

**Thesis Title** Add later I forgot what title I submitted for my thesis along with my graduation request

**Author Name:** Hongda Li

A thesis presented for the degree of: ????

Department Name: Department of Applied Mathematics  
University Name: University of Washington

### Abstract

This thesis is expository in nature and we intend to discuss the Conjugate Gradient without pre-conditioning together with the Lanczos iteration in thorough details. We derive the Conjugate Gradient algorithm using the Conjugate Direction method and projector. We derive Lanczos iterations as the symmetric case of Arnoldi iterations. We analyze the convergence rate of Conjugate Gradient in exact arithmetic and prove the equivalence between the Conjugate Gradient algorithm and the Lanczos iterations. In the end, we analyze the behaviors of Lanczos iterations and Conjugate Gradient under floating point arithmetic, illustrate some of their behaviors under floating point arithmetic using numerical experiments.

# Contents

<b>1</b>	<b>Notations</b>	<b>4</b>
<b>2</b>	<b>Introduction</b>	<b>4</b>
<b>3</b>	<b>Foundations</b>	<b>6</b>
3.1	The Basics	6
3.1.1	Krylov Subspace	6
3.1.2	Projectors	7
3.2	Subspace Projection Methods	7
3.2.1	Prototype Algorithm	7
3.2.2	Energy Norm Minimization using Gradient	8
3.3	Deriving Conjugate Gradient from Conjugate Directions	9
3.3.1	CG Objective and Framework	10
3.3.2	Using the Projector	11
3.3.3	Method of Conjugate Directions	12
3.3.4	Properties of CDM	13
3.3.5	Conjugate Gradient	13
3.3.6	CG and Krylov Subspace	16
3.4	Arnoldi Iterations and Lanczos	17
3.4.1	The Arnoldi Iterations	17
3.4.2	Arnoldi Produces Orthogonal Basis for Krylov Subspace	18
3.4.3	The Lanczos Iterations	19
<b>4</b>	<b>Analysis of Conjugate Gradient and Lanczos Iterations</b>	<b>21</b>
4.1	Conjugate Gradient and Matrix Polynomial	21
4.1.1	Termination Conditions of CG	23
4.2	Convergence Rate of CG under Exact Arithmetic	23
4.2.1	Uniformly Distributed Eigenvalues	23
4.2.2	One Outlier Eigenvalue	25
4.3	From Conjugate Gradient to Lanczos	27
4.3.1	The Base Case	29
4.3.2	The Inductive Case	29
4.3.3	Fixing the Sign	31
4.4	From Lanczos to Conjugate Gradient	31
4.4.1	Matching the Residual and Conjugate Vectors	32
4.4.2	Matching the $a_k, b_k$ in CG	34
<b>5</b>	<b>Effects of Floating Point Arithmetic</b>	<b>36</b>
5.1	Partial Orthogonalization and Full Orthogonalization	36
5.2	Relative Errors of CG Under Floating Points	37
5.2.1	Experiments	37
5.3	Paige's Floating Point Analysis	38
5.3.1	Bounding the Relative Residuals	38
5.3.2	Paige's Theorem and Backwards Analysis of CG	39

5.4	Ghost Eigenvalues and Losing Orthogonality . . . . .	41
5.4.1	Ghost Eigenvalues Experiments . . . . .	42
5.4.2	Losing Orthogonality on Ritz Vectors . . . . .	44
5.5	The Ritz Vector's View and Another Paige's Theorem . . . . .	44
5.6	Forward Error Analysis . . . . .	46
<b>6</b>	<b>Appendix</b>	<b>47</b>
6.1	Useful Lemmas . . . . .	47
6.2	Theorems, Propositions . . . . .	47

# 1 Notations

1.  $\text{ran}(A) := \{Ax : \forall x \in \mathbb{R}^n\}$ ,  $A \in \mathbb{R}^{m \times n}$ , The range of a matrix.
2.  $(A)_{i,j}$ : The element in  $i$  th row and  $j$ th column of the matrix  $A$ .
3.  $(A)_{i:i',j:j'}$ : The submatrix whose top left corner is the  $(i,j)$  element in matrix  $A$ , and whose' right bottom corner is the  $(i',j')$  element in the matrix  $A$ . The notation is similar to MATLAB's rules for indexing.
4.  $\forall 0 \leq j \leq k$ : under certain context it indicates the range for an index:  $j = 0, 1, \dots, k-1, k$
5. Boldface  $\mathbf{0}$  denotes the zero vector or matrix, depending on the context it can be either a zero row/column vector, or a zero matrix.
6. The  $\hat{\cdot}$  decorator is reserved for denoting the unit vector of some non zero vector. For example  $\hat{x} := x/\|x\|, x \neq \mathbf{0}$ .
7.  $p_k(A|w)$  denotes the matrix polynomial  $\sum_{j=0}^k w_j A^j$ .

# 2 Introduction

The Conjugate Gradient method is an iterative method used for solving linear systems which date back to the period when computers were programmed using punched cards. It didn't receive much attention at the start but was revised and reappeared as a method for solving large sparse linear systems decades later, becoming the best option for positive definite linear systems that are sparse and large and, by extension, for optimizing strongly convex functions as well. In this thesis, we discuss Conjugate Gradient without pre-conditioning by deriving it and analyzing it along with Lanczos Iterations, a closely related algorithm for symmetric eigenproblems. Finally, we use their connections to analyze their behaviours under floating-point arithmetic. The thesis will require some background in numerical linear algebra for the best understanding.

In the first section, we introduce the projectors and Krylov subspace as important mathematical objects for the subspace projections method, the frameworks for subspace projection methods, and the Lanczos Iteration as a symmetric case of the Arnoldi Iterations. At the end of the section, we proceed and derive the conjugate gradient method using only these ideas and concepts. In the second section, we analyze the behaviors of Conjugate gradient and Lanczos Iterations, including the terminations conditions for both algorithm, their convergence rate, and the equivalence between them. The goal is to show how they can be the same and have similar properties and how the connections between them can spark other methods for solving a symmetric indefinite linear system. And in the final parts of the second section, we derive the convergence bound for the conjugate gradient algorithm.

In the third section, we use numerical experiment to better understand the algorithm behaviors under floating point arithmetic. We show how floating points affect the Lanczos algorithm more rigorously and how it may be fixed and mitigated, consequently improving the conjugate gradient algorithm.

**Acknowledgement** Thanks to Professor Greenbaum for taking the time to review the first draft of this thesis and for her help with my research throughout the year. Thanks to my old undergraduate classmate and friend who helped with fixing the typos on the later drafts.

## 3 Foundations

In this section, we go over the foundations of Conjugate Gradient and the Lanczos Algorithm. We introduce the important ideas at the beginning, and then we proceed to prove the conjugate gradient algorithm via the method conjugate directions, and then we state the Lanczos Iterations as a symmetric case of the Arnoldi Iterations.

### 3.1 The Basics

In this subsection we go over some basics concepts and mathematical entities that are important to the Subspace Projection methods in general.

#### 3.1.1 Krylov Subspace

**Definition 1** (Krylov Subspace).

$$\mathcal{K}_k(A|b) = \text{span}(b, Ab, A^2b, \dots, A^{k-1}b)$$

Observe that, for every element in the subspace, it's a matrix polynomial multiplied by  $b$ , we write it as  $p_{k-1}(A|w)b$ , a  $k-1$  degree matrix polynomial where  $A$  is the matrix and  $w$  is a vector denoting the coefficients.

**Definition 2** (The Grade of Krylov Subspace). The grade of the Krylov subspace is the smallest  $k$  such that  $\mathcal{K}_k(A|v)$  as a basis is linearly dependent, denoted as  $\text{grade}(A|v)$ .

The word grade of a Krylov Subspace is used in Y. Saad's work (**CITATION NEEDED**) in reference to Krylov Subspace. And once the grade is reached, the Krylov Subspace becomes an invariant subspace to the matrix  $A$ . For a proof, refers to [Krylov Subspace Grade Invariant Theorem \(6.1\)](#) in the appendix.

**Proposition 3.1** (When the Grade is Reached). Assuming that matrix  $A$  is diagonalizable,  $A = V\Lambda V^{-1}$ , then  $\text{grade}(A|u)$  is the number of unique  $\lambda_i$  such that  $(V^{-1}u)_i$  is non-zero.

*Proof.* Let  $k$  be the grade, then it's the minimum number such that the matrix polynomial of  $A$  multiplied by  $u$  equals to the zero vector, and it's nontrivial.

$$\mathbf{0} = \sum_{j=0}^k w_j A^j u \tag{3.1.1}$$

$$\mathbf{0} = V \sum_{j=0}^k w_j \Lambda^j V^{-1} u \tag{3.1.2}$$

$$\forall i \quad 0 = \sum_{j=0}^k w_j \lambda_i^j (V^{-1}u)_i \tag{3.1.3}$$

When non-trivial solution exists for some  $w_j \neq 0$ , then a polynomial will have to interpolate  $\lambda_i$  where  $(V^{-1}u)_i \neq 0$ . Therefore, the minimum  $k$  is the number of unique  $\lambda_i$  such that  $(V^{-1}u)_i \neq 0$   $\square$

### 3.1.2 Projectors

**Definition 3.** A matrix  $P$  is a projector when  $P^2 = P$ , we call this property idempotent.

There are 2 types of projectors, oblique and orthogonal projectors. A projector is orthogonal projector when it's Hermitian, it's oblique when it's not Hermitian.

**Proposition 3.2** (Projector Complementary). The projector  $I - P$  projects onto the null space of  $P$  and vice versa.

$$\text{ran}(P) = \text{null}(I - P) \quad (3.1.4)$$

$$\text{ran}(I - P) = \text{null}(P) \quad (3.1.5)$$

The proof is immediate from the definition. For more coverage of facts, refer to Trefethen's Book on Numerical Linear Algebra(CITATION HERE).

## 3.2 Subspace Projection Methods

Let  $\mathcal{K}, \mathcal{L}$  be 2 subspaces.  $\mathcal{K}$  can be viewed as a subspace where we choose our solutions for the  $x$  for the linear system  $Ax = b$ , and  $\mathcal{L}$  is a subspace where we orthogonalize our residual  $r = b - Ax$  against  $\mathcal{L}$ . This is a description of this framework:

$$\text{choose } \tilde{x} \in x_0 + \mathcal{K} \text{ s.t: } b - A\tilde{x} \perp \mathcal{L} \quad (3.2.1)$$

We look for a  $\tilde{x}$  in the affine linear subspace  $x_0 + \mathcal{K}$  such that it's perpendicular to the subspace  $\mathcal{L}$ , or, equivalently, minimizing the projection onto the subspace  $\mathcal{L}$ . The above conditions can be expressed using matrix.

$$\text{Let } V \in \mathbb{R}^{n \times m} \text{ be a basis for: } \mathcal{K} \quad (3.2.2)$$

$$\text{Let } W \in \mathbb{R}^{n \times m} \text{ be a basis for: } \mathcal{L} \quad (3.2.3)$$

Substituting the matrices to the above set of conditions:

$$\tilde{x} = x_0 + Vy \quad (3.2.4)$$

$$\text{choose } x \text{ s.t: } b - A\tilde{x} \perp \text{ran}(W) \quad (3.2.5)$$

$$\implies W^T(b - Ax_0 - AVy) = \mathbf{0} \quad (3.2.6)$$

$$W^T r_0 - W^T AVy = \mathbf{0} \quad (3.2.7)$$

$$W^T AVy = W^T r_0 \quad (3.2.8)$$

### 3.2.1 Prototype Algorithm

And from here, we can define a simple prototype algorithm using this framework.

While not converging :

Increase Span for:  $\mathcal{K}, \mathcal{L}$

Choose:  $V, W$  for  $\mathcal{K}, \mathcal{L}$

$$y := (W^T AV)^{-1} W^T r_0 \quad (3.2.9)$$

$$x := x + Vy$$

$$r := r_0 - AVy$$

Each time, we increase the span of the subspace  $\mathcal{K}, \mathcal{L}$ , which gives us more space to choose the solution  $x$ , and more space to reduce the residual vector  $r$ . This idea is incredibly flexible, and we will see in later part where it reduces to a more concrete algorithm. Finally, when  $\mathcal{K} = \mathcal{L}$ , this is referred to as a Petrov Galerkin's Conditions (**TODO CHECK THIS**).

