequently, the eigenvectors of $B^{-1}A$ and $B^{-1}C$ (respectively AB^{-1} and CB^{-1}) are identical, and the eigenvalues are related via a simple shift of 1. However, since $B^{-1}C$ and CB^{-1} are of rank $m \leq n$, their eigenvalue distribution has a smaller tail than $B^{-1}A$ and AB^{-1} respectively. As will be shown in the theoretical analysis proposed in Section 4.4, this means that applying the algorithms to $B^{-1}C$ and CB^{-1} is expected to yield more accurate approximations than applying them to $B^{-1}A$ and AB^{-1} respectively. This fact will be further detailed in the numerical experiments provided in Section 4.5, where we consider test matrices with such a structure.

4.4 Average-case analysis

Let us now propose a theoretical analysis of the algorithms presented in Section 4.3. In this section, the random matrix Ω is assumed to be a Gaussian matrix, thus denoted by $G \in \mathbb{R}^{n\times p}$, such that $G \sim \mathcal{N}(0, \mathbf{Cov}(G))$. We propose an average-case analysis of the methods based on low-rank approximation errors. The proposed results directly stem from the general analysis presented in Chapter 3. In this regard, let us begin with defining the appropriate norms. Let $A, \mathsf{W} \in \mathbb{R}^{n\times n}$ be two matrices with W symmetric positive definite. We define the following norms

$$\|A\|_{2,\mathsf{W}} = \max_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Ax\|_{\mathsf{W}}}{\|x\|_{\mathsf{W}}} \quad \text{and} \quad \|A\|_{F,\mathsf{W}} = \operatorname{tr}\left(A^{\mathsf{T}}\mathsf{W}A\mathsf{W}\right).$$

Those are respectively weighted spectral and Frobenius norms. Whenever possible, we will use the shortcut $\|A\|_{2,F,\mathsf{W}}$ to denote either the weighted spectral or Frobenius norm. Those norms are related to the standard spectral and Frobenius norms as

$$\|A\|_{2,F,\mathsf{W}} = \|\mathsf{W}^{\frac{1}{2}}A\mathsf{W}^{-\frac{1}{2}}\|_{2,F}. \tag{4.21}$$

We separate the analysis between the methods for the initial GEP and the ones for the transformed GEP. Let us begin with analyzing the methods for the initial GEP, that is Algorithms 4.1 and 4.2. We analyze the randomized methods for the GEP in initial form in Section 4.4.1, and the ones for the GEP with basis transformation in Section 4.4.2 As will be shown, the analysis for Algorithms 4.3 and 4.4 can be deduced from the one of Algorithms 4.1 and 4.2. As a consequence, we put more effort in the details for the first analysis.

4.4.1 Probabilistic analysis of the randomized methods for the generalized eigenvalue problem in initial form

Let us consider the eigenvalue decomposition $B^{-1}AV = V\Lambda$, where $V \in \mathbb{R}^{n\times n}$ is a ΥB-orthonormal matrix containing the eigenvectors and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ with $\lambda_1 \geq \cdots \geq \lambda_n$ the corresponding eigenvalues. For a given $k \in \{1, n\}$, we define the following partitioning

$$B^{-1}A = \begin{bmatrix} V_k & \underline{V}_k \end{bmatrix} \begin{bmatrix} \Lambda_k & \\ & \underline{\Lambda}_k \end{bmatrix} \begin{bmatrix} V_k^\mathsf{T} \\ \underline{V}_k^\mathsf{T} \end{bmatrix} \Upsilon B, \tag{4.22}$$

with $V_k \in \mathbb{R}^{n\times k}$, $\underline{V}_k \in \mathbb{R}^{n\times(n-k)}$, $\Lambda_k \in \mathbb{R}^{k\times k}$ and $\underline{\Lambda}_k \in \mathbb{R}^{(n-k)\times(n-k)}$. Let us also set $A_k = BV_k\Lambda_kV_k^\mathsf{T}\Upsilon B$ and $\underline{A}_k = B\underline{V}_k\underline{\Lambda}_k\underline{V}_k^\mathsf{T}\Upsilon B$ so that $B^{-1}A = B^{-1}A_k + B^{-1}\underline{A}_k$. Also, for a given k, we define the eigenvalue ratios as

$$\gamma_i = \frac{\lambda_{k+1}}{\lambda_i}, \quad 1 \leq i \leq k. \tag{4.23}$$

The following proposition is a variant of the Eckart-Young theorem [30], which provides an expression for the optimal rank k approximation of $B^{-1}A$. Besides justifying the definitions of the weighted norms, it will be at the core of the forthcoming analysis.

Proposition 4.6. *Let $\Upsilon \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix. Let $A, B \in \mathbb{R}^{n\times n}$ be Υ-symmetric matrices such that ΥB is symmetric positive definite. Then one has*

$$\min_{\substack{Z\in\mathbb{R}^{n\times k} \\ \mathrm{rank}(Z)=k}} \|[I_n - \pi_{\Upsilon B}(Z)]B^{-1}A\|_{2,F,\Upsilon B} = \|\underline{\Lambda}_k\|_{2,F}$$

where the optimal value is attained for $\mathcal{R}(Z) = \mathcal{R}(V_k)$.

Proof. Using the relation between the weighted and standard norms highlighted in (4.21), we have

$$\|[I_n - \pi_{\Upsilon B}(Z)]B^{-1}A\|_{2,F,\Upsilon B} = \|(\Upsilon B)^{\frac{1}{2}}[I_n - \pi_{\Upsilon B}(Z)]B^{-1}A(\Upsilon B)^{-\frac{1}{2}}\|_{2,F}.$$

Then,

$$\begin{aligned}(\Upsilon B)^{\frac{1}{2}}[I_n - \pi_{\Upsilon B}(Z)]B^{-1}A(\Upsilon B)^{-\frac{1}{2}} &= (\Upsilon B)^{\frac{1}{2}}[I_n - Z(Z^\mathsf{T}\Upsilon BZ)^{-1}Z^\mathsf{T}\Upsilon B]B^{-1}A(\Upsilon B)^{-\frac{1}{2}} \\ &= [I_n - \pi((\Upsilon B)^{\frac{1}{2}}Z)](\Upsilon B)^{\frac{1}{2}}B^{-1}A(\Upsilon B)^{-\frac{1}{2}} \\ &= \widehat{A} - \pi((\Upsilon B)^{\frac{1}{2}}Z)\widehat{A},\end{aligned}$$

where $\widehat{A} = (\Upsilon B)^{\frac{1}{2}}B^{-1}A(\Upsilon B)^{-\frac{1}{2}}$. Then, from the eigenvalue decomposition of $B^{-1}A$ we have

$$\widehat{A} = (\Upsilon B)^{\frac{1}{2}}V\Lambda V^\mathsf{T}(\Upsilon B)^{\frac{1}{2}}. \tag{4.24}$$

The ΥB-orthonormality of V implies that $(\Upsilon B)^{\frac{1}{2}}V$ is orthonormal. Relation (4.24) is then simply the standard SVD of $\widehat{A}$. Applying [13, Theorem 2.2.11], we obtain that the solution of the minimization problem is

$$\pi((\Upsilon B)^{\frac{1}{2}}Z)\widehat{A} = (\Upsilon B)^{\frac{1}{2}}V_k\Lambda_kV_k^\mathsf{T}(\Upsilon B)^{\frac{1}{2}} = \pi((\Upsilon B)^{\frac{1}{2}}V_k)\widehat{A},$$

that is $\mathcal{R}(Z) = \mathcal{R}(V_k)$, with the corresponding optimal value being $\|\underline{\Lambda}_k\|_{2,F}$. □

Since we are looking for approximations of the dominant eigenvectors, Proposition 4.6 allows us to consider $\|[I_n - \pi_{\Upsilon B}(Z)]B^{-1}A\|_{2,F,\Upsilon B}$ as an indirect measure of the approximate eigenvector accuracy. The closest this quantity is to $\|\underline{\Lambda}_k\|_{2,F}$ the better Z approximates V_k. This is where

the low rank approximation error meets the approximate eigenvector error. This is relevant in our situation since the algorithms proposed in Section 4.3 are dedicated to the approximation of the dominant eigenmodes.

To analyze Algorithms 4.1 and 4.2, we observe that they both rely on the same search space, that is $Z = (B^{-1}A)^q G$. They only differ in the way they extract the approximations from Z. However, the general error analysis from Chapter 3 is uniquely based on the search space. Consequently, the following analysis will be similar for Algorithms 4.1 and 4.2, although their theoretical foundations are different. Said differently, our theoretical analysis is blind to the extraction phase, which is a second limitation.

From elementary properties of Gaussian vectors, Z is Gaussian with covariance matrix $\mathbf{Cov}(Z) = (B^{-1}A)^q \mathbf{Cov}(G)(A^\mathsf{T}B^{-\mathsf{T}})^q$. The theoretical analysis is based on a block partitioning of the matrix $V^\mathsf{T}\Upsilon B\,\mathbf{Cov}(Z)\,\Upsilon BV$. For a given integer $k \in \{1, \dots, p\}$, we define

$$\begin{aligned} V^\mathsf{T}\Upsilon B\,\mathbf{Cov}(Z)\,\Upsilon BV &= \begin{bmatrix} V_k^\mathsf{T} \\ \underline{V}_k^\mathsf{T} \end{bmatrix} \Upsilon B\,\mathbf{Cov}(Z)\,\Upsilon B \begin{bmatrix} V_k & \underline{V}_k \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{Cov}_k(Z) & \mathbf{Cov}_{\perp,k}(Z)^\mathsf{T} \\ \mathbf{Cov}_{\perp,k}(Z) & \underline{\mathbf{Cov}}_k(Z) \end{bmatrix}, \end{aligned}$$

with

$$\begin{aligned} \mathbf{Cov}_k(Z) &= V_k^\mathsf{T}\Upsilon B\,\mathbf{Cov}(Z)\,\Upsilon BV_k \in \mathbb{R}^{k\times k} \\ \mathbf{Cov}_{\perp,k}(Z) &= \underline{V}_k^\mathsf{T}\Upsilon B\,\mathbf{Cov}(Z)\,\Upsilon BV_k \in \mathbb{R}^{(n-k)\times k}, \\ \underline{\mathbf{Cov}}_k(Z) &= \underline{V}_k^\mathsf{T}\Upsilon B\,\mathbf{Cov}(Z)\,\Upsilon B\underline{V}_k \in \mathbb{R}^{(n-k)\times(n-k)}. \end{aligned} \tag{4.25}$$

Similarly, and for reasons that will be made clear later, we also introduce the following block partitioning of $V^\mathsf{T}\Upsilon B\,\mathbf{Cov}(G)\,\Upsilon BV$

$$\begin{aligned} V^\mathsf{T}\Upsilon B\,\mathbf{Cov}(G)\,\Upsilon BV &= \begin{bmatrix} V_k^\mathsf{T} \\ \underline{V}_k^\mathsf{T} \end{bmatrix} \Upsilon B\,\mathbf{Cov}(G)\,\Upsilon B \begin{bmatrix} V_k & \underline{V}_k \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{Cov}_k(G) & \mathbf{Cov}_{\perp,k}(G)^\mathsf{T} \\ \mathbf{Cov}_{\perp,k}(G) & \underline{\mathbf{Cov}}_k(G) \end{bmatrix}. \end{aligned}$$

Let us define $\Omega_k = V_k^\mathsf{T}\Upsilon BZ$ and $\underline{\Omega}_k = \underline{V}_k^\mathsf{T}\Upsilon BZ$. Using the properties of Gaussian vectors we have $\Omega_k \sim \mathcal{N}(0, \mathbf{Cov}_k(Z))$ and $\underline{\Omega}_k \sim \mathcal{N}(0, \underline{\mathbf{Cov}}_k(Z))$. Deriving the theoretical bounds necessitates to first express the different reduced covariance matrices defined in (4.25), as long as the conditional covariance matrix of $\underline{\Omega}_k$ with respect to Ω_k. This is the object of Lemma 4.8.

Remark 4.7. We point out that, although the matrices Ω_k and $\underline{\Omega}_k$ are centered, the conditional law of $\underline{\Omega}_k$ with respect to Ω_k also follows a Gaussian distribution which is not necessarily centered [67, Theorem 1.2.11].

Lemma 4.8. *Let $\Upsilon \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix. Let $A, B \in \mathbb{R}^{n\times n}$ be Υ-symmetric matrices such that ΥB is symmetric positive definite. Let $G \in \mathbb{R}^{n\times p}$ be a Gaussian matrix such that $G \sim \mathcal{N}(0, \mathbf{Cov}(G))$ satisfying $2 < p \leq \min\{\mathrm{rank}(A), \mathrm{rank}(\mathbf{Cov}(G))\}$. Let $Z = (B^{-1}A)^q G \in \mathbb{R}^{n\times p}$, with $q \in \mathbb{N}$. Then for any given integer $k \in \{1, \dots, p\}$, one has*

$$\begin{aligned} \mathbf{Cov}_k(Z) &= \Lambda_k^q\,\mathbf{Cov}_k(G)\,\Lambda_k^q \\ \mathbf{Cov}_{\perp,k}(Z) &= \underline{\Lambda}_k^q\,\mathbf{Cov}_{\perp,k}(G)\,\Lambda_k^q \\ \underline{\mathbf{Cov}}_k(Z) &= \underline{\Lambda}_k^q\,\underline{\mathbf{Cov}}_k(G)\,\underline{\Lambda}_k^q. \end{aligned}$$

Furthermore, set $\Omega_k = V_k^\mathsf{T}\Upsilon BZ$ and $\underline{\Omega}_k = \underline{V}_k^\mathsf{T}\Upsilon BZ$. If the covariance matrix $\mathbf{Cov}_k(G)$ is nonsingular, then the random matrix $\underline{\Omega}_k$ conditioned by Ω_k follows a Gaussian distribution of

covariance matrix

$$\mathbf{Cov}(\underline{\Omega}_k \mid \Omega_k) = \underline{\Lambda}_k^q \left(\underline{\mathbf{Cov}}_k(G) / \mathbf{Cov}_k(G) \right) \underline{\Lambda}_k^q.$$

Proof. From the eigenvalue decomposition of $B^{-1}A$ we have $(B^{-1}A)^q = V\Lambda^q V^{\mathsf{T}} \Upsilon B$, which yields

$$V^{\mathsf{T}} \Upsilon B \, \mathbf{Cov}(Z) \, \Upsilon B V = \Lambda^q V^{\mathsf{T}} \Upsilon B \, \mathbf{Cov}(G) \Upsilon B V \Lambda^q,$$

where we have used the ΥB-orthonormality of V. The expression for $\mathbf{Cov}_k(Z)$, $\mathbf{Cov}_{\perp,k}(Z)$ and $\underline{\mathbf{Cov}}_k(Z)$ then follows similarly. For the conditional distribution, the conditional covariance matrix can be expressed in terms of a Schur complement (Lemma 3.9) as

$$\begin{aligned} \mathbf{Cov}\left(\underline{\Omega}_k \mid \Omega_k\right) &= \underline{\mathbf{Cov}}_k(Z) / \mathbf{Cov}_k(Z) \\ &= \underline{\Lambda}_k^q \left(\underline{\mathbf{Cov}}_k(G) / \mathbf{Cov}_k(G) \right) \underline{\Lambda}_k^q. \end{aligned}$$

□

We are now ready to state the main theorems. Theorems 4.9 and 4.10 address the average-case error in weighted Frobenius and spectral norms respectively. Those are obtained by application of Theorems 3.18 and 3.20 from Chapter 3, respectively. We propose error bounds for

$$\|[I_n - \pi_{\Upsilon B}(Z)]B^{-1}A\|_{2,F,\Upsilon B} - \|[I_n - \pi_{\Upsilon B}(Z)]B^{-1}\underline{A}_k\|_{2,F,\Upsilon B}.$$

Since $\|[I_n - \pi_{\Upsilon B}(Z)]B^{-1}\underline{A}_k\|_{2,F,\Upsilon B} \leq \|\underline{\Lambda}_k\|_{2,F}$, it is clear that the proposed bounds are also bounds for

$$\|[I_n - \pi_{\Upsilon B}(Z)]B^{-1}A\|_{2,F,\Upsilon B} - \|\underline{\Lambda}_k\|_{2,F}.$$

Therefore, our bounds imply bounds for this latter quantity, which are generally preferred in the randomized numerical linear algebra community.

Theorem 4.9 (Average-case analysis in weighted Frobenius norm)**.** *Let $\Upsilon \in \mathbb{R}^{n\times n}$ be symmetric positive definite and $A, B \in \mathbb{R}^{n\times n}$ be Υ-symmetric matrices such that ΥB is symmetric positive definite. Let $G \in \mathbb{R}^{n\times p}$ be a Gaussian matrix such that $G \sim \mathcal{N}(0, \mathbf{Cov}(G))$ with $2 < p \leq \min\{\mathrm{rank}(A), \mathrm{rank}(\mathbf{Cov}(G)\}$. Let $Z = (B^{-1}A)^q G \in \mathbb{R}^{n\times p}$, with $q \in \mathbb{N}$ and let $\pi_{\Upsilon B}(Z)$ denote the ΥB-orthogonal projection onto the vector space spanned by the columns of Z. Let $\varphi : x \mapsto x/\sqrt{1+x^2}$ for $x \geq 0$. Then, if $\mathbf{Cov}_k(G)$ is nonsingular, we have for all $k \in \{1, \ldots, p-2\}$*

$$\mathbb{E}\left[\|[I_n - \pi_{\Upsilon B}(Z)]B^{-1}A\|_{F,\Upsilon B} - \|[I_n - \pi_{\Upsilon B}(Z)]B^{-1}\underline{A}_k\|_{F,\Upsilon B} \right] \leq \min\left\{ \sqrt{\alpha_k}, \sqrt{k}\,\varphi\left(\frac{\sqrt{\beta_k}}{\sqrt{k}}\right)\lambda_1 \right\},$$

where

$$\begin{aligned} \alpha_k &= \|\underline{\Lambda}_k^q\|_F^2 \, \frac{c_k(G)^2}{\lambda_k^{2(q-1)}} + \|\underline{\Lambda}_k^q\|_F^2 \, \delta_k(G) \frac{\|\Lambda_k^{-(q-1)}\|_F^2}{p-k-1} \\ \beta_k &= \|\underline{\Lambda}_k^q\|_F^2 \, \frac{c_k(G)^2}{\lambda_k^{2q}} + \|\underline{\Lambda}_k^q\|_F^2 \, \delta_k(G) \frac{\|\Lambda_k^{-q}\|_F^2}{p-k-1}, \end{aligned}$$

with $\mathbf{Cov}_k(G) = V_k^{\mathsf{T}} \Upsilon B \, \mathbf{Cov}(G) \, \Upsilon B V_k$, $\underline{\mathbf{Cov}}_k(G) = \underline{V}_k^{\mathsf{T}} \Upsilon B \, \mathbf{Cov}(G) \, \Upsilon B \underline{V}_k$, $\mathbf{Cov}_{\perp,k}(G) = \underline{V}_k^{\mathsf{T}} \Upsilon B \, \mathbf{Cov}(G) \, \Upsilon B V_k$, $\delta_k(G) = \|\,\mathbf{Cov}_k(G)^{-1}\|_2 \|\,\underline{\mathbf{Cov}}_k(G) / \mathbf{Cov}_k(G)\|_2$, *and* $c_k(G) = \|\,\mathbf{Cov}_{\perp,k}(G)\;\mathbf{Cov}_k(G)^{-1}\|_2$.

Proof. Let us define $\widehat{A} = (\Upsilon B)^{\frac{1}{2}} B^{-1} A (\Upsilon B)^{\frac{1}{2}}$. We recall that the eigenvalue decomposition of $B^{-1}A$ corresponds to the SVD of $\widehat{A}$ (4.6), and that from (4.21) one has

$$\|[I_n - \pi_S(Z)]B^{-1}A\|_{F,\Upsilon B} = \|[I_n - \pi((\Upsilon B)^{\frac{1}{2}} Z)]\widehat{A}\|_F,$$

Applying Theorem 3.18 then yields

$$\mathbb{E}\left[\, \|[I_n - \pi((\Upsilon B)^{\frac{1}{2}} Z]\widehat{A}\|_F - \|[I_n - \pi((\Upsilon B)^{\frac{1}{2}} Z]\underline{\widehat{A}}_k\|_F \,\right] \leq \min\left\{ \sqrt{a_k},\ \sqrt{k}\ \varphi\left(\frac{1}{\sqrt{k}}\ \sqrt{b_k}\right) \lambda_1 \right\},$$

where

$$a_k = \|\operatorname{\mathbf{Cov}}_{\perp,k}(Z)[\operatorname{\mathbf{Cov}}_k(Z)]^{-1}\Lambda_k\|_F^2 + \frac{\|\operatorname{\mathbf{Cov}}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_F^2 \|(\Lambda_k^{\mathsf{T}}[\operatorname{\mathbf{Cov}}_k(Z)]^{-1}\Lambda_k)^{\frac{1}{2}}\|_F^2}{p-k-1}, \tag{4.26}$$

$$b_k = \|\operatorname{\mathbf{Cov}}_{\perp,k}(Z)[\operatorname{\mathbf{Cov}}_k(Z)]^{-1}\|_F^2 + \frac{\|\operatorname{\mathbf{Cov}}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_F^2 \|[\operatorname{\mathbf{Cov}}_k(Z)]^{-\frac{1}{2}}\|_F^2}{p-k-1}. \tag{4.27}$$

Step 1: Bounding $\|\operatorname{\mathbf{Cov}}_{\perp,k}(Z)[\operatorname{\mathbf{Cov}}_k(Z)]^{-1}\|_F^2$ and $\|\operatorname{\mathbf{Cov}}_{\perp,k}(Z)[\operatorname{\mathbf{Cov}}_k(Z)]^{-1}\Lambda_k\|_F^2$. One gets from Lemma 4.8 that

$$\operatorname{\mathbf{Cov}}_{\perp,k}(Z)[\operatorname{\mathbf{Cov}}_k(Z)]^{-1} = \underline{\Lambda}_k^q \operatorname{\mathbf{Cov}}_{\perp,k}(G)\ \operatorname{\mathbf{Cov}}_k(G)^{-1}\Lambda_k^{-q}.$$

Taking the squared Frobenius norm and applying the submultiplicativity, one gets

$$\|\operatorname{\mathbf{Cov}}_{\perp,k}(Z)[\operatorname{\mathbf{Cov}}_k(Z)]^{-1}\|_F^2 \leq \|\underline{\Lambda}_k^q\|_F^2 \|\operatorname{\mathbf{Cov}}_{\perp,k}(G)\ \operatorname{\mathbf{Cov}}_k(G)^{-1}\|_2^2 \|\Lambda_k^{-q}\|_2^2.$$

Note that $\|\Lambda_k^{-q}\|_2^2 = \lambda_k^{-2q}$ and defining $c_k(G) = \|\operatorname{\mathbf{Cov}}_{\perp,k}(G)\ \operatorname{\mathbf{Cov}}_k(G)^{-1}\|_2$, we obtain

$$\|\operatorname{\mathbf{Cov}}_{\perp,k}(Z)[\operatorname{\mathbf{Cov}}_k(Z)]^{-1}\|_F^2 \leq c_k(G)^2\, \frac{\|\underline{\Lambda}_k^q\|_F^2}{\lambda_k^{2q}}. \tag{4.28}$$

Similarly, we obtain

$$\|\operatorname{\mathbf{Cov}}_{\perp,k}(Z)[\operatorname{\mathbf{Cov}}_k(Z)]^{-1}\Lambda_k\|_F^2 \leq c_k(G)^2\, \frac{\|\underline{\Lambda}_k^q\|_F^2}{\lambda_k^{2(q-1)}}. \tag{4.29}$$

Step 2: Bounding $\|\operatorname{\mathbf{Cov}}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_F^2$. From Lemma 4.8 we have

$$\operatorname{\mathbf{Cov}}\left(\underline{\Omega}_k \mid \Omega_k\right) = \underline{\Lambda}_k^q \left(\underline{\operatorname{\mathbf{Cov}}}_k(G)/\operatorname{\mathbf{Cov}}_k(G)\right)\underline{\Lambda}_k^q.$$

Then, using the partial ordering and taking the trace yields

$$\|\operatorname{\mathbf{Cov}}\left(\underline{\Omega}_k \mid \Omega_k\right)^{\frac{1}{2}}\|_F^2 \leq \|\underline{\operatorname{\mathbf{Cov}}}_k(G)/\operatorname{\mathbf{Cov}}_k(G)\|_2\, \|\underline{\Lambda}_k^q\|_F^2. \tag{4.30}$$

Step 3: Bounding $\|[\operatorname{\mathbf{Cov}}_k(Z)]^{-\frac{1}{2}}\|_F^2$ and $\|(\Lambda_k^{\mathsf{T}}[\operatorname{\mathbf{Cov}}_k(Z)]^{-1}\Lambda_k)^{\frac{1}{2}}\|_F^2$. From Lemma 4.8, one gets

$$[\operatorname{\mathbf{Cov}}_k(Z)]^{-1} = \Lambda_k^{-q} \operatorname{\mathbf{Cov}}_k(G)^{-1}\Lambda_k^{-q} \preccurlyeq \|\operatorname{\mathbf{Cov}}_k(G)^{-1}\|_2 \Lambda_k^{-2q}.$$

Taking the trace yields

$$\|[\operatorname{\mathbf{Cov}}_k(Z)]^{-\frac{1}{2}}\|_F^2 \leq \|\operatorname{\mathbf{Cov}}_k(G)^{-1}\|_2 \|\Lambda_k^{-q}\|_F^2. \tag{4.31}$$

In a similar way, we obtain

$$\|(\Lambda_k^{\mathsf{T}}[\operatorname{\mathbf{Cov}}_k(Z)]^{-1}\Lambda_k)^{\frac{1}{2}}\|_F^2 \leq \|\operatorname{\mathbf{Cov}}_k(G)^{-1}\|_2 \|\Lambda_k^{-(q-1)}\|_F^2. \tag{4.32}$$

Summary. Plugging (4.29), (4.30) and (4.32) into (4.26), we obtain

$$a_k \leq c_k(G)^2 \, \frac{\|\underline{\Lambda}_k^q\|_F^2}{\lambda_k^{2(q-1)}} + \delta_k(G) \, \frac{\|\underline{\Lambda}_k^q\|_F^2 \|\Lambda_k^{-(q-1)}\|_F^2}{p-k-1} = \alpha_k,$$

where we have used that $\delta_k(G) = \|\,\mathbf{Cov}_k(G)^{-1}\|_2 \|\,\underline{\mathbf{Cov}}_k(G)/\,\mathbf{Cov}_k(G)\|_2$. Similarly, plugging (4.28), (4.30) and (4.31) into (4.27), we obtain

$$b_k \leq c_k(G)^2 \, \frac{\|\underline{\Lambda}_k^q\|_F^2}{\lambda_k^{2q}} + \delta_k(G) \, \frac{\|\underline{\Lambda}_k^q\|_F^2 \|\Lambda_k^{-q}\|_F^2}{p-k-1} = \beta_k.$$

To complete the proof, we notice that $\varphi : x \mapsto x/\sqrt{1+x^2}$ is monotonically increasing. □

Theorem 4.10 (Average-case analysis in weighted spectral norm)**.** *Let $\Upsilon \in \mathbb{R}^{n\times n}$ be symmetric positive definite and $A, B \in \mathbb{R}^{n\times n}$ be Υ-symmetric matrices such that ΥB is symmetric positive definite. Let $G \in \mathbb{R}^{n\times p}$ be a Gaussian matrix such that $G \sim \mathcal{N}(0, \mathbf{Cov}(G))$ with $2 < p \leq \min\{\mathrm{rank}(A), \mathrm{rank}(\mathbf{Cov}(G))\}$. Let $Z = (B^{-1}A)^q G \in \mathbb{R}^{n\times p}$, with $q \in \mathbb{N}$ and let $\pi_{\Upsilon B}(Z)$ denote the ΥB-orthogonal projection onto the vector space spanned by the columns of Z. Let $\varphi : x \mapsto x/\sqrt{1+x^2}$ for $x \geq 0$. Then, if $\mathbf{Cov}_k(G)$ is nonsingular, we have for all $k \in \{1, \ldots, p-2\}$*

$$\mathbb{E}\left[\, \|[I_n - \pi_{\Upsilon B}(Y)]B^{-1}A\|_{2,\Upsilon B} - \|[I_n - \pi_{\Upsilon B}(Y)]B^{-1}\underline{A}_k\|_{2,\Upsilon B} \,\right] \leq \min\left\{c_k,\ \varphi(d_k)\lambda_1\right\},$$

where

$$c_k = \lambda_{k+1}\,\gamma_k^{q-1}\,c_k(G) + \sqrt{\delta_k(G)}\left(\lambda_{k+1}^q \, \frac{\|\Lambda_k^{-(q-1)}\|_F}{\sqrt{p-k-1}} + \frac{\|\underline{\Lambda}_k^q\|_F}{\lambda_k^{q-1}} \, \frac{e\sqrt{p}}{p-k}\right)$$

$$d_k = \gamma_k^q\,c_k(G) + \sqrt{\delta_k(G)}\left(\lambda_{k+1}^q \, \frac{\|\Lambda_k^{-q}\|_F}{\sqrt{p-k-1}} + \frac{\|\underline{\Lambda}_k^q\|_F}{\lambda_k^q} \, \frac{e\sqrt{p}}{p-k}\right)$$

with $\mathbf{Cov}_k(G) = V_k^{\mathsf{T}}\Upsilon B\,\mathbf{Cov}(G)\,\Upsilon B V_k$, $\underline{\mathbf{Cov}}_k(G) = \underline{V}_k^{\mathsf{T}}\Upsilon B\,\mathbf{Cov}(G)\,\Upsilon B\underline{V}_k$, $\mathbf{Cov}_{\perp,k}(G) = \underline{V}_k^{\mathsf{T}}\Upsilon B\,\mathbf{Cov}(G)\,\Upsilon B V_k$, $\delta_k(G) = \|\,\mathbf{Cov}_k(G)^{-1}\|_2 \|\,\underline{\mathbf{Cov}}_k(G)/\,\mathbf{Cov}_k(G)\|_2$, *and* $c_k(G) = \|\,\mathbf{Cov}_{\perp,k}(G)\;\mathbf{Cov}_k(G)^{-1}\|_2$.