**Remark 3.2.1** (Projector in the Prototype Algorithm). One may proceed to find a projector in the above prototype algorithm.

$$y = (W^T A v)^{-1} W^T r_0 \quad (3.2.10)$$

$$A V y = A V (W^T A V)^{-1} W^T r_0 \quad (3.2.11)$$

$$x = x_0 + A V y \quad (3.2.12)$$

$$b - A x = b - A x_0 - A V y \quad (3.2.13)$$

$$r = r_0 - A V (W^T A V)^{-1} W^T r_0 \quad (3.2.14)$$

In here,  $A V (W^T A V)^{-1} W^T$  is a projector, assuming the invertibility of  $W^T A V$ .

### 3.2.2 Energy Norm Minimization using Gradient

Other times, iterative method will choose to build up a subspace for each step with a subspace generator, and build up the solution on this expanding subspace, but with the additional objective of minimizing the residual under some norm. Assuming that the vector  $x \in x_0 + \mathcal{K}$ , and we want to minimize the residual under a norm induced by positive definite operator  $B$ . Let it be the case that the columns of matrix  $K$  span subspace  $\mathcal{K}$  with  $\dim(\mathcal{K}) = k$ , then one may consider using gradient as a more direct approach instead of projector.

$$\min_{x \in x_0 + \mathcal{K}} \|b - A x\|_B^2 \quad (3.2.15)$$

$$= \min_{w \in \mathbb{R}^k} \|b - A(x_0 + K w)\|_B^2 \quad (3.2.16)$$

$$= \min_{w \in \mathbb{R}^k} \|r_0 - A K w\|_B^2 \quad (3.2.17)$$

We take the derivative of it and set the derivative to zero to look for the minimum, skipping the proof that the derivative of  $\nabla_x [\frac{1}{2} \|x\|_A^2] = A x$  and just apply this formula for computing the gradient  $\nabla_x [f(Ax)] = A^T \nabla_x [f(x)]$  where  $f(x)$  is a mapping from  $\mathbb{R}^n$  to  $\mathbb{R}$ .

$$\nabla_w [\|r_0 - A K w\|_B^2] = \mathbf{0} \quad (3.2.18)$$

$$(A K)^T B (r_0 - A K w) = \mathbf{0} \quad (3.2.19)$$

$$(A K)^T B r_0 - (A K)^T B A K w = \mathbf{0} \quad (3.2.20)$$

$$(A K)^T B r_0 = (A K)^T B A K w \quad (3.2.21)$$

The above formulation is powerful. We used gradient instead of projector for the simplicity of the argument. One can derive the same using orthogonal projector to minimize the equivalent 2 norm of the energy norm, but the math is a bit more tedious. However, this minimization objective is minimizing the residual, which is fine for deriving subspace methods such as the GMRes, or the Minres and Orthomin, however, for the sake of the conjugate gradient, we have to consider the alternative. Let's this be a proposition that we proceed to prove.



**Proposition 3.3** (Conditions for Minimum Error Under Energy Norm). Here, we let matrix  $B$  be positive definite so that it can induce a norm, we let  $K$  be a matrix whose columns form a basis for  $\mathcal{K}$ , we let  $e_k$  denotes the error, given by:  $A^{-1}b - x_k$ , and we let  $r_k$  denotes the residual given as  $b - Ax_k$ .

$$\min_{x_k \in x_0 + \mathcal{K}} \|A^{-1}b - x\|_B^2 \iff K^T B e_0 - K^T B K w = \mathbf{0} \quad (3.2.22)$$

Next, we proceed to prove it and explain its interpretations and importance:

*Proof.*

$$\min_{x \in x_0 + \mathcal{K}} \|A^{-1}b - x\|_B^2 = \min_{x \in \mathbb{R}} \|A^{-1}b - x_0 - Kw\|_B^2 \quad (3.2.23)$$

$$= \min_{x \in \mathbb{R}^k} \|e_0 - Kw\|_B^2 \quad (3.2.24)$$

To attain the minimum of the norm, we take the derivative and set it to be zero, giving us:

$$\mathbf{0} = \nabla_w [\|e_0 - Kw\|_B^2] \quad (3.2.25)$$

$$= \nabla_w [e_0 - Kw]^T B (e_0 - Kw) \quad (3.2.26)$$

$$= 2K^T B (e_0 - Kw) \quad (3.2.27)$$

$$\implies K^T B e_0 - K^T B K w = \mathbf{0} \quad (3.2.28)$$

□

This conditions (3.2.28) is implicitly describing the objective of a Preconditioned Conjugate Gradient algorithm, where  $B$  is the  $E^{-1}AE^{-T}$  matrix with  $M = E^{-1}E^{-T}$  a symmetric positive definite matrix approximating the inverse of matrix  $A$ . Let's refrain from such a digression and instead let's set  $B$  to  $A$ , so that it's equivalent to the Energy Norm minimization of Conjugate Gradient, giving us this condition:

$$K^T A A^{-1} r_0 - K^T A K w = \mathbf{0} \quad (3.2.29)$$

$$K^T r_0 - K^T A K w = \mathbf{0} \quad (3.2.30)$$

Here, we just made the substitution of  $e_0 = A^{-1}r_0$ , and  $B = A$  to (3.2.28). Later, we will see how this condition is related to oblique Projector, similar to how an Orthogonal Projector is able to minimize the 2-Norm of the residual.

### 3.3 Deriving Conjugate Gradient from Conjugate Directions

By the time this is being written, it's been 70 years since the first time Conjugate Gradient algorithm is proposed by Hestenes and Stiefel back in 1952(**CITATION NEEDED**). Upon their first discussion of the algorithm, numerous perspectives were explored. Three of the most important ideas are using Conjugate Directions, minimizing the energy norm of the error of the linear system and coming up with an update of the conjugate vectors using the residual vector at the current iteration. Here, we use the exact same idea, but we diverge from Hestenes and Stiefel's approach in favor of using the oblique projector and the

subspace orthogonality conditions to derive it. The ideas are rehashed and in the end we point out its relations to Krylov Subspace, which ultimately, leads to other important new ideas related to solving symmetric indefinite systems that are not yet present back in 1952. In most time under classroom settings or textbook, the relations of Conjugate Gradient, Lanczos Iterations and Krylov Subspaces are discussed together to explain some of the more important properties of the algorithm so that we can move on and talk about other things. In this section of the paper, we take a different approach that is similar in spirit to a course note by Shewchuk (**CITATION NEEDED**) where we focus more on that inspirations behind the algorithm and avoid mentioning Lanczos Algorithms and Krylov Subspace until the very end, after CG is derived.

### 3.3.1 CG Objective and Framework

We introduce the algorithm as an attempt to minimize the energy norm of the error for a system of linear equations  $Ax = b$ , and we make the assumptions:

- 1) The matrix  $A$  is symmetric positive definite.
- 2) Assume another matrix  $P_k = [p_0 \ p_1 \ \cdots \ p_{k-1}]$  as a matrix whose columns is a basis for  $x_k$ .

Let's consider the following objective of minimizing the energy norm of the error over a subspace.

$$\min_{w \in \mathbb{R}^k} \|A^{-1}b - (x_0 + P_k w)\|_A^2 \iff P_k^T r_0 = P_k^T A P_k w \quad (3.3.1)$$

Refer back to [Energy Norm Minimization using Gradient \(3.2.2\)](#) for how to obtain the above minimization objective. Using the matrix from the [subspace projection method \(3.2\)](#) where  $W, V$  are both  $P_k$ , we reformulate the norm minimization conditions as:

$$\text{choose: } x \in x_0 + \text{ran}(P_k) \text{ s.t.: } b - Ax \perp \text{ran}(P_k) \quad (3.3.2)$$

Take note that the link between a norm minimization and an equivalent subspace orthogonality conditions don't guarantee to happen for other subspace projection methods, for example the FOM and Bi-Lanczos Methods are orthogonalizations method that doesn't link directly to a norm minimization objective (**CITATION NEEDED**).

To solve for  $w$ , we wish to make  $P_k^T A P_k$  to be an easy-to-solve matrix. Let the easy-to-solve matrix be a diagonal matrix and hence we let  $P_k$  be a *matrix whose columns are A-Orthogonal vectors*.

$$P_k^T A P_k = D_k \text{ where: } (D_k)_{i,i} = \langle p_{i-1}, A p_{i-1} \rangle \quad (3.3.3)$$

$$P_k r_0 = P_k^T A P_k w = D_k w \quad (3.3.4)$$

$$w = D_k^{-1} P_k^T r_0 \quad (3.3.5)$$

The idea here is: accumulate vectors  $p_j$  into the matrix  $P_k$  and then iteratively improve the solution  $x_k$  by reducing the error denote as  $e_k$  defined as  $A^{-1}b - x_k$ . Then, we derive the

following expression for  $x_k$  and the residual  $r_k = b - Ax_k$ :

$$\begin{cases} x_k = x_0 + P_k D_k^{-1} P_k^T r_0 \\ r_k = r_0 - AP_k D_k^{-1} P_k^T r_0 \\ P_k^T AP_k = D_k \end{cases} \quad (3.3.6)$$

Let this algorithm be the prototype.

### 3.3.2 Using the Projector

Here, we consider the above prototype algorithm. Please observe that  $AP_k D_k^{-1} P_k$  is a projector, and so is  $P_k D_k^{-1} P_k^T A$ .

*Proof.*

$$AP_k D_k^{-1} P_k^T (AP_k D_k^{-1} P_k^T) = AP_k D_k^{-1} P_k^T AP_k D_k^{-1} P_k^T \quad (3.3.7)$$

$$= AP_k D_k^{-1} D_k D_k^{-1} P_k^T \quad (3.3.8)$$

$$= AP_k D_k^{-1} P_k^T \quad (3.3.9)$$

$$P_k D_k^{-1} P_k^T A (P_k D_k^{-1} P_k^T A) = P_k D_k^{-1} D_k D_k^{-1} P_k^T A \quad (3.3.10)$$

$$= P_k D_k^{-1} P_k^T A \quad (3.3.11)$$

□

Both matrices are indeed projectors. Please take note that they are not Hermitian, which would mean that they are not orthogonal projector, hence, oblique projectors. For notational convenience, we denote  $\bar{P}_k = P_k D_k^{-1} P_k^T$ ; then these 2 projectors are:

$$AP_k D_k^{-1} P_k^T = A\bar{P}_k \quad (3.3.12)$$

$$P_k D_k^{-1} P_k^T A = \bar{P}_k A \quad (3.3.13)$$

One immediate consequence is:

$$\text{ran}(I - A\bar{P}_k) \perp \text{ran}(P_k) \quad (3.3.14)$$

$$\text{ran}(I - \bar{P}_k A) \perp \text{ran}(AP_k) \quad (3.3.15)$$

*Proof.*

$$P_k^T (I - A\bar{P}_k) = P_k^T - P_k^T A\bar{P}_k \quad (3.3.16)$$

$$= P_k^T - D_k D_k^{-1} P_k^T \quad (3.3.17)$$

$$= \mathbf{0} \quad (3.3.18)$$

$$(AP_k)^T (I - \bar{P}_k A) = P_k^T A - P_k^T A\bar{P}_k A \quad (3.3.19)$$

$$= P_k^T A - P_k^T AP_k D_k^{-1} P_k^T A \quad (3.3.20)$$

$$= P_k^T A - P_k^T A \quad (3.3.21)$$

$$= \mathbf{0} \quad (3.3.22)$$

□

Using the properties of the oblique projector, we can prove two facts about this simple norm minimization method we developed:

**Proposition 3.4** (Residuals are Orthogonal to  $P_k$ ).

$$r_k = r_0 - A\bar{P}_k r_0 = (I - A\bar{P}_k)r_0 \quad (3.3.23)$$

$$\implies r_k \perp \text{ran}(P_k) \quad (3.3.24)$$

**Proposition 3.5** (Generating A Orthogonal Vectors). Given any set of basis vector, for example  $\{u_k\}_{i=0}^{n-1}$ , one can generate a set of A-Orthogonal vectors from it. More specifically:

$$p_k = (I - \bar{P}_k A)u_k \quad (3.3.25)$$

$$\text{span}(p_k) \perp \text{ran}(AP_k) \quad (3.3.26)$$

For above propositions, we used the immediate consequence of the range of these oblique projectors.