Proof. The proof follows similar arguments and is based on Theorem 3.20. □

Let us now consider the particular case where we change the quantity of interest, and rather consider

$$\|[I_n - \pi_{\Upsilon B}(Z)]B^{-1}A\|_{2,F,\Upsilon B} - \|\underline{\Lambda}_k\|_2.$$

In this case we can derive an improved bound for the weighted spectral norm, derived from Theorem 3.21 and stated in Theorem 4.11.

Theorem 4.11 (Average-case analysis in weighted spectral norm, improved bound)**.** *Let $\Upsilon \in \mathbb{R}^{n\times n}$ be symmetric positive definite and $A, B \in \mathbb{R}^{n\times n}$ be Υ-symmetric matrices such that ΥB is symmetric positive definite. Let $G \in \mathbb{R}^{n\times p}$ be a Gaussian matrix such that $G \sim \mathcal{N}(0, \mathbf{Cov}(G))$ with $2 < p \leq \min\{\mathrm{rank}(A),\ \mathrm{rank}(\mathbf{Cov}(G))\}$. Let $Z = (B^{-1}A)^q G \in \mathbb{R}^{n\times p}$, with $q \in \mathbb{N}$ and let $\pi_{\Upsilon B}(Z)$ denote the ΥB-orthogonal projection onto the vector space spanned by the columns of Z. Let $\varphi : x \mapsto x/\sqrt{1+x^2}$ for $x \geq 0$. Then, if $\mathbf{Cov}_k(G)$ is nonsingular, we have for all $k \in \{1, \ldots, p-2\}$*

$$\mathbb{E}\left[\, \|[I_n - \pi_{\Upsilon B}(Y)]B^{-1}A\|_{2,\Upsilon B} - \|\underline{\Lambda}_k\|_2 \,\right] \leq \min\left\{\widehat{c}_k,\ \varphi(\widehat{d}_k)\sqrt{\lambda_1^2 - \lambda_{k+1}^2}\right\},$$

where

$$\widehat{c}_k = \lambda_{k+1}\,\gamma_\ell^{q-1}\sqrt{1-\gamma_\ell^2}\,c_k(G) + \sqrt{\delta_k(G)}\left(\sum_{i=1}^{k}\gamma_i^{2q-1}\sqrt{1-\gamma_i^2}\right)^{\frac{1}{2}}\frac{\lambda_{k+1}}{\sqrt{p-k-1}}$$
$$+\sqrt{1-\gamma_\ell^2}\,\frac{\|\underline{\Lambda}_k^q\|_F}{\lambda_\ell^{q-1}}\,\frac{e\sqrt{p}}{p-k}$$
$$\widehat{d}_k = \gamma_k^q\,c_k(G) + \sqrt{\delta_k(G)}\left(\lambda_{k+1}^q\,\frac{\|\Lambda_k^{-q}\|_F}{\sqrt{p-k-1}} + \frac{\|\underline{\Lambda}_k^q\|_F}{\lambda_k^q}\,\frac{e\sqrt{p}}{p-k}\right)$$

with $\mathbf{Cov}_k(G) = V_k^{\mathsf{T}}\Upsilon B\,\mathbf{Cov}(G)\,\Upsilon B V_k$, $\underline{\mathbf{Cov}}_k(G) = \underline{V}_k^{\mathsf{T}}\Upsilon B\,\mathbf{Cov}(G)\,\Upsilon B\underline{V}_k$, $\mathbf{Cov}_{\perp,k}(G) = \underline{V}_k^{\mathsf{T}}\Upsilon B\,\mathbf{Cov}(G)\,\Upsilon B V_k$, $\delta_k(G) = \|\,\mathbf{Cov}_k(G)^{-1}\|_2\|\,\underline{\mathbf{Cov}}_k(G)/\,\mathbf{Cov}_k(G)\|_2$, $c_k(G) = \|\,\mathbf{Cov}_{\perp,k}(G)\;\mathbf{Cov}_k(G)^{-1}\|_2$ *and* $\widehat{\Lambda}_k = (\Lambda_k^2 - \lambda_{k+1}^2 I_k)^{\frac{1}{2}}$.

$$\ell = \begin{cases} 1 & \text{if} \quad \dfrac{q}{q-1}\lambda_{k+1}^2 \geq \lambda_1^2 \quad \text{or} \quad q = 1 \\ \arg\max\left\{\dfrac{\sqrt{1-\gamma_i^2}}{\lambda_i^{q-1}}, \dfrac{\sqrt{1-\gamma_{i+1}^2}}{\lambda_{i+1}^{q-1}}\right\} & \text{if} \quad \dfrac{q}{q-1}\lambda_{k+1}^2 \in [\lambda_i, \lambda_{i+1}] \\ k & \text{if} \quad \dfrac{q}{q-1}\lambda_{k+1}^2 \leq \lambda_k^2 \end{cases}.$$

Proof. Applying Theorem 3.21, it remains to compute $\|\Lambda_k^{-q}\widehat{\Lambda}_k\|_2$ and $\|(\widehat{\Lambda}_k\Lambda_k^{-2q}\widehat{\Lambda}_k)^{\frac{1}{2}}\|_2$ but simple algebraic manipulations show that they are actually equal. One has

$$\widehat{\Lambda}_k\Lambda_k^{-2q}\widehat{\Lambda}_k = \operatorname{diag}\left(\frac{\lambda_i^2 - \lambda_{k+1}^2}{\lambda_i^{2q}}\right), \quad 1 \leq i \leq k,$$

which suggests to introduce the map $\psi : x \mapsto (x - \lambda_{k+1}^2)/x^q$. A simple calculation shows that the extremum is reached for $x = \frac{q}{q-1}\lambda_{k+1}^2$ if $q \neq 1$. Consequently, the spectral norm is such that

$$\|\Lambda_k^{-q}\widehat{\Lambda}_k\|_2 = \frac{\sqrt{\lambda_\ell^2 - \lambda_{k+1}^2}}{\lambda_\ell^q} = \frac{\sqrt{1-\gamma_\ell^2}}{\lambda_\ell^{q-1}}$$

with ℓ defined as in the theorem. □

4.4.2 Probabilistic analysis of the methods for the generalized eigenvalue problem with basis transformation

A probabilistic analysis for Algorithms 4.3 and 4.4 addressing (4.12) can be directly deduced from the previous analysis. The next proposition states a result in this sense.

Proposition 4.12. *Let $\Upsilon \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix and $A, B \in \mathbb{R}^{n\times n}$ be Υ-symmetric matrices such that ΥB is symmetric positive definite. Then for any full column matrix rank matrix $Z \in \mathbb{R}^{n\times k}$ one has*

$$\|[I_n - \pi_{\Upsilon B}(Z)]B^{-1}A\|_{2,F,\Upsilon B} = \|[I_n - \pi_{\Upsilon B^{-1}}(BZ)]AB^{-1}\|_{2,F,\Upsilon B^{-1}}, \tag{4.33}$$

Proof. First, we note that $[I_n - \pi_{\Upsilon B}(Z)]B^{-1}A = B^{-1}[I_n - B\pi_{\Upsilon B}(Z)B^{-1}]A$. Then, we notice that

$$\begin{aligned} B\pi_{\Upsilon B}(Z)B^{-1} &= BZ(Z^\mathsf{T}\Upsilon BZ)^{-1}Z^\mathsf{T}\Upsilon \\ &= BZ(Z^\mathsf{T}B^\mathsf{T}B^{-\mathsf{T}}\Upsilon BZ)^{-1}Z^\mathsf{T}B^\mathsf{T}(B^{-\mathsf{T}})\Upsilon \\ &= \pi_{\Upsilon B^{-1}}(BZ), \end{aligned}$$

where we have used the Υ symmetry of B to write $B^{-\mathsf{T}}\Upsilon = \Upsilon B^{-1}$. We complete the proof by noticing that for any matrix $M \in \mathbb{R}^{n\times n}$, one has by definition of the weighted norms that

$$\|B^{-1}M\|_{2,F,\Upsilon B} = \|MB^{-1}\|_{2,F,\Upsilon B^{-1}}.$$

□

Hence, from Proposition 4.12, one can deduce bounds for $\|[I_n - \pi_{\Upsilon B^{-1}}(BZ)]AB^{-1}\|_{2,F,\Upsilon B^{-1}}$ where $Z = (B^{-1}A)^q G$ with $G \in \mathbb{R}^{n\times p}$ a Gaussian matrix such that $G \sim \mathcal{N}(0, \mathbf{Cov}(G))$. Since the methods addressing the GEP with basis transformation are based on matrices of the form $Z = (AB^{-1})^q G'$, taking $G' = BG$ and remarking that $\mathbf{Cov}(G') = B\,\mathbf{Cov}(G)B^\mathsf{T}$ allows us to state the new theorems. It then remains to use the relation between the eigenvectors of $B^{-1}A$ and AB^{-1} to conclude. In this regard, we simply state the three analogous theorems without further details.

Theorem 4.13 (Average-case analysis in weighted Frobenius norm)**.** *Let $\Upsilon \in \mathbb{R}^{n\times n}$ be symmetric positive definite and $A, B \in \mathbb{R}^{n\times n}$ be Υ-symmetric matrices such that ΥB is symmetric positive definite. Let $G \in \mathbb{R}^{n\times p}$ be a Gaussian matrix such that $G \sim \mathcal{N}(0, \mathbf{Cov}(G))$ with $2 < p \leq \min\{\mathrm{rank}(A), \mathrm{rank}(\mathbf{Cov}(G))\}$. Let $Z = (AB^{-1})^q G \in \mathbb{R}^{n\times p}$, with $q \in \mathbb{N}$ and let $\pi_{\Upsilon B^{-1}}(Z)$ denote the ΥB^{-1}-orthogonal projection onto the vector space spanned by the columns of Z. Let $\varphi : x \mapsto x/\sqrt{1+x^2}$ for $x \geq 0$.*

Then, if $\mathbf{Cov}_k(G)$ is nonsingular, we have for all $k \in \{1, \ldots, p-2\}$

$$\mathbb{E}\left[\, \|[I_n - \pi_{\Upsilon B^{-1}}(BZ)]AB^{-1}\|_{F,\Upsilon B^{-1}} - \|[I_n - \pi_{\Upsilon B^{-1}}(BZ)]\underline{A}_k B^{-1}\|_{F,\Upsilon B^{-1}} \,\right] \quad \leq$$
$$\min\left\{ \sqrt{\alpha_k}, \sqrt{k}\varphi\left(\frac{\sqrt{\beta_k}}{\sqrt{k}}\right)\lambda_1 \right\},$$

where

$$\alpha_k = \|\underline{\Lambda}_k^q\|_F^2 \,\frac{c_k(G)^2}{\lambda_k^{2(q-1)}} + \|\underline{\Lambda}_k^q\|_F^2\,\delta_k(G)\frac{\|\Lambda_k^{-(q-1)}\|_F^2}{p-k-1}$$
$$\beta_k = \|\underline{\Lambda}_k^q\|_F^2 \,\frac{c_k(G)^2}{\lambda_k^{2q}} + \|\underline{\Lambda}_k^q\|_F^2\,\delta_k(G)\frac{\|\Lambda_k^{-q}\|_F^2}{p-k-1},$$

with $\mathbf{Cov}_k(G) = U_k^\mathsf{T}\Upsilon B^{-1}\,\mathbf{Cov}(G)\,\Upsilon B^{-1}U_k$, and $\underline{\mathbf{Cov}}_k(G) = \underline{U}_k^\mathsf{T}\Upsilon B^{-1}\,\mathbf{Cov}(G)\,\Upsilon B^{-1}\underline{U}_k$, and $\mathbf{Cov}_{\perp,k}(G) = \underline{U}_k^\mathsf{T}\Upsilon B^{-1}\,\mathbf{Cov}(G)\,\Upsilon B^{-1}U_k$, and $c_k(G) = \|\,\mathbf{Cov}_{\perp,k}(G)\;\mathbf{Cov}_k(G)^{-1}\|_2$, and $\delta_k(G) = \|\,\mathbf{Cov}_k(G)^{-1}\|_2\|\,\underline{\mathbf{Cov}}_k(G)/\,\mathbf{Cov}_k(G)\|_2$.

Theorem 4.14 (Average-case analysis in weighted spectral norm)**.** *Let $\Upsilon \in \mathbb{R}^{n\times n}$ be symmetric positive definite and $A, B \in \mathbb{R}^{n\times n}$ be Υ-symmetric matrices such that ΥB is symmetric positive definite. Let $G \in \mathbb{R}^{n\times p}$ be a Gaussian matrix such that $G \sim \mathcal{N}(0, \mathbf{Cov}(G))$ with $2 < p \leq \min\{\mathrm{rank}(A), \mathrm{rank}(\mathbf{Cov}(G))\}$. Let $Z = (AB^{-1})^q G \in \mathbb{R}^{n\times p}$, with $q \in \mathbb{N}$ and let $\pi_{\Upsilon B^{-1}}(Z)$*

denote the ΥB^{-1}-orthogonal projection onto the vector space spanned by the columns of Z. Let $\varphi : x \mapsto x/\sqrt{1+x^2}$ for $x \geq 0$.

Then, if $\mathbf{Cov}_k(G)$ is nonsingular, we have for all $k \in \{1, \ldots, p-2\}$

$$\mathbb{E}\left[\|[I_n - \pi_{\Upsilon B^{-1}}(BZ)]AB^{-1}\|_{2,\Upsilon B^{-1}} - \|[I_n - \pi_{\Upsilon B^{-1}}(BZ)]\underline{A}_k B^{-1}\|_{2,\Upsilon B^{-1}} \right] \quad \leq \\ \min\left\{c_k,\ \varphi(d_k)\lambda_1\right\},$$

where

$$c_k = \lambda_{k+1}\,\gamma_k^{q-1}\,c_k(G) + \sqrt{\delta_k(G)}\left(\lambda_{k+1}^q\,\frac{\|\Lambda_k^{-(q-1)}\|_F^2}{\sqrt{p-k-1}} + \frac{\|\underline{\Lambda}_k^q\|_F}{\lambda_k^{q-1}}\,\frac{e\sqrt{p}}{p-k}\right)$$
$$d_k = \gamma_k^q\,c_k(G) + \sqrt{\delta_k(G)}\left(\lambda_{k+1}^q\,\frac{\|\Lambda_k^{-q}\|_F^2}{\sqrt{p-k-1}} + \frac{\|\underline{\Lambda}_k^q\|_F}{\lambda_k^q}\,\frac{e\sqrt{p}}{p-k}\right)$$

with $\mathbf{Cov}_k(G) = U_k^\mathsf{T}\Upsilon B^{-1}\,\mathbf{Cov}(G)\,\Upsilon B^{-1}U_k$, and $\underline{\mathbf{Cov}}_k(G) = \underline{U}_k^\mathsf{T}\Upsilon B^{-1}\,\mathbf{Cov}(G)\,\Upsilon B^{-1}\underline{U}_k$, and $\mathbf{Cov}_{\perp,k}(G) = \underline{U}_k^\mathsf{T}\Upsilon B^{-1}\,\mathbf{Cov}(G)\,\Upsilon B^{-1}U_k$, and $c_k(G) = \|\,\mathbf{Cov}_{\perp,k}(G)\;\mathbf{Cov}_k(G)^{-1}\|_2$, and $\delta_k(G) = \|\,\mathbf{Cov}_k(G)^{-1}\|_2\|\,\underline{\mathbf{Cov}}_k(G)/\,\mathbf{Cov}_k(G)\|_2$.

Theorem 4.15 (Average-case analysis in weighted spectral norm, improved bound)**.** *Let $\Upsilon \in \mathbb{R}^{n\times n}$ be symmetric positive definite and $A, B \in \mathbb{R}^{n\times n}$ be Υ-symmetric matrices such that ΥB is symmetric positive definite. Let $G \in \mathbb{R}^{n\times p}$ be a Gaussian matrix such that $G \sim \mathcal{N}(0, \mathbf{Cov}(G))$ with $2 < p \leq \min\left\{\mathrm{rank}(A),\ \mathrm{rank}(\mathbf{Cov}(G))\right\}$. Let $Z = (B^{-1}A)^q G \in \mathbb{R}^{n\times p}$, with $q \in \mathbb{N}$ and let $\pi_{\Upsilon B}(Z)$ denote the ΥB-orthogonal projection onto the vector space spanned by the columns of Z. Let $\varphi : x \mapsto x/\sqrt{1+x^2}$ for $x \geq 0$.*

Then, if $\mathbf{Cov}_k(G)$ is nonsingular, we have for all $k \in \{1, \ldots, p-2\}$

$$\mathbb{E}\left[\|[I_n - \pi_{\Upsilon B^{-1}}(BZ)]AB^{-1}\|_{2,\Upsilon B^{-1}} - \|\underline{\Lambda}_k\|_2 \right] \leq \min\left\{\widehat{c}_k,\ \varphi(\widehat{d}_k)\sqrt{\lambda_1^2 - \lambda_{k+1}^2}\right\},$$

where

$$\widehat{c}_k = \lambda_{k+1}\,\gamma_\ell^{q-1}\sqrt{1-\gamma_\ell^2}\,c_k(G) + \sqrt{\delta_k(G)}\left(\sum_{i=1}^{k}\gamma_i^{2q-1}\sqrt{1-\gamma_i^2}\right)^{\frac{1}{2}}\,\frac{\lambda_{k+1}}{\sqrt{p-k-1}} \\ + \sqrt{1-\gamma_\ell^2}\,\frac{\|\underline{\Lambda}_k^q\|_F}{\lambda_\ell^{q-1}}\,\frac{e\sqrt{p}}{p-k}$$
$$\widehat{d}_k = \gamma_k^q\,c_k(G) + \sqrt{\delta_k(G)}\left(\lambda_{k+1}^q\,\frac{\|\Lambda_k^{-q}\|_F}{\sqrt{p-k-1}} + \frac{\|\underline{\Lambda}_k^q\|_F}{\lambda_k^q}\,\frac{e\sqrt{p}}{p-k}\right)$$

with $\mathbf{Cov}_k(G) = U_k^\mathsf{T}\Upsilon B^{-1}\,\mathbf{Cov}(G)\,\Upsilon B^{-1}U_k$, and $\underline{\mathbf{Cov}}_k(G) = \underline{U}_k^\mathsf{T}\Upsilon B^{-1}\,\mathbf{Cov}(G)\,\Upsilon B^{-1}\underline{U}_k$, and $\mathbf{Cov}_{\perp,k}(G) = \underline{U}_k^\mathsf{T}\Upsilon B^{-1}\,\mathbf{Cov}(G)\,\Upsilon B^{-1}U_k$, and $c_k(G) = \|\,\mathbf{Cov}_{\perp,k}(G)\;\mathbf{Cov}_k(G)^{-1}\|_2$, and $\delta_k(G) =$

$\| \mathbf{Cov}_k(G)^{-1} \|_2 \| \underline{\mathbf{Cov}}_k(G) / \mathbf{Cov}_k(G) \|_2$, *and* $\widehat{\Lambda}_k = (\Lambda_k^2 - \lambda_{k+1}^2)^{\frac{1}{2}}$ *and*

$$\ell = \begin{cases} 1 & \text{if} \quad \dfrac{q}{q-1}\lambda_{k+1}^2 \geq \lambda_1^2 \quad \text{or} \quad q = 1 \\ \arg\max \left\{ \dfrac{\sqrt{1-\gamma_i^2}}{\lambda_i^{q-1}}, \dfrac{\sqrt{1-\gamma_{i+1}^2}}{\lambda_{i+1}^{q-1}} \right\} & \text{if} \quad \dfrac{q}{q-1}\lambda_{k+1}^2 \in [\lambda_i, \lambda_{i+1}] \\ k & \text{if} \quad \dfrac{q}{q-1}\lambda_{k+1}^2 \leq \lambda_k^2 \end{cases}.$$

4.4.3 Discussion on the proposed error bounds

Let us discuss the bounds obtained in Theorems 4.9 and 4.10. By definition of the Schur complement, one has,

$$\underline{\mathbf{Cov}}_k(G) / \mathbf{Cov}_k(G) \preccurlyeq \underline{\mathbf{Cov}}_k(G),$$

which implies that $\| \underline{\mathbf{Cov}}_k(G) / \mathbf{Cov}_k(G) \|_2 \leq \| \underline{\mathbf{Cov}}_k(G) \|_2$. Here, we note that the matrix $\underline{\mathbf{Cov}}_k(G)$ is a principal submatrix of $V^\mathsf{T} \Upsilon B \, \mathbf{Cov}(G) \, \Upsilon B V$. If we denote by $\mu_1 \geq \cdots \geq \mu_n$ the eigenvalues of $V^\mathsf{T} \Upsilon B \, \mathbf{Cov}(G) \, \Upsilon B V$, then applying the Cauchy's interlacing theorem [72, Theorem 10.1.1] yields

$$\mu_{k+1} \leq \| \underline{\mathbf{Cov}}_k(G) \|_2 \leq \mu_1.$$

A similar argument holds for $\mathbf{Cov}_k(G)$, yielding

$$\frac{1}{\mu_k} \leq \| \mathbf{Cov}_k(G)^{-1} \|_2 \leq \frac{1}{\mu_n}.$$

Furthermore, we note that $V^\mathsf{T} \Upsilon B \, \mathbf{Cov}(G) \, \Upsilon B V$ and $(\Upsilon B)^{\frac{1}{2}} \, \mathbf{Cov}(G) \, (\Upsilon B)^{\frac{1}{2}}$ share the same eigenvalues since V is ΥB-orthonormal. Altogether, we obtain the following bound

$$1 \leq \frac{\mu_{k+1}}{\mu_k} \leq \delta_k(G) \leq \frac{\mu_1}{\mu_n} = \kappa_2 \left((\Upsilon B)^{\frac{1}{2}} \, \mathbf{Cov}(G) \, (\Upsilon B)^{\frac{1}{2}} \right).$$

These inequalities are loose. Consequently, it is not clear how the 2-norm condition number of $(\Upsilon B)^{\frac{1}{2}} \, \mathbf{Cov}(G) \, (\Upsilon B)^{\frac{1}{2}}$ actually affects the approximation accuracy. This phenomenon has already been highlighted in [78]. Nevertheless, taking $\mathbf{Cov}(G) = (\Upsilon B)^{-1}$ yields $\delta_k(G) = 1$, which is the optimal value.

Let us now discuss the term $c_k(G)$. We propose an interpretation of $c_k(G)$ in terms of angles between subspaces. Let us define $F_k = \mathbf{Cov}(G)^{\frac{1}{2}} T V_k \, \mathbf{Cov}_k(G)^{-\frac{1}{2}}$ and $G_k = \mathbf{Cov}(G)^{\frac{1}{2}} \Upsilon B \underline{V}_k$ so that,

$$c_k(G) = \| \mathbf{Cov}_k(G)^{-\frac{1}{2}} F_k^\mathsf{T} G_k \|_2 \leq \| \mathbf{Cov}_k(G)^{-\frac{1}{2}} \|_2 \| F_k^\mathsf{T} G_k \|_2.$$

Then, because F_k is orthonormal, one has from the definition of principal angles (see Section 2.1.3) that

$$\| F_k^\mathsf{T} G_k \|_2 \leq \mathbf{cos}\left(F_k, G_k\right)^2 \| G_k \|_2^2 = \mathbf{cos}\left(F_k, G_k\right)^2 \| \underline{\mathbf{Cov}}_k(G) \|_2.$$

Using (4.4.3) we finally obtain

$$0 \leq c_k(G) \leq \mathbf{cos}\left(F_k, G_k\right)^2 \, \kappa_2 \left((\Upsilon B)^{\frac{1}{2}} \, \mathbf{Cov}(G) \, (\Upsilon B)^{\frac{1}{2}} \right).$$

In particular, we observe again that taking $\mathbf{Cov}(G) = (\Upsilon B)^{-1}$ yields the optimal value $c_k(G) = 0$ since V_k and $\underline{V}_k$ are ΥB-conjugate.

Remark 4.16. When $\mathbf{Cov}(G) = (\Upsilon B)^{-1}$, we can show that the analysis corresponds to the analysis of the standard randomized SVD of $(\Upsilon B)^{\frac{1}{2}} B^{-1} A (\Upsilon B)^{\frac{1}{2}}$, for which the error bounds are indeed tighter.

Let us rapidly discuss the sensitivity of the bounds with respect to the number of random samples p, and the number of subspace iterations q. In all the bounds, there is a term which does not depend on p. This term stems from the statistical dependence encountered in the general analysis in Chapter 3. Consequently, when p becomes large, the bounds do not approach zero, while the approximation error will necessarily approach zero. Let us now look at the behavior with respect to q. Observing that

$$\frac{\|\underline{\Lambda}_k^q\|_F^2}{\lambda_k^{2q}} = \sum_{i=k+1}^{\mathrm{rank}(A)} \left(\frac{\lambda_i}{\lambda_k}\right)^{2q} \leq (\mathrm{rank}(A) - k)\, \gamma_k^{2q},$$

and

$$\|\underline{\Lambda}_k^q\|_F^2 \|\Lambda_k^{-q}\|_F^2 \leq k(\mathrm{rank}(A) - k)\gamma_k^{2q},$$

we obtain that the terms α_k and β_k from Theorem 4.9 are both $O(\gamma_k^{2q})$. Consequently, they approach zero as γ_k^{2q}, implying that the expected low rank approximation error in weighted Frobenius norm approaches zero too. Analogous arguments yield that the terms c_k and d_k in Theorem 4.10 are both $O(\gamma_k^q)$, which yields analogous consequences.

4.4.4 Comparison with prior error bounds

An average case analysis has been proposed in [80, Theorem 1] to analyze a single-pass randomized method for generalized symmetric eigenvalue problems. In our context, it corresponds to $\Upsilon = \mathbf{Cov}(G) = I_n$ and $q = 1$. The analysis provides a bound in weighted spectral norm. We briefly propose a comparison with this bound.