### 3.3.3 Method of Conjugate Directions

So far, we have this particular scheme of solving the optimization problem, coupled with the way to computing the solution  $x_k$  at each step, and the residual at each step, while also getting the residual vector at each step too. However, it would be great if we can accumulate on the same subspace  $P_k$  and look for a chance to reuse the computational results from the previous iterations of the algorithm:

**Definition 4** (Method of Conjugate Directions).

$$\begin{cases} \bar{P}_k = P_k D_k^{-1} P_k^T \\ x_k = x_0 + \bar{P}_k r_0 \\ r_k = (I - A\bar{P}_k)r_0 \\ P_k^T A P_k = D_k \\ p_k = (I - \bar{P}_k A)u_k \quad \{u_i\}_{i=0}^{n-1} \text{ is a Basis} \end{cases} \quad (3.3.27)$$

With the assistance of a set of basis vector that span the whole space, this algorithm is possible to achieve the objective. Take note that we can accumulate the solution for  $x_k$  accumulatively, instead of computing the whole projector process, we have the choice to update it recursively as the newest  $p_k$  vector is introduced at that step. Let's Call this formulation of the algorithm: *Conjugate Direction Method* (CDM).

**Remark 3.3.1.** This CDM method is nothing new, in the original paper from Hestenes and Stiefel back in 1952(**CITATION NEEDED**), they commented on the method of Conjugate Direction, for each choice of basis  $\{u_i\}_{i=1}^n$  there resides a unique algorithm. If one were to choose the basis to be the set of standard basis vector, then the resulting algorithm will be the equivalent of a Gaussian Elimination.

**Remark 3.3.2** (Geometric Intuition of CDM). What is happening geometrically is that the A-Orthogonal vectors are orthogonal if described under the alternative eigenspace. Intuitively, one should think of a high-dimensional sphere that sits along some orthogonal basis,

and the transformation of  $A$  is stretching and rotating sphere, along with the orthogonal axis, resulting in a new ellipsoid in a different orientation; when the transformation is applied, the orthogonal coordinate inside the sphere got stretched along with it, and now these axes had become A-orthogonal vectors. Tracing along the direction of these vectors will ensure minimum redundancy of search directions.

### 3.3.4 Properties of CDM

Here we set up several useful lemma and propositions that can derive the short recurrences of A-Orthogonal vectors

**Proposition 3.6.**

$$p_{k+j}^T r_k = p_{k+j}^T r_0 \quad \forall 0 \leq j \leq n - k \quad (3.3.28)$$

*Proof.*

$$p_{k+j}^T r_k = p_{k+j}^T (I - A\bar{P}_k) r_0 \quad (3.3.29)$$

$$= (p_{k+j}^T - p_{k+j}^T A\bar{P}_k) r_0 \quad (3.3.30)$$

$$= p_{k+j}^T r_0 \quad (3.3.31)$$

This is true because the vector  $p_{k+j}$ , a conjugate vector in the future is orthogonal to all previous conjugate vectors.  $\square$

As a consequence of the above proposition, we can obtain a short recurrence for the residuals and the solution  $x$  of CDM:

**Proposition 3.7** (CDM Recurrence).

$$r_k - r_{k-1} = r_0 - A\bar{P}_k r_0 - (r_0 - A\bar{P}_{k-1} r_0) \quad (3.3.32)$$

$$= A\bar{P}_k r_0 - A\bar{P}_{k-1} r_0 \quad (3.3.33)$$

$$= -A p_{k-1} \frac{\langle p_{k-1}, r_0 \rangle}{\langle p_{k-1}, A p_{k-1} \rangle} \quad (3.3.34)$$

$$\implies x_k - x_{k-1} = p_{k-1} \frac{\langle p_{k-1}, r_0 \rangle}{\langle p_{k-1}, A p_{k-1} \rangle} \quad (3.3.35)$$

$$\text{def: } a_{k-1} := \frac{\langle p_{k-1}, r_0 \rangle}{\langle p_{k-1}, A p_{k-1} \rangle} = \frac{\langle p_{k-1}, r_{k-1} \rangle}{\langle p_{k-1}, A p_{k-1} \rangle} \quad (3.3.36)$$

On (3.3.36) we used proposition 3.6. We define the value of  $a_{k-1}$ , and in above, we have 2 equivalent representation. Please take note that, this proposition remains true for the future CG algorithm that we are going to develop.

### 3.3.5 Conjugate Gradient

Now, consider the case where, the set of basis vector:  $\{u\}_{i=0}^{n-1}$  to be the residual vector generated from the CDM itself. Then there are a series of new added lemmas that are true. However, this is where things started to get exciting, because a short recurrence for  $p_k$  during each iteration arised and residuals are all orthogonal. We wish to proceed to prove that part.

**Lemma 3.3.1.**

$$\langle p_{k+j}, Ap_k \rangle = \langle r_k, Ap_{k+j} \rangle = \langle p_{k+j}, Ar_k \rangle \quad \forall 0 \leq j \leq n-k \quad (3.3.37)$$

*Proof.*

$$p_{k+j}Ap_k = p_{k+j}^T Ar_k - p_{k+j}^T A\bar{P}_k Ar_k \quad \forall 0 \leq j \leq n-k \quad (3.3.38)$$

$$= p_{k+j}^T Ar_k \quad (3.3.39)$$

$$\langle p_{k+j}, Ap_k \rangle = \langle r_k, Ap_{k+j} \rangle = \langle p_{k+j}, Ar_k \rangle \quad (3.3.40)$$

On the first line we invoked the CDM algorithms definition back in (3.3.27), replacing  $u_k$  into  $r_k$ , hence  $p_k = (I - \bar{P}_k A)r_k$ , which is then substituted to on line (3.3.38).  $\square$

**Lemma 3.3.2.**

$$\langle r_k, p_k \rangle = \langle r_k, r_k \rangle \quad (3.3.41)$$

*Proof.*

$$\langle r_k, p_k \rangle \quad (3.3.42)$$

$$= \langle r_k, r_k \rangle - \langle r_k, \bar{P}_k Ar_k \rangle \quad (3.3.43)$$

$$= \langle r_k, r_k \rangle \quad (3.3.44)$$

From (3.3.42) to (3.3.43) we used  $p_k = (I - \bar{P}_k A)r_k$ , in addition we also make use of [Residual are Orthogonal \(3.4\)](#) from (3.3.43) to (3.3.44).  $\square$

**Proposition 3.8** (CG Generates Orthogonal Residuals).

$$\langle r_k, r_j \rangle = 0 \quad \forall 0 \leq j \leq k-1 \quad (3.3.45)$$

Let this above claim be inductively true then consider the following proof:

*Proof.*

$$r_{k+1} = r_k - a_k Ap_k \quad (3.3.46)$$

$$\implies \langle r_{k+1}, r_k \rangle = \langle r_k, r_k \rangle - a_k \langle r_k, Ap_k \rangle \quad (3.3.47)$$

$$= \langle r_k, r_k \rangle - \frac{\langle r_k, r_k \rangle}{\langle p_k, Ap_k \rangle} \langle r_k, Ap_k \rangle \quad (3.3.48)$$

$$= 0 \quad (3.3.49)$$

The first line is from the recurrence of CDM residuals, and then next we make use of  $a_k$  from [\(CDM Recurrence \(3.7\)\)](#) together with [Lemma 3.3.1](#). Next we consider:

$$p_j = (I - \bar{P}_j A)r_j \quad \forall 0 \leq j \leq k-1 \quad (3.3.50)$$

$$\implies r_j = p_j + \bar{P}_j Ar_j \quad (3.3.51)$$

$$r_k = (I - A\bar{P}_k)r_0 \quad (3.3.52)$$

$$r_k \perp \text{ran}(P_k) \implies \langle r_k, r_j \rangle = \langle r_k, p_j + \bar{P}_j Ar_j \rangle = 0 \quad (3.3.53)$$

The second line(3.3.51) is a result of the first line (3.3.50) rearranged. Here we again make use of the projector  $I - A\bar{P}_k$ . The last line (3.3.53) is using the second line 3.3.51. The base case of the argument is simple, because  $p_0 = r_0$ , and by the property of the projector,  $\langle r_1, r_0 \rangle = 0$ . The theorem is now proven.  $\square$

**Proposition 3.9** (CG Recurrences).

$$p_k = r_k + b_{k-1}p_{k-1} \quad b_{k-1} = \frac{\|r_k\|_2^2}{\|r_{k-1}\|_2^2} \quad (3.3.54)$$

*Proof.* The proof is direct and we start with the definition of CDM, which is given as:

$$p_k = (I - \bar{P}_k A)r_k \quad (3.3.55)$$

$$r_k - \bar{P}_k A r_k = r_k - P_k D_k^{-1} P_k^T A r_k \quad (3.3.56)$$

$$= r_k - P_k D_k^{-1} (A P_k)^T r_k \quad (3.3.57)$$

Observe:

$$(A P_k)^T r_k = \begin{bmatrix} \langle p_0, A r_k \rangle \\ \langle p_1, A r_k \rangle \\ \vdots \\ \langle p_{k-1}, A r_k \rangle \end{bmatrix} \quad (3.3.58)$$

Next, we can make use of [lemma 3.3.1](#) to get rid of  $A r_k$ :

$$\langle p_j, A r_k \rangle \quad \forall 0 \leq j \leq k-2 \quad (3.3.59)$$

$$\langle p_j, A r_k \rangle = \langle r_k, A p_j \rangle \quad (3.3.60)$$

$$= \langle r_k, a_j^{-1} (r_j - r_{j+1}) \rangle \quad (3.3.61)$$

$$= a_j^{-1} \langle r_k, (r_j - r_{j+1}) \rangle = 0 \quad (3.3.62)$$

The second line is also using the property that the matrix  $A$  is symmetric, the third line is using the recurrence of the residual established for CDM ([CDM Recurrences \(Proposition 3.7\)](#)), and the last line is true for all  $0 \leq j \leq k-2$  by the orthogonality of the residual proved in [CG Generates Orthogonal Residuals \(Proposition 3.8\)](#). Therefore we have:

$$(A P_k)^T r_k = \begin{bmatrix} \langle p_0, A r_k \rangle \\ \langle p_1, A r_k \rangle \\ \vdots \\ \langle p_{k-1}, A r_k \rangle \end{bmatrix} = a_{k-1}^{-1} \langle r_k, (r_{k-1} - r_k) \rangle \xi_k \quad (3.3.63)$$

Take note that the vector  $\xi_k$  is the  $k$  th standard basis vector in  $\mathbb{R}^k$ , keep in mind that  $r_k \perp r_{k-1}$  as well. Using these facts we can simplify the expression for  $p_k$  into:

$$p_k = r_k - P_k D_k^{-1} (A P_k)^T r_k \quad (3.3.64)$$

$$= r_k - P_k D_k^{-1} a_{k-1}^{-1} (\langle r_k, (r_{k-1} - r_k) \rangle) \xi_k \quad (3.3.65)$$

$$= r_k - \frac{a_{k-1}^{-1} \langle -r_k, r_k \rangle}{\langle p_{k-1}, A p_{k-1} \rangle} p_k \quad (3.3.66)$$

$$= r_k + \frac{a_{k-1}^{-1} \langle r_k, r_k \rangle}{\langle p_{k-1}, A p_{k-1} \rangle} p_k \quad (3.3.67)$$

$$= r_k + \left( \frac{\langle r_{k-1}, r_{k-1} \rangle}{\langle p_{k-1}, A p_{k-1} \rangle} \right)^{-1} \frac{\langle r_k, r_k \rangle}{\langle p_{k-1}, A p_{k-1} \rangle} p_k \quad (3.3.68)$$

$$= r_k + \frac{\langle r_k, r_k \rangle}{\langle r_{k-1}, r_{k-1} \rangle} p_k \quad (3.3.69)$$

We make use of the definition for  $a_{k-1}$  for the CDM algorithm ([proposition 3.7](#) together with [lemma 3.3.2](#)). At this point, we have proven the short CG recurrences for  $p_k$ .  $\square$

Up until this point we have proven the usual of conjugate gradient proposed by Hestenes & Stiefel (**CITATION NEEDED**), we started with the minimization objective and the properties of  $P_k$ , then we define a recurrence for the residual (Simultaneously the solution  $x_k$ ), and the A-Orthogonal vectors using a basis as assistance for the generations process. Next, we make the key changes of the assistance basis, making it equal to the set of residuals vector generated from the algorithm itself; after some proofs, we uncovered the exact same parameters found in most of the definitions of the CG algorithm:

**Definition 5** (CG).

$$p^{(0)} = b - Ax^{(0)} \quad (3.3.70)$$

$$\text{For } i = 0, 1, \dots \quad (3.3.71)$$

$$\begin{aligned} a_i &= \frac{\|r^{(i)}\|^2}{\|p^{(i)}\|_A^2} \\ x^{(i+1)} &= x^{(i)} + a_i p^{(i)} \\ r^{(i+1)} &= r^{(i)} - a_i A p^{(i)} \\ b_i &= \frac{\|r^{(j+1)}\|_2^2}{\|r^{(i)}\|_2^2} \\ p^{(i+1)} &= r^{(i+1)} + b_i p^{(i)} \end{aligned} \quad (3.3.72)$$

That is the algorithm, stated with all the iteration number listed as a super script inside parenthesis. Which is equivalent to what we have proven for the CG.

### 3.3.6 CG and Krylov Subspace

The conjugate Gradient Algorithm is actually a CG, a special case of the algorithm we derived at the start of the excerpt. The full algorithm can be seemed by the short recurrence for the residual and the conjugation vector. This part is trivial. Next, we want to show the relations to the Krylov Subspace, which only occurs for the CG algorithm.

**Proposition 3.10.**

$$p_k \in \mathcal{K}_{k+1}(A|r_0) \quad (3.3.73)$$

$$r_k \in \mathcal{K}_{k+1}(A|r_0) \quad (3.3.74)$$

*Proof.* The base case is trivial and it's directly true from the definition of CG:  $r_0 \in \mathcal{K}_1(A|r_0)$ ,  $p_0 = r_0 \in \mathcal{K}_1(A|r_0)$ . Next, we inductively assume that  $r_k \in \mathcal{K}_{k+1}(A|r_0)$ ,  $p_k \in \mathcal{K}_{k+1}(A|r_0)$ , then we consider:

$$r_{k+1} = r_k - a_k A p_k \quad (3.3.75)$$

$$\in r_k + A \mathcal{K}_{k+1}(A|r_0) \quad (3.3.76)$$

$$\in r_k + \mathcal{K}_{k+2}(A|r_0) \quad (3.3.77)$$

$$r_k \in \mathcal{K}_{k+1}(A|r_0) \subseteq \mathcal{K}_{k+2}(A|r_0) \quad (3.3.78)$$

$$\implies r_{k+1} \in \mathcal{K}_{k+2}(A|r_0) \quad (3.3.79)$$



At the same time the update of  $p_k$  would assert the property that:

$$p_{k+1} = r_{k+1} + b_k p_k \quad (3.3.80)$$

$$\in r_{k+1} + \mathcal{K}_{k+1}(A|r_0) \quad (3.3.81)$$

$$\in \mathcal{K}_{k+2}(A|r_0) \quad (3.3.82)$$

This is true because  $r_{k+1}$  is already a member of the expanded subspace  $\mathcal{K}_{k+2}(A|r_0)$ . And from this formulation of the algorithm, we can update the Petrov Galerkin's Conditions to be:

**Theorem 1** (CG and Krylov Subspace).

$$\text{choose: } x_k \in x_0 + \mathcal{K}_k(A|r_0) \text{ s.t. } r_k \perp \mathcal{K}_k(A|r_0) \quad (3.3.83)$$

Take note that,  $\text{ran}(P_k) = \mathcal{K}_k(A|r_0)$  because the index starts with zero. The above formulations gives theoretical importance for the Conjugate Gradient Algorithm.  $\square$

## 3.4 Arnoldi Iterations and Lanczos

In this section, we introduce another important algorithm: The Lanczos Algorithm. However, to give more context for the discussion, the Arnoldi iteration is considered as well and it's used to emphasize that Lanczos Iterations is just Arnoldi but with the matrix  $A$  being a symmetric matrix. Finally we make the link between Lanczos Iterations and Krylov Subspace, which will inevitably linked back to CG and plays an important role for the analysis of CG.

### 3.4.1 The Arnoldi Iterations

We first define the Arnoldi Algorithm, and then we proceed to derive it using the idea of orthogonal projector. Next, we discuss a special case of the Arnoldi Iteration: the Lanczos Algorithm, which is just Arnoldi applied to a symmetric matrix. And such algorithm will inherit the properties of the Arnoldi Iterations.

Before stating the algorithm, I would like to point out the interpretations of the algorithm and its relations to Krylov Subspace. Consider a matrix of Hessenberg Form:

$$\tilde{H}_k = \begin{bmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,k} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,k} \\ & \ddots & & \vdots \\ & & h_{k,k-1} & h_{k,k} \\ & & & h_{k+1,k} \end{bmatrix} \quad (3.4.1)$$

We initialize the orthogonal projector with the vector  $q_1$ , which is  $q_1 q_1^H$ , next, we apply the linear operator  $A$  on the current range of the projector:  $Aq_1$ , then, we orthogonalize it against  $q$ . Let the projection of  $Aq_1$  onto  $I - q_1 q_1^H$  be  $h_{1,2} q_2$ , and let the projection onto  $q_1 q_1^H$

be  $h_{1,1}$ . This completes the first column of  $H_k$ , we do this recursively. Please allow me to demonstrate:

$$(\tilde{H}_k)_{2,1}q_2 = (I - q_1q_1^H)Aq_1 \quad (3.4.2)$$

$$(\tilde{H}_k)_{1,1}q_1 = q_1q_1^H Aq_1 \quad (3.4.3)$$

$$Q_2 := [q_1 \quad q_2] \quad (3.4.4)$$

$$(\tilde{H}_k)_{3,2}q_3 = (I - Q_2Q_2^H)Aq_2 \quad (3.4.5)$$

$$(\tilde{H}_k)_{1:2,2} = Q_2Q_2^H Aq_2 \quad (3.4.6)$$

$$Q_3 := [q_1 \quad q_2 \quad q_3] \quad (3.4.7)$$

$$\vdots \quad (3.4.8)$$

$$Q_j := [q_1 \quad q_2 \quad \cdots \quad q_j] \quad (3.4.9)$$

$$(\tilde{H}_k)_{j+1,j}q_{j+1} = (I - Q_jQ_j^H)Aq_j \quad (3.4.10)$$

$$(\tilde{H}_k)_{1:j,j} = Q_jQ_j^H Aq_j \quad (3.4.11)$$

$$\vdots \quad (3.4.12)$$

$$Q_k := [q_1 \quad q_2 \quad \cdots \quad q_k] \quad (3.4.13)$$

$$(\tilde{H}_k)_{k+1,k}q_{k+1} = (I - Q_kQ_k^H)Aq_k \quad (3.4.14)$$

$$(\tilde{H}_k)_{1:k,k} = Q_kQ_k^H Aq_k \quad (3.4.15)$$

Reader please observe that  $Q_k$  is going to be orthogonal because how at the start,  $q_1q_1^H$  and  $I - q_1q_1^H$  are orthogonal projectors. As a consequence, we can express the recurrences of the subspace vector in matrix form:

$$AQ_k = Q_{k+1}\tilde{H}_k \quad (3.4.16)$$

$$Q_k^H AQ_k = H_k \quad (3.4.17)$$

And here, we explicitly define  $H_k$  to be the principal submatrix of  $\tilde{H}_k$ . Reader please immediately observe that, if  $A$  is symmetric, then it has to be the case that  $Q_k^H AQ_k$  is also symmetric, which will make  $H_k$  to be symmetric as well, which implies that  $H_k$  will be a symmetric tridiagonal Matrix. And under that assumption, we can develop the Lanczos Algorithm. Instead of orthogonalizing against all previous vectors, we have the option to simply orthogonalize against the previous  $q_k, q_{k-1}$  vector. And we can reuse the sub-diagonal elements for  $q_{k-1}$ ; giving us the Lanczos Algorithm.

### 3.4.2 Arnoldi Produces Orthogonal Basis for Krylov Subspace

One important observation reader should make about the idea of Arnoldi Iteration is that, during each iteration, the matrix  $Q_k$  spans the same range as  $\mathcal{K}_k(A|q_1)$ .

**Proposition 3.11.**

$$\text{ran}(Q_k) = \mathcal{K}_k(A|q_1) \quad (3.4.18)$$

*Proof.* The base case is simple:  $q_1 \in \mathcal{K}_1(A|q_1)$ , inductively assuming the proposition is true, using the polynomial property of Krylov Subspace we consider:

$$\begin{aligned}
& Q_k \in \mathcal{K}_k(A|q_1) \\
\iff & w_k^+ : \exists p_k(A|w_k^+)q_1 = q_k \\
& \implies Aq_k = Ap_k(A|w_k^+)q_1 \in \mathcal{K}_{k+1}(A|w_k^+) \\
& q_{k+1} \in \mathcal{K}_{k+1}(A|q_1) \\
& \implies \text{ran}(Q_{k+1}) = \mathcal{K}_{k+1}(A|q_1)
\end{aligned}$$

The Arnoldi Algorithm terminates if the value  $h_{k+1,k}$  is set to be zero. This is the case because the normalization process is dividing by  $h_{k+1,k}$  to get  $q_{k+1}$ . This only happens when  $Aq_k \in \text{ran}(Q_k)$ ; because  $h_{k+1,k}$  is given by the projector of  $I - Q_k Q_k^H$  applied to  $Aq_k$  and the null space of this projector is  $\text{ran}(Q_k)$ , resulting in  $h_{k+1,k} = 0$ .  $\square$

**Remark 3.4.1** (Arnoldi Produces Minimal Monic in Krylov Subspace). The characteristic polynomial of  $H$ , minimizes  $\|p(A|w)q_1\|_2$ . For more information, Trefethen has a coverage on the topic in his works(**CITATION NEEDED**). The minimization property in Arnoldi translate to Lanczos Iterations as well.