Let $\sigma_1, \ldots, \sigma_n$ denote the singular values of $B^{-\frac{1}{2}} A$, with $\sigma_1 \geq \cdots \geq \sigma_n$. If $Z = B^{-1} A \Omega$, then, using our notations, Theorem 1 in [80] states that

$$\mathbb{E}\left[\, \|[I_n - \pi_B(Z)] B^{-1} A\|_{2,B} \,\right] \leq 2\sqrt{\|B^{-1}\|_2} \left(\left[1 + \frac{\sqrt{k}}{\sqrt{p-k-1}} \right] \sigma_{k+1} + \frac{e\sqrt{p}}{p-k} \left(\sum_{j=k+1}^{n} \sigma_j^2 \right)^{\frac{1}{2}} \right).$$

Here, we notice that the factor two in the right-hand side has been omitted in the statement of [80, Theorem 1], which was corrected by the authors in [78, p.16]. Since $\|[I_n - \pi_B(Y)] B^{-1} \underline{A}_k\|_B \leq \|\underline{\Lambda}_k\|_2 = \lambda_{k+1}$ Theorem 4.10 implies

$$\mathbb{E}\left[\, \|\!|[I_n - \pi_B(Y)] B^{-1} A|\!\|_2 \,\right] \leq \lambda_{k+1} + c_k.$$

Using the bounds proposed in Section 4.4.3, namely $\delta_k(G) \leq \kappa_2(B)$, $c_k(G) \leq \kappa_2(B)$ and $\kappa_2(B)^{-1} \leq 1$, we obtain

$$\mathbb{E}\left[\, \|[I_n - \pi_B(Z)] B^{-1} A\|_{2,B} \,\right] \leq \sqrt{\kappa_2(B)} \left(\left[2 + \frac{\sqrt{k}}{\sqrt{p-k-1}} \right] \lambda_{k+1} + \frac{e\sqrt{p}}{p-k} \left(\sum_{j=k+1}^{n} \lambda_j^2 \right)^{\frac{1}{2}} \right).$$

The difference in the multiplicative term is then explained by noticing that

$$\sigma_j \leq \|B^{\frac{1}{2}}\|_2 \lambda_j = \sqrt{\|B\|_2} \lambda_j, \quad 1 \leq j \leq n.$$

Consequently, when expressed in terms of eigenvalues, the bound from Theorem 4.10 is almost twice as good as the one derived in [80]. We numerically illustrate this fact on test problems later on. In weighted Frobenius norm, to the best of our knowledge, Theorem 4.9 is new.

4.5 Numerical experiments

We divide the numerical experiments in three different parts. In Section 4.5.1, we propose a comparison between our theoretical bounds and the one from [80, Theorem 1], on test matrices inspired from [78]. Then, in Section 4.5.2 we focus on a three-dimensional variational data assimilation test case to illustrate numerically the equivalence with algorithms from [80] highlighted above, and to investigate the accuracy of the approximate eigenpairs of our algorithms.

4.5.1 Error bounds in expectation versus the state-of-the-art

Here we propose a comparison between the bounds obtained in Theorem 4.10 and the ones given in [80, Theorem 1], along with an illustration of the bounds from Theorem 4.9. This comes as a complement to the theoretical comparison proposed in Section 4.4.4. We recall that performing such a comparison implies that we choose $\Upsilon = I_n$. For the test matrices, we consider a fixed matrix A, and we propose two different choices for B. We define

$$A = U \operatorname{diag}\left(1, 1/2^2, \ldots, 1/n^2\right) U^{\mathsf{T}},$$

with $U \in \mathbb{R}^{n\times n}$ an orthogonal matrix obtained from the QR factorization of a standard Gaussian matrix. Matrices whose eigenvalue distribution has a polynomial decay are frequently used as test cases in the randomized linear algebra literature (see [87] for instance). For B, we consider two cases:

- `Min_ij`: $B \in \mathbb{R}^{n\times n}$ such that $B_{i,j} = \min\{i, j\}$ for all $1 \leq i, j \leq n$,
- `Rand`: $B \in \mathbb{R}^{n\times n}$ such that $B = V \operatorname{diag}\left(1, 1/2^d, \ldots, 1/n^d\right) V^{\mathsf{T}}$ with $V \in \mathbb{R}^{n\times n}$ an orthogonal matrix obtained from the QR factorization of a standard Gaussian matrix, with d such that $\kappa_2(B) = 10^3$, that is, $d = \ln(10^3)/\ln(n)$.

Those are inspired from the MATLAB matrix gallery, and have been used in [78] to define inner-products to test algorithms for the GSVD. We choose $n = 500$ and $k = 10$. Figures 4.1 show the bounds for

$$\|[I_n - \pi_B(Z)]B^{-1}A\|_{2,F,B} - \|\underline{\Lambda}_k\|_{2,F,B}, \tag{4.34}$$

for a number of samples p varying from 12 to 110. Here, we point out that to the best of our knowledge, there are no available bounds in weighted Frobenius norm. In weighted spectral norm, it seems that the only available result is given in [80, Theorem 1] and corresponds to $q = 1$. We denote by `SLK` this bound.

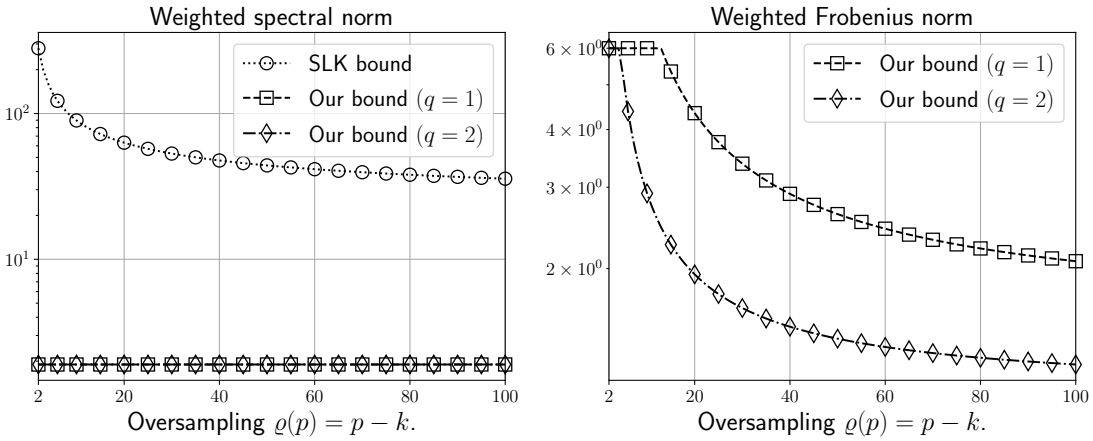
In Figure 4.1, our bound in weighted spectral norm and $q = 1$ is tighter than the one from `SLK`. The overall behavior of the bounds is then comparable to the one detailed in Chapter 3, so we will not further discuss this aspect. Following Section 4.4.3, we provide the values of $c_k(G)$ and $\delta_k(G)$ in Table 4.3. We observe that $\delta_k(G)$ is actually quite close to $\kappa_2(B)$. $c_k(G)$ remains moderate, meaning that this term, which does not approach zero as p increases is relatively small compared to the remaining terms. It is also interesting to remark that despite the large value of $\delta_k(G)$ for `Min_ij`, the resulting bounds are closer to zero than for `Rand`. The difference can possibly be explained by the difference in the eigenvalue distributions.

4.5.2 Application to a 3D-Var data assimilation problem

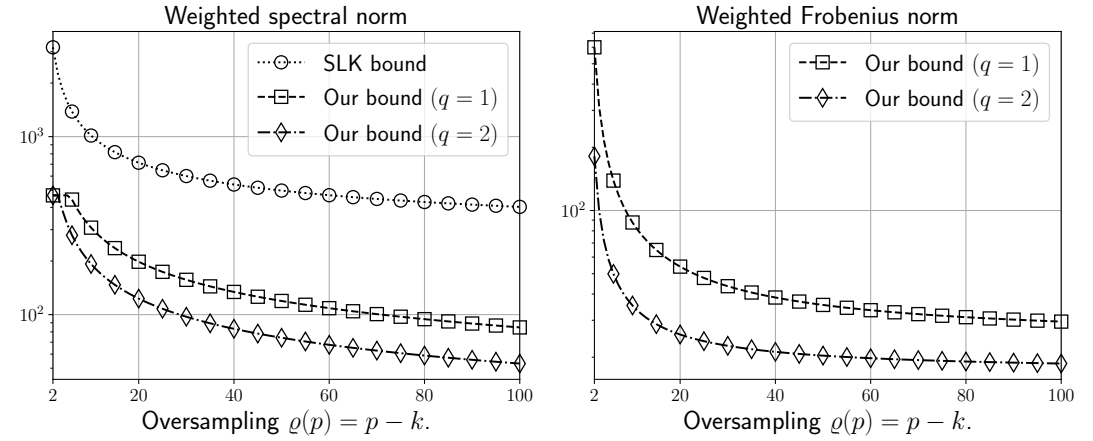
We consider an application with matrices from a data assimilation problem. Data assimilation is a general framework where observations of a dynamical system are used to determine its true

	$c_k(G)$	$\delta_k(G)$	$\kappa_2(B)$	γ_k
`Min_ij`	$4.45 \cdot 10^1$	$2.38 \cdot 10^5$	$4.06 \cdot 10^5$	0.782
`Rand`	$3.02 \cdot 10^0$	$5.34 \cdot 10^2$	$1.00 \cdot 10^3$	0.847

Table 4.3: Values for the constants appearing in the bounds in Theorems 4.9 and 4.10 when computed on the test problems `Min_ij` and `Rand`.



(a) Results for `Min_ij`.



(b) Results for `Rand`.

Figure 4.1: Comparison of different bounds for the error quantity (4.34) for two different matrices B: `Min_ij` (top) and `Rand` (bottom), in weighted spectral norm (left) and Frobenius norm (right).

underlying state. Its variational approach can be framed as a weighted nonlinear least-squares problems as introduced in Section 2.4.3, yielding matrices of the form

$$A_{\text{3D-Var}} = \Gamma_b^{-1} + H^{\mathsf{T}} \Gamma_o^{-1} H. \tag{4.35}$$

Our objective in this section is to investigate the performance of our algorithms when computing eigenpairs of either $\Gamma_b A_{\text{3D-Var}}$ or $A_{\text{3D-Var}} \Gamma_b$. As already highlighted in Section 2.4.3, such eigenvalue problems are notably needed to construct spectral LMP as in (2.19).

In this section, we consider a three-dimensional variational data assimilation problem (3D-Var). The covariance matrix Γ_b is defined as a discretized diffusion operator, with standard deviation $\sigma_b = 1.0$, and the observation error covariance matrix is defined as $\Gamma_o = \sigma_o^2 I_m$, with σ_o the corresponding standard deviation. The observation operator $\mathcal{H}$ is a selection operator. With $n = 1000$ state variables, we consider two different settings

- `LowObs`: $m = 100$ and $\sigma_o = 10^{-2}$,
- `HighObs`: $m = 400$ and $\sigma_o = 2 \times 10^{-2}$.

The sensitivity analysis regarding the number of observations is critical since it strongly impacts the eigenvalue distribution of the preconditioned matrix $\Gamma_b A_{\text{3D-Var}}$. Consequently, following the theoretical results obtained above, the randomized methods are expected to behave differently on both test cases. The parameters have been selected so as to yield a condition number of approximately 10^4 for both $\Gamma_b A_{\text{3D-Var}}$ and $A_{\text{3D-Var}}\Gamma_b$. The eigenvalue distributions of `LowObs` and `HighObs` are shown in Figure 4.2.

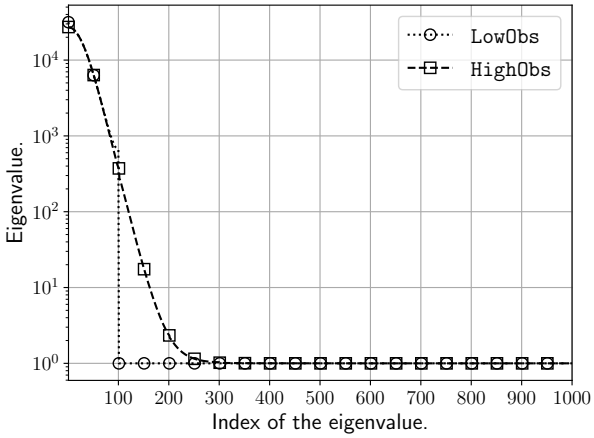


Figure 4.2: Eigenvalue distributions of the two 3D-Var test problems, `LowObs` and `HighObs`.

We split the numerical experiments in two parts. First, we numerically confirm that Algorithms 4.1 and 4.2 provide equivalent implementations of [80, Algorithms 6, 7 and 8]. We mentioned in Section 4.3.6 the mathematical equivalence. Hence, our interest here is to show that our implementation, *which does not require applications of* B, produces numerically equivalent approximations. In a second part, we investigate the accuracy of the approximate eigenvectors and eigenvalues.

Equivalence with Saibaba, Lee and Kitanidis' algorithms

We recall that in [80], the authors proposed three randomized methods to address the generalized Hermitian eigenvalue problem involving the matrix pencil $\{A, B\}$, with A symmetric and B symmetric positive definite, denoted by `SLK_6`, `SLK_7` and `SLK_8` hereafter. The setting of Algorithms 4.1 and 4.2 used to recover those methods in exact arithmetic is given in Table 4.2.

Quantities of interest. Let $\lambda_1 \geq \cdots \geq \lambda_k \in \mathbb{R}$ denote the k dominant eigenvalues of $B^{-1}A$ and $v_1, \ldots, v_k \in \mathbb{R}^n$ denote the corresponding eigenvectors. Let $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_k \in \mathbb{R}$, and $\widetilde{v}_1, \ldots, \widetilde{v}_k \in \mathbb{R}^n$ denote approximate eigenvalues and eigenvectors of $B^{-1}A$.

To measure the difference between the approximate eigenvalues, we consider the average relative distance with the exact eigenvalues, that is

$$\Delta_k^\lambda = \frac{1}{k} \sum_{j=1}^{k} \frac{|\widetilde{\lambda}_j - \lambda_j|}{\lambda_j}. \tag{4.36}$$

For the eigenvectors, we compute

$$\Delta_k^v = \frac{1}{k} \|\pi_B(V_k)[I_n - \pi_B(\widetilde{V}_k)]\|_{F,B}, \tag{4.37}$$

where $V_k = [v_1 \dots v_k] \in \mathbb{R}^{n\times k}$ and $\widetilde{V}_k = [\widetilde{v}_1 \dots \widetilde{v}_k] \in \mathbb{R}^{n\times k}$. If $\sigma_1 \geq \dots \geq \sigma_k$ denote the k largest singular values of $\pi_B(V_k)(I_n - \pi_B(\widetilde{V}_k))$, then $\sigma_i = \sin(\theta_i)$ with θ_i being the principal canonical angles between $\mathcal{R}(V_k)$ and $\mathcal{R}(\widetilde{V}_k)$ measured in the B-inner product (see Section 2.1.3). Therefore Δ_k^v represents the average sine of the principal canonical angles.

Finally, we also compare the B-orthonormality of the obtained approximate eigenvectors via

$$\Delta_k^\perp = \|\widetilde{V}_k^\mathsf{T} B \widetilde{V}_k - I_k\|_2. \tag{4.38}$$

This is relevant because the applications of B in `SLK_6`, `SLK_7` and `SLK_8` are explicitly required to ensure the B-orthonormality of the provided approximate eigenvectors. Consequently, *since our algorithms no longer require them*, it is important to verify whether the B-orthonormality is altered or not.

Results. Algorithms 4.1 and 4.2 are applied to the matrix pencil $(A_{3\text{D}}, \Gamma_b^{-1})$ with $\Upsilon = I_n$. All the methods are compared using the same standard Gaussian matrix $\Omega \in \mathbb{R}^{n\times p}$. We set $k = 20$ and $p = 40$ random samples, which corresponds to an oversampling of 20. The results are presented in Table 4.4.

Keeping 3 digits for Δ_{20}^λ and Δ_{20}^v does not allow us to highlight any difference between the provided approximate eigenvectors and eigenvalues. Consequently, the proposed implementation seems to be numerically consistent with the mathematical equivalence property. Plus, we observe that the B-orthonormality is of comparable quality, since the values of $\Delta_{20}^\perp$ are of similar order of magnitude. This is even more convincing knowing that in our 3D-Var setting, $\kappa_2(\Gamma_b) \approx 10^9$. Consequently, whenever the inverse of B can be applied accurately, approximations of eigeninformation can be obtained without (potentially expensive) applications of B.

Accuracy of the proposed algorithms

We now investigate the performance of our algorithms on the 3D-Var preconditioned matrices. We handle the methods for the GEP in initial form and with basis transformation separately. Here, we focus on the behavior in terms of approximate eigenvector and eigenvalue quality.

We study similar quantities of interest as in (4.36) and (4.37). For the eigenvalues, if $\lambda_1, \dots, \lambda_k$ denote the k largest eigenvalues of $\Gamma_b^{-1} A_{3\text{D-Var}}$, and $\widetilde{\lambda}_1, \dots, \widetilde{\lambda}_k$ the k approximate eigenvalues, then we compute

$$\delta_j = \frac{|\widetilde{\lambda}_j - \lambda_j|}{\lambda_j}, \quad 1 \leq j \leq k. \tag{4.39}$$

The quantity of interest to measure the accuracy of eigenvectors depends on whether we consider the methods for the GEP in initial form and with basis transformation. Accordingly, they are defined in the corresponding section.

To account for the randomness, we apply each algorithm 100 times with statistically independent standard Gaussian matrices Ω_i. We then perform a statistical analysis on the corresponding computed quantities of interest, focusing on the empirical mean and standard deviation.

	Algorithm	Δ_{20}^{λ}	Δ_{20}^{v}	$\Delta_{20}^{\perp}$
LowObs ($m = 100$)	SLK_6	$8.53 \cdot 10^{-1}$	$2.12 \cdot 10^{-1}$	$5.76 \cdot 10^{-8}$
	Algorithm 4.1	$8.53 \cdot 10^{-1}$	$2.12 \cdot 10^{-1}$	$5.85 \cdot 10^{-8}$
	SLK_7	1.00	$2.12 \cdot 10^{-1}$	$5.74 \cdot 10^{-8}$
	Algorithm 4.2	1.00	$2.12 \cdot 10^{-1}$	$5.88 \cdot 10^{-8}$
	SLK_8	$7.09 \cdot 10^{-2}$	$8.76 \cdot 10^{-2}$	$2.60 \cdot 10^{-7}$
	Algorithm 4.2	$7.09 \cdot 10^{-2}$	$8.76 \cdot 10^{-2}$	$2.19 \cdot 10^{-7}$
HighObs ($m = 400$)	SLK_6	$6.31 \cdot 10^{-1}$	$1.91 \cdot 10^{-1}$	$1.11 \cdot 10^{-7}$
	Algorithm 4.1	$6.31 \cdot 10^{-1}$	$1.91 \cdot 10^{-1}$	$1.16 \cdot 10^{-7}$
	SLK_7	1.00	$1.93 \cdot 10^{-1}$	$1.10 \cdot 10^{-7}$
	Algorithm 4.2	1.00	$1.93 \cdot 10^{-1}$	$1.14 \cdot 10^{-7}$
	SLK_8	$6.89 \cdot 10^{-2}$	$8.67 \cdot 10^{-2}$	$2.46 \cdot 10^{-7}$
	Algorithm 4.2	$6.89 \cdot 10^{-2}$	$8.67 \cdot 10^{-2}$	$2.28 \cdot 10^{-7}$

Table 4.4: Comparison between SLK_6, SLK_7, SLK_8 and our corresponding algorithms following Table 4.2 in terms of approximate eigenvalues (Δ_{20}^{λ}), eigenvectors (Δ_{20}^{v}), and B-orthonormality ($\Delta_{20}^{\perp}$). The different quantities of interest are defined in (4.36), (4.37) and (4.38) respectively.

In the following, the algorithms compute $k = 20$ approximate eigenpairs. We consider two values for the number of random samples, namely $p = 40$, 60, corresponding to an oversampling of 20 and 40, respectively. We consider the settings LowObs and HighObs to illustrate the behavior of the algorithms regarding different eigenvalue distributions.

Methods for the GEP in initial form. Let us first analyze the algorithms for the initial GEP, that is Algorithms 4.1 and 4.2, applied to $\{H^{\mathsf{T}}\Gamma_o^{-1}H, \Gamma_b^{-1}\}$. The different settings we investigate are detailed in Table 4.5. Our targeted applications are of large scale, meaning that the operators must be parsimoniously applied. Consequently, we focus on settings limited to two applications of both $H^{\mathsf{T}}\Gamma_o^{-1}H$ and Γ_b to blocks of p vectors (see the last two columns of Table 4.1).

To measure the accuracy of the obtained approximate eigenvectors, we again rely on subspace angles. Let $v_1, \ldots, v_k \in \mathbb{R}^n$ denote the corresponding eigenvectors associated to the k largest eigenvalues of $\Gamma_b A_{\text{3D-Var}}$, and $\widetilde{v}_1, \ldots, \widetilde{v}_k \in \mathbb{R}^n$ the corresponding approximate eigenvectors. Let us define $V_k = [v_1 \ \ldots \ v_k] \in \mathbb{R}^{n\times k}$ and $\widetilde{V}_k = [\widetilde{v}_1 \ \ldots \ \widetilde{v}_k] \in \mathbb{R}^{n\times k}$. As a quantity of interest, we compute the k largest singular values $\sigma_1 \geq \cdots \geq \sigma_k$ of $\pi_{\Gamma_b^{-1}}(V_k)[I_n - \pi_{\Gamma_b^{-1}}(\widetilde{V}_k)]$. Again, we have

$$\sigma_j = \sin(\theta_j), \quad 1 \leq j \leq k, \tag{4.40}$$

with θ_j being the principal canonical angles between $\mathcal{R}(V_k)$ and $\mathcal{R}(\widetilde{V}_k)$ measured in the Γ_b^{-1} inner product.

Figures 4.3 and 4.4 show the results for the approximate eigenvalues and eigenvectors respectively. Figure 4.3 reveals that the hierarchy in terms of accuracy strictly follows the number of applications of the different operators. The effect of the oversampling is significant, but seems to benefit more to LeftDir_1 (diamond) and LeftInv_2 (circle) than LeftInv_1 (square), especially in the HighObs setting. Surprisingly, the different eigenvalue distributions seem to alter very slightly the accuracy of LeftDir_1 and LeftInv_2. By contrast, in the HighObs, the performance of LeftInv_1 are very poor. Thus, LeftInv_1 seem to be unappropriated if one is interested in getting accurate approximation eigenvalues.

	Algorithm	q	$H^\mathsf{T}\Gamma_o^{-1}H$	Γ_b
`LeftDir_1`	4.1	1	2	1
`LeftInv_1`	4.2	1	1	1
`LeftInv_2`	4.2	2	2	2

Table 4.5: Settings of Algorithms 4.1 and 4.2. The last two columns recall the number of applications of the operators to a block of p vectors for each setting.

The behavior in terms of eigenvectors (Figure 4.4) is, in a large extent, similar to the one in terms of eigenvalues. From a broad perspective, the differences between $\sin(\theta_1)$ and $\sin(\theta_k)$ show that, for all the methods, very well converged approximate eigenvectors coexist with poorly converged ones, even for the more expensive methods. Also, it is noticeable that although `LeftInv_1` poorly performs on `HighObs` in terms of eigenvalues, the obtained eigenvectors are of similar accuracy as `LeftDir_1` for $p = 40$. This means that the additional applications of $H^\mathsf{T}\Gamma_o^{-1}H$ required for the latter improves the approximate eigenvalues quality far more than the eigenvectors one. By contrast, the superiority of `LeftInv_2` over `LeftInv_1` means that the additional application of Γ_b performed in `LeftInv_2` has a strong positive impact on the approximate eigenvectors accuracy.

In conclusion, although `LeftInv_1` might be the only affordable method in concrete applications, it must be considered solely if accurate eigenvalues are not of primary interest. Also, it is clear that both operators do not impact the overall accuracy in a symmetric way. For the 3D-Var problem, additional applications of Γ_b improve the approximation quality significantly more than additional applications of $H^\mathsf{T}\Gamma_o^{-1}H$. If an additional application of Γ_b is affordable, an option to improve the accuracy of the algorithms would be to apply them with an initial random matrix of the form $\Omega = \Gamma_b G$, with $G \in \mathbb{R}^{n\times p}$ a standard Gaussian instead. We let this for future work.

Methods for the GEP with basis transformation Now, let us consider the methods for the GEP with basis transformation and analyze the performance of Algorithms 4.3 and 4.4 applied to $H^\mathsf{T}\Gamma_o^{-1}H\Gamma_b$. The settings we study are detailed in Table 4.6. Again, we limit our study to the variants requiring at most two applications of $H^\mathsf{T}\Gamma_o^{-1}H$, to be consistent with the computational constraint of targeted applications.

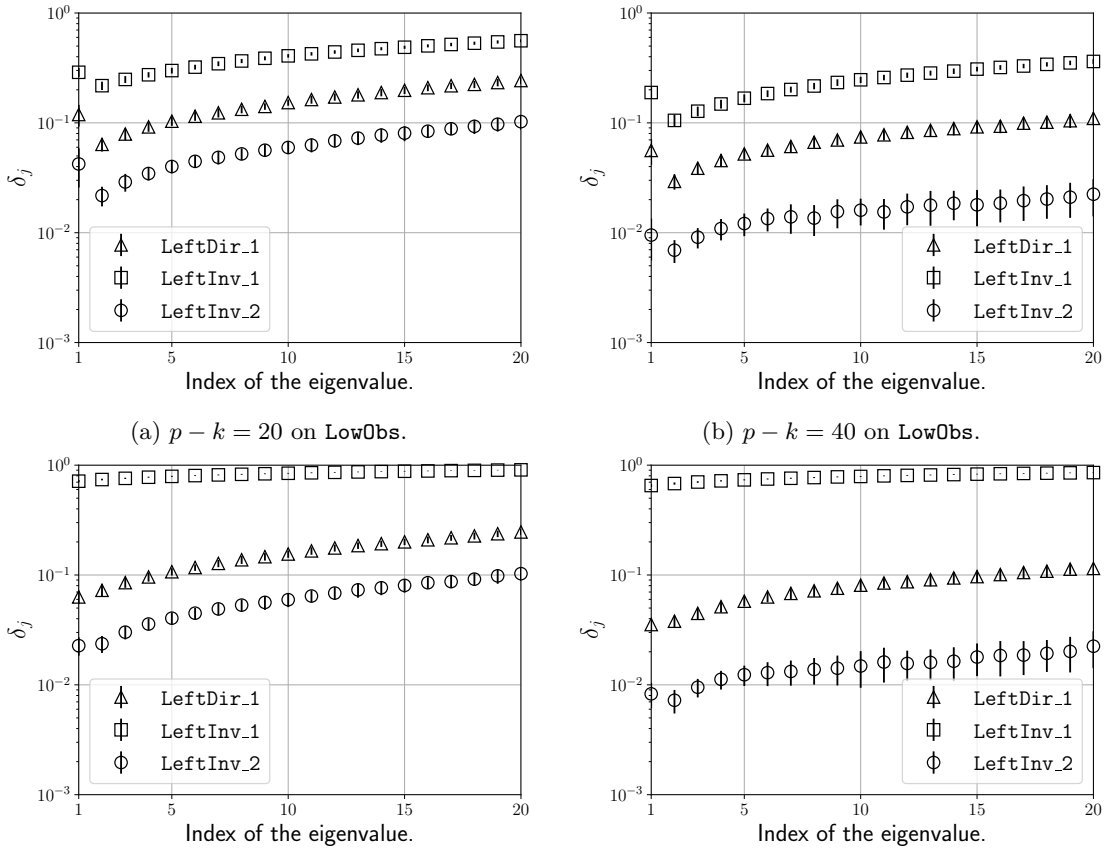
Let $u_1, \dots, u_k \in \mathbb{R}^n$ denote the eigenvectors associated to the k largest eigenvalues of $A_{\text{3D-Var}}\Gamma_b$, and $\widetilde{u}_1, \dots, \widetilde{u}_k \in \mathbb{R}^n$ the corresponding approximate eigenvectors. Let us define $U_k = [u_1 \ \dots \ u_k] \in \mathbb{R}^{n\times k}$ and $\widetilde{U}_k = [\widetilde{u}_1 \ \dots \ \widetilde{u}_k] \in \mathbb{R}^{n\times k}$. This time, we measure the approximate eigenvector accuracy by computing the k largest singular values $\sigma_1 \geq \cdots \geq \sigma_k$ of $\pi_{\Gamma_b}(U_k)[I_n - \pi_{\Gamma_b}(\widetilde{V}_k)]$, where

$$\sigma_j = \sin(\phi_j), \quad 1 \leq j \leq k, \tag{4.41}$$

with ϕ_j being the principal canonical angles between $\mathcal{R}(U_k)$ and $\mathcal{R}(\widetilde{U}_k)$ measured in the Γ_b inner product.