### 3.4.3 The Lanczos Iterations

**Definition 6** (Lanczos Iterations).

$$\text{Given arbitrary: } q_1 \text{ s.t: } \|q_1\| = 1 \quad (3.4.19)$$

$$\text{set: } \beta_0 = 0 \quad (3.4.20)$$

$$\text{For } j = 1, 2, \dots \quad (3.4.21)$$

$$\tilde{q}_{j+1} := Aq_j - \beta_{j-1}q_{j-1}$$

$$\alpha_j := \langle q_j, \tilde{q}_{j+1} \rangle$$

$$\tilde{q}_{j+1} \leftarrow \tilde{q}_{j+1} - \alpha_j q_j \quad (3.4.22)$$

$$\beta_j = \|\tilde{q}_{j+1}\|$$

$$q_{j+1} := \tilde{q}_{j+1}/\beta_j$$

Here, let it be the case that  $H_k$  is a symmetric tridiagonal Matrix with  $\alpha_i$  on the diagonal,  $\beta_i$  on the sub and super diagonal; the Lanczos is Arnoldi, but we make use of the symmetric properties to orthogonalize  $Aq_j$  against  $q_{j-1}$  using  $\beta_{j-1}$ , and in this case, each iteration only consists of one vector inner product. Note that another equivalent algorithm where I tweaked

it to handle the base case of  $T_k$  being a  $1 \times 1$  matrix can be phrased in the following way:

$$\begin{aligned}
&\text{Given arbitrary: } q_1 \text{ s.t: } \|q_1\| = 1 \\
&\alpha_1 := \langle q_1, Aq_1 \rangle \\
&\beta_0 := 0 \\
&\text{Memorize : } Aq_1 \\
&\text{For } j = 1, 2, \dots \\
&\quad \tilde{q}_{j+1} := Aq_j - \beta_{j-1}q_{j-1} \tag{3.4.23} \\
&\quad \tilde{q}_{j+1} \leftarrow \tilde{q}_{j+1} - \alpha_j q_j \\
&\quad \beta_j = \|\tilde{q}_{j+1}\| \\
&\quad q_{j+1} := \tilde{q}_{j+1}/\beta_j \\
&\quad \alpha_{j+1} := \langle q_{j+1}, Aq_{j+1} \rangle \\
&\quad \text{Memorize: } Aq_{j+1}
\end{aligned}$$

The algorithm generates the following 2 matrices,  $Q_k$  which is orthogonal and it spans  $\mathcal{K}_k(A|q_1)$ , and a symmetric tridiagonal Matrix:

$$Q_k = [q_1 \quad q_2 \quad \dots \quad q_k] \tag{3.4.24}$$

$$T_k = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_{k-1} \\ & & \beta_{k-1} & \alpha_k \end{bmatrix} \tag{3.4.25}$$

Similar to the recurrence from the Arnoldi Algorithm, the Lanczos also create a recurrence between  $Aq_k$  and  $Q_k$  and  $q_{k+1}$ , but the recurrence is shorter sothat it simply makes use of the previous 2 vectors:

**Theorem 2** (Lanczos Recurrences).

$$AQ_k = Q_k T_k + \beta_k q_{k+1} \xi_k^T = Q_{k+1} \tilde{T}_k \tag{3.4.26}$$

$$\implies Aq_k = \beta_{j-1}q_{j-1} + \alpha_j q_j + \beta_j q_{j+1} \quad \forall 2 \leq j \leq k \tag{3.4.27}$$

$$\implies Aq_1 = \alpha_1 q_1 + \beta_1 q_2 \tag{3.4.28}$$

Often time, we refer the  $k \times k$  symmetric tridiagonal matrix generated from Iterative Lanczos as  $T_k$ . Finally; I wish to make the following important remark about the algorithm for later use. Given a matrix  $A$  and an initial vector  $q_1$ , The Lanczos algorithm produces an irreducible symmetric tridiagonal Matrix that has unique eigenvalues. The proof for the fact that any symmetric tridiagonal Matrices with Non-zeros on the sub/super diagonal must have unique non-zero eigenvalues is skipped. What we can immediately show here is the fact that Lanczos Algorithm will produce such a matrix.

**Proposition 3.12** (Lanczos Termination Conditions). The Lanczos Iteration produces a symmetric tridiagonal Matrix that has no zero element on its super and sub-diagonal, and if

$\beta_k$  is zero, then the algorithm must terminate, and  $k$  would equal to  $\text{grade}(A|q_1)$ , the grade of the Krylov Subspace.

*Proof.* It's true because the  $\beta_k$  in the Lanczos is equivalent to  $h_{k+1,k}$ . It's been discussed previously that if  $h_{k+1,1} = 0$  for the Arnoldi's Iteration, then the Krylov Subspace  $\mathcal{K}_k(A|q_1)$  became an invariant subspace under  $A$ , and in that sense, the algorithm has to terminate due to a divide by zero error.  $\square$

**Remark 3.4.2** (Minimal Polynomial from Lanczos Iterations). The characteristic polynomial of  $T_k$  has a special minimization property. Here recall [remark 3.4.1](#), we make use of the minimization property of the characteristic polynomial of the Hessenberg matrix from the Arnoldi Iterations. Under Lanczos iterations the matrix  $H_k$  becomes the tridiagonal  $T_k$ . Since matrix  $A$  is symmetric, we consider its eigendecomposition in the form:  $A = V\Lambda V^T$ , we let  $\bar{p}_k(x)$  denote the characteristic polynomial of matrix  $T_k$ , then using the 2 norm minimization properties we have:

$$\min_{p_k: \text{monic}} \|p(A)q_1\|_2 \quad (3.4.29)$$

$$= \|\bar{p}_k(A)q_1\|_2 \quad (3.4.30)$$

$$= \|V\bar{p}_k(\Lambda)V^Tq_1\|_2 \quad (3.4.31)$$

$$= \|\bar{p}_k(\Lambda)V^Tq_1\|_2 \quad (3.4.32)$$

$$= \min_{i=1, \dots, n} \|\bar{p}_n(\lambda_i)(V^Tq_1)_i\|_2 \quad (3.4.33)$$

The last line is saying the conditions that the characteristic polynomial for  $T_k$  from the Lanczos iterations is minimizing the weighted sum at the eigenvalues of the matrix  $A$ . A direct consequence of this is a more refined terminations conditions than [proposition 3.12](#) for the Lanczos iterations, giving the information for initial vector  $q_1$ . Under exact arithmetic, the number of iterations underwent by Lanczos equals to the unique number of eigenvalues  $\lambda_i$  where  $(V^Tq_1)_i \neq 0$ .

## 4 Analysis of Conjugate Gradient and Lanczos Iterations

In this section, we state the termination conditions for the Lanczos iterations and the CG algorithm we developed using the property of Krylov Subspace. This is just a thorough discussion about these 2 algorithm and applying the foundations and it's not following any references.

### 4.1 Conjugate Gradient and Matrix Polynomial

One important result of the optimization objective listed [CG and Krylov Subspace](#) is the connections to matrix polynomial of  $A$  and Conjugate Gradient. More specifically we consider the following proposition:

**Proposition 4.1** (CG Relative Energy Error).

$$x_k \in \mathcal{K}(A|r_0)w + x_0 \quad (4.1.1)$$

$$\frac{\|e_k\|_A^2}{\|e_0\|_A^2} = \min_{w \in \mathbb{R}^k} \|(1 + Ap_k(A|w))A^{1/2}e_0\|_2^2 \leq \min_{p_k: p_k(0)=1} \max_{x \in [\lambda_{\min}, \lambda_{\max}]} |p_k(x)| \quad (4.1.2)$$

Here we use the notation  $e_k = A^{-1}b - x_k$  to denotes the error vector.

*Proof.*

$$\|e_k\|_A^2 = \min_{x_k \in x_0 + \mathcal{K}_k(A|r_0)} \|x^+ - x_k\|_A^2 \quad (4.1.3)$$

$$x_k \in x_0 + \mathcal{K}_k(A|r_0) \implies e_k = e_0 + p_{k-1}(A|w)r_0 \quad (4.1.4)$$

$$\implies = \min_{w \in \mathbb{R}^k} \|e_0 + p_{k-1}(A|w)r_0\|_A^2 \quad (4.1.5)$$

$$= \min_{w \in \mathbb{R}^k} \|e_0 + Ap_{k-1}(A|w)e_0\|_A^2 \quad (4.1.6)$$

$$= \min_{w \in \mathbb{R}^k} \|A^{1/2}(I + Ap_{k-1}(A|w))e_0\|_2^2 \quad (4.1.7)$$

$$\leq \min_{w \in \mathbb{R}^k} \|I + Ap_{k-1}(A|w)\|_2^2 \|e_0\|_A^2 \quad (4.1.8)$$

$$= \min_{w \in \mathbb{R}^k} \left( \max_{i=1, \dots, n} |1 + \lambda_i p_{k-1}(\lambda_i|w)|^2 \right) \|e_0\|_A^2 \quad (4.1.9)$$

$$\leq \min_{w \in \mathbb{R}^k} \left( \max_{x \in [\lambda_{\min}, \lambda_{\max}]} |1 + \lambda_i p_{k-1}(\lambda_i|w)|^2 \right) \|e_0\|_A^2 \quad (4.1.10)$$

$$= \min_{p_k: p_k(0)=1} \max_{x \in [\lambda_{\min}, \lambda_{\max}]} |p_k(x)|^2 \|e_0\|_A^2 \quad (4.1.11)$$

$$\implies \frac{\|e_k\|_A}{\|e_0\|_A} \leq \min_{p_k: p_k(0)=1} \max_{x \in [\lambda_{\min}, \lambda_{\max}]} |p_k(x)| \quad (4.1.12)$$

(4.1.3) is the Error Energy norm minimization objective of CG, we proceed with writing up the affine subspace where  $x_k$  is from:  $x_0 + \mathcal{K}_k(A|r_0)$  at (4.1.4), putting Krylov subspace in terms of a matrix polynomial multiplied by  $r_0$  and then use  $A^{-1}b$  to subtract both side to get the expression for  $e_k$ . From the (4.15) line to the (4.16), we use the fact that  $r_0 = Ae_0$ , allowing us to extract out a factor  $A$ .

Next, from (4.1.6) to (5.1.7), we use the fact that every symmetric definite matrix  $A$  has the factorization of  $A^{1/2}A^{1/2}$  where  $A^{1/2}$  is also a symmetric definite matrix. After that we moved the  $A^{1/2}$  to  $e_0$  to get  $\|e_0\|_A^2$  from (4.1.7) to (4.1.8), the matrix polynomial part is left with the 2-norm. From (4.1.8) to (4.1.9) we use the eigendecomposition of  $A$  which is diagonalizable with a unitary transform, giving us the form of  $Q\Lambda Q^T = A$  where  $Q$  is an Unitary Matrix and diagonals of  $\Lambda$  are the eigenvalues of  $A$ . Allow me to explain:

$$\|I + Ap_{k-1}(A|w)\|_2^2 = \|Q(I + \Lambda p_{k-1}(\Lambda|w))Q^T\|_2^2 \quad (4.1.13)$$

$$= \|I + \Lambda p_{k-1}(\Lambda|w)\|_2^2 \quad (4.1.14)$$

$$= \max_{i=1, \dots, n} |1 + \lambda_i p_{k-1}(\lambda_i|w)|^2 \quad (4.1.15)$$

Where, the 2-norm of a diagonal matrix  $\Lambda$  is just its biggest diagonal element. And then we relax the conditions for  $\lambda_i$  by reducing it to be some element in the interval between the

minimum and the maximum of the eigenvalues for the matrix  $A$  (from (4.1.9) to (4.1.10)). Finally, please notice that we use an monic  $p_{k+1}(x)$  at the end to simplify things.  $\square$

The above results will be useful for proving the convergence of CG.

**Remark 4.1.1** (The CG Relative Energy Error Norm is Tight). The bound is tight refers to the fact that the matrix is SPD, therefore  $\|Ax\| \leq \|A\|\|x\|$  is tight, which justifies that (4.1.10) is tight in the sense that for any iteration  $k$ , we can choose an initial vector  $e_0$  such that the equality is achieved. (4.1.12) can still be tight if we have the freedom to choose the eigenvalues of the matrix  $A$ . However, the bound is rarely tight if the initial error vector  $e_0$  and the matrix  $A$  is fixed.

#### 4.1.1 Termination Conditions of CG

**Proposition 4.2** (Termination of CG). For all initial guess, the maximum iterations underwent by the CG algorithm is the number of unique eigenvalues for the matrix  $A$ .

This result is direct from (4.1.8), the CG algorithm terminates when a polynomial that interpolates all the unique eigenvalue is found. This bound is true for all initial guess, and sometimes for some given  $e_0$ , the terminations can come with fewer iterations.