Figures 4.5 and 4.6 show the results for the approximate eigenvalues and eigenvectors respectively. We point out that the apparent variability of the obtained approximations materialized by the vertical bars is a simple distortion due to the logarithmic scale. The apparent greater performance is a consequence of the fact that the tested variants are globally more expensive than the ones for the GEP in initial form. In this regard, we note that no single-pass variant exists to address the GEP with basis transformation. For the rest, analogous comments to the previous section can be made. Let us therefore focus on comparing both approaches.

The targeted eigenvalues are the same for the two GEP formulations, making the comparison relevant. In particular, we observe that `LeftInv_2` and `RightDir_1` require the same amount of



(a) $p - k = 20$ on `LowObs`.

(b) $p - k = 40$ on `LowObs`.

(c) $p - k = 20$ on `HighObs`.

(d) $p - k = 40$ on `HighObs`.

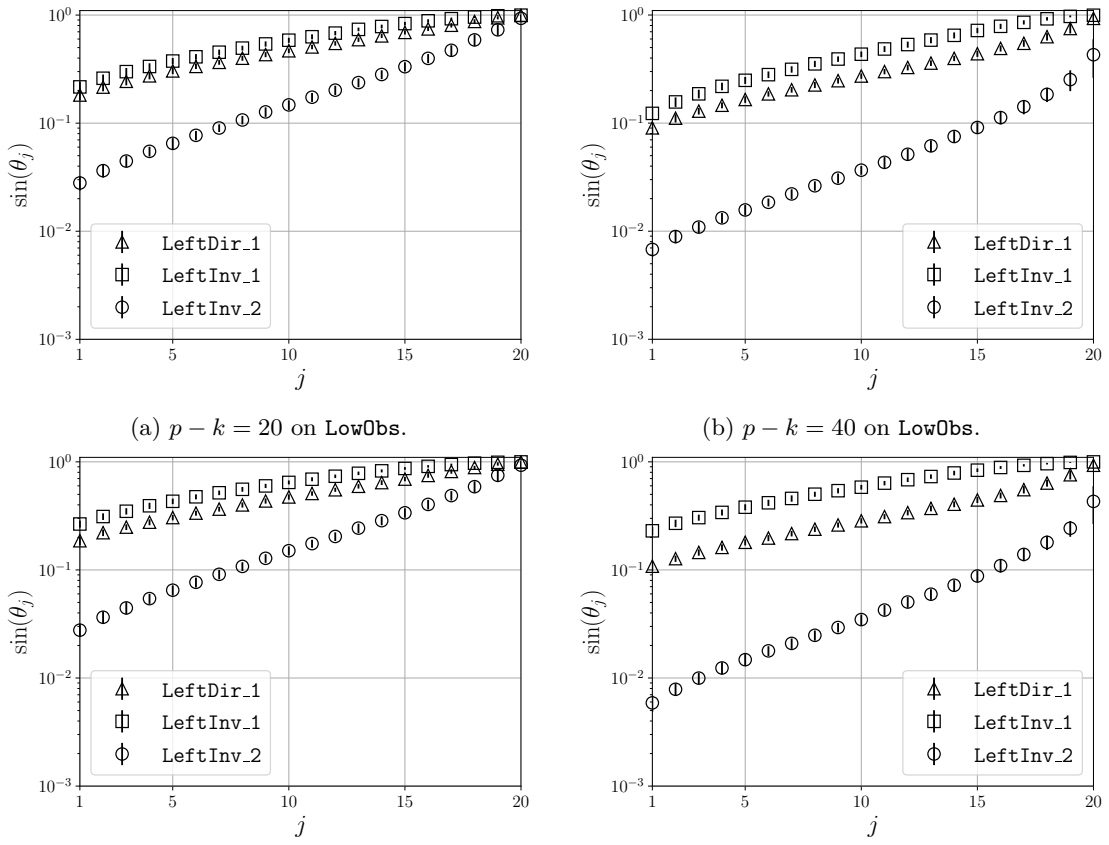
Figure 4.3: Accuracy of the algorithms for approximating dominant eigenvalues of $\{A_{\text{3D-Var}}, \Gamma_b^{-1}\}$ for the `LowObs` (top) and `HighObs` (bottom) test case with δ_j as in (4.39). The algorithms extract $k = 20$ approximate eigenpairs using either $p = 40$ (left) or $p = 60$ (right) Gaussian random samples.

	Algorithm	q	$H^\mathsf{T}\Gamma_o^{-1}H$	Γ_b
`RightDir_1`	4.3	1	2	2
`RightInv_1`	4.4	1	1	2
`RightInv_2`	4.4	2	2	3

Table 4.6: Settings of Algorithms 4.3 and 4.4. The last two columns recall the number of applications of the operators to a block of p vectors for each setting.

applications of $H^\mathsf{T}\Gamma_o^{-1}H$ and Γ_b (see Tables 4.5 and 4.6). However, the approximations obtained with `RightDir_1` are roughly one order of magnitude more accurate than the one of `LeftInv_2`, for both $p = 40$ and $p = 60$, and both `LowObs` and `HighObs`. Again, this illustrates that the two operators in the GEP do not have a symmetric role when using randomized algorithms. For the 3D-Var test case we consider, it thus seems that approximating eigenpairs of $H^\mathsf{T}\Gamma_o^{-1}H\Gamma_b$ can be performed more efficiently than approximating ones from $\Gamma_b H^\mathsf{T}\Gamma_o^{-1}H$.

It is also interesting to compare `RightInv_1` with `LeftDir_1`. The first one uses one addi-



(a) $p - k = 20$ on `LowObs`.

(b) $p - k = 40$ on `LowObs`.

(c) $p - k = 20$ on `HighObs`.

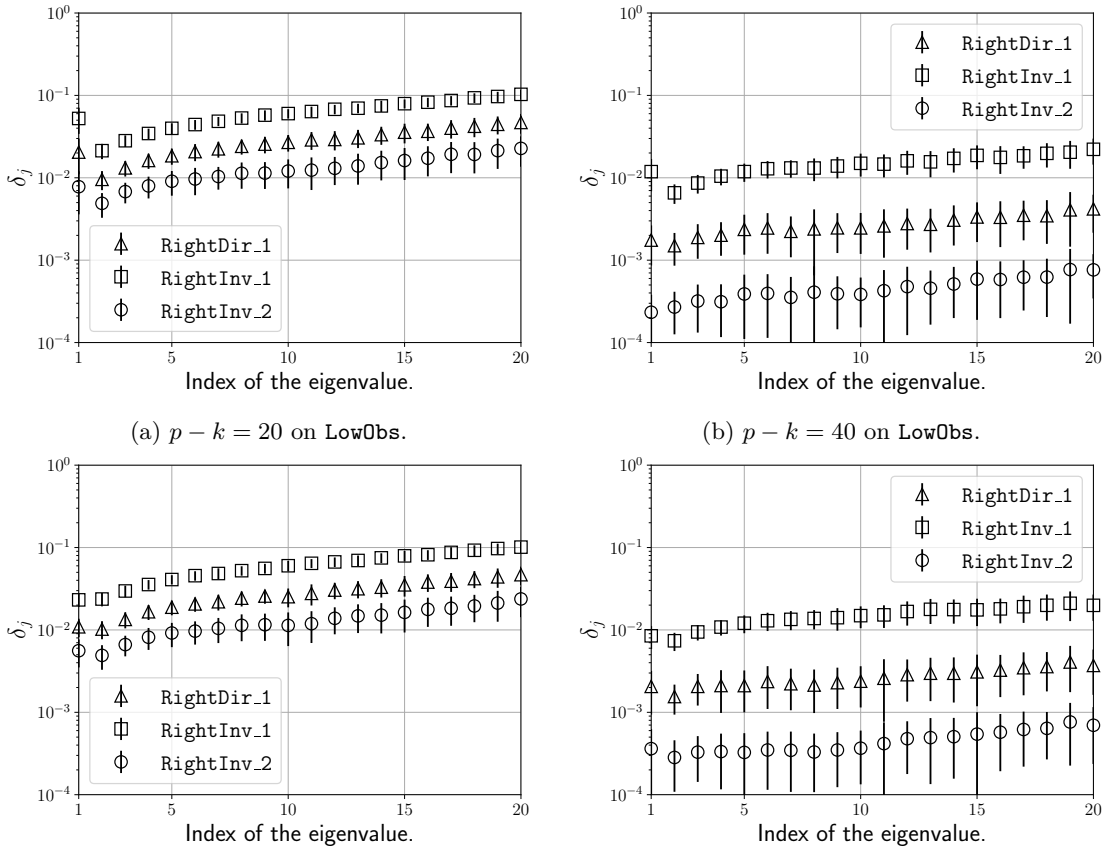
(d) $p - k = 40$ on `HighObs`.

Figure 4.4: Accuracy of the algorithms for approximating dominant eigenvectors of $\{A_{\text{3D-Var}}, \Gamma_b^{-1}\}$ for the `LowObs` (top) and `HighObs` (bottom) test case with $\sin(\theta_j)$ as in (4.40). The algorithms extract $k = 20$ approximate eigenpairs using either $p = 40$ (left) or $p = 60$ (right) Gaussian random samples.

tional application of Γ_b while the second one uses one additional application of $H^\mathsf{T}\Gamma_o^{-1}H$. The superiority of `RightInv_1` thus appears as another clue that Γ_b plays a more important role than $H^\mathsf{T}\Gamma_o^{-1}H$ to get accurate eigenvalue approximations. This statement remains true for approximate eigenvectors, as revealed in Figure 4.6. Indeed, `RightDir_1` and `RightInv_1` differ only from one application of $H^\mathsf{T}\Gamma_o^{-1}H$ but yield almost equivalent approximate eigenvectors.

4.6 Conclusions and perspectives

In this chapter, we have derived randomized methods to address the solution of two related generalized eigenvalue problems. The proposed algorithms are based on the Rayleigh-Ritz method, which provides both a rigorous and flexible framework to derive approximate eigenpairs. In particular, our algorithms generalize prior contributions from [80] and [24]. For the former, it turns out that our approach yields equivalent results while enjoying a cheaper implementation. Based on the general analysis developed in Chapter 3, an average-case error analysis of the new algorithms is proposed in both weighted Frobenius and spectral norms. This analysis gives insights regarding the number of random samples, the number of subspace iterations, and



(a) $p - k = 20$ on `LowObs`.

(b) $p - k = 40$ on `LowObs`.

(c) $p - k = 20$ on `HighObs`.

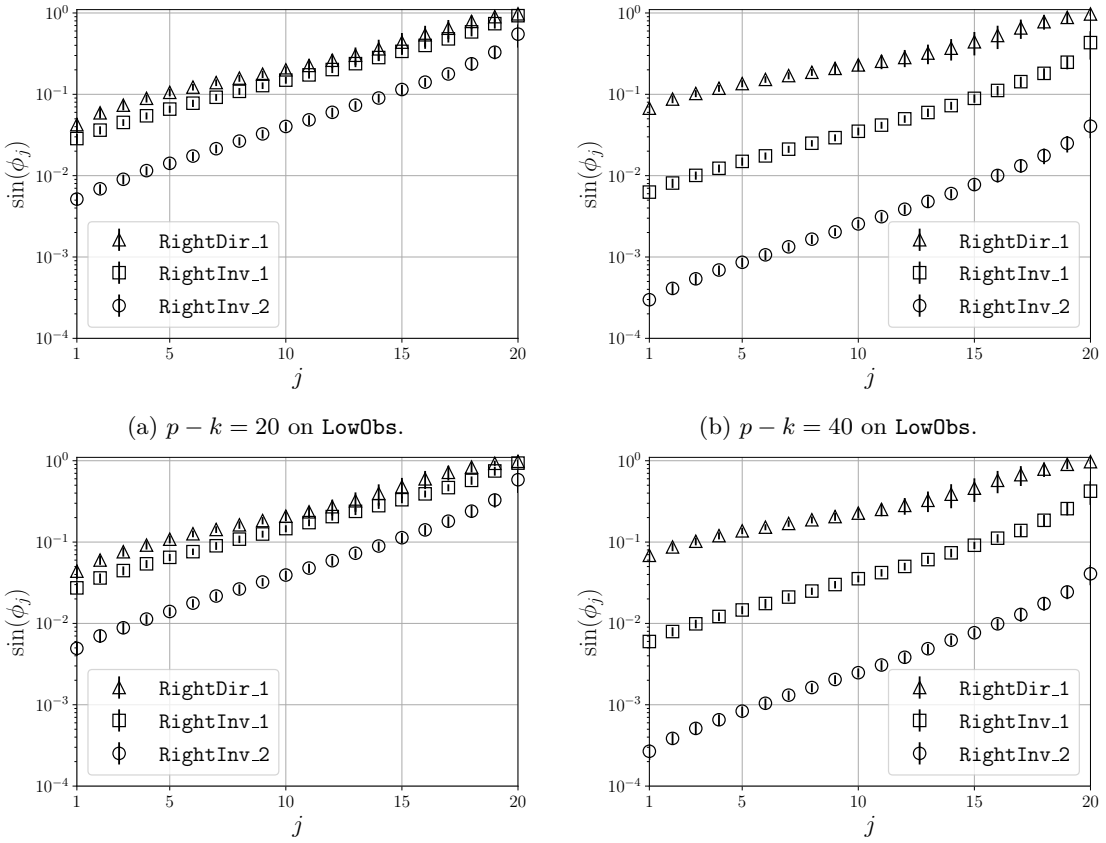
(d) $p - k = 40$ on `HighObs`.

Figure 4.5: Accuracy of the algorithms for approximating dominant eigenvalues of $\{A_{\text{3D-Var}}\Gamma_b, I_n\}$ for the `LowObs` (top) and `HighObs` (bottom) test case with δ_j as in (4.39). The algorithms extract $k = 20$ approximate eigenpairs using either $p = 40$ (left) or $p = 60$ (right) Gaussian random samples.

the optimal covariance matrix for the Gaussian sample matrix. The analysis in spectral norm generalizes and improves the state-of-the-art bounds from [80], while the analysis in Frobenius norm is new. Finally, numerical experiments on a three-dimensional variational data assimilation problem allowed us to demonstrate the potential of the proposed methods.

In this chapter, we have let open several questions. First, we have pointed out in Section 4.4 that our theoretical analysis is not able to distinguish between the different extraction processes because it only considers the form of the search space. This leads to a common analysis for Algorithm 4.1 and 4.2 (resp. Algorithms 4.3 and 4.4). To account for the extraction phase, a refined theoretical analysis should handle random matrices of the form $(B^{-1}A)^q\Omega W$ (resp. $(AB^{-1})^q\Omega W$) with $\Omega \in \mathbb{R}^{n\times p}$ a Gaussian matrix and $W \in \mathbb{R}^{p\times k}$ the matrix of change of basis associated to the reduced eigenvalue problems. We remark that this problem has already been identified in [53, Section 9.4], where the authors questioned the way to theoretically handle the truncation in the RSVD, but has not yet been addressed.

Another important point that must be investigated concerns the implementation. We have pointed out in Section 4.3.5 that the QR factorization and the algorithm to solve the projected eigenvalue problem are critical. In this regard, it is crucial to identify the appropriate algorithms



(a) $p - k = 20$ on `LowObs`.

(b) $p - k = 40$ on `LowObs`.

(c) $p - k = 20$ on `HighObs`.

(d) $p - k = 40$ on `HighObs`.

Figure 4.6: Accuracy of the algorithms for approximating dominant eigenvectors of $\{A_{\text{3D-Var}}\Gamma_b, I_n\}$ for the `LowObs` (top) and `HighObs` (bottom) test case with $\sin(\phi_j)$ as in (4.41). The algorithms extract $k = 20$ approximate eigenpairs using either $p = 40$ (left) or $p = 60$ (right) Gaussian random samples.

to obtain both a numerically efficient and stable algorithm. In particular, for large scale problems, efficient algorithms for the QR factorization must be considered, such as parallel communication-avoiding methods [27, 37, 38, 41, 95].

CHAPTER 5

Randomized preconditioning for weighted nonlinear least-squares problems

Contents

Abstract

In this chapter, we propose a class of randomized spectral Limited Memory Preconditioners (LMP) for the Preconditioned Conjugate Gradient method (PCG) when solving a weighted nonlinear least-squares problem with the Gauss-Newton method. We focus on two particular variants of the PCG dedicated to the solution of the linearized subproblem: the inverse-free PCG (PCGIF) and the Augmented Restricted PCG (RPCG). For both methods, specific LMP formulations have been proposed, and spectral variants have been derived in the literature.

Then, we propose randomized spectral LMPs where the exact eigenpairs are replaced by approximations obtained using randomized methods. These randomized methods are adaptations of the algorithms introduced in Chapter 4. For both Krylov subspace methods, we propose two algorithms depending on the availability of a preconditioner.

In [50], relations between the LMP for the PCGIF and the RPCG have been identified so that the produced iterates are mathematically equivalent. We propose such a relation for the proposed randomized spectral LMPs. This relation gives insight on new strategies that can be used to construct more efficient randomized spectral LMP for the PCGIF.

Finally, we investigate the performance of the proposed randomized LMPs on a four-dimensional variational data assimilation (4D-Var) problem. The randomized preconditioners are compared to the exact spectral LMPs and to the Ritz LMPs. The obtained results are encouraging and open a number of interesting perspectives.

5.1 Introduction

In this chapter, we are interested in solving the sequence of linear systems arising from the solution of the weighted nonlinear least-squares problem with the Gauss-Newton method. Accordingly, let us consider the solution of

$$\underbrace{\left(\Gamma_b^{-1} + H_j^\mathsf{T}\Gamma_o^{-1}H_j\right)}_{=A_j} s_j = \underbrace{\Gamma_b^{-1}(x_c - x_j) + H_j^\mathsf{T}\Gamma_o^{-1}d_j}_{=b_j}, \quad j \geq 1. \tag{5.1}$$

We recall that the solution of (5.1) is obtained using the PCG (Algorithm 2.2) with Γ_b as a first-level preconditioner. When eigeninformation related to the dominant eigenmodes of $\Gamma_b A_j$ is available, then the LMP is rather used as a preconditioner.

In concrete applications, exact eigeninformation is never available, and cannot be computed using dedicated eigensolvers because of the prohibitive computational cost. Instead, a widespread strategy consists in using Ritz pairs computed as described in Section 2.2.5 from the previous application of the PCG. This strategy has proven to perform well in data assimilation contexts [89]. However, it has some limitations. First, the number of approximate eigenpairs is always determined by the number of PCG iterations performed to solve the previous linear system. Second, let us assume that we compute the Ritz pairs from the application of the PCG on the $(j-1)$-th linear system with preconditioner M_{j-1}. Let $V_{j-1} \in \mathbb{R}^{n\times k}$ be the matrix containing the Ritz vectors and $\Lambda_{j-1}\mathbb{R}^{k\times k}$ the diagonal matrix containing the corresponding Ritz values. Then these matrices satisfy

$$M_{j-1}A_{j-1}V_{j-1} \approx V_{j-1}\Lambda_{j-1}.$$

But, it is not obvious that V_{j-1} and Λ_{j-1} will also be such that

$$M_j A_j V_{j-1} \approx V_{j-1}\Lambda_{j-1}.$$

Thus for the Ritz pairs to be relevant, we must first assume that $A_{j-1} \approx A_j$. Then, to account for the change of preconditioner, there are two possibilities. The first one consists in using the same preconditioner for all the linear systems, that is $M_{j-1} = M_j = \Gamma_b$. The second one imposes to use for M_j a LMP such that

$$M_j = \left[I_n - Q_j A_j\right] M_{j-1} \left[I_n - A_j Q_j\right] + Q_j, \tag{5.2}$$

where $Q_j = V_{j-1}(V_{j-1}^\mathsf{T} A V_{j-1})^{-1} V_{j-1}^\mathsf{T}$. In this case, M_j is relevant if V_{j-1} contains eigeninformation related to $M_{j-1}A_j$ and we fall back on the assumption $A_{j-1} \approx A_j$. With this second strategy, the preconditioner increases in complexity along the sequence, and requires an increased cost in terms of storage and computational applications. For these reasons, the approach generally considered is to compute the Ritz pairs only for the first system, and to use them in the LMP for all the next Gauss-Newton steps.

In this chapter, we propose an alternative solution based on randomized algorithms. We propose a class of randomized LMPs where the approximate eigenpairs are obtained using adaptations of the randomized algorithms introduced in Chapter 4.

5.1.1 Related research

Randomized methods for solving (5.1) have already been introduced in Section 2.4. First, the RIOT (Algorithm 2.7) method introduced in [17] and tested in [16] where the Nyström method (Algorithm 2.5) is used to compute a low rank approximation of $\Gamma_b^{1/2} A_j \Gamma_b^{1/2}$ and directly approximate the solution of $A_j x_j = b_j$. We refer the reader to Section 2.4.4 for more details. In this approach, the PCG is replaced by a fully parallel randomized approach. In general, this technique has proven to perform well [16], except in certain situations where convergence issues have been observed.

Therefore, the use of the PCG is convenient to maintain theoretical guarantees on the obtained approximate solutions. In this regard, the authors in [24] have proposed to use randomized approximate eigenvalue decompositions to construct randomized LMPs for the PCG. In addition, based on the Nyström method, they have introduced the Ritzit method (Algorithm 2.6) which has proven to perform well on the Lorenz 95 model.

Recently, the authors in [33] have proposed a theoretical analysis of the Nyström LMP. Their theoretical analysis demonstrates that those preconditioners can perform well whenever the eigenvalues of the symmetric positive definite matrix are quickly decaying [33, Theorem 5.1]. This result confirms that such randomized LMPs are particularly adapted to variational data assimilation, as the eigenvalues distribution of $\Gamma_b A_j$ does satisfy this condition.

5.1.2 Contributions

The manifest drawback of both the RIOT and Ritzit method is that they all rely on the availability of a factorization for Γ_b. This factorization allows to use Γ_b as a split preconditioner, and consequently, to apply randomized methods dedicated to symmetric positive definite problems. However, such factorization is neither always available, nor computationally affordable. In this case, to the best of our knowledge, no randomized approach has been proposed, and this chapter intends to bridge this gap.

In this regard, we propose randomized LMPs for two Krylov subspace methods dedicated to the solution of (5.1) when $\Gamma_b^{1/2}$ is not available. The first method has been developed to address the solution of (5.1) when Γ_b^{-1} is also not available. In this case, applying PCG to (5.1) with $M_j = \Gamma_b$ is not possible, as applying A_j (step 7 of Algorithm 2.2) implies to apply Γ_b^{-1}. Consequently, a dedicated solver named PCG Inverse-Free (PCGIF) has been proposed in [50, Section 3.1] (Approach (C)). It relies on right preconditioning, and thus modifies the inner-product in consequence (see Section 2.2.3). Adapted formulations of the LMP [50, Lemma 3.1] and its spectral variant [50, Equation 3.80] have also been proposed.

The second approach we consider is based on the dual formulation of (2.17) [68, Section 12.9]. This approach was initially motivated by the fact that the dual space is of dimension m, yielding linear systems of size $m \times m$ instead of $n \times n$ for the primal space. This involves obvious advantages from a computational and storage perspective in the case where $m \ll n$. The variant of the PCG we consider, named the Augmented Restricted PCG (RPCG) [48, Algorithm 5], has been proposed to guarantee that the iterates obtained with the dual method are mathematically equivalent to the ones from the primal approach. Similarly, expressions for the LMP and its spectral variant have also been proposed to precondition the (Augmented) RPCG [50].

In this chapter, we thus propose and study randomized spectral LMPs for both the PCGIF (Algorithms 5.4 and 5.5) and Augmented RPCG (Algorithms 5.6 and 5.7). As will be shown, constructing the spectral LMP for these Krylov subspace methods requires to compute eigenpairs of generalized eigenvalue problems with non-Euclidean inner products. In this context, we adapt the algorithms proposed in Chapter 4, to substitute the expensive computation of exact eigenpairs with the computation of approximate eigenpairs using randomized methods. Then, in Theorems 5.9 and 5.10, we show the equivalence between the randomized spectral LMP for the PCGIF and the RPCG that ensures the mathematical equivalence of the iterates. Finally, we investigate the performance of the proposed randomized LMPs on a four-dimensional variationonal data assimilation problem.

5.2 Preliminaries

In this preliminary section, we present the PCGIF [50] and the Augmented RPCG [48] methods.

5.2.1 Solving the linearized subproblem in the primal space

A method for solving (5.1) when Γ_b^{-1} is not available has been proposed in [50, Section 3.1]. As mentioned in the introduction, this inverse-free approach is based on right preconditioning (5.1) with Γ_b, that is considering the change of variable $\bar{s}_j = \Gamma_b^{-1} s_j$, so that (5.1) becomes

$$A_j \Gamma_b \bar{s}_j = b_j. \tag{5.3}$$

The new system matrix, denoted by $\bar{A}_j = A_j \Gamma_b = I_n + H_j^\mathsf{T} \Gamma_o^{-1} H_j \Gamma_b$ no longer contains Γ_b^{-1}. However, $\bar{A}_j$ is now symmetric with respect to the Γ_b inner product. Consequently, (5.3) can be solved by applying the PCG in the Γ_b-inner product, which yields the PCG *Inverse-Free* (PCGIF) proposed in [50, Algorithm 3.4]. In this context, preconditioning (5.3) is also perfectly possible, but the preconditioner $\bar{M}_j$ must also be Γ_b-symmetric.

It remains to cope with the right-hand side, whose expression reads

$$b_j = \Gamma_b^{-1}(x_c - x_j) + H_j^\mathsf{T} \Gamma_o^{-1} d_j.$$

It is actually possible to form the right-hand side without requiring Γ_b^{-1}. Remembering that the Gauss-Newton iterates satisfy $x_{j+1} = x_j + \Gamma_b \bar{s}_j$, we can write by induction

$$x_j = x_1 + \Gamma_b \left(\sum_{k=1}^{j-1} \bar{s}_k \right).$$

If the optimization process is initialized with $x_1 = x_c$, then rearranging the terms yields

$$\Gamma_b^{-1}(x_c - x_j) = -\sum_{k=1}^{j-1} \bar{s}_k,$$

meaning that b_j can be computed without requiring Γ_b^{-1} as

$$b_j = -\sum_{k=1}^{j-1} \bar{s}_k + H_j^\mathsf{T} \Gamma_o^{-1} d_j,$$

given that the solutions from the previous linear systems are stored. Let us take a moment to discuss whether $x_1 = x_c$ is restrictive or not. By definition, the center vector is a relevant estimate of the nonlinear regression solution. Consequently, starting the optimization process with $x_1 = x_c$ is both relevant and done in practice. If there were a better estimate for the solution, then x_c would be modified to that better estimate.

Let us discuss under which conditions applying the PCG to (5.1) with preconditioner M_j and applying the PCGIF to (5.3) with $\bar{M}_j$ yields identical Gauss-Newton iterates. In this regard, for a given j, let us denote $\bar{s}_i$ the i-th iterate of the PCGIF when applied to solve the j-th linear system (5.3), and $\bar{r}_i = b_j - \bar{A}_j \bar{s}_i$ the corresponding i-th residual. It can been shown (see [50, Approach (C) Section 3.1]) that under the conditions

$$\begin{cases} r_0 = \bar{r}_0 \\ M_j = \Gamma_b \bar{M}_j \end{cases}, \tag{5.4}$$

then the quantities in the PCG (Algorithm 2.2) and in the PCGIF (Algorithm 5.1) satisfy [50, Relation 3.15] for all $i \geq 0$

$$\begin{cases} r_i = \bar{r}_i \\ s_i = \Gamma_b \bar{s}_i \end{cases}. \tag{5.5}$$

For the first condition, since we focus on the PCG with a zero initial guess, one has $r_0 = b_j$. Consequently, since (5.1) and (5.3) share the same right-hand side, one can verify that the first condition in (5.4) imposes a zero initial guess for the PCGIF. For the second condition on the preconditioner, we can readily see that preconditioning the PCG with $M_j = \Gamma_b$ is equivalent to using $\bar{M}_j = I_n$ in the PCGIF. In this regard, the PCGIF integrates this preconditioner by default.

The resulting algorithm is presented in Algorithm 5.1. It is notable that getting a viable implementation requires some care. In particular, maintaining one application of $H_j^\mathsf{T} \Gamma_o^{-1} H_j$, Γ_b and $\bar{M}_j$ per iteration requires to introduce two additional recurrence vectors, namely $\bar{w}_i$ and $\bar{t}_i$. For the stopping criterion, we monitor the convergence with the relative decrease in the residual $\bar{r}_i$ measured in the $\Gamma_b \bar{M}_j$ norm. This criterion is classical, and can easily be computed using the relation

$$\|\bar{r}_i\|_{\Gamma_b \bar{M}_j} = \sqrt{\bar{\rho}_i},$$

where $\bar{\rho}_i$ is obtained at step 18 of Algorithm 5.1. Plus, we observe that if M_j and $\bar{M}_j$ satisfy (5.4), then using (5.5) we have

$$\|\bar{r}_i\|^2_{\Gamma_b \bar{M}_j} = \bar{r}_i^\mathsf{T} \Gamma_b \bar{M}_j \bar{r}_i = \bar{r}_i^\mathsf{T} M_j \bar{r}_i = r_i M_j r_i = \|r_i\|^2_{M_j}.$$

Consequently, if the conditions in (5.4) are satisfied, the monitoring of the convergence for the PCG and the PCGIF is also compatible, in the sense that the algorithms will be stopped at the same moment.

Preconditioning the PCGIF with LMPs

Let us now recall LMP formulas for the PCGIF, that is, LMPs for linear systems that are Γ_b-symmetric. This problem has been addressed in [50, Lemma 3.1], and we recall the result for completeness.

Algorithm 5.1: PCGIF to solve $A_j s_j = b_j$ [50, Algorithm 3.4].

Input: Operators $\Gamma_o^{-1}, H_j, \Gamma_b, H_j^\mathsf{T}$, preconditioner $\bar{M}_j$ which is Γ_b-symmetric, misfit vector $d_j \in \mathbb{R}^m$, solutions $\bar{s}_1, \ldots, \bar{s}_{j-1} \in \mathbb{R}^n$ of the previous $j-1$ linear system, tolerance $\varepsilon > 0$.

$\bar{s}_0 = 0$
$\bar{r}_0 = -\sum_{k=1}^{j-1} \bar{s}_k + H_j^\mathsf{T} \Gamma_o^{-1} d_j$
$\bar{z}_0 = \bar{M}_j \bar{r}_0$
$\bar{p}_0 = \bar{z}_0$
$\bar{w}_0 = \Gamma_b \bar{z}_0$
$\bar{t}_0 = \bar{w}_0$
$\bar{\rho}_0 = \bar{r}_0^\mathsf{T} \bar{w}_0$
while *convergence is not reached* **do**
  $\bar{q}_i = \bar{p}_i + H_j^\mathsf{T} \Gamma_o^{-1} H_j \bar{t}_i$
  $\alpha_i = \bar{\rho}_i / \bar{q}_i^\mathsf{T} \bar{t}_i$
  $\bar{s}_{i+1} = \bar{s}_i + \alpha_i \bar{p}_i$
  $\bar{r}_{i+1} = \bar{r}_i - \alpha_i \bar{q}_i$
  $\bar{z}_{i+1} = \bar{M}_j \bar{r}_{i+1}$
  $\bar{w}_{i+1} = \Gamma_b \bar{z}_{i+1}$
  if $\|\bar{r}_{i+1}\|_{\Gamma_b \bar{M}_j} \leq \varepsilon \|\bar{r}_0\|_{\Gamma_b \bar{M}_j}$ **then**
    Stop the method.
  end
  $\bar{\rho}_{i+1} = \bar{r}_{i+1}^\mathsf{T} \bar{w}_{i+1}$
  $\beta_i = \bar{\rho}_{i+1} / \bar{\rho}_i$
  $\bar{p}_{i+1} = \bar{z}_{i+1} + \beta_i \bar{p}_i$
  $\bar{t}_{i+1} = \bar{w}_{i+1} + \beta_i \bar{t}_i$
end

Output: Final iterate $\bar{s}_f$ such that $s_j = \Gamma_b \bar{s}_f$.

Proposition 5.1 ([50], Lemma 3.1)**.** *Let $\bar{A}, \bar{M} \in \mathbb{R}^{n \times n}$ be two Γ_b-symmetric positive definite matrices. Let $\bar{S} \in \mathbb{R}^{n \times k}$ be any full column rank matrix. Then, if we denote $\bar{Q} = \bar{S}(\bar{S}^\mathsf{T} \Gamma_b \bar{A} \bar{S})^{-1} \bar{S}^\mathsf{T} \Gamma_b$, then*

$$C = \left[I_n - \bar{Q}\bar{A}\right] \bar{M} \left[I_n - \bar{A}\bar{Q}\right] + \bar{Q},$$

is a Γ_b-symmetric matrix, and $\Gamma_b C$ is symmetric positive definite.

As for the standard LMP, it is possible to obtain a simplified expression when $\bar{S}$ contains eigeninformation. The obtained spectral LMP has been derived in [50, Relation 3.80] and is recalled in the following proposition.

Proposition 5.2. *Let $\bar{A}, \bar{M} \in \mathbb{R}^{n \times n}$ be two Γ_b-symmetric matrices. Let $\bar{S} \in \mathbb{R}^{n \times k}$ be a matrix whose columns are distinct eigenvectors of $\bar{M}\bar{A}$ and $\bar{\Lambda} \in \mathbb{R}^{k \times k}$ a diagonal matrix containing the corresponding eigenvalues. Then, the matrix C introduced in Lemma 5.1 takes the form*

$$C_{sp} = \bar{M} + \bar{S}\left(\bar{\Lambda}^{-1} - I_k\right) \bar{S}^\mathsf{T} \Gamma_b. \tag{5.6}$$

Proof. The proof follows from immediate algebraic manipulations. First, by assumption one has the identity $\bar{M}\bar{A}\bar{S} = \bar{S}\bar{\Lambda}$ and the fact that $\bar{S}^\mathsf{T} \Gamma_b \bar{M}^{-1} \bar{S} = I_k$. From this, we obtain $\bar{Q} = \bar{S}\bar{\Lambda}^{-1}\bar{S}^\mathsf{T}\Gamma_b$ and $\bar{M}\bar{A}\bar{Q} = \bar{S}\bar{S}^\mathsf{T}\Gamma_b$. □

Remark 5.3. In practice, we would store both $\bar{S}$ and $\Gamma_b\bar{S}$ so that applying C_{sp} does not require applying Γ_b. We will notice that deriving an approximate eigenvalue decomposition via randomized methods actually allows us to obtain both $\bar{S}$ and $\Gamma_b\bar{S}$ at the same cost.

Now, the natural question is to determine the conditions under which the matrix C as in Proposition 5.1 and P as in (2.19) satisfy the second condition of (5.4), that is,

$$P = \Gamma_b C.$$

Those have already been highlighted in [50, Lemma 3.1], and one can indeed verify that if

$$\begin{cases} A = \bar{A}\Gamma_b^{-1} \\ S = \Gamma_b\bar{S} \\ M = \Gamma_b\bar{M} \end{cases}, \tag{5.7}$$

then P and C do satisfy the condition (5.4). Indeed, under these conditions, one has

$$\begin{aligned} QA &= S(S^\mathsf{T}AS)^{-1}S^\mathsf{T}A \\ &= \Gamma_b\bar{S}(\bar{S}^\mathsf{T}\Gamma_b\bar{A}\Gamma_b^{-1}\Gamma_b\bar{S})^{-1}\bar{S}^\mathsf{T}\Gamma_b\bar{A}\Gamma_b^{-1} \\ &= \Gamma_b\bar{Q}\bar{A}\Gamma_b^{-1}, \end{aligned}$$

Plugging this relation into (2.19) and using that $M = \Gamma_b\bar{M}$ yields the result. We observe that the system matrices in (5.1) and (5.3) actually satisfy the first condition by definition, since $\bar{A}_j = A_j\Gamma_b$. Also, we have already highlighted that $M = \Gamma_b\bar{M}$ is for instance satisfied when $M = \Gamma_b$ and $\bar{M} = I_n$.

5.2.2 Solving the linearized subproblem in the dual space

Let us now introduce the second approach, based on the dual formulation. The main objective of this approach is to exploit the structure of the problem to alleviate both the computational costs and memory requirements of the inner-iterations. Let us start from (5.1), that is

$$\left(\Gamma_b^{-1} + H_j^\mathsf{T}\Gamma_o^{-1}H_j\right)s_j = \Gamma_b^{-1}(x_c - x_j) + H_j^\mathsf{T}\Gamma_o^{-1}d_j.$$

This system can be rewritten

$$\left(\Gamma_b^{-1} + H_j^\mathsf{T}\Gamma_o^{-1}H_j\right)(s_j - x_c + x_j) = H_j^\mathsf{T}\Gamma_o^{-1}\left(d_j - H_j(x_c - x_j)\right).$$

Consequently, the analytic expression of the increment reads

$$s_j = x_c - x_j + \left(\Gamma_b^{-1} + H_j^\mathsf{T}\Gamma_o^{-1}H_j\right)^{-1}H_j^\mathsf{T}\Gamma_o^{-1}\left(d_j - H_j(x_c - x_j)\right).$$

By applying the Sherman-Morrison-Woodbury (3.3), we can obtain the equivalent expression

$$s_j = x_c - x_j + \Gamma_b H_j^\mathsf{T}\left(R + H_j\Gamma_b H_j^\mathsf{T}\right)^{-1}(d_j - H_j(x_c - x_j)),$$

which suggests that the increment can be obtained by rather solving the linear system

$$\left(R + H_j\Gamma_b H_j^\mathsf{T}\right)\widehat{s}_j = d_j - H_j(x_c - x_j), \tag{5.8}$$

which is of order m instead of n, and then retrieve the initial solution via

$$s_j = x_c - x_j + \Gamma_b H_j^\mathsf{T} \widehat{s}_j.$$

This is the main idea behind the dual approach, but some precautions have to be taken. In concrete applications, (5.1) is not solved accurately, because it is neither needed nor affordable. Consequently, a few inner loop iterations are computed, hopefully minimizing as much as possible the quadratic function (2.17). The fact is that the PCG algorithm precisely solves (5.1) by minimizing (2.17). However, applying the PCG on (5.8) will not minimize the same quadratic function, which can lead to very poor results when few inner loop iterations are performed. This was the main motivation behind the Augmented Restricted PCG (RPCG) proposed in [48]. To ensure that the inner loop iterates in the dual space are indeed minimizing the primal quadratic (2.17), one must rather solve

$$\left(I_m + \Gamma_o^{-1} H_j \Gamma_b H_j^\mathsf{T}\right) \widehat{s}_j = \Gamma_o^{-1} \left(d_j - H_j(x_c - x_j)\right), \tag{5.9}$$

with the PCG in the inner product induced by $H_j \Gamma_b H_j^\mathsf{T}$. Indeed, if we denote $\widehat{A}_j = I_m + \Gamma_o^{-1} H_j \Gamma_b H_j^\mathsf{T}$, then $\widehat{A}_j$ is $H_j \Gamma_b H_j^\mathsf{T}$-symmetric. We see here the analogy with the inverse-free approach. A preconditioner $\widehat{M}_j$ can also be used, given that it is $H_j \Gamma_b H_j^\mathsf{T}$-symmetric too. To maintain the equivalence with the iterates obtained with the primal approach, it has been shown in [50, p.70] that the following two assumptions must be satisfied

$$\begin{cases} r_0 = H_j^\mathsf{T} \widehat{r}_0 \\ M_j H_j^\mathsf{T} = \Gamma_b H_j^\mathsf{T} \widehat{M}_j \end{cases}, \tag{5.10}$$

where $\widehat{r}_0$ denotes the initial residual in the dual space. Under these conditions, it can then be shown [48, Section 3.2] that for all $i \geq 0$ one has

$$\begin{cases} r_i = H_j^\mathsf{T} \widehat{r}_i \\ s_i = \Gamma_b H_j^\mathsf{T} \widehat{s}_i \end{cases}. \tag{5.11}$$

Once again, we observe that the second condition in (5.10) is satisfied when $M_j = \Gamma_b$ and $\widehat{M}_j = I_m$, which means that the method in the dual space also integrates the preconditioner Γ_b in its structure. However, in our case, since $r_0 = b_j = \Gamma_b^{-1}(x_c - x_j) + H_j^\mathsf{T} \Gamma_o^{-1} d_j$, it is not clear how to obtain $\widehat{r}_0$ such that (5.10) is satisfied. The only favorable case is for the first Gauss-Newton step, where $x_1 = x_c$ implies that $r_0 = H_1^\mathsf{T} \Gamma_o^{-1} d_1$, yielding $\widehat{r}_0 = \Gamma_o^{-1} d_1$. Here, we notice that $\widehat{r}_0$ is actually equal to the right-hand side in (5.8) for $j = 1$.

From this particular situation, we obtain a first algorithm, called the Restricted PCG (RPCG), summarized in Algorithm 5.2. Again, the proposed implementation involves two additional recurrence vectors $\widehat{w}_i$ and $\widehat{t}_i$ to ensure that the operators $H_j \Gamma_b H_j^\mathsf{T}$, Γ_o^{-1} and $\widehat{M}_j$ are applied only once per iteration. For the stopping criterion, the method is stopped whenever a sufficient relative decrease in the residual measured in the $H_j \Gamma_b H_j^\mathsf{T} \widehat{M}_j$ norm is observed. This is the natural quantity to monitor as it is easily accessible using

$$\|\widehat{r}_i\|_{H_j \Gamma_b H_j^\mathsf{T} \widehat{M}_j} = \sqrt{\widehat{\rho}_i},$$

where $\widehat{\rho}_i$ is obtained at step 18 of Algorithm 5.2. Plus, if M_j is a preconditioner for the standard PCG satisfying (5.10), then using (5.11) we have

$$\|\widehat{r}_i\|^2_{H_j \Gamma_b H_j^\mathsf{T} \widehat{M}_j} = \widehat{r}_i^\mathsf{T} H_j \Gamma_b H_j^\mathsf{T} \widehat{M}_j \widehat{r}_i = \widehat{r}_i^\mathsf{T} H_j M_j H_j^\mathsf{T} \widehat{r}_i = r_i M_j r_i = \|r_i\|^2_{M_j}.$$

Algorithm 5.2: RPCG to solve $A_1 s_1 = b_1$ [48, Algorithm 5]

Input: Operators $\Gamma_o^{-1}, H_1, \Gamma_b, H_1^\mathsf{T}$, preconditioner $\bar{M}_1$ which is $H_1^\mathsf{T}\Gamma_b H_1$-symmetric, misfit vector $d_1 \in \mathbb{R}^m$, tolerance $\varepsilon > 0$.

$\widehat{s}_0 = 0$
$\widehat{r}_0 = \Gamma_o^{-1} d_1$
$\widehat{z}_0 = \widehat{M}_1 \widehat{r}_0$
$\widehat{p}_0 = \widehat{z}_0$
$\widehat{w}_0 = H_1 \Gamma_b H_1^\mathsf{T} \widehat{z}_0$
$\widehat{t}_0 = \widehat{w}_0$
$\widehat{\rho}_0 = \widehat{r}_0^\mathsf{T} \widehat{w}_0$
while *convergence is not reached* **do**
  $\widehat{q}_i = \widehat{p}_i + \Gamma_o^{-1} \widehat{t}_i$
  $\alpha_i = \widehat{\rho}_i / \widehat{q}_i^\mathsf{T} \widehat{t}_i$
  $\widehat{s}_{i+1} = \widehat{s}_i + \alpha_i \widehat{p}_i$
  $\widehat{r}_{i+1} = \widehat{r}_i - \alpha_i \widehat{q}_i$
  $\widehat{z}_{i+1} = \widehat{M}_1 \widehat{r}_{i+1}$
  $\widehat{w}_{i+1} = H_1 \Gamma_b H_1^\mathsf{T} \widehat{z}_{i+1}$
  if $\|\widehat{r}_{i+1}\|_{H_1 \Gamma_b H_1^\mathsf{T} \widehat{M}_1} \leq \varepsilon \|\widehat{r}_0\|_{H_1 \Gamma_b H_1^\mathsf{T} \widehat{M}_1}$ **then**
    Stop the method.
  end
  $\widehat{\rho}_{i+1} = \widehat{r}_{i+1}^\mathsf{T} \widehat{w}_{i+1}$
  $\beta_i = \widehat{\rho}_{i+1} / \widehat{\rho}_i$
  $\widehat{p}_{i+1} = \widehat{z}_{i+1} + \beta_{i+1} \widehat{p}_i$
  $\widehat{t}_{i+1} = \widehat{w}_{i+1} + \beta_{i+1} \widehat{t}_i$
end

Output: Final iterate $\widehat{s}_f$ such that $s_1 = \Gamma_b H_1^\mathsf{T} \widehat{s}_1$.

Consequently, this stopping criterion is compatible with the one of the PCG, in the sense that both solvers will stop at the same time.

To overcome the problem incurred by the condition $r_0 = H_j^\mathsf{T} \widehat{r}_0$, the authors in [48] proposed an alternative formulation of the RPCG. The idea is to consider augmented matrices so that the condition becomes naturally satisfied by the corresponding augmented residual. Accordingly, let us define

$$\underline{H}_j = \begin{bmatrix} H_j \\ (x_c - x_j)^\mathsf{T} \Gamma_b^{-1} \end{bmatrix}, \quad \underline{R}^{-1} = \begin{bmatrix} \Gamma_o^{-1} & \\ & 0 \end{bmatrix}, \quad \underline{d}_j = \begin{bmatrix} \Gamma_o^{-1} d_j \\ 1 \end{bmatrix}. \tag{5.12}$$

Now, it is clear that

$$r_0 = \Gamma_b^{-1}(x_c - x_j) + \Gamma_o^{-1} d_j = \underline{H}_j^\mathsf{T} \underline{d}_j,$$

meaning that the augmented initial residual to choose is simply $\underline{d}_j$. The Augmented RPCG can then be derived following the steps as for the RPCG, basically replacing the matrices by their augmented counterparts. However, the vectors in the Augmented PCG are now of size $m + 1$, and the procedure is summarized in Algorithm 5.3. Here, we note that an alternative version is proposed in [48, Algorithm 8] where there is no longer reference to the augmented matrices. However, this implies to explicitly write partitioning of vectors, which degrades the readability and does not bring additional information. For the stopping criterion, the natural quantity to

look at remains the residual in the $\underline{H}_j\Gamma_b\underline{H}_j^\mathsf{T}\widehat{\underline{M}}_j$-norm, since it can be computed using

$$\|\widehat{r}_{i+1}\|_{\underline{H}_j\Gamma_b\underline{H}_j^\mathsf{T}\widehat{\underline{M}}_j} = \sqrt{\widehat{\rho}_i}.$$

However, here there is no longer a compatibility with the stopping criterion of the PCG.

Algorithm 5.3: Augmented RPCG to solve $A_j s_j = b_j$ with $j \geq 1$ [48, Algorithm 7]

Input: Operators $\Gamma_o^{-1}, \underline{H}_j, \Gamma_b, \underline{H}_j^\mathsf{T}$, preconditioner $\widehat{\underline{M}}_j$ which is $\underline{H}_j^\mathsf{T}\Gamma_b\underline{H}_j$-symmetric, misfit vector $d_j \in \mathbb{R}^m$, tolerance $\varepsilon > 0$.

$\widehat{s}_0 = 0$
$\widehat{r}_0 = \underline{d}_j$
$\widehat{z}_0 = \widehat{\underline{M}}_j\widehat{r}_0$
$\widehat{p}_0 = \widehat{z}_0$
$\widehat{w}_0 = \underline{H}_j\Gamma_b\underline{H}_j^\mathsf{T}\widehat{z}_0$
$\widehat{t}_0 = \widehat{w}_0$
$\widehat{\rho}_0 = \widehat{r}_0^\mathsf{T}\widehat{w}_0$
while *convergence is not reached* **do**
  $\widehat{q}_i = \widehat{p}_i + \underline{R}^{-1}\widehat{t}_i$
  $\alpha_i = \widehat{\rho}_i / \widehat{q}_i^\mathsf{T}\widehat{t}_i$
  $\widehat{s}_{i+1} = \widehat{s}_i + \alpha_i\widehat{p}_i$
  $\widehat{r}_{i+1} = \widehat{r}_i - \alpha_i\widehat{q}_i$
  $\widehat{z}_{i+1} = \widehat{\underline{M}}_j\widehat{r}_{i+1}$
  $\widehat{w}_{i+1} = \underline{H}_j\Gamma_b\underline{H}_j^\mathsf{T}\widehat{z}_{i+1}$
  if $\|\widehat{r}_{i+1}\|_{\underline{H}_j\Gamma_b\underline{H}_j^\mathsf{T}\widehat{\underline{M}}_j} \leq \varepsilon\,\|\widehat{r}_0\|_{\underline{H}_j\Gamma_b\underline{H}_j^\mathsf{T}\widehat{\underline{M}}_j}$ **then**
    Stop the method.
  end
  $\widehat{\rho}_{i+1} = \widehat{r}_{i+1}^\mathsf{T}\widehat{w}_{i+1}$
  $\beta_i = \widehat{\rho}_{i+1}/\widehat{\rho}_i$
  $\widehat{p}_{i+1} = \widehat{z}_{i+1} + \beta_{i+1}\widehat{p}_i$
  $\widehat{t}_{i+1} = \widehat{w}_{i+1} + \beta_{i+1}\widehat{t}_i$
end

Output: Final iterate $\widehat{s}_f$ such that $s_j = \Gamma_b\underline{H}_j^\mathsf{T}\widehat{s}_f$.

LMP for the dual approach

We are again interested in finding LMP formulas for the preconditioner $\widehat{M}_j$ and $\widehat{\underline{M}}_j$ appearing in Algorithms 5.2 and 5.3. The problem has been solved for the RPCG in [50] and we present the result in Proposition 5.4. We discuss the case of the Augmented RPCG afterward.

Proposition 5.4 ([50], Lemma 4.1)**.** *Let $\widehat{A}, \widehat{M} \in \mathbb{R}^{m\times m}$ be two $H_j\Gamma_b H_j^\mathsf{T}$-symmetric matrices. Let $\widehat{S} \in \mathbb{R}^{m\times k}$ be any full column rank matrix and $\widehat{Q}_j = \widehat{S}(\widehat{S}^\mathsf{T}H_j\Gamma_b H_j^\mathsf{T}\widehat{A}\widehat{S})^{-1}\widehat{S}^\mathsf{T}H_j\Gamma_b H_j^\mathsf{T}$. Then*

$$D_j = \left[I_m - \widehat{Q}_j\widehat{A}\right]\widehat{M}\left[I_m - \widehat{A}\widehat{Q}_j\right] + \widehat{Q}_j,$$

is a $H_j\Gamma_b H_j^\mathsf{T}$-symmetric matrix, and $H_j\Gamma_b H_j^\mathsf{T}D_j$ is symmetric positive definite.

Provided that we consider the appropriate eigenvalue problem, a simplified variant for D_j in Proposition 5.4 can be derived. This spectral variant is detailed in the next proposition.

Proposition 5.5. *Let $\widehat{A}, \widehat{M} \in \mathbb{R}^{m\times m}$ be two $H_j\Gamma_b H_j^\mathsf{T}$-symmetric matrices. Let $\widehat{S} \in \mathbb{R}^{m\times k}$ be a matrix whose columns are distinct eigenvectors of $\widehat{M}\widehat{A}$ and $\widehat{\Lambda} \in \mathbb{R}^{k\times k}$ a diagonal matrix containing the corresponding eigenvalues. Then the matrix D introduced in Lemma 5.4 takes the form*

$$D_{j,sp} = \widehat{M} + \widehat{S}\left(\widehat{\Lambda}^{-1} - I_k\right)\widehat{S}^\mathsf{T} H_j\Gamma_b H_j^\mathsf{T}. \tag{5.13}$$

Proof. The proof follows from immediate algebraic manipulations. By assumption one has $\widehat{M}\widehat{A}\widehat{S} = \widehat{S}\widehat{\Lambda}$ and the fact that $\widehat{S}^\mathsf{T} H_j\Gamma_b H_j^\mathsf{T}\widehat{M}^{-1}\widehat{S} = I_k$. □

Remark 5.6. Again, we will consider methods that provide both $\widehat{S}$ and $H_j\Gamma_b H_j^\mathsf{T}\widehat{S}$ so that applying D_{sp} will not require to apply $H_j\Gamma_b H_j^\mathsf{T}$.

Let us get interested in the conditions under which a LMP D for the RPCG defined in Proposition 5.1 and a standard LMP P as in (2.10) satisfy the compatibility condition in (5.10). One can verify that if

$$\begin{cases} A\Gamma_b H_j^\mathsf{T} = H_j^\mathsf{T}\widehat{A} \\ S = \Gamma_b H_j^\mathsf{T}\widehat{S} \\ MH_j^\mathsf{T} = \Gamma_b H_j^\mathsf{T}\widehat{M} \end{cases}, \tag{5.14}$$

then P and D do satisfy $PH_j^\mathsf{T} = \Gamma_b H_j^\mathsf{T} D$.

Regarding the operators in (5.1) and (5.9) one remarks that

$$\begin{aligned} A_j\Gamma_b H_j^\mathsf{T} &= (\Gamma_b^{-1} + H_j^\mathsf{T}\Gamma_o^{-1}H_j)\Gamma_b H_j^\mathsf{T} \\ &= H_j^\mathsf{T} + H_j^\mathsf{T}\Gamma_o^{-1}H_j\Gamma_b H_j^\mathsf{T} \\ &= H_j^\mathsf{T}(I_m + \Gamma_o^{-1}H_j\Gamma_b H_j^\mathsf{T}) \\ &= H_j^\mathsf{T}\widehat{A}_j. \end{aligned}$$

Consequently, the matrices in (5.1) and (5.9) satisfy the condition on the operators. The condition $MH_j^\mathsf{T} = \Gamma_b H_j^\mathsf{T}\widehat{M}$ is for instance also satisfied for $M = \Gamma_b$ and $\widehat{M} = I_m$.

Let us now discuss on LMPs for the Augmented LMP. It turns out that the only formulations available in [50] are obtained imposing the equivalence between the Augmented RPCG and RPCG when possible. However, as we have already discussed, this would only be possible for the first Gauss-Newton step. When this equivalence does not hold, it is not clear how to use a preconditioner $\widehat{M}$ for the RPCG into a preconditioner $\underline{\widehat{M}}$ for the Augmented RPCG. Consequently, we propose a formulation inspired from Proposition 5.5. This is motivated by the analogy between the two solvers, and we consider the augmented system matrix

$$\underline{\widehat{A}}_j = I_{m+1} + \underline{R}^{-1}\underline{H}_j\Gamma_b\underline{H}_j^\mathsf{T} \in \mathbb{R}^{(m+1)\times(m+1)}.$$

Definition 5.7. Let $\underline{\widehat{A}}, \underline{\widehat{M}} \in \mathbb{R}^{(m+1)\times(m+1)}$ be two $\underline{H}_j\Gamma_b\underline{H}_j^\mathsf{T}$-symmetric. Let $\widehat{S} \in \mathbb{R}^{(m+1)\times k}$ be any full column rank matrix and let us denote $\widehat{Q}_j = \widehat{S}(\widehat{S}^\mathsf{T}\underline{H}_j\Gamma_b\underline{H}_j^\mathsf{T}\underline{\widehat{A}}\widehat{S})^{-1}\widehat{S}^\mathsf{T}\underline{H}_j\Gamma_b\underline{H}_j^\mathsf{T}$. Then we define the LMP for the Augmented RPCG as

$$\underline{D}_j = \left[I_m - \widehat{Q}_j\underline{\widehat{A}}\right]\underline{\widehat{M}}\left[I_m - \underline{\widehat{A}}\widehat{Q}_j\right] + \widehat{Q}_j,$$

We remark that $\underline{D}_j$ is $\underline{H}_j\Gamma_b\underline{H}_j^\mathsf{T}$-symmetric and that $\underline{H}_j\Gamma_b\underline{H}_j^\mathsf{T}\underline{D}_j$ is symmetric positive definite by analogous arguments as in Proposition 5.4. Similarly, we define the spectral variant in the next proposition.