## 4.2 Convergence Rate of CG under Exact Arithmetic

In this section we make heavy use of Greenbaum's Analysis for convergence rate of the algorithm. The core idea is to use a Chebyshev Polynomial to establish a bound and it's applicable when the linear operator has extremely high dimension and we limit the number of iterations to  $k$  where  $k$  is much smaller than  $n$ , the size of the matrix. We will follow Greenbaum's Analysis but with some more details.

#### 4.2.1 Uniformly Distributed Eigenvalues

**Theorem 3** (CG Convergence Rate). The relative error squared measured over the energy norm is bounded by:

$$\frac{\|e_k\|_A}{\|e_0\|_A} \leq 2 \left( \frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1} \right)^k \quad (4.2.1)$$

Where  $k$  is the number of iterations, and  $e_k = A^{-1}b - x_k$ , the upper bound is the most general and it's able to bound the convergence given  $\lambda_{\min}, \lambda_{\max}$  of the operator  $A$ . The bound is loose if there are some kind of clustering of the eigenvalue of matrix  $A$ , and the bound would be tighter given that  $k \ll n$  and the eigenvalues of  $A$  are evenly spread out on the spectrum.

Before the proof, I need to point out the analysis draws inspiration from the interpolating mononic polynomial for the spectrum of the matrix  $A$ , and we make use of the Inf Norm minimization property of the Chebyshev Polynomial. Here, we order all the eigenvalues of matrix  $A$  so that  $\lambda_1, \lambda_n$  denotes the maximum and the minimum eigenvalues for  $A$ .

*Proof.* We start by adapting the Chebyshev Polynomial to the convex hull of the spectrum for matrix  $A$ , while also making it monic:

$$T_k(x) = \arg \min_{p \in \mathcal{P}_k} \max_{x \in [-1, 1]} |p(x)| \quad (4.2.2)$$

$$p_k(x) := \frac{T_k(\varphi(x))}{T_k(\varphi(0))} \quad \text{where: } \varphi(x) := \frac{2x - \lambda_1 - \lambda_n}{\lambda_n - \lambda_1} \quad (4.2.3)$$

$$\implies p_k(x) = \arg \min_{\substack{p \in \mathcal{P}_k \\ \text{s.t.: } p(0)=1}} \max_{x \in [\lambda_1, \lambda_n]} |p(x)| \quad (4.2.4)$$

At this point, we have defined a new polynomial  $p_k$  that minimizes the inf norm over the convex hull of the eigenvalues and it's monic. Note that here we use  $T_k$  for the type T Chebyshev Polynomial of degree  $k$  and it's not the tridiagonal symmetric matrix from Lanczos iterations. Next, we use the property that the range of the Chebyshev is bounded within the interval  $[-1, 1]$  to obtain inequality:

$$\forall x \in [\lambda_1, \lambda_n] : \left| \frac{T_k(\varphi(x))}{T_k(\varphi(0))} \right| \leq \left| \frac{1}{T_k(\varphi(0))} \right| \quad (4.2.5)$$

Next, our objective is to find any upper bound for the quantities on the RHS in relations to the Condition number for matrix  $A$  and the degree of the Chebyshev polynomial. Firstly observe that  $\varphi(0) < -1$ ,  $\varphi(0) \notin [\lambda_1, \lambda_n]$ , because all Eigenvalues are larger than zero, therefore it's out of the range of the Chebyshev polynomial and we need to find the actual value of it by considering alternative form of Chebyshev T for values outside of the  $[-1, 1]$ :

$$T_k(x) = \cosh(k \operatorname{arccosh}(z)) \quad \forall z \geq 1 \quad (4.2.6)$$

$$\implies T_k(\cosh(\zeta)) = \cosh(k\zeta) \quad z := \cosh(\zeta) \quad (4.2.7)$$

We need to match the form of the expression  $T_k(\varphi(0))$  with the expression of the form  $T_k(\cosh(\zeta))$  given the freedom of varying  $\zeta$ . To do that we consider a substitution of  $\zeta = \ln(y)$ , so that we only need to match  $\varphi(0)$  with the form  $(y + y^{-1})/2$ , which is just a quadratic equation.

$$\varphi(0) = \cosh(\zeta) = \cosh(\ln(y)) \quad \ln(y) := \zeta \quad (4.2.8)$$

$$\text{recall: } \cosh(x) = (\exp(-x) + \exp(x))/2 \quad (4.2.9)$$

$$\implies \cosh(\ln(y)) = (y + y^{-1})/2 \quad (4.2.10)$$

$$\varphi(0) = (y + y^{-1})/2 \quad (4.2.11)$$

Recall the definition of  $\varphi(x)$  and then simplifies:

$$\begin{aligned} \varphi(0) &= \frac{-\lambda_n - \lambda_1}{\lambda_n - \lambda_1} \\ &= \frac{-\lambda_n/\lambda_1 - 1}{\lambda_n/\lambda_1 - 1} \\ &= -\frac{\lambda_n/\lambda_1 + 1}{\lambda_n/\lambda_1 - 1} \\ \implies \varphi(0) &= -\frac{\kappa + 1}{\kappa - 1} \end{aligned}$$



Our objective is now simple. We know what  $\varphi(0)$  is, we want it to form match with  $\cosh(\ln(y))$ , and hence we simply solve for  $y$ :

$$-\frac{\kappa+1}{\kappa-1} = \frac{1}{2}(y+y^{-1}) \quad (4.2.12)$$

$$y = \frac{\sqrt{\kappa} \pm 1}{\sqrt{\kappa} \mp 1} \quad (4.2.13)$$

It's a quadratic and we solved it. The above  $\pm, \mp$  are correlated, meaning that they are of opposite sign, which gives us 2 roots for the quadratic expression. Now, given the hyperbolic form for  $\varphi(0)$ , we can substitute and get the value of  $T_k(\varphi(0))$  in terms of  $y$  and then  $\kappa$ :

$$\varphi(0) = \frac{1}{2}(y+y^{-1}) \quad (4.2.14)$$

$$\implies T_k(\varphi(0)) = T_k(\cosh(\ln(y))) \quad (4.2.15)$$

$$= \cosh(k \ln(y)) \quad (4.2.16)$$

$$= (y^k + y^{-k})/2 \quad (4.2.17)$$

Then, substituting the value of  $y$ , and invert the quantity we have:

$$\frac{1}{T_k(\varphi(0))} = 2(y^k + y^{-k})^{-1} \quad (4.2.18)$$

$$= 2 \left( \left( \frac{\sqrt{\kappa} \pm 1}{\sqrt{\kappa} \mp 1} \right)^k + \left( \frac{\sqrt{\kappa} \mp 1}{\sqrt{\kappa} \pm 1} \right)^{-k} \right)^{-1} \quad (4.2.19)$$

$$= 2 \left( \underbrace{\left( \frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1} \right)^k}_{>1} + \underbrace{\left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^{-k}}_{<1} \right)^{-1} \quad (4.2.20)$$

$$\leq 2 \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \quad (4.2.21)$$

Which completes the proof. Recall from the previous discussion for the squared of the relative error, we have:

$$\frac{\|e_k\|_A}{\|e_0\|_A} \leq \min_{p_{k+1}: p_{k+1}(0)=1} \max_{x \in [\lambda_1, \lambda_n]} |p_{k+1}(x)| \leq 2 \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \quad (4.2.22)$$

□

#### 4.2.2 One Outlier Eigenvalue

Using the derived theorem, we can extend it to other type of distributions of eigenvalues. Imagine one of the extreme case where some matrices that have one group of eigenvalues that are close together and one single eigenvalue that is far away from the cluster. In that case, we can use Chebyshev differently by focusing its minimizing power across the clustered eigenvalues and use a simple polynomial to interpolate the outlier eigenvalue. Consider the following proposition:

**Proposition 4.3** (Big Outlier CG Convergence Rate). If, there exists a  $\lambda_n$  that is much later than all previous  $n - 1$  eigenvalues for the matrix  $A$ , then a tighter convergence bound that being only parameterized by the range of clustered eigenvalues can be obtained and it is:

$$\frac{\|e^{(k)}\|_A}{\|e^{(0)}\|_A} \leq 2 \left( \frac{\sqrt{\kappa_{n-1}} - 1}{\sqrt{\kappa_{n-1}} + 1} \right)^{k-1} \quad \kappa_{n-1} = \frac{\lambda_{n-1}}{\lambda_1} \quad (4.2.23)$$

Reader please observe that, the outlier eigenvalue  $\lambda_n$  plays a smaller role in determining the convergence rate of the algorithm compare to the previous bound.

*Proof.* Here, we wish to show that a more focused use of the Chebyshev will introduce a better convergence rate for the Conjugate Gradient. We define the notation for the adapted  $k$ -th degree Chebyshev Polynomial over an closed interval:  $[a, b]$  as:

$$\hat{T}_{[a,b]}^{(k)}(x) := T_k \left( \frac{2x - b - a}{b - a} \right) \quad (4.2.24)$$

Next, we consider the following polynomial:

$$p_k(x) := \frac{\hat{T}_{[\lambda_1, \lambda_{n-1}]}^{(k-1)}(x)}{\hat{T}_{[\lambda_1, \lambda_{n-1}]}^{(k-1)}(0)} \left( \frac{\lambda_n - x}{x} \right) \quad (4.2.25)$$

Where, we use an  $k - 1$  degree polynomial for the clustered eigenvalues, and then we multiply that by a linear function  $(\lambda_n - z)/\lambda_n$  which is zero at right boundary  $\lambda_n$  and it's less than one at the left boundary  $\lambda_1$ . Next, observe the following facts about the above polynomials:

$$\frac{\lambda_n - z}{\lambda_n} \in [0, 1] \quad \forall z \in [\lambda_1, \lambda_n] \quad (4.2.26)$$

$$|p_k(x)| \leq \left| \frac{\hat{T}_{[\lambda_1, \lambda_{n-1}]}^{(k-1)}(x)}{\hat{T}_{[\lambda_1, \lambda_{n-1}]}^{(k-1)}(0)} \frac{\lambda_n - z}{\lambda_n} \right| \leq \frac{1}{\left| \hat{T}_{[\lambda_1, \lambda_{n-1}]}^{(k-1)}(0) \right|} \quad (4.2.27)$$

As a result, we can apply the convergence rate we proven for the uniform case, giving us:

$$T_{[\lambda_1, \lambda_{n-1}]}^{(k-1)}(0) = \left| T_{k-1} \left( \frac{-\lambda_{n-1} - \lambda_1}{\lambda_{n-1} - \lambda_1} \right) \right| \quad (4.2.28)$$

$$= \frac{1}{2} (y^{k-1} + y^{-(k-1)}) \quad (4.2.29)$$

$$\text{where: } y = \frac{\sqrt{\kappa_{n-1}} + 1}{\sqrt{\kappa_{n-1}} - 1}, \kappa_{n-1} = \frac{\lambda_{n-1}}{\lambda_1} \quad (4.2.30)$$

Substituting the value for  $y$  we obtain the bound:

$$\frac{\|e_k\|_A}{\|e_0\|_A} \leq 2 \left( \frac{\sqrt{\kappa_{n-1}} - 1}{\sqrt{\kappa_{n-1}} + 1} \right)^{k-1} \quad (4.2.31)$$

□

Another case that is worth considering is when there is one eigenvalue that is smaller than all the other eigenvalues which are clustered at a way larger value than it, by which I mean the value of  $\lambda_1$  is much smaller than all other eigenvalues and the other eigenvalues are clustered close together in an interval uniformly.

**Proposition 4.4** (Small Outlier CG Convergence Rate). The convergence rate is:

$$\frac{\|e_k\|_A}{\|e_0\|_A} \leq 2 \left( \frac{\lambda_n - \lambda_1}{\lambda_1} \right) \left( \frac{\sqrt{\kappa_0} - 1}{\sqrt{\kappa_0} + 1} \right)^{k-1} \quad (4.2.32)$$

Where  $\kappa_0$  is  $\lambda_n/\lambda_1$ .