Proposition 5.8. *Let $\widehat{\underline{A}}, \widehat{\underline{M}} \in \mathbb{R}^{(m+1)\times(m+1)}$ be two $\underline{H}_j \Gamma_b \underline{H}_j^{\mathsf{T}}$-symmetric. Let $\widehat{S} \in \mathbb{R}^{(m+1)\times k}$ be a matrix whose columns are distinct eigenvectors of $\widehat{\underline{M}}\widehat{\underline{A}}$ and $\widehat{\Lambda} \in \mathbb{R}^{k\times k}$ a diagonal matrix containing the corresponding eigenvalues. Then the matrix D introduced in Definition 5.7 takes the form*

$$\underline{D}_{j,sp} = \widehat{\underline{M}} + \widehat{S}\left(\widehat{\Lambda}^{-1} - I_k\right)\widehat{S}^{\mathsf{T}}\underline{H}_j\Gamma_b\underline{H}_j^{\mathsf{T}}. \tag{5.15}$$

Proof. The proof follows from immediate algebraic manipulations. By assumption one has $\widehat{\underline{M}}\widehat{\underline{A}}\widehat{S} = \widehat{S}\widehat{\Lambda}$ and the fact that $\widehat{S}^{\mathsf{T}}\underline{H}_j\Gamma_b\underline{H}_j^{\mathsf{T}}\widehat{\underline{M}}^{-1}\widehat{S} = I_k$. □

5.3 Randomized spectral limited memory preconditioners

Let us now propose randomized variants for the LMP presented in the previous section. Since exact eigeninformation is generally out of reach from a computational viewpoint, the general idea is to construct the spectral variants with approximate eigenpairs instead of exact ones. It is important to highlight that doing so, we loose the properties of the exact spectral LMP as described in [47, Section 3]. However, it turns out that using the spectral LMP formula with approximate eigenpairs has proven to perform well in the context of nonlinear regressions [89]. In this thesis, the approximations are obtained using variants of the randomized algorithms introduced in Chapter 4 that are adapted to the structure of the problem.

5.3.1 A class of randomized spectral limited memory preconditioners for the inverse-free preconditioned conjugate gradient method

According to Proposition 5.2, we know that the spectral LMP for the PCGIF considers eigenpairs associated to the eigenvalue problem

$$\bar{M}\bar{A}\,v = \lambda\,v,$$

where $\bar{M}$ and $\bar{A}$ are two Γ_b-symmetric matrices. This corresponds to the GEP in initial form introduced in Section 4.3.2 and consequently, we can use Algorithms 4.1 and 4.2 of Chapter 4. Having applications in data assimilation in mind, viable algorithms should be parsimonious in the applications of the different matrices. In this regard, we focus on Algorithm 4.2 of Chapter 4 with $q = 1$ (see Table 4.1). The implementation of Algorithm 4.2 cannot be straightforwardly used for several reasons. First, the particular structure of the matrices can be exploited to avoid useless applications of operators. Then, Algorithm 4.2 is designed to provide the approximate eigenvevectors V and eigenvalues Λ only. However, we have already mentioned that to construct a spectral LMP as in (5.6), it is also required to compute $\Gamma_b V$. Consequently, the construction of the reduced eigenvalue problem, at the root of Algoritihm 4.2 must be done differently to satisfy this constraint. The resulting algorithm is presented in Algorithm 5.4.

Let us now present an alternative approach in the case $\bar{M} = I_n$. We recall that this is equivalent to preconditioning the PCG with Γ_b. This situation occurs for instance at the first Gauss-Newton iteration, where no other preconditioner than Γ_b is available. In this case, the eigenvalue problem of interest becomes

$$\bar{A}\,v = \lambda\,v, \tag{5.16}$$

where we recall that $\bar{A} = A\Gamma_b = I_n + H_j^{\mathsf{T}}\Gamma_o^{-1}H_j\Gamma_b$. It is then possible to further exploit the structure of the operator, and to rather consider

$$H_j^{\mathsf{T}}\Gamma_o^{-1}H_j\Gamma_b\,v = \lambda'\,v. \tag{5.17}$$

This strategy has already been considered in Chapter 4, in particular for the numerical experiments related to the 3D-Var test problems. The eigenvectors are identical, and the eigenvalues

Algorithm 5.4: Construction of a randomized spectral LMP for the PCGIF in the general case.

Input: Matrices $\Gamma_b, H_j, \Gamma_o^{-1}, H_j^\mathsf{T}$, preconditioner $\bar{M}$ which is Γ_b-symmetric, number of random samples $1 \leq p \leq m$, number of approximate eigenpairs $1 \leq k \leq p$ to provide.

1 Draw a random matrix $\Omega \in \mathbb{R}^{n\times p}$
2 Compute $V = \Gamma_b\Omega \in \mathbb{R}^{n\times p}$
3 Compute $X = \Omega + H_j^\mathsf{T}\Gamma_o^{-1}H_jV \in \mathbb{R}^{n\times p}$
4 Perform the thin QR factorization $X = QR$ and set $X = Q$
5 Form $T = R^{-\mathsf{T}}V^\mathsf{T}X \in \mathbb{R}^{p\times p}$
6 Compute $V = \bar{M}X \in \mathbb{R}^{n\times p}$ and $Z = \Gamma_bV \in \mathbb{R}^{n\times p}$
7 Form $\Phi = Z^\mathsf{T}X \in \mathbb{R}^{p\times p}$
8 Solve the generalized Hermitian eigenvalue problem $TW = \Phi W\Theta$ with $W \in \mathbb{R}^{p\times p}$ a Φ-orthogonal matrix and $\Theta \in \mathbb{R}^{p\times p}$ a diagonal matrix with the eigenvalues sorted in **increasing** order
9 Remove the last $p-k$ columns of W and Θ
10 Remove the last $p-k$ rows of Θ
11 Set $V = VW \in \mathbb{R}^{n\times k}$, $Z = ZW \in \mathbb{R}^{n\times k}$ and $\Lambda = \Theta^{-1}$

Output: Matrices $V, Z \in \mathbb{R}^{n\times k}$ and $\Lambda \in \mathbb{R}^{k\times k}$ such that $\bar{M}(I_n + H_j^\mathsf{T}\Gamma_o^{-1}H_j\Gamma_b)V \approx V\Lambda$ and $Z = \Gamma_bV$ with $V^\mathsf{T}\Gamma_b\bar{M}^{-1}V = I_k$ and Λ diagonal.

are simply shifted by one. This new formulation offers several advantages. First, given the theoretical bounds obtained in Theorem 4.9 and 4.10, we know that the performance of the randomized methods depend on the tail of the eigenvalue distribution. Here, $H_j^\mathsf{T}\Gamma_o^{-1}H_j\Gamma_b$ is a matrix of rank $m \leq n$. Consequently, $n-m$ of its eigenvalues are zero, meaning that the tail of $H_j^\mathsf{T}\Gamma_o^{-1}H_j\Gamma_b$ is much smaller than the one of $I_n + H_j^\mathsf{T}\Gamma_o^{-1}H\Gamma_b$. Subtracting the identity basically turns the cluster of eigenvalues at 1 into a cluster of same size at 0.

Here, (5.17) corresponds to the GEP with basis transformation (4.2), meaning that we can rather consider using Algorithms 4.3 and 4.4. Again, we focus on the variant that only requires one application of $H_j^\mathsf{T}\Gamma_o^{-1}H_j$, that is, Algorithm 4.4 with $q = 1$. An implementation in this current context is proposed in Algorithm 5.5.

In terms of computational cost, Algorithms 5.4 and 5.5 require to apply $H_j\Gamma_o^{-1}H_j^\mathsf{T}$ to a block of p vectors (step 3) only once. This step can strongly benefit from parallelization, as all the vectors are available simultaneously. Similarly, two applications of Γ_b to a block of p vectors are performed in steps 2 and 6 and could enjoy the same benefit. The thin QR factorization can be done using specific methods for tall and skinny matrices. The QR factorization can be performed using efficient parallel methods adapted to tall and skinny matrices of very large size [82]. For the memory requirements, we point out that three blocks of p vectors must be kept to perform Algorithm 5.4 and two for Algorithm 5.5. The computational costs and memory requirements of both algorithms are presented in Table 5.1. They are perfectly equivalent in terms of number of applications of the operators, except of course for $\bar{M}$.

5.3.2 A class of randomized spectral limited memory preconditioners for the restricted and augmented restricted preconditioned conjugate gradient method

We follow the same idea presented in Section 5.3.1. We intend to use randomized methods to approximate eigeninformation, that will be used to construct an approximate spectral LMP as

Algorithm 5.5: Construction of a randomized spectral LMP for the PCGIF in the case $\bar{M} = I_n$.

Input: Matrices $\Gamma_b, H_j, \Gamma_o^{-1}, H_j^\mathsf{T}$, number of random samples $1 \leq p \leq m$, number of approximate eigenpairs $1 \leq k \leq p$ to provide.

1 Draw a random matrix $\Omega \in \mathbb{R}^{n\times p}$
2 Compute $V = \Gamma_b \Omega \in \mathbb{R}^{n\times p}$
3 Compute $X = H_j^\mathsf{T} \Gamma_o^{-1} H_j V \in \mathbb{R}^{n\times p}$
4 Perform the thin QR factorization $X = QR$ and set $X = Q$
5 Form $T = R^{-\mathsf{T}} V^\mathsf{T} X \in \mathbb{R}^{p\times p}$
6 Compute $Z = \Gamma_b X \in \mathbb{R}^{n\times p}$
7 Form $\Phi = Z^\mathsf{T} X \in \mathbb{R}^{p\times p}$
8 Solve the generalized Hermitian eigenvalue problem $TW = \Phi W \Theta$ with $W \in \mathbb{R}^{p\times p}$ a Φ-orthogonal matrix and $\Theta \in \mathbb{R}^{p\times p}$ a diagonal matrix with the eigenvalues sorted in **increasing** order
9 Remove the last $p-k$ columns of W and Θ
10 Remove the last $p-k$ rows of Θ
11 Set $V = XW \in \mathbb{R}^{n\times k}$, $Z = ZW \in \mathbb{R}^{n\times k}$ and $\Lambda = \Theta^{-1} + I_k$

Output: Matrices $V, Z \in \mathbb{R}^{n\times k}$ and $\Lambda \in \mathbb{R}^{k\times k}$ such that $(I_n + H_j^\mathsf{T} \Gamma_o^{-1} H_j \Gamma_b) V \approx V\Lambda$ and $Z = \Gamma_b V$ with $V^\mathsf{T} \Gamma_b V = I_k$ and Λ diagonal.

	H_j	Γ_o^{-1}	H_j^T	Γ_b	$\bar{M}$	Storage
Algorithm 5.4	1	1	1	2	1	$3np$
Algorithm 5.5	1	1	1	2	-	$2np$

Table 5.1: Number of applications of the matrices to a block of p vectors required for Algorithms 5.4 and 5.5, along with the memory requirements.

in (5.13) and (5.15). Here we focus on constructing a randomized spectral LMP for the RPCG. Indeed, since we have defined the LMP for the Augmented PCG by analogy with the one for the RPCG, it is clear that transposing the arguments will be straightforward.

According to Proposition 5.5, constructing an approximate spectral LMP for the RPCG requires to compute approximate solutions of the eigenvalue problem

$$\widehat{M}\widehat{A}\, v = \lambda\, v, \tag{5.18}$$

where $\widehat{M}$ and $\widehat{A}$ are $H_j \Gamma_b H_j^\mathsf{T}$-symmetric matrices and $\widehat{A} = I_m + \Gamma_o^{-1} H_j \Gamma_b H_j^\mathsf{T}$. Again, we adapt Algorithms 4.2 with $q = 1$ of Section 4.3.2. Given the corresponding computational cost detailed in Table 4.1, it is clear that we cannot avoid the curse of applying $H_j \Gamma_b H_j^\mathsf{T}$ to a block of p vectors twice. The first one occurs when applying $\widehat{A}$, and the second one when applying the inner product matrix. Therefore, using parallel methods to compute both series of matrix-vector products seems even more critical for the dual approach. We will notice in Section 5.3.3 that there is an alternative to partly mitigate this extra cost. Algorithm 5.6 presents an adapted version of Algorithm 4.2 to address the eigenvalue problem (5.18).

Likewise the inverse-free approach, if $\widehat{M} = I_m$, then one may rather study the eigenvalue problem

$$\Gamma_o^{-1} H_j \Gamma_b H_j^\mathsf{T}\, v = \lambda'\, v,$$

which we interpret as the eigenvalues of the GEP with basis transformation. In this case, algorithms from Section 4.3.3 can be applied. Following Table 4.1, Algorithm 4.4 with $q = 1$ is less

Algorithm 5.6: Construction of a randomized spectral LMP for the RPCG in the general case.

Input: Matrices $\Gamma_b, H_j, \Gamma_o^{-1}, H_j^\mathsf{T}$, preconditioner $\widehat{M}$ which is $H_j\Gamma_b H_j^\mathsf{T}$-symmetric, number of random samples $1 \leq p \leq m$, number of approximate eigenpairs $1 \leq k \leq p$ to provide.

1 Draw a random matrix $\Omega \in \mathbb{R}^{m\times p}$
2 Compute $V = H_j\Gamma_b H_j^\mathsf{T}\Omega \in \mathbb{R}^{m\times p}$
3 Compute $X = \Omega + \Gamma_o^{-1}V \in \mathbb{R}^{m\times p}$
4 Perform the thin QR factorization $X = QR$ and set $X = Q$
5 Form $T = R^{-\mathsf{T}}V^\mathsf{T}X \in \mathbb{R}^{p\times p}$
6 Compute $V = \widehat{M}X \in \mathbb{R}^{m\times p}$ and $Z = H_j\Gamma_b H_j^\mathsf{T}V \in \mathbb{R}^{m\times p}$
7 Form $\Phi = Z^\mathsf{T}X \in \mathbb{R}^{p\times p}$
8 Solve the generalized Hermitian eigenvalue problem $TW = \Phi W\Theta$ with $W \in \mathbb{R}^{p\times p}$ a Φ-orthogonal matrix and $\Theta \in \mathbb{R}^{p\times p}$ a diagonal matrix with the eigenvalues sorted in **increasing** order
9 Remove the last $p-k$ columns of W and Θ
10 Remove the last $p-k$ rows of Θ
11 Set $V = VW \in \mathbb{R}^{m\times k}$, $Z = ZW \in \mathbb{R}^{m\times k}$ and $\Lambda = \Theta^{-1}$

Output: Matrices $V, Z \in \mathbb{R}^{m\times k}$ and $\Lambda \in \mathbb{R}^{k\times k}$ such that $\widehat{M}(I_m + \Gamma_o^{-1}H_j\Gamma_b H_j^\mathsf{T})V \approx V\Lambda$ and $Z = H_j\Gamma_b H_j^\mathsf{T}V$ with $V^\mathsf{T}H_j\Gamma_b H_j^\mathsf{T}\widehat{M}^{-1}V = I_k$ and Λ diagonal.

ressource consuming and an implementation adapted to the context is given in Algorithm 5.7. The computational costs are detailed in Table 5.2.

Here, we note that Algorithms 5.6 and 5.7 require two applications of $H_j\Gamma_b H_j^\mathsf{T}$ to a block of p vectors (steps 2 and 6), and one with Γ_o^{-1} (step 3). Consequently, H_j and H_j^T are used twice in Algorithms 5.6 and 5.7 than in Algorithms 5.4 and 5.5. In data assimilation, applying those operators is more expensive than applying Γ_b and Γ_o^{-1}. This implies that the construction of randomized spectral LMP for the RPCG is more costly than for the PCGIF. However, in terms of memory requirements, Algorithms 5.6 and 5.7 require the storage of three and two blocks of p vectors of size m respectively. Consequently, if $m \ll n$, then the storage can be significantly less critical. The computational costs and memory requirements are summarized in Table 5.2.

For the Augmented RPCG, two algorithms can also be obtained following the same arguments. It only requires to replace H_j, H_j^T and R by their augmented counterparts defined in (5.12), and to draw a random matrix of size $(m+1)\times p$ instead of $m \times p$. Hence, it seemed not necessary to provide two additional algorithms.

5.3.3 Equivalence between the primal and dual approaches

For both the PCGIF and the RPCG, we have highlighted conditions on the initial residuals and on the preconditioners ((5.4) and (5.10) respectively) under which applying the PCGIF (resp. the RPCG) produces mathematically equivalent iterates as the ones obtained when applying the PCG. The condition on the initial residuals has already been settled, and we have also shown conditions on the different LMPs ((5.7) and (5.14) respectively) so that the condition on the preconditioners is satisfied. Our objective in this section is to determine under which condition the randomized spectral LMPs proposed for the inverse-free and dual approach satisfy the conditions given in (5.4) and (5.10).

Let us first relate the inverse-free and the dual approach directly to one another. Let C_{sp} be

Algorithm 5.7: Construction of a randomized spectral LMP for the RPCG in the case $\widehat{M} = I_m$.

Input: Matrices $\Gamma_b, H_j, \Gamma_o^{-1}, H_j^\mathsf{T}$, number of random samples $1 \leq p \leq m$, number of approximate eigenpairs $1 \leq k \leq p$ to provide.

1 Draw a random matrix $\Omega \in \mathbb{R}^{m\times p}$
2 Compute $Z = H\Gamma_b H_j^\mathsf{T}\Omega \in \mathbb{R}^{m\times p}$
3 Compute $V = \Gamma_o^{-1} Z \in \mathbb{R}^{m\times p}$
4 Perform the thin QR factorization $V = QR$ and set $V = Q$
5 Form $T = R^{-\mathsf{T}} Z^\mathsf{T} V$
6 Compute $Z = H\Gamma_b H_j^\mathsf{T} V \in \mathbb{R}^{m\times p}$
7 Form $\Phi = Z^\mathsf{T} V$
8 Solve the generalized Hermitian eigenvalue problem $TW = \Phi W \Theta$ with $W \in \mathbb{R}^{p\times p}$ a Φ-orthogonal matrix and $\Theta \in \mathbb{R}^{p\times p}$ a diagonal matrix with the eigenvalues sorted in **increasing** order
9 Remove the last $p-k$ columns of W and Θ
10 Remove the last $p-k$ rows of Θ
11 Set $V = VW \in \mathbb{R}^{m\times k}$, $Z = ZW \in \mathbb{R}^{m\times k}$ and $\Lambda = \Theta^{-1} + I_k$

Output: Matrices $V, Z \in \mathbb{R}^{m\times k}$ and $\Lambda \in \mathbb{R}^{k\times k}$ such that $(I_m + \Gamma_o^{-1} H_j \Gamma_b H_j^\mathsf{T})V \approx V\Lambda$ and $Z = H_j\Gamma_b H_j^\mathsf{T} V$ with $V^\mathsf{T} H_j \Gamma_b H_j^\mathsf{T} V = I_k$ and Λ diagonal.

	H_j	Γ_o^{-1}	H_j^T	Γ_b	$\widehat{M}$	Storage
Algorithm 5.6	2	1	2	2	1	$3mp$
Algorithm 5.7	2	1	2	2	-	$2mp$

Table 5.2: Number of applications of the matrices to a block of p vectors required for Algorithms 5.6 and 5.7, along with the memory requirements.

as in Proposition 5.2 and D_{sp} as in Proposition 5.5. Combining (5.4) and (5.10) implies that preconditioning PCGIF with C_{sp} and RPCG with D_{sp} produces equivalent iterates if they satisfy

$$C_{\text{sp}} H_j^\mathsf{T} = H_j^\mathsf{T} D_{\text{sp}}.$$

From relations (5.7) and (5.14), we obtain that this is satisfied when

$$\begin{cases} \bar{S} = H_j^\mathsf{T}\widehat{S} \\ \bar{\Lambda} = \widehat{\Lambda} \\ \bar{M} H_j^\mathsf{T} = H_j^\mathsf{T}\widehat{M} \end{cases}, \tag{5.19}$$

Indeed, in this case, one has

$$\begin{aligned} C_{\text{sp}} H_j^\mathsf{T} &= \bar{M} H_j^\mathsf{T} + \bar{S}\left(\bar{\Lambda}^{-1} - I_k\right)\bar{S}^\mathsf{T}\Gamma_b H_j^\mathsf{T} \\ &= H_j^\mathsf{T}\widehat{M} + H_j^\mathsf{T}\widehat{S}\left(\bar{\Lambda}^{-1} - I_k\right)\widehat{S}^\mathsf{T} H_j \Gamma_b H_j^\mathsf{T} \\ &= H_j^\mathsf{T}[I_m + \widehat{S}\left(\bar{\Lambda}^{-1} - I_k\right)\widehat{S}^\mathsf{T} H_j \Gamma_b H_j^\mathsf{T}] = H_j^\mathsf{T} D_{\text{sp}}. \end{aligned}$$

The question now is under which conditions the relations (5.19) are satisfied when C_{sp} and D_{sp} are no longer exact spectral LMP, but randomized spectral LMPs constructed in Algorithms 5.4

and 5.6 respectively. Since $\bar{M}$ and $\widehat{M}$ are parameters of Algorithm 5.4 and Algorithm 5.6 respectively, we are going to assume that they do satisfy (5.19). We state the compatibility condition in Proposition 5.9.

Proposition 5.9. *Let C_{rand} denote the spectral LMP introduced in Proposition 5.2 constructed using approximate eigenpairs provided by Algorithm 5.4 with $\bar{M}$ and $\bar{\Omega} \in \mathbb{R}^{n\times p}$. Let D_{rand} be the spectral LMP introduced in Proposition 5.5 constructed using the approximate eigenpairs provided by Algorithm 5.6 with $\widehat{M}$ and $\widehat{\Omega} \in \mathbb{R}^{m\times p}$. Then if $\bar{M}H_j^\mathsf{T} = H_j^\mathsf{T}\widehat{M}$ and $\bar{\Omega} = H_j^\mathsf{T}\widehat{\Omega}$, then C_{rand} and D_{rand} satisfy*

$$C_{rand}H_j^\mathsf{T} = H_j^\mathsf{T} D_{rand}.$$

Proof. The proof is rather technical and requires to go back to the theoretical derivation of the algorithms presented in Section 4.3. Let $\bar{V}, \bar{Z} \in \mathbb{R}^{n\times k}$ and $\bar{\Lambda} \in \mathbb{R}^{k\times k}$ denote the outputs of Algorithm 5.4, and $\widehat{V}, \widehat{Z} \in \mathbb{R}^{m\times k}$ and $\widehat{\Lambda} \in \mathbb{R}^{k\times k}$ denote the outputs of Algorithm 5.6. One therefore has

$$\begin{aligned} C_{\text{rand}} &= \bar{M} + \bar{V}\left(\bar{\Lambda}^{-1} - I_k\right)\bar{Z}^\mathsf{T} \\ \text{and} \quad D_{\text{rand}} &= \widehat{M} + \widehat{V}\left(\widehat{\Lambda}^{-1} - I_k\right)\widehat{Z}^\mathsf{T}. \end{aligned}$$

Since by assumption $\bar{M}H_j^\mathsf{T} = H_j^\mathsf{T}\widehat{M}$, following (5.19) it only remains to verify that $\bar{V} = H_j^\mathsf{T}\widehat{V}$ and $\bar{\Lambda} = \widehat{\Lambda}$.

Let us denote $\bar{A} = I_n + H_j^\mathsf{T}\Gamma_o^{-1}H_j\Gamma_b$ and $\widehat{A} = I_m + \Gamma_o^{-1}H_j\Gamma_b H_j^\mathsf{T}$. Let us focus on $\bar{V}$, the arguments will be analogous for $\widehat{V}$. Algorithm 5.4 is a particular case of Algorithm 4.2 with $q = 1$. Consequently, one has $\bar{V} = \bar{M}\bar{A}\bar{\Omega}\bar{W}$ with $\bar{W} \in \mathbb{R}^{p\times k}$ and $\bar{\Lambda}$ containing eigenvectors and eigenvalues of the reduced eigenvalue problem given whose expression can be obtained from (4.11). In our case, its expression reads

$$\bar{\Omega}^\mathsf{T}\Gamma_b\bar{A}\bar{\Omega}\,\bar{w} = \bar{\lambda}\,\bar{\Omega}^\mathsf{T}\Gamma_b\bar{A}\bar{M}\bar{A}\bar{\Omega}\,\bar{w}. \tag{5.20}$$

Similarly, one has $\widehat{V} = \widehat{M}\widehat{A}\widehat{\Omega}\widehat{W}$ with $\widehat{W} \in \mathbb{R}^{p\times k}$ and $\widehat{\Lambda}$ containing eigenvectors and eigenvalues of the reduced eigenvalue problem

$$\widehat{\Omega}^\mathsf{T}H_j\Gamma_b H_j^\mathsf{T}\widehat{A}\widehat{\Omega}\,\widehat{w} = \widehat{\lambda}\,\widehat{\Omega}^\mathsf{T}H_j\Gamma_b H_j^\mathsf{T}\widehat{A}\widehat{M}\widehat{A}\widehat{\Omega}\,\widehat{w}. \tag{5.21}$$

Now, since one has $\bar{\Omega} = H_j^\mathsf{T}\widehat{\Omega}$ by assumption, recalling that $\bar{A}H_j^\mathsf{T} = H_j^\mathsf{T}\widehat{A}$ we obtain

$$\bar{\Omega}^\mathsf{T}\Gamma_b\bar{A}\bar{\Omega} = \widehat{\Omega}^\mathsf{T}H_j\Gamma_b\bar{A}H_j^\mathsf{T}\widehat{\Omega} = \widehat{\Omega}^\mathsf{T}H_j\Gamma_b H_j^\mathsf{T}\widehat{A}\widehat{\Omega}.$$

For analogous reasons we have

$$\bar{\Omega}^\mathsf{T}\Gamma_b\bar{A}\bar{M}\bar{A}\bar{\Omega} = \widehat{\Omega}^\mathsf{T}H_j\Gamma_b H_j^\mathsf{T}\widehat{A}\widehat{M}\widehat{A}\widehat{\Omega}.$$

Therefore, the reduced eigenvalue problems (5.20) and (5.21) are identical. This implies that $\bar{W} = \widehat{W}$ and $\bar{\Lambda} = \widehat{\Lambda}$. Accordingly, one has

$$\bar{V} = \bar{M}\bar{A}\bar{\Omega}\bar{W} = H_j^\mathsf{T}\widehat{M}\widehat{A}\widehat{\Omega}\widehat{W} = H_j^\mathsf{T}\widehat{V},$$

and the result. □

Similarly, we can derive an equivalence condition relating the randomized spectral LMP obtained with Algorithms 5.5 and 5.7. It is stated in Proposition 5.10. We note that in this case, one trivially has $\bar{M}H_j^\mathsf{T} = H_j^\mathsf{T}\widehat{M}$ since $\bar{M} = I_n$ and $\widehat{M} = I_m$.

Proposition 5.10. *Let C_{rand} be the randomized spectral LMP as introduced in Proposition 5.2 constructed using the approximations provided by Algorithm 5.5 with $\bar{\Omega} \in \mathbb{R}^{n\times p}$. Let D_{rand} be the randomized spectral LMP as introduced in Proposition 5.5 constructed using the approximations provided by Algorithm 5.7 and $\widehat{\Omega} \in \mathbb{R}^{m\times p}$. Then if $\bar{\Omega} = H_j^\mathsf{T}\widehat{\Omega}$, C_{rand} and D_{rand} satisfy*

$$C_{rand}H_j^\mathsf{T} = H_j^\mathsf{T}D_{rand}.$$

Proof. The proof follows similar arguments as in Proposition 5.9. □

An interesting consequence can be drawn from Proposition 5.9. As mentioned above, Algorithms 5.6 and 5.7 require two additional applications of H_j^T and H_j to a block of p vectors compared to Algorithms 5.4 and 5.5, which can be costly. According to Proposition 5.9, applying the PCGIF with approximate spectral LMP C_{rand} constructed with matrices of the form $H_j^\mathsf{T}\Omega$ with $\Omega \in \mathbb{R}^{m\times k}$ would yield mathematically equivalent iterates. Although we apply a primal method, that is in the n-dimensional space, it would only require one extra application of H_j^T, thus saving one application of H.