*Proof.*

$$w(z) := \frac{\lambda_1 - z}{\lambda_1} \quad (4.2.33)$$

$$p_k(z) := w(z) \left( \frac{\hat{T}_{[\lambda_2, \lambda_n]}^{(k-1)}(z)}{\hat{T}_{[\lambda_2, \lambda_n]}^{(k-1)}(0)} \right) \quad (4.2.34)$$

$$\implies \max_{x \in [\lambda_2, \lambda_n]} |w(x)| = \frac{\lambda_n - \lambda_1}{\lambda_1} \quad (4.2.35)$$

In this case, the maximal value of the linear function  $w$  is achieved via  $x = \lambda_1$ , and the absolute value swapped the sign of the function. Therefore, we have:

$$|p_k(x)| = \left| w(x) \frac{\hat{T}_{[\lambda_2, \lambda_n]}^{(k-1)}(x)}{\hat{T}_{[\lambda_2, \lambda_n]}^{(k-1)}(0)} \right| \quad (4.2.36)$$

$$\leq \left| \frac{w(x)}{\hat{T}_{[\lambda_2, \lambda_n]}^{(k-1)}(0)} \right| \quad (4.2.37)$$

$$\leq \left| \left( \frac{\lambda_n - \lambda_1}{\lambda_1} \right) \hat{T}_{[\lambda_2, \lambda_n]}^{(k-1)}(0) \right| \quad (4.2.38)$$

$$\implies \leq \left( \frac{\lambda_n - \lambda_1}{\lambda_1} \right) 2 \left( \frac{\sqrt{\kappa_0} + 1}{\sqrt{\kappa_0} - 1} \right)^{k-1} \quad (4.2.39)$$

We applied the Chebyshev Bound theorem proved in the previous part. And  $\kappa_0 = \lambda_n/\lambda_1$ , and that is the maximal bound for the absolute value of the polynomial.

Take notice that it's not immediately clear which type of outlier eigenvalue make the convergence better or worse, but in this case, the weight  $w(x)$  introduces a term that grows inversely proportional to  $\lambda_1$ .  $\square$

### 4.3 From Conjugate Gradient to Lanczos

Up until this point of discussion, we had been brewing the fact that, the Iterative Lanczos Algorithm and the Conjugate gradient algorithm are pretty much the same thing. From the previous discussion we can observe that:

- 1.) Both Lanczos and CG terminates when the grade of Krylov subspace is reached. For Lanczos it's  $\mathcal{K}_k(A|q_1)$  and for CG it's  $\mathcal{K}_k(A|r_0)$ .
- 2.) Both Lanczos and CG generates orthogonal vectors, for Lanczos they are the  $q_i$  vector and for CG they are the  $r_i$  vectors.

These 2 properties in particular, is hinting at an equivalence between the residual vectors  $r_j$  from CG and the orthogonal vectors  $q_j$  from Lanczos. However, it's also not entirely obvious because CG is derived using CDM in the first section, and yet it doesn't make use of any orthogonal projector. Further more, one might also notice that Iterative Lanczos are for General Symmetric Matrices while CG is only for positive definite matrices. To see how everything ties together, we have to go both directions way to show the connections between these 2 iterative algorithms, which are what this section and the subsequent section about. Here, we refer Lanczos vectors as the sequence of  $q_j$  generated by the Iterative Lanczos Algorithm.

For this subsection, our objective is to establish the equivalence between the parameters from the Lanczos algorithm:  $\alpha_k, \beta_k, q_k$  and the  $a_j, b_j, r_j$  from the conjugate gradient algorithm. We establish it by going from the conjugate gradient to the Lanczos Algorithm.

**Proposition 4.5.** The residual and the Lanczos vectors have the following relations:

$$q_1 = \hat{r}_0 \tag{4.3.1}$$

$$q_2 = -\hat{r}_1 \tag{4.3.2}$$

$$\vdots \tag{4.3.3}$$

$$q_j = (-1)^{j+1} \hat{r}_{j+1} \tag{4.3.4}$$

Here,  $\hat{r}_j := r_j / \|r_j\|$  and we can fill in the Lanczos Tridiagonal matrix using the CG parameters to obtain the tridiagonalization of the Positive definite matrix  $A$ :

$$\begin{cases} \alpha_{j+1} = \frac{1}{a_j} + \frac{b_{j-1}}{a_{j-1}} & \forall 1 \leq j \leq k-1 \\ \beta_j = \frac{\sqrt{b_{j-1}}}{a_{j-1}} & \forall 2 \leq j \leq k-2 \\ \alpha_1 = a_0^{-1} \\ \beta_1 = \frac{\sqrt{b_0}}{\alpha_0} \end{cases} \tag{4.3.5}$$

Where  $\alpha_j$  for  $1 \leq j \leq n-1$  are the diagonal of the tridiagonal matrix  $T_k$  generated by Lanczos, and  $\beta_j$  for  $2 \leq j \leq k-2$  are the lower and upper subdiagonals of the matrix  $T_k$ .

We will break the proof into several parts. Firstly we address the base case, and then we address the inductive case to establish the parameters between the Tridiagonal matrix and  $a_k, b_k$ , finally we resolve the sign problem between the Lanczos vectors and the residual vectors.

### 4.3.1 The Base Case

Right from the start of the CG iteration we have:

$$r_0 = p_0 \quad (4.3.6)$$

$$r_1 = r_0 - a_0 A r_0 \quad (4.3.7)$$

$$A r_0 = a_0^{-1} (r_0 - r_1) \quad (4.3.8)$$

$$A r_0 = \frac{\|r_0\|_A^2}{\|r_0\|^2} (r_0 - r_1) \quad (4.3.9)$$

Consider substituting  $r_0 = \|r_0\|q_1, r_1 = -\|r_1\|q_2$ , then:

$$A\|r_0\|q_1 = \frac{\|r_0\|_A^2}{\|r_0\|^2} (\|r_0\|q_1 + \|r_1\|q_2) \quad (4.3.10)$$

$$= \frac{\|r_0\|_A^2}{\|r_0\|^2} \|r_0\|q_1 + \frac{\|r_1\|}{\|r_0\|} q_2 \quad (4.3.11)$$

And from this relation, using the Lanczos recurrence theorem would imply that  $\alpha_1 = a_0^{-1}$ ;  $\beta_1 = \frac{\sqrt{b_0}}{\alpha_0}$ . So far so good, we have shown that there is an equivalence between the Lanczos and the CG for the first iterations of the CG algorithm.

### 4.3.2 The Inductive Case

**Lemma 4.3.1.** Inductively we wish to show the relation that:

$$\begin{cases} \alpha_{j+1} = \frac{1}{a_j} + \frac{b_{j-1}}{a_{j-1}} & \forall 1 \leq j \leq n-1 \\ \beta_j = \frac{\sqrt{b_{j-1}}}{a_{j-1}} & \forall 2 \leq j \leq n-2 \end{cases} \quad (4.3.12)$$

*Proof.* We start by considering:

$$r_j = r_{j-1} - a_{j-1} A p_{j-1} \quad (4.3.13)$$

$$= r_{j-1} - a_{j-1} A (r_{j-1} + b_{j-2} p_{j-1}) \quad (4.3.14)$$

$$= r_{j-1} - a_{j-1} A r_{j-1} - a_{j-1} b_{j-2} A p_{j-1} \quad (4.3.15)$$

We make use of the recurrence asserted by the CG algorithm, giving us:

$$r_{j-1} = r_{j-1} - a_{j-2} A p_{j-1} \quad (4.3.16)$$

$$r_{j-1} - r_{j-1} = a_{j-2} A p_{j-1} \quad (4.3.17)$$

$$A p_{j-1} = a_{j-2}^{-1} (r_{j-2} - r_{j-1}) \quad (4.3.18)$$

Here, we can substitute the results of for the term  $A p_{j-1}$ , and then we can express the

recurrence of residual purely in terms of residual. Consider:

$$r_j = r_{j-1} - a_{j-1}Ar_{j-1} - a_{j-1}b_{j-2}Ap_{j-2} \quad (4.3.19)$$

$$= r_{j-1} - a_{j-1}Ar_{j-1} - \frac{a_{j-1}b_{j-2}}{a_{j-2}}(r_{j-2} - r_{j-1}) \quad (4.3.20)$$

$$= \left(1 + \frac{a_{j-1}b_{j-2}}{a_{j-2}}r_{j-1}\right) - a_{j-1}Ar_{j-1} - \frac{a_{j-1}b_{j-2}}{a_{j-2}}r_{j-2} \quad (4.3.21)$$

$$a_{j-1}Ar_{j-1} = \left(1 + \frac{a_{j-1}b_{j-2}}{a_{j-2}}r_{j-1}\right) - \frac{a_{j-1}b_{j-2}}{a_{j-2}}r_{j-2} \quad (4.3.22)$$

$$Ar_{j-1} = \left(\frac{1}{a_{j-1}} + \frac{b_{j-2}}{a_{j-2}}\right)r_{j-1} + \frac{r_j}{a_{j-1}} - \frac{b_{j-2}}{a_{j-2}}r_{j-2} \quad (4.3.23)$$

Finally, we increment the index  $j$  by one for notational convenience, and therefore we establish the following relations between the residuals of the conjugate gradient algorithm:

$$Ar_j = \left(\frac{1}{a_j} + \frac{b_{j-1}}{a_{j-1}}\right)r_j + \frac{r_{j+1}}{a_j} - \frac{b_{j-1}}{a_{j-1}}r_{j-1} \quad (4.3.24)$$

Reader please observe that, this is somewhat similar to the recurrence relations between the Lanczos vectors, however it's failing to match the sign, at the same time, it's not quiet matching the form of the recurrence of  $\beta_k$  from the Lanczos algorithm. To match it, we need the coefficients of  $r_{j-1}$  and  $r_{j+1}$  to be in the same form, parameterized by the same iterations parameter:  $j$ . To do that, consider the doing this:

$$q_{j+1} := \frac{r_j}{\|r_j\|} \quad (4.3.25)$$

$$q_j := -\frac{r_{j-1}}{\|r_{j-1}\|} \quad \text{Note: This is Negative} \quad (4.3.26)$$

$$q_{j+2} := \frac{r_{j+1}}{\|r_{j+1}\|} \quad (4.3.27)$$

$$\implies A\|r_j\|q_{j+1} = \left(\frac{1}{a_j} + \frac{b_{j-1}}{a_{j-1}}\right)\|r_j\|q_{j+1} + \frac{\|r_{j+1}\|q_{j+2}}{a_j} + \frac{b_{j-1}\|r_{j-1}\|}{a_{j-1}}q_j \quad (4.3.28)$$

$$Aq_{j+1} = \left(\frac{1}{a_j} + \frac{b_{j-1}}{a_{j-1}}\right)q_{j+1} + \frac{\|r_{j+1}\|}{a_j\|r_j\|}q_{j+2} + \frac{b_{j-1}\|r_{j-1}\|}{a_{j-1}\|r_j\|}q_j \quad (4.3.29)$$

Recall that parameters from Conjugate Gradient,  $\sqrt{b_j} = \|r_{j+1}\|/\|r_j\|$ , and  $a_j = \frac{\|r_j\|^2}{\|p_j\|_A^2}$ , and we can use the substitution to match the coefficients for  $q_{j+2}$  and  $q_j$ , giving us:

$$\frac{\|r_{j+1}\|}{a_j\|r_j\|} = \frac{1}{a_j}\sqrt{b_j} \quad (4.3.30)$$

$$\frac{b_{j-1}\|r_{j-1}\|}{a_{j-1}\|r_j\|} = \frac{b_{j-1}}{a_{j-1}}\frac{1}{\sqrt{b_{j-1}}} = \frac{\sqrt{b_{j-1}}}{a_{j-1}} \quad (4.3.31)$$

$$\implies \begin{cases} \alpha_{j+1} = \frac{1}{a_j} + \frac{b_{j-1}}{a_{j-1}} & \forall 1 \leq j \leq n-1 \\ \beta_j = \frac{\sqrt{b_{j-1}}}{a_{j-1}} & \forall 2 \leq j \leq n-2 \end{cases} \quad (4.3.32)$$