5.4 Application to variational data assimilation

In this section, we propose an application to variational data assimilation. The variational formulation of data assimilation problems takes the form (2.16), and is generally solved using the Gauss-Newton method. We split the numerical experiments in two distinct parts. First, we study on the same instructional 3D-Var problem as in Chapter 4 the eigenvalue distribution of the matrices when using the proposed randomized spectral LMPs as preconditioners. This is intended to illustrate the performance of the randomized spectral LMPs, and in particular to compare the randomized variants to the exact spectral LMP. Then, we propose an application to the Lorentz 95 model, which is a classical 4D-Var problem often used as a benchmark problem. In this second part, we focus on the performance of the randomized preconditioners in terms of improvements in the convergence of the Gauss-Newton method.

5.4.1 Eigenvalue distribution of the preconditioned matrix

In this section, we use the same 3D-Var test matrix $A_{\text{3D-Var}}$ as in Chapter 4, whose expression reads

$$A_{\text{3D-Var}} = \Gamma_b^{-1} + H_j^\mathsf{T}\Gamma_o^{-1}H_j,$$

which we have considered in two different settings, denoted by `LowObs` and `HighObs`. Here, we drop the subscript for convenience. The corresponding matrices associated with the PCGIF and RPCG are denoted by $\bar{A} = I_n + H_j^\mathsf{T}\Gamma_o^{-1}H_j\Gamma_b$ and $\widehat{A} = I_m + \Gamma_o^{-1}H_j\Gamma_bH_j^\mathsf{T}$, respectively. In this section, we focus on the case $\bar{M} = I_n$ and $\widehat{M} = I_m$. We do not investigate the performance of the randomized spectral LMP for the Augmented RPCG since there is no augmentation to perform.

Let C_{rand} denote a randomized spectral LMP constructed using either Algorithm 5.4 or 5.5, and C_{sp} denote the spectral LMP as in Proposition 5.2 constructed using exact dominant eigenpairs of $\bar{A}$. Similarly, let D_{rand} denote a randomized spectral LMP constructed using either Algorithm 5.6 or 5.7, and D_{sp} denote the spectral LMP as in Proposition 5.5 constructed using exact dominant eigenpairs of $\widehat{A}$.

Our objective in this section is to investigate the eigenvalue distribution of $C_{\text{rand}}\bar{A}$ and $D_{\text{rand}}\widehat{A}$ compared to the one of $C_{\text{sp}}\bar{A}$ and $D_{\text{sp}}\widehat{A}$. The latter will thus serve as a reference, since our randomized approaches aim at substituting them. We will thus observe how the accuracy of the approximate eigenpairs investigated in Section 4.5.2 impacts the eigenvalue distribution of the preconditioned matrices.

Inverse-free primal space approach

Let us first present the results for the inverse-free approach. To evaluate the performance of the randomized spectral LMP C_{rand} compared to the exact spectral LMP C_{sp}, we compute

$$\Delta_j^{\lambda}(C_{\text{rand}}) = \frac{|\lambda_j(C_{\text{sp}}\bar{A}) - \lambda_j(C_{\text{rand}}\bar{A})|}{\lambda_j(C_{\text{sp}}\bar{A})}, \quad 1 \leq j \leq n. \tag{5.22}$$

To account for the randomness, we perform a statistical analysis on (5.22). We apply the randomized algorithms 100 times, with independent draws of standard Gaussian matrices $\Omega_i \in \mathbb{R}^{n\times p}$, from which we obtain independent randomized spectral LMPs $C_{\text{rand},i}$. In particular, we focus on the empirical mean and standard deviation of $\Delta_j^{\lambda}(C_{\text{rand}})$.

We denote `C_General` the randomized LMP constructed using Algorithm 5.4 and `C_M=In` the one constructed using Algorithm 5.5. Since two-level preconditioners are expected to improve over the first-level preconditioner, it seems relevant to also study $\Delta_j^{\lambda}(I_n)$, that we denote by `No_LMP`. The methods are applied with $k = 20$, and $p = 40$ and 60, that is an oversampling of 20 and 40 respectively. Results are presented in Figure 5.1.

We remark that the difference between `C_General` and `C_M=In` is only noticeable for small eigenvalues, and in the `HighObs` setting. To decrease the condition number, which is related to approximating the largest eigenvalues, then they should perform equally. The effect of the oversampling is particularly visible, and allows us to gain approximately one order of magnitude for the dominant eigenvalues. When $p - k = 40$ one has $\Delta_j^{\lambda} \leq 10^{-2}$, which means that the condition number obtained with the randomized spectral LMP should be very close to the one obtained with C_{sp}.

Dual space approach

For the dual approach, we consider the quantity of interest

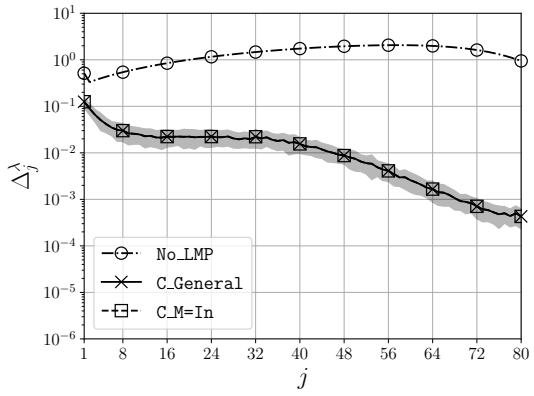
$$\Delta_j^{\lambda}(D_{\text{rand}}) = \frac{|\lambda_j(D_{\text{sp}}\widehat{A}) - \lambda_j(D_{\text{rand}}\widehat{A})|}{\lambda_j(D_{\text{sp}}\widehat{A})}, \quad 1 \leq j \leq n. \tag{5.23}$$

We perform the same statistical analysis as in the previous section, to compute the empirical mean and standard deviation of (5.23). We denote by `D_General` the randomized LMP constructed using Algorithm 5.6 and by `D_M=In` the one constructed using Algorithm 5.7. We denote by `No_LMP` the quantity $\Delta_j^{\lambda}(I_n)$. The algorithms are again applied with $k = 20$, and $p = 40$ and 60. Results are presented in Figure 5.2.

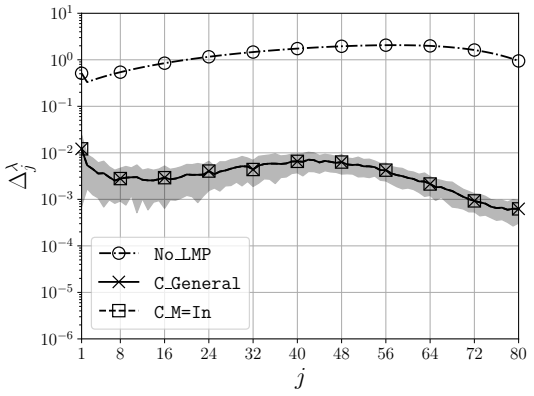
The obtained results are very similar to the ones of Figure 5.1. Consequently, we can expect the randomized spectral LMP for the RPCG to perform well too. By extension, the one related to the Augmented RPCG may have similar characteristics.

5.4.2 A 4D-Var application: The Lorenz 95 model

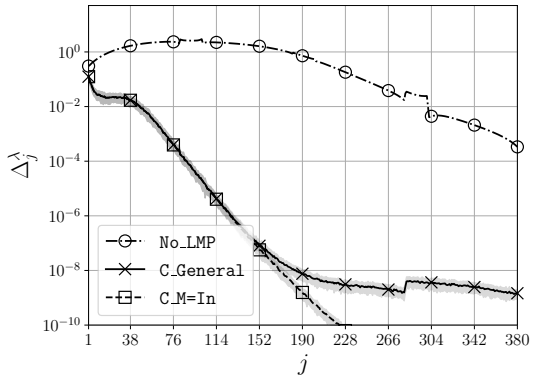
We study the effect of the proposed randomized spectral LMP in terms of convergence for the Gauss-Newton method within a strong-constraint 4D-Var data assimilation problem. We rapidly introduce the mathematical material, although it is very similar to 3D-Var, there exist differences that are worth being highlighted. Let $x_0 \in \mathbb{R}^n$ denote the state vector of the system at the time t_0. Prior information on the true state is gathered in the background vector, denoted by $x_c \in \mathbb{R}^n$. This background state is not known with certainty, and a standard way to model the uncertainty is to consider that the background state variables are noisy with a Gaussian noise with zero mean and covariance matrix $\Gamma_b \in \mathbb{R}^{s\times s}$, i.e. $x_0 \sim \mathcal{N}(x_c, \Gamma_b)$. Here we consider that we have a series of observations denoted by $y_0, \dots y_N \in \mathbb{R}^m$ made at the different times $t_0, \dots, t_N$. The observations are related to the system state via the observation operator, $\mathcal{H} : \mathbb{R}^s \times \mathbb{R} \to \mathbb{R}^m$,



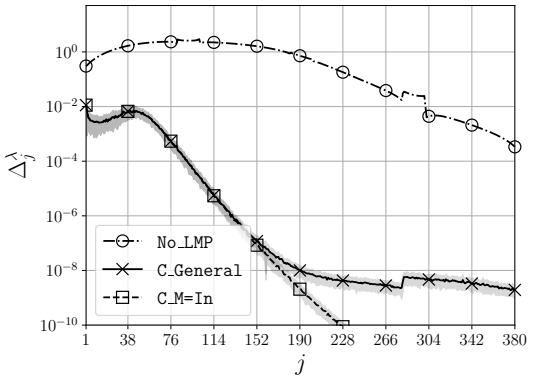
(a) $p - k = 20$ on `LowObs`.



(b) $p - k = 40$ on `LowObs`.



(c) $p - k = 20$ on `HighObs`.



(d) $p - k = 40$ on `HighObs`.

Figure 5.1: Relative distance (5.22) with respect to the eigenvalues of $C_{\text{sp}}\bar{A}$ for the `LowObs` (top) and `HighObs` (bottom) settings. Randomized spectral LMPs are built using Algorithms 5.4 or 5.5 applied with $k = 20$ and $p = 40$ (left) and $p = 60$ (right).

which is often nonlinear. Given the true initial state x_0, the state at the different time steps is then computed using a model operator denoted by $\mathcal{M}$, such that $x_{i+1} = \mathcal{M}(x_i)$. In the strong-constraint 4D-Var variant, we consider that the model is perfect, meaning that we do not account for model errors. Finally, we assume that all the observations are noisy, with Gaussian noise associated to the covariance matrix $R_i \in \mathbb{R}^{m\times m}$, i.e. for $0 \leq i \leq N$ one has $y_i = \mathcal{H}(x_i, t_i) + \xi_i$ with $\xi_i \sim \mathcal{N}(\mathbf{0}, R_i)$. Here we note that we have implictly considered that the observation errors are assumed uncorrelated in time.

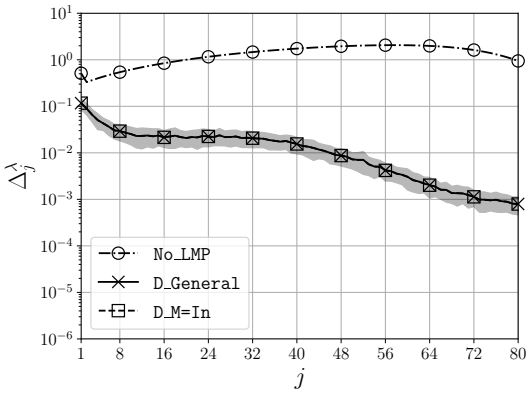
Estimating the true initial state x_0 can then be done by minimizing the following functional

$$J(x_0) = \frac{1}{2}\|x_0 - x_c\|^2_{\Gamma_b^{-1}} + \frac{1}{2}\sum_{i=0}^{N} \|y_i - \mathcal{H}(x_i, t_i)\|^2_{R_i^{-1}} \tag{5.24}$$

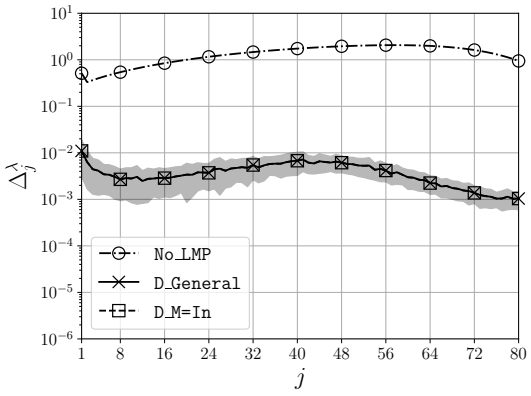
subject to $x_{i+1} = \mathcal{M}(x_i)$. Let $H_i \in \mathbb{R}^{m\times n}$ and $M_i \in \mathbb{R}^{n\times n}$ be the linearized of $\mathcal{H}$ and $\mathcal{M}$ around (x_i, t_i) respectively. For a given Gauss-Newton iterate p_j, the new iterate p_{j+1} is computed as $p_{j+1} = p_j + s_j$ with the increment s_j being the minimizer of the quadratic cost function

$$\widetilde{J}(s_j) = \frac{1}{2}\|HL^{-1}s_j - d_j\|^2_{\Gamma_o^{-1}} + \frac{1}{2}\|s_j + p_j - x_c\|^2_{\Gamma_b^{-1}}, \tag{5.25}$$

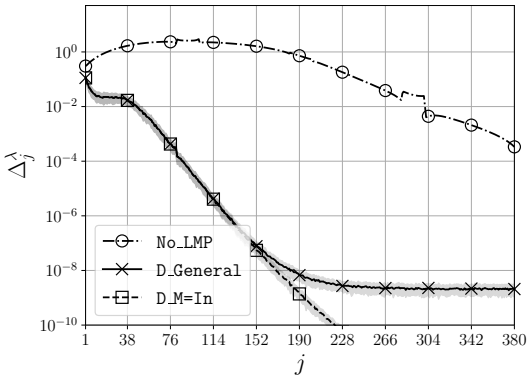
where $R = \text{diag}(R_0, \ldots, R_N)$. The matrix H is such that $H = \text{diag}(H_0, \ldots, H_N) \in \mathbb{R}^{(N+1)m\times(N+1)n}$



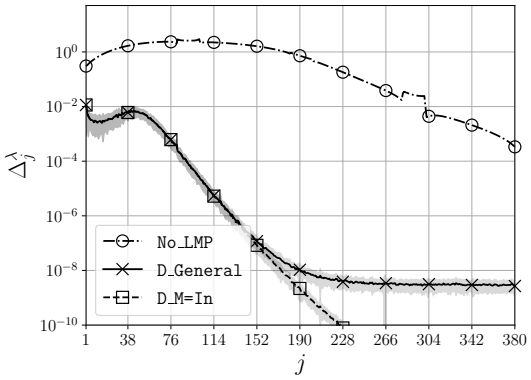
(a) $p - k = 20$ on `LowObs`.



(b) $p - k = 40$ on `LowObs`.



(c) $p - k = 20$ on `HighObs`.



(d) $p - k = 40$ on `HighObs`.

Figure 5.2: Relative distance (5.23) with respect to the eigenvalues of $D_{\text{sp}}\widehat{A}$ for the `LowObs` (top) and `HighObs` (bottom) settings. Randomized spectral LMPs are built using Algorithms 5.6 or 5.7 applied with $k = 20$ and $p = 40$ (left) and $p = 60$ (right).

and the matrix $L \in \mathbb{R}^{(N+1)n \times (N+1)n}$ is defined as

$$L^{-1} = \begin{bmatrix} I_n & & & \\ M_0 & I_n & & \\ & \ddots & \ddots & \\ & & M_{N-1} & I_n \end{bmatrix}.$$

Lorenz 95 model

In the Lorenz 95 model, the evolution of the state vector $x \in \mathbb{R}^n$ components, denoted by $X_1, \ldots, X_n$ is governed by a set of n coupled ordinary differential equations

$$\frac{dX_l}{dt} = -X_{l-2}X_{l-1} + X_{l-1}X_{l+1} - X_l + F, \quad 1 \leq l \leq n$$

where we impose periodic boundary conditions, namely $X_{-1} = X_{n-1}$, $X_0 = X_n$ and $X_{n+1} = X_1$. The constant F is a parameter of the problem which is generally set to $F = 8$. The equations are integrated using a fourth-order Runge-Kutta scheme [20, Chapter 3], with time step of 0.025. The matrix Γ_b is a discretized diffusion operator with standard deviation $\sigma_b = 1.0$. We consider

$R_0 = \cdots = R_N = \sigma_r^2 I_m$ with $\sigma_r = 0.2$. We consider $n = 500$ and $N = 24$, implying operators of size up to 12500×12500. We study the following settings:

- `Obs_1`: 20 evenly distributed observations of the state variables are made at 6 evenly distributed time steps, for a total of 120 observations ($\approx 1\%$ of observations),
- `Obs_10`: 70 evenly distributed observations of the state variables are made at 18 evenly distributed time steps, for a total of 1260 observations ($\approx 10\%$ of observations),
- `Obs_20`: 140 evenly distributed observations of the state variables are made at 18 evenly distributed time steps, for a total of 2520 observations ($\approx 20\%$ of observations).

This will allow us to study how the number of observations affects the performance of the randomized spectral LMP.

Preconditioning strategies

Let us describe the preconditioning strategies we analyze.

- `No_LMP`. In this first strategy, we apply the PCG (Algorithm 2.2) with $M_j = \Gamma_b$ for all j, that is we do not use the LMP. This is intended to illustrate a worst case scenario to study the impact of the LMPs. If the proposed randomized LMPs bring no improvements to Γ_b, then they should peform similar to this strategy.
- `Exact_Eigs`. Here, by contrast, we consider the ideal case where the spectral LMP (2.19) is constructed using exact eigenpairs. Thus, for each Gauss-Newton step, we construct beforehand the spectral LMP as in (2.19) with the dominant eigenpairs of $\Gamma_b A_j$ obtained via a dedicated eigensolver. Then, we apply the PCG (Algorithm 2.2) with the spectral LMP as a preconditioner. This approach is relevant to study since the randomized approaches aim at approximating the spectral LMP.
- `Ritz`. Then, we consider a more realistic intermediate strategy, widely used in concrete applications, and relying on Ritz approximations. The Ritz pairs are computed only once, after the application of the PCG (Algorithm 2.2) on the first Gauss-Newton step using $M_1 = \Gamma_b$. Then, they are used to construct LMP in standard form (2.10) for all the next linear systems. In particular, this implies that the LMP for the j-th system is constructed using approximate eigenpairs of $\Gamma_b A_1$.
- `Rand_InvFree`. For this randomized strategy, the linear systems are solved using the PCGIF (Algorithm 5.1). For each system, we set $\widehat{M}_j = I_n$ and thus construct a randomized spectral LMP as in (5.6) using approximate eigenpairs obtained with Algorithms 5.5.
- `Rand_Dual`. For the dual space method, we distinguish between $j = 1$ and $j > 1$. For the first Gauss-Newton step, the system is solved using the RPCG (Algorithm 5.2). Therefore, setting $\widehat{M}_j = I_m$, we construct a randomized spectral LMP as in (5.13) using Algorithm 5.7. Then for $j > 1$, the linear systems are solved using the Augmented RPCG (Algorithm 5.3). Consequently, the randomized LMP are constructed as in (5.15).

For the randomized strategies, we have deliberately limited ourselves to simple approaches, that is without any first-level preconditioner. We let this for future work. In the strategies `Exact_Eigs`, `Ritz`, `Rand_InvFree` and `Rand_Dual`, we consider the same number of vectors, namely $k = 30$. For the randomized methods, we use an oversampling of 20, that is we apply Algorithms 5.5 and 5.7 with $p = 50$, as suggested in [80].

Results

Let us now present the results. For each setting (`Obs_1`, `Obs_10` and `Obs_20`) we have performed 6 Gauss-Newton steps, which was generally sufficient to achieve convergence. The PCGIF and Augmented RPCG were both applied using a tolerance of $\varepsilon = 10^{-4}$. This ensures that when the Krylov subspace methods are stopped, the final iterates will be relatively similar between all the proposed approaches. This guarantees that the different strategies go through the same Gauss-Newton steps, and therefore encounter the same sequence of linear systems.

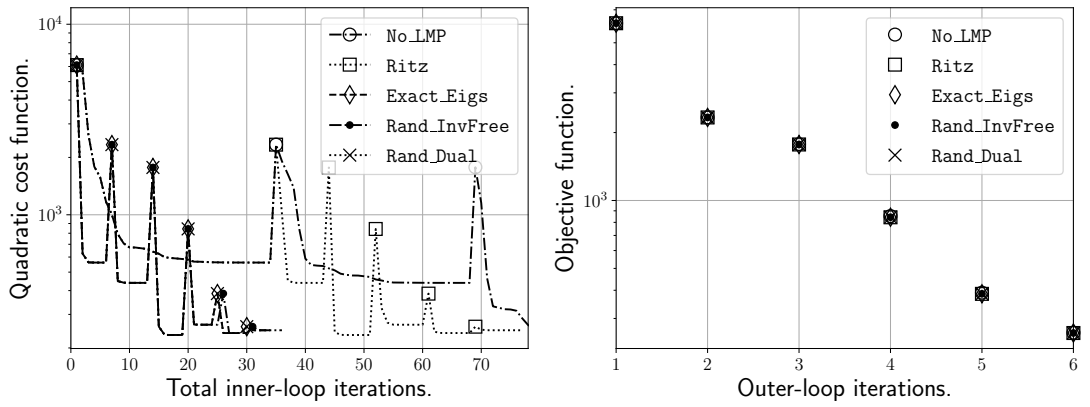
Results are presented in Figure 5.3, where both the evolution of the quadratic cost function (left) and of the true objective function (right) are shown. For the quadratic cost function, we have cropped to focus more on the strategies with LMPs. For all the three settings, there are almost no noticeable differences between `Exact_Eigs` (square), `Rand_InvFree` (dot) and `Rand_Dual` (cross). This illustrates that the randomized approaches are very well capturing the behavior of the exact spectral LMP, while being computed at a fewer cost. The `Ritz` (square) strategy performs actually well, but is penalized by the first Gauss-Newton where no LMP is used. This is even more visible when looking at the number of inner iterations per Gauss-Newton step shown in Figure 5.4. In Figure 5.4, one can clearly notice that from the second Gauss-Newton step, the Ritz LMP tends to yield similar results as `Exact_Eigs`, `Rand_InvFree` and `Rand_Dual`. On `Obs_20`, `Ritz` slightly outperforms the other strategies, which may be due to the fact that Ritz approximations also account for the right-hand side via the Krylov subspace.

In addition, we notice that the dual and inverse-free randomized strategies perform similarly. This seems to be another illustration of the phenomenon highlighted in Section 4.5.2. Indeed, the dual approach requires additional applications of H_j and H_j^T compared to the inverse-free approach. And as already highlighted when studying the approximate eigenpairs accuracy, it seems that additional applications of H_j and H_j^T lead to very few improvements for variational data assimilation problems. Consequently, we do not observe a significant gain with the dual approach compared to the inverse-free approach, although it is more costly regarding the applications of H_j and H_j^T. However, although there is no benefit in terms of number of inner-iterations, the dual methods remain cheaper in terms of storage and arithmetic costs.

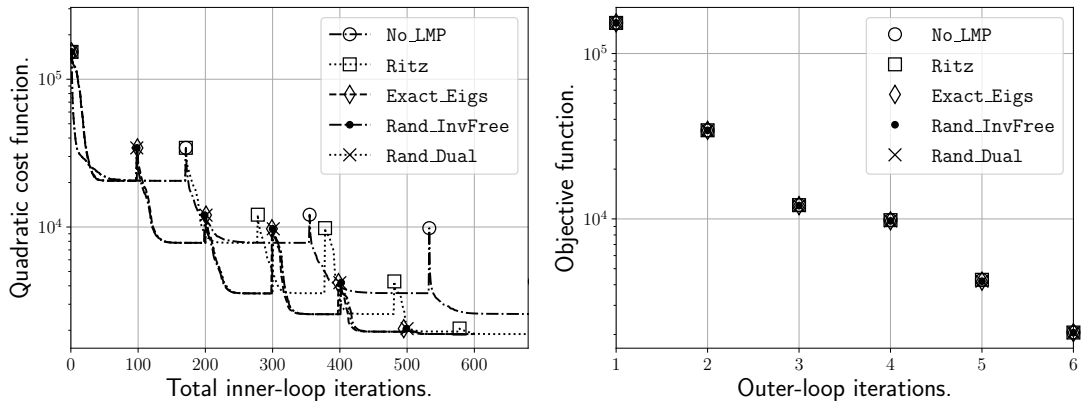
5.5 Conclusions and perspectives

In this chapter, we have proposed classes of randomized spectral LMP adapted to both the PCGIF and the Augmented RPCG. In this regard, we have used an adaptation of the algorithms proposed in Chapter 4, that further exploits the structure of variational data assimilation problems. For each Krylov subspace method, we have proposed two different variants of randomized spectral LMP depending on the availability of a first-level preconditioner. We have also identified conditions under which the randomized spectral LMPs for the PCGIF and the RPCG produce mathematically equivalent iterates. Finally, a first numerical illustration on a 3D-Var problem allowed us to demonstrate that the randomized spectral LMPs are indeed accurate approximations of exact spectral LMPs. Then, the improvements in terms of convergence of the optimization procedure, the randomized spectral LMP performs similarly as the exact spectral LMP. It is also comparable to the Ritz LMP, but can be computed from the beginning while the Ritz LMP necessitates a first Gauss-Newton step without LMP.

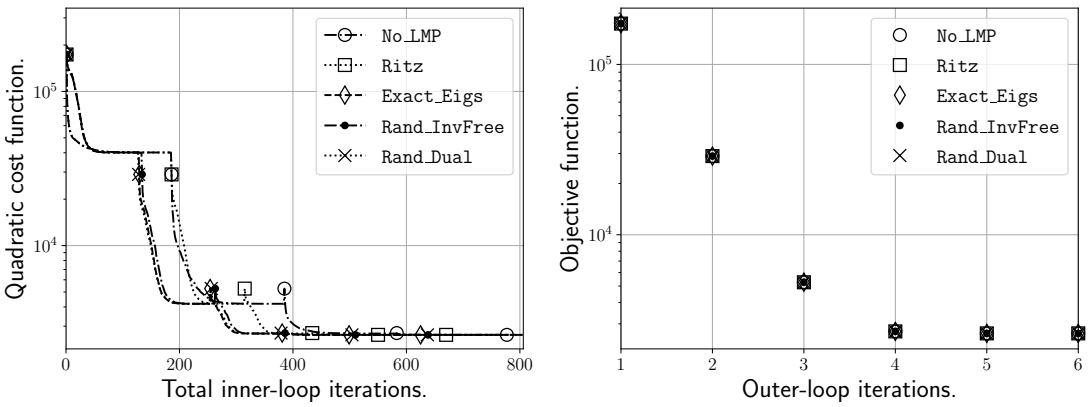
One objective for future research on this topic is to propose and study more sophisticated randomized approaches adapted to the solution of the sequence. Here, we have only considered the elementary approach where a new randomized LMP is constructed in the beginning of each Gauss-Newton step assuming there is no first-level preconditioner. Nevertheless, the algorithms proposed in this chapter allow us to construct randomized spectral LMPs also when such a first-level preconditioner is available. Consequently, a strategy that must be studied would consist in building randomized LMPs on top of each other, that is the new randomized spectral LMP is



(a) `Obs_1`.



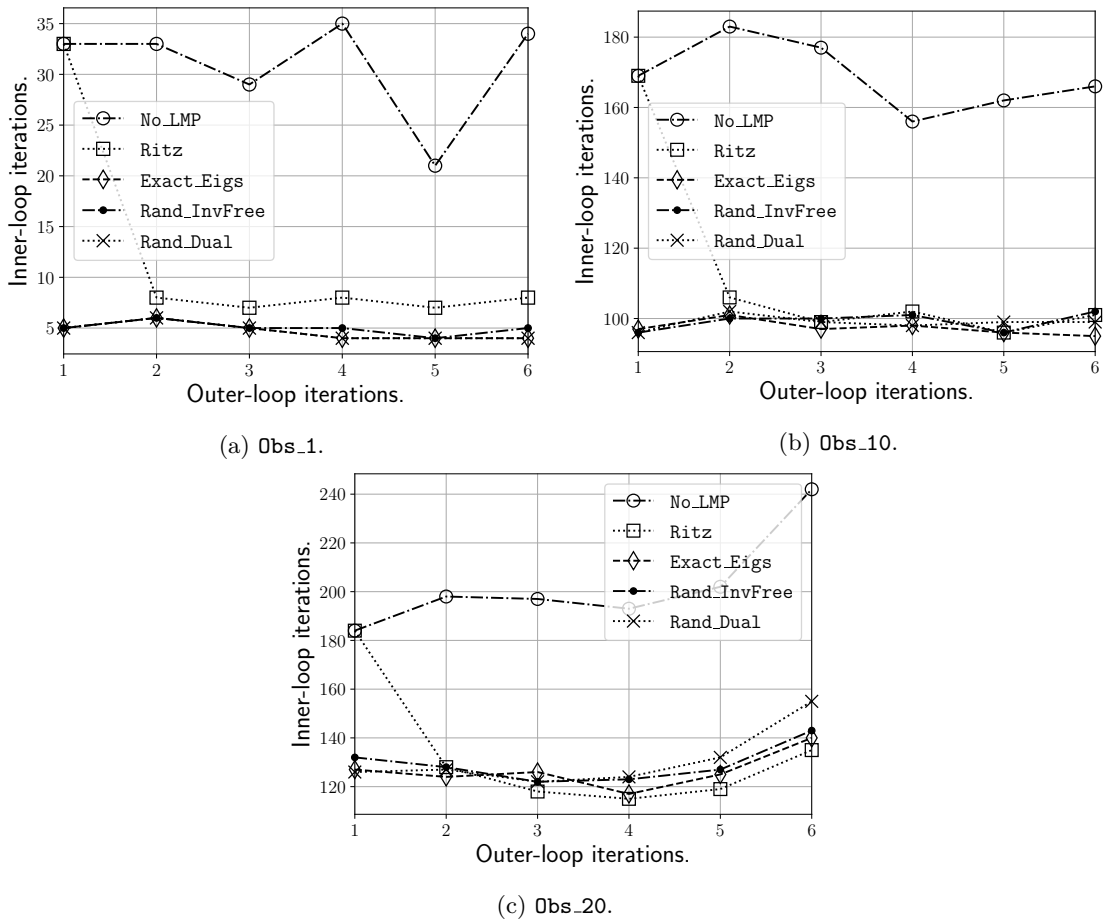
(b) `Obs_10`.



(c) `Obs_20`.

Figure 5.3: Convergence of the Gauss-Newton method on the `Obs_1` (top), `Obs_10` (middle), and `Obs_20` (bottom) settings. Behaviors of the quadratic cost function (left) and of the true objective function (right) are shown.

constructed using the previous one as a first-level preconditioner. The j-th LMP would then carry eigeninformation related to all the previous systems, which could lead to further convergence improvements.



(a) `Obs_1`.

(b) `Obs_10`.

(c) `Obs_20`.

Figure 5.4: Number of inner iterations per Gauss-Newton step on the `Obs_1` (top), `Obs_10` (middle), and `Obs_20` (bottom) settings.

Also, it seems relevant to study the potential combinations between randomized and deterministic approaches such as the Ritz LMP. Ritz LMP has been proposed for the PCGIF and the Augmented RPCG in [50] (see Section 3.1.2 and 4.4.1, respectively). An advantage of the Ritz LMP is that the approximate eigenpairs are obtained using Krylov subspaces, and thus integrate the right-hand side. This can improve the convergence rate in the first iterations of the PCG, which is desirable in practical applications where few iterations are performed. There are several possibilities to perform such a combination. A first approach would consist in primarily constructing the Ritz LMP, and then consider it as the first-level preconditioner in Algorithms 5.4 or 5.6. Accordingly, the j-th preconditioner will carry information on both the previous Krylov subspace, and complementary eigeninformation obtained with the randomized methods. Another possibility is to use the Ritz vectors to draw the random matrix Ω in the orthogonal subspace, which is equivalent to consider a particular covariance matrix for Ω. In both cases, the objective is to use randomized methods to obtain eigeninformation that is complementary to the Ritz one.

Chapter 6

Conclusions and perspectives

Conclusions. In this thesis, we have proposed several contributions to finally address the variational data assimilation problem using randomized numerical linear algebra methods.

In Chapter 3, we have proposed a general error analysis of the randomized low rank approximation error. This analysis extends the existing work to Gaussian matrices with general covariance mean and non-trivial mean term. The proposed generalization enables to analyze a larger class of randomized methods based on randomized subspace iterations. An application to the analysis of the randomized singular value decomposition has demonstrated that the proposed bounds improve over the reference bounds in [53].

Secondly, in Chapter 4 we have developed algorithms to address two generalized eigenvalue problems that notably arise in variational data assimilation. The proposed methods are fairly general and, in particular, allowed us to recover existing algorithms from [80, 79, 24]. Then, we have derived an average case analysis of the algorithms using the general analysis from Chapter 3. Finally, we have illustrated the performance of our algorithms in terms of eigenpair accuracy on a three-dimensional variational data assimilation problem.

Finally, in Chapter 5, we have adapted the algorithms from Chapter 4 to design a new class of randomized spectral limited memory preconditioners. Those randomized preconditioners are designed for two particular Krylov subspace methods adapted to variational data assimilation: the inverse-free and the augmented restricted preconditioned conjugate gradient methods. We have presented two variants of randomized spectral limited memory preconditioners depending on the availability of a first-level preconditioner. Then, we have identified relations between the randomized spectral limited memory preconditioners for the inverse-free and the restricted preconditioned conjugate gradient to ensure mathematically equivalent iterates. The performance of the proposed randomized spectral limited memory preconditioners have then been illustrated on a toy four-dimensional variational data assimilation problem. The obtained results show that the randomized preconditioners yield similar performance as the conjugate gradient method preconditioned by the exact spectral LMP, which opens interesting perspectives.

Perspectives. The proposed randomized methods have proven to perform well either for approximating eigenpairs or preconditioning Krylov subspace methods within variational data assimilation. In both the theoretical analysis and numerical experiments we have focused on the Gaussian distribution. However, in computationally intensive contexts, structured distributions [63, 93] (such as subsampled trigonometric transforms [94] and sparse random distributions [22]) are generally preferred. In particular, random sparse matrices can be particularly adapted for large scale data assimilation problems since they are cheaper to generate and store, and thus be beneficial to the overall arithmetic cost. Consequently, a future research direction should investigate how the randomized algorithms presented in this thesis behave when the random sample

matrix is drawn from a sparse distribution. As pointed out in [70], randomized algorithms are relatively insensitive to the underlying distribution meaning that we should numerically observe similar performance. Theoretically, deriving a randomized low rank approximation analysis as in Chapter 3 for sparse distributions would also be of great interest. It seems clear that such an analysis cannot be deduced using similar arguments as for the Gaussian case. Consequently, we will need to employ refined theoretical tools. Matrix concentration inequalities [85] seem also relevant in this context.

In Chapter 5, we have proposed a numerical illustration to determine how well the randomized spectral limited memory preconditioners approximate the corresponding exact spectral one. This empirically appears as a consequence of the theoretical analysis proposed in Chapter 4. In a recent paper, the authors proposed a theoretical analysis [33, Section 2] of a class of randomized preconditioners for symmetric positive definite matrices when using the shifted Nyström approximation [86, Algorithm 3]. These preconditioners are based on the spectral limited memory preconditioner given in (2.19). Their analysis relates the error between the randomized and exact spectral limited memory preconditioner to the low rank approximation error [63, Section 11]. Consequently, it would be interesting to investigate possible extensions of their results to the analysis of the randomized preconditioners proposed in this thesis.

The algorithms proposed in Chapter 4 can be applied to general matrices. Consequently, our randomized methods can be used to compute approximate eigenpairs of symmetric indefinite matrices. In combination with appropriate limited memory preconditioner formulations [46, 64], this would allow us to address indefinite formulations appearing in variational data assimilation [32]. In such cases, the truncation performed at the end of each algorithm should rather focus on the largest eigenvalues in absolute value. Assuming that the randomized methods will accurately capture these extremal eigenmodes, it would be interesting to consider the combination with an harmonic Ritz analysis which is known to provide accurate interior eigenpairs. We also note that in this case our theoretical analysis will have to be adapted.

In operational variational data assimilation problems, the Ritz limited memory preconditioner is the method of choice. An important aspect that should be investigated is how to combine the Ritz pairs that are available almost for free and the randomized methods proposed in this thesis. The objective would be to use the randomized methods to obtain complementary information and thus to take advantage of both the deterministic and randomized aspects. A first option would be to use the Ritz limited memory preconditioner as the first preconditioner in the algorithms presented in Chapter 4. Thus, the randomized method will act as an additional layer of preconditioning on top of the deterministic two-level Ritz preconditioner. Another possibility is to use available information within the covariance matrix of the random sample matrix Ω. For instance, one can draw the random sample matrix in the orthogonal of the subspace spanned by the available Ritz vectors, or use deflated operators as a covariance matrix. In these cases, one must be careful with the interpretation of the resulting approximate eigenpairs. Especially, it will be crucial to clearly identify to which operator the approximate eigeninformation corresponds and use it accordingly.

The randomized preconditioning strategies investigated in Chapter 5 are elementary and in particular, they do not really consider the aspects related to the sequence. A natural extension would then be to study more sophisticated strategies that are specifically adapted to the sequence. The main objective here is to construct the j-th preconditioner taking into account the $(j-1)$ previous solutions. This implies to determine how to use the already available information and also to which operator apply the randomized methods on. It turns out that the flexibility of the algorithms in Chapters 4 and 5 allows us to consider several options. A first one would be to combine multiplicatively the randomized limited memory preconditioners. In this case, the algorithms in Chapter 5 will be applied with the $(j-1)$-th preconditioner as the first level preconditioner of the j-th preconditioner. This will obviously increase the complexity of the preconditioner along the sequence, but the aggregated information may yield significant

improvements in the overall solution process. Another possibility is to construct several randomized preconditioners in parallel and combine them additively. In this case, the challenge is to ensure that the preconditioners provide different information to maximize the improvements.

Bibliography

[1] D. Achlioptas, Z. S. Karnin, and E. Liberty, *Near-optimal entrywise sampling for data matrices*, in Advances in Neural Information Processing Systems, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, eds., vol. 26, Curran Associates, Inc., 2013, pp. 1565–1573.

[2] D. Achlioptas and F. Mcsherry, *Fast computation of low-rank matrix approximations*, Journal of the ACM, 54 (2007), pp. 9–es.

[3] M. Altaf, M. El Gharamti, A. Heemink, and I. Hoteit, *A reduced adjoint approach to variational data assimilation*, Computer Methods in Applied Mechanics and Engineering, 254 (2013), pp. 1–13.

[4] W. E. Arnoldi, *The principle of minimized iterations in the solution of the matrix eigenvalue problem*, Quarterly of Applied Mathematics, 9 (1951), pp. 17–29.

[5] M. Asch, M. Bocquet, and M. Nodet, *Data Assimilation: Methods, Algorithms, and Applications*, no. 11 in Fundamentals of Algorithms, SIAM, Society for Industrial and Applied Mathematics, Philadelphia, 2016.

[6] H. Avron, P. Maymounkov, and S. Toledo, *Blendenpik: Supercharging LAPACK's Least-Squares Solver*, SIAM Journal on Scientific Computing, 32 (2010), pp. 1217–1236.

[7] O. Axelsson and I. Kaporin, *On the sublinear and superlinear rate of convergence of conjugate gradient methods*, Numerical Algorithms, 25 (2000), pp. 1–22.

[8] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, eds., *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, Society for Industrial and Applied Mathematics, Jan. 2000.

[9] P. Bauer, T. Quintino, N. Wedi, A. Bonanni, M. Chrust, W. Deconinck, M. Diamantakis, P. Düben, S. English, J. Flemming, P. Gillies, I. Hadade, J. Hawkes, M. Hawkins, O. Iffrig, C. Kühnlein, M. Lange, P. Lean, O. Marsden, A. Müller, S. Saarinen, D. Sarmany, M. Sleigh, S. Smart, P. Smolarkiewicz, D. Thiemert, G. Tumolo, C. Weihrauch, C. Zanna, and P. Maciel, *The ECMWF Scalability Programme: Progress and Plans*, (2020).

[10] B. Beckermann and A. B. J. Kuijlaars, *Superlinear Convergence of Conjugate Gradients*, SIAM Journal on Numerical Analysis, 39 (2001), pp. 300–329.

[11] S. Bellavia, V. De Simone, D. di Serafino, and B. Morini, *Efficient Preconditioner Updates for Shifted Linear Systems*, SIAM Journal on Scientific Computing, 33 (2011), pp. 1785–1809.

[12] S. Bellavia, B. Morini, and M. Porcelli, *New updates of incomplete LU factorizations and applications to large nonlinear systems*, Optimization Methods and Software, 29 (2014), pp. 321–340.

[13] Å. Björck, *Numerical Methods in Matrix Computations*, no. 59 in Texts in Applied Mathematics, Springer International Publishing : Imprint: Springer, Cham, first ed., 2015.

[14] N. Boullé and A. Townsend, *A generalization of the randomized singular value decomposition*, arXiv:2105.13052 [cs, math, stat], (2022).

[15] ——, *Learning Elliptic Partial Differential Equations with Randomized Linear Algebra*, Foundations of Computational Mathematics, (2022).

[16] N. Bousserez, J. J. Guerrette, and D. K. Henze, *Enhanced parallelization of the incremental 4D-Var data assimilation algorithm using the Randomized Incremental Optimal Technique*, Quarterly Journal of the Royal Meteorological Society, 146 (2020), pp. 1351–1371.

[17] N. Bousserez and D. K. Henze, *Optimal and scalable methods to approximate the solutions of large-scale Bayesian problems: Theory and application to atmospheric inversion and data assimilation*, Quarterly Journal of the Royal Meteorological Society, 144 (2018), pp. 365–390.

[18] C. Boutsidis, M. W. Mahoney, and P. Drineas, *An Improved Approximation Algorithm for the Column Subset Selection Problem*, in Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, Jan. 2009, pp. 968–977.

[19] A. Buluc, T. G. Kolda, S. M. Wild, M. Anitescu, A. DeGennaro, J. Jakeman, C. Kamath, R. Kannan, M. E. Lopes, P.-G. Martinsson, K. Myers, J. Nelson, J. M. Restrepo, C. Seshadhri, D. Vrabie, B. Wohlberg, S. J. Wright, C. Yang, and P. Zwart, *Randomized Algorithms for Scientific Computing (RASC)*, arXiv:2104.11079 [cs], (2021), pp. None, 1807223.

[20] J. C. Butcher, *Numerical Methods for Ordinary Differential Equations*, John Wiley & Sons, Ltd, Chichester, UK, July 2016.

[21] A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen, *Data assimilation in the geosciences: An overview of methods, issues, and perspectives*, WIREs Climate Change, 9 (2018).

[22] M. B. Cohen, *Nearly Tight Oblivious Subspace Embeddings by Trace Inequalities*, in Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, Jan. 2016, pp. 278–287.

[23] R. Daley, *Atmospheric Data Analysis*, no. 2 in Cambridge Atmospheric and Space Science Series, Cambridge University Press, Cambridge, first ed., 1999.

[24] I. Daužickaitė, A. S. Lawless, J. A. Scott, and P. J. Leeuwen, *Randomised preconditioning for the forcing formulation of weak-constraint 4D-Var*, Quarterly Journal of the Royal Meteorological Society, 147 (2021), pp. 3719–3734.

[25] C. Davis, *Separation of Two Linear Subspaces*, Acta Sci. Math. (Szeged), 19 (1958), pp. 172–187.

[26] T. A. Davis, *Direct Methods for Sparse Linear Systems*, Society for Industrial and Applied Mathematics, Jan. 2006.

[27] J. Demmel, L. Grigori, M. Hoemmen, and J. Langou, *Communication-optimal Parallel and Sequential QR and LU Factorizations*, SIAM Journal on Scientific Computing, 34 (2012), pp. A206–A239.

[28] P. Drineas and I. C. F. Ipsen, *Low-Rank Matrix Approximations Do Not Need a Singular Value Gap*, SIAM Journal on Matrix Analysis and Applications, 40 (2019), pp. 299–319.

[29] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós, *Faster least squares approximation*, Numerische Mathematik, 117 (2011), pp. 219–249.

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

[31] X. Feng and Z. Zhang, *The rank of a random matrix*, Applied Mathematics and Computation, 185 (2007), pp. 689–694.

[32] M. Fisher, S. Gratton, S. Gürol, Y. Trémolet, and X. Vasseur, *Low rank updates in preconditioning the saddle point systems arising from data assimilation problems*, Optimization Methods and Software, 33 (2018), pp. 45–69.

[33] Z. Frangella, J. A. Tropp, and M. Udell, *Randomized Nyström Preconditioning*, arXiv:2110.02820 [cs, math], (2021).

[34] M. A. Freitag, *Numerical linear algebra in data assimilation*, GAMM, 43 (2020).

[35] M. A. Freitag and D. L. Green, *A low-rank approach to the solution of weak constraint variational data assimilation problems*, Journal of Computational Physics, 357 (2018), pp. 263–281.

[36] M. A. Freitag and A. Spence, *Convergence of inexact inverse iteration with application to preconditioned iterative solves*, BIT Numerical Mathematics, 47 (2007), pp. 27–44.

[37] T. Fukaya, R. Kannan, Y. Nakatsukasa, Y. Yamamoto, and Y. Yanagisawa, *Shifted Cholesky QR for Computing the QR Factorization of Ill-Conditioned Matrices*, SIAM Journal on Scientific Computing, 42 (2020), pp. A477–A503.

[38] T. Fukaya, Y. Nakatsukasa, Y. Yanagisawa, and Y. Yamamoto, *CholeskyQR2: A Simple and Communication-Avoiding Algorithm for Computing a Tall-Skinny QR Factorization on a Large-Scale Parallel System*, in 2014 5th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems, New Orleans, LA, USA, Nov. 2014, IEEE, pp. 31–38.

[39] A. Gaul, *Recycling Krylov Subspace Methods for Sequences of Linear Systems – Analysis and Applications*, PhD thesis, Technische Universität Berlin, Berlin, 2014.

[40] M. Ghil and P. Malanotte-Rizzoli, *Data Assimilation in Meteorology and Oceanography*, in Advances in Geophysics, vol. 33, Elsevier, 1991, pp. 141–266.

[41] A. Gittens, A. Devarakonda, E. Racah, M. Ringenburg, L. Gerhardt, J. Kottalam, J. Liu, K. Maschhoff, S. Canon, J. Chhugani, P. Sharma, J. Yang, J. Demmel, J. Harrell, V. Krishnamurthy, M. W. Mahoney, and Prabhat, *Matrix factorizations at scale: A comparison of scientific data analytics in spark and C+MPI using three case studies*, in 2016 IEEE International Conference on Big Data (Big Data), Washington DC,USA, Dec. 2016, IEEE, pp. 204–213.

[42] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins Series in the Mathematical Sciences, Johns Hopkins university press, Baltimore London, third ed., 1996.

[43] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins Studies in the Mathematical Sciences, The Johns Hopkins University Press, Baltimore, fourth ed., 2013.

[44] R. M. Gower and P. Richtárik, *Randomized Quasi-Newton Updates Are Linearly Convergent Matrix Inversion Algorithms*, SIAM Journal on Matrix Analysis and Applications, 38 (2017), pp. 1380–1409.

[45] S. Gratton, A. S. Lawless, and N. K. Nichols, *Approximate Gauss–Newton Methods for Nonlinear Least Squares Problems*, SIAM Journal on Optimization, 18 (2007), pp. 106–132.

[46] S. Gratton, S. Mercier, N. Tardieu, and X. Vasseur, *Limited memory preconditioners for symmetric indefinite problems with application to structural mechanics*, Numerical Linear Algebra with Applications, 23 (2016), pp. 865–887.

[47] S. Gratton, A. Sartenaer, and J. Tshimanga, *On A Class of Limited Memory Preconditioners For Large Scale Linear Systems With Multiple Right-Hand Sides*, SIAM Journal on Optimization, 21 (2011), pp. 912–935.

[48] S. Gratton and J. Tshimanga, *An observation-space formulation of variational assimilation using a restricted preconditioned conjugate gradient algorithm*, Quarterly Journal of the Royal Meteorological Society, 135 (2009), pp. 1573–1585.

[49] M. Gu, *Subspace Iteration Randomization and Singular Value Problems*, SIAM Journal on Scientific Computing, 37 (2015), pp. A1139–A1173.

[50] S. Gürol, *Solving Regularized Nonlinear Least-Squares Problem in Dual Space with Application to Variational Data Assimilation*, PhD thesis, Université de Toulouse, Toulouse, June 2013.

[51] A. Gut, *An Intermediate Course in Probability*, Springer Texts in Statistics, Springer, Dordrecht New York, NY, second ed., 2009.

[52] W. Hackbusch, *Hierarchical Matrices: Algorithms and Analysis*, no. 49 in Hierarchical Matrices, Springer Berlin Heidelberg, Berlin, Heidelberg, first ed., 2015.

[53] N. Halko, P.-G. Martinsson, and J. A. Tropp, *Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions*, SIAM Review, 53 (2011), pp. 217–288.

[54] N. P. Halko, *Randomized Methods for Computing Low-Rank Approximations of Matrices*, PhD thesis, University of Colorado, Boulder, CO, United States, 2012.

[55] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, Journal of Research of the National Bureau of Standards, 49 (1952), pp. 409–435.

[56] N. J. Higham, *Functions of Matrices: Theory and Computation*, Society for Industrial and Applied Mathematics, Philadelphia, 2008.

[57] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge ; New York, second ed., 2012.

[58] A. V. Knyazev and M. E. Argentati, *Principal Angles between Subspaces in an A-Based Scalar Product: Algorithms and Perturbation Estimates*, SIAM Journal on Scientific Computing, 23 (2002), pp. 2008–2040.

[59] C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, Journal of Research of the National Bureau of Standards, 45 (1950), pp. 255–282.

[60] C. Lanczos, *Solution of systems of linear equations by minimized iterations*, Journal of Research of the National Bureau of Standards, 49 (1952), p. 33.

[61] M. W. Mahoney, *Randomized Algorithms for Matrices and Data*, Foundations and Trends® in Machine Learning, 3 (2010), pp. 123–224.

[62] J. Mandel, *Balancing domain decomposition*, Communications in Numerical Methods in Engineering, 9 (1993), pp. 233–241.

[63] P.-G. Martinsson and J. A. Tropp, *Randomized numerical linear algebra: Foundations and algorithms*, Acta Numerica, 29 (2020), pp. 403–572.

[64] S. Mercier, S. Gratton, N. Tardieu, and X. Vasseur, *A new preconditioner update strategy for the solution of sequences of linear systems in structural mechanics: Application to saddle point problems in elasticity*, Computational Mechanics, 60 (2017), pp. 969–982.

[65] G. Meurant, *The Lanczos and Conjugate Gradient Algorithms: From Theory to Finite Precision Computations*, no. 19 in Software, Environments, and Tools, SIAM, Philadelphia, 2006.

[66] J. L. Morales and J. Nocedal, *Automatic Preconditioning by Limited Memory Quasi-Newton Updating*, SIAM Journal on Optimization, 10 (2000), pp. 1079–1096.

[67] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, Wiley Series in Probability and Mathematical Statistics, Wiley, New York, 1982.

[68] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Series in Operation Research and Financial Engineering, Springer, New York, NY, second ed., 2006.

[69] M. A. Olshanskii and E. E. Tyrtyšnikov, *Iterative Methods for Linear Systems: Theory and Applications*, Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, 2014.

[70] S. Oymak and J. A. Tropp, *Universality laws for randomized dimension reduction, with applications*, Information and Inference: A Journal of the IMA, 7 (2018), pp. 337–446.

[71] C. C. Paige and M. Wei, *History and generality of the CS decomposition*, Linear Algebra and its Applications, 208–209 (1994), pp. 303–326.

[72] B. N. Parlett, *The Symmetric Eigenvalue Problem*, no. 20 in Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, 1998.

[73] J. W. Pearson and J. Pestana, *Preconditioners for Krylov subspace methods: An overview*, GAMM-Mitteilungen, 43 (2020).

[74] V. Rokhlin, A. Szlam, and M. Tygert, *A Randomized Algorithm for Principal Component Analysis*, SIAM Journal on Matrix Analysis and Applications, 31 (2010), pp. 1100–1124.

[75] Y. Saad, *Iterative Methods for Sparse Linear Systems*, SIAM, Philadelphia, second ed., 2003.

[76] ——, *Iterative methods for linear systems of equations: A brief historical journey*, arXiv:1908.01083 [math], (2019).

[77] A. K. Saibaba, *Randomized Subspace Iteration: Analysis of Canonical Angles and Unitarily Invariant Norms*, SIAM Journal on Matrix Analysis and Applications, 40 (2019), pp. 23–48.

[78] A. K. Saibaba, J. Hart, and B. van Bloemen Waanders, *Randomized algorithms for generalized singular value decomposition with application to sensitivity analysis*, Numerical Linear Algebra with Applications, 28 (2021).

[79] A. K. Saibaba and P. K. Kitanidis, *Fast computation of uncertainty quantification measures in the geostatistical approach to solve inverse problems*, Advances in Water Resources, 82 (2015), pp. 124–138.

[80] A. K. Saibaba, J. Lee, and P. K. Kitanidis, *Randomized algorithms for generalized Hermitian eigenvalue problems with application to computing Karhunen–Loève expansion*, Numerical Linear Algebra with Applications, 23 (2016), pp. 314–339.

[81] J. M. Tang, R. Nabben, C. Vuik, and Y. A. Erlangga, *Comparison of Two-Level Preconditioners Derived from Deflation, Domain Decomposition and Multigrid Methods*, Journal of Scientific Computing, 39 (2009), pp. 340–370.

[82] A. E. Tomás and E. S. Quintana-Ortí, *Tall-and-skinny QR factorization with approximate Householder reflectors on graphics processors*, The Journal of Supercomputing, 76 (2020), pp. 8771–8786.

[83] A. Toselli and O. Widlund, *Domain Decomposition Methods: Algorithms and Theory*, no. 34 in Springer Series in Computational Mathematics, Springer, Berlin, 2005.

[84] L. N. Trefethen and D. Bau, *Numerical Linear Algebra*, Society for Industrial and Applied Mathematics, Philadelphia, 1997.

[85] J. A. Tropp, *An Introduction to Matrix Concentration Inequalities*, arXiv:1501.01571 [cs, math, stat], (2015).

[86] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher, *Fixed-Rank Approximation of a Positive-Semidefinite Matrix from Streaming Data*, arXiv:1706.05736 [cs, stat], (2017).

[87] ——, *Practical Sketching Algorithms for Low-Rank Matrix Approximation*, SIAM Journal on Matrix Analysis and Applications, 38 (2017), pp. 1454–1485.

[88] U. Trottenberg, C. W. Oosterlee, and A. Schüller, *Multigrid*, Academic Press, San Diego, 2001.

[89] J. Tshimanga, S. Gratton, A. T. Weaver, and A. Sartenaer, *Limited-memory preconditioners, with application to incremental four-dimensional variational data assimilation*, Quarterly Journal of the Royal Meteorological Society, 134 (2008), pp. 751–769.

[90] A. van der Sluis and H. A. van der Vorst, *The rate of convergence of Conjugate Gradients*, Numerische Mathematik, 48 (1986), pp. 543–560.

[91] C. F. Van Loan, *Generalizing the Singular Value Decomposition*, SIAM Journal on Numerical Analysis, 13 (1976), pp. 76–83.

[92] A. J. Wathen, *Preconditioning*, Acta Numerica, 24 (2015), pp. 329–376.

[93] D. P. Woodruff, *Sketching as a Tool for Numerical Linear Algebra*, Foundations and Trends® in Theoretical Computer Science, 10 (2014), pp. 1–157.

[94] F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert, *A fast randomized algorithm for the approximation of matrices*, Applied and Computational Harmonic Analysis, 25 (2008), pp. 335–366.

[95] Y. Yamamoto, Y. Nakatsukasa, Y. Yanagisawa, and T. Fukaya, *Roundoff error analysis of the CholeskyQR2 algorithm in an oblique inner product*, JSIAM Letters, 8 (2016), pp. 5–8.

[96] S. Zhang and P. Wu, *High Accuracy Low Precision QR Factorization and Least Square Solver on GPU with TensorCore*, arXiv:1912.05508 [cs], (2019).

[97] P. Zhu and A. V. Knyazev, *Angles between subspaces and their tangents*, Journal of Numerical Mathematics, 21 (2013), pp. 325–340.