Take notes that the form is now matched, but the expression for  $\alpha_{j+1}$  has an extra  $b_{j-1}/a_{j-1}$ , to resolve that, we take the audacity to make  $b_0$  so that it's consistent with the base case.  $\square$

### 4.3.3 Fixing the Sign

We can't take the triumph yet; we need to take a more careful look into the sign between  $q_j$  the Lanczos Vector and its equivalence residual:  $r_{j-1}$  in CG. Here, I want to point out the fact that, there are potentially 2 substitution possible for the above derivation for the inductive case and regardless of which one we use, it would still preserve the correctness for the proof. By which I mean the following substitutions would have both made it work:

$$\begin{cases} q_{j+1} := \pm \frac{r_j}{\|r_j\|} \\ q_j := \mp \frac{r_{j-1}}{\|r_{j-1}\|} \\ q_{j+2} := \pm \frac{r_{j+1}}{\|r_{j+1}\|} \end{cases} \quad (4.3.33)$$

Under the context, the operations  $\pm, \mp$  are correlated, choose a sign for one, the other must be of opposite sign. In this case both substitutions work the same because multiplying the equation by negative one would give the same equality, and we can always multiply by and other negative sign to get it back. The key here is that, the sign going from  $q_j$  to the next  $q_{j-1}$  will have to alternate. To find out precisely which one it is, we consider the base case for the Lanczos Vectors and Residuals:

$$q_1 = \hat{r}_0 \quad (4.3.34)$$

$$q_2 = -\hat{r}_1 \quad (4.3.35)$$

$$\vdots \quad (4.3.36)$$

$$q_j = (-1)^{j+1} \hat{r}_{j+1} \quad (4.3.37)$$

And at this point, we have established the equivalence going from the Conjugate Gradient algorithm to the Lanczos Algorithm. And the moral of the story is, CG is a special case of applying the Lanczos Iterations with  $q_1 = r_0$  to a positive definite matrix. However, something is still off and one can ask the following questions to inquiry further, leading us to more discussion between these 2 algorithms.

- 1.) How are the solutions  $x_k$  generated by CG relates to the Lanczos Iterations?
- 2.) How are the A-Orthogonal vectors  $p_k$  from CG relates to Lanczos?

**Remark 4.3.1** (A Better Terminations Conditions for CG). The derivation hinted at a better terminations conditions for the CG algorithm. Because CG is equivalent to the Lanczos Iterations initialized with  $q_1 = r_0$ , and we can directly apply from [proposition 3.4.2](#) to get the precise number of iterations of CG under exact arithmetic given  $r_0$ , improving the bound we got from [proposition 4.2](#).

## 4.4 From Lanczos to Conjugate Gradient

In this section, we show the equivalence of the Conjugate Gradient and Lanczos iterations by deriving the CG using the Lanczos.

The goal is to show that CG is a special case of Lanczos. In the end we discuss the key that it can inspire solvers for symmetric indefinite systems of linear equations.

We start of by considering the LU decomposition of the tridiagonal matrix of Lanczos and us its entries to express the scalars  $a_k, b_k$  in CG. In addition, we express the conjugate vectors  $p_k$  as a short recurrence of the Lanczos vectors  $q_k$ .

#### 4.4.1 Matching the Residual and Conjugate Vectors

In this section, we go from the Iterative Lanczos algorithm to the Conjugate Gradient algorithm and we seek to establish a link between the solution  $x_k, p_k$  from CG with the Lanczos vectors and the symmetric tridiagonal matrix generated by Lanczos. This section will also play an important role for the backwards analysis for the floating point behaviors for the CG algorithm in the later parts of the paper. At the end, we highlight some of the hidden insights that this particular derivation of equivalence leads to, and how it inspires algorithms that directly solves symmetric indefinite systems.

**Proposition 4.6** (Lanczos Vectors and Residuals). The  $Q_k$  is the orthogonal matrix generated by Lanczos Iteration. To match the Krylov Subspace generated by the Lanczos iterations and CG, we initialize  $q_1 = \hat{r}_0$ , then the following relationship between Lanczos and CG occurs between their parameters:

$$\begin{cases} y_k = T_k^{-1} \beta \xi_1 \\ x_k = x_0 + Q_k y_k \\ r_k = -\beta_k \xi_k^T y_k q_{k+1} \end{cases} \quad (4.4.1)$$

The quantities  $\alpha, \beta$  are the diagonal and the sub or super diagonal of the matrix  $T_k$  from the Iterative Lanczos Algorithm, and  $r_k$  is the residual from the Conjugate Gradient Algorithm, and  $Q_k$  is the orthogonal matrix generated from the Lanczos Algorithm. For notations, we use  $\xi_i$  to denote the  $i$ th canonical basis vector.  $\beta$  without the subscript denotes  $\|r_0\|$ .

*Proof.* To start recall that the Lanczos Algorithm Asserts the following recurrences:

$$AQ_k = Q_{k+1} \begin{bmatrix} T_k \\ \beta_k \xi_k^T \end{bmatrix} \quad (4.4.2)$$

Recall that the Conjugate Gradient algorithm takes the guesses from the affine span of  $x_0 + \mathcal{K}_k(A|r_0)$ , from section [CG and Krylov Subspace 3.3.6](#) we know that:  $p_k \in \mathcal{K}_{k+1}(A|r_0)$ , the matrix  $P_k, Q_k$  spans the same subspace, and that means:

$$x_{k+1} = x_0 + Q_k y_k \quad (4.4.3)$$

$$r_{k+1} = r_0 - AQ_k y_k \quad (4.4.4)$$

$$Q_k^H r_{k+1} = Q_k^H r_0 - Q_k^H A Q_k y_k \quad (4.4.5)$$

$$\implies 0 = \beta \xi_1 - T_k y_k \quad (4.4.6)$$

$$y_k = T_k^{-1} \beta \xi_1 \quad (4.4.7)$$



Now to get the residual we simply consider:

$$r_{k+1} = r_0 - AQ_k y_k \quad (4.4.8)$$

$$= r_0 - AQ_k T_k^{-1} \beta \xi_1 \quad (4.4.9)$$

$$\implies = \beta q_1 - AQ_k T_k^{-1} \beta \xi_1 \quad (4.4.10)$$

$$= \beta q_1 - Q_{k+1} \begin{bmatrix} T_k \\ \beta_k \xi_k^T \end{bmatrix} T_k^{-1} \beta \xi_1 \quad (4.4.11)$$

$$= \beta q_1 - (Q_k T_k + \beta_k q_{k+1} \xi_k^T) T_k^{-1} \beta \xi_1 \quad (4.4.12)$$

$$= \beta q_1 - (Q_k \beta \xi_1 + \beta_k q_{k+1} \xi_{k+1} T_k^{-1} \beta \xi_1) \quad (4.4.13)$$

$$= -\beta_k q_{k+1} \xi_k^T T_k^{-1} \beta \xi_1 \quad (4.4.14)$$

On the third line we recall the fact that  $q_1 = \hat{r}_0$  which initialized the Krylov Subspace for the Lanczos Iteration. At the 4th line, we make use of the Lanczos Vector recurrences and we simply substituted it.

By observing the fact that  $\xi_k^T T_k^{-1} \xi_1$  the  $(k, 1)$  element of the matrix  $T_k^{-1}$  which is a scalar, we can conclude that the residual from CG and the Lanczos vector are scalar multiple of each other, therefore,  $r_k$  from the CG must be orthogonal as well.  $\square$

**Proposition 4.7** (Lanczos Vectors and Conjugate Vectors). The  $P_k$  matrix as derived in the CG algorithm can be related to the Lanczos iterations by the formula:

$$P_k = Q_k U_k^{-1} \quad (4.4.15)$$

Where  $T_k = L_k U_k$ , representing the LU decomposition of the tridiagonal matrix  $T_k$  from the Lanczos Iterations. Because of the tridiagonal nature of the matrix  $T_k$ ,  $L_k$  will be a unit bi-diagonal matrix and  $U_k$  will be an upper bi-diagonal matrix.

*Proof.* To prove it, we start by considering the  $x_k$  at step k of the iterations:

$$x_k = x_0 + Q_k y_k \quad (4.4.16)$$

$$= x_0 + Q_k T_k^{-1} \beta \xi_1 \quad (4.4.17)$$

$$= x_0 + Q_k U_k^{-1} L_k^{-1} \beta \xi_1 \quad (4.4.18)$$

$$= x_0 + P_k L_k^{-1} \beta \xi_1 \quad (4.4.19)$$

So far we have written the solution vector  $x_k$ . Next, we are going to prove that the matrix  $P_k$  indeed consists of vectors that are A-orthogonal. To show that we consider:

$$P_k^T A P_k \quad (4.4.20)$$

$$= (Q_k U_k^{-1})^T A Q_k U_k^{-1} \quad (4.4.21)$$

$$= U_k^{-T} Q_k^T A Q_k U_k^{-1} \quad (4.4.22)$$

$$= U_k^{-T} T_k U_k^{-1} \quad (4.4.23)$$

$$= U_k^{-T} L_k \quad (4.4.24)$$

Reader please observe that  $U_k$  is upper triangular, therefore, its inverse is also upper triangular, therefore,  $U_k^{-T}$  is lower triangular, and because  $L_k$  is also lower triangular, their

product is a lower triangular matrix, and therefore, the resulting matrix above is lower triangular, however, given that  $P_k^T A P_k$  is symmetric, therefore,  $U_k^{-T} L_k$  will have to be symmetric as well, and a matrix that is lower triangular and symmetric has to be diagonal. Therefore, the columns of  $P_k$  are conjugate vectors.  $\square$

#### 4.4.2 Matching the $a_k, b_k$ in CG

Similar to how we can generate the tridiagonal matrix for the Lanczos iterations with  $q_1 = \hat{r}_0$ , we can also generate the parameters  $a_k, b_k$  in the CG algorithm using parameters from the Lanczos Iterations. To achieve it, one can simply build up the recurrences for the  $y_k$  vectors using the elements from the  $L_k, U_k$  matrix which comes from LU decomposition of the  $T_k$  matrix. This will come at the expense of losing some degree of accuracy because it's equivalent to doing the LU decomposition of  $T_k$  without pivoting, but it comes at the advantage computing  $\xi_k^T T_k^{-1} \xi_1$  with as little efforts as possible. Let's take a look.

For discussion in this section, we briefly switch the indexing and let it start counting from one instead of zero.

$$P_k = [p_1 \ p_2 \ \cdots \ p_k] \quad Q_k = [q_1 \ q_2 \ \cdots \ q_k] \quad (4.4.25)$$

Assuming the matrix  $A$  is invertible and using the Cauchy Interlace Theorem,  $T_k$  is invertible, we consider the LU decomposition of the symmetric tridiagonal matrix:

$$T_k = L_k U_k = \begin{bmatrix} 1 & & & \\ l_1 & 1 & & \\ & \ddots & \ddots & \\ & & l_{k-1} & 1 \end{bmatrix} \begin{bmatrix} u_1 & \beta_1 & & \\ & u_2 & \beta_2 & \\ & & \ddots & \beta_{k-1} \\ & & & u_k \end{bmatrix} \quad (4.4.26)$$

Reader should agree with some considerations that the upper diagonal of  $U_k$  are indeed the same as the upper diagonal of the symmetric tridiagonal matrix  $T_k$ . And recall the expression for  $x_k$  from the previous section, we have:

$$x_k = x_0 + P_k L_k^{-1} \beta \xi_1 \quad (4.4.27)$$

$$x_k - x_{k-1} = P_k L_k^{-1} \beta \xi_1 - P_{k-1} L_{k-1}^{-1} \beta \xi_1 \quad (4.4.28)$$

$$= P_k \beta (L_k^{-1})_{:,1} - P_{k-1} \beta (L_{k-1}^{-1})_{:,1} \quad (4.4.29)$$

$$= \beta (L_k^{-1})_{k,1} P_k \quad (4.4.30)$$

$$\implies x_k = x_{k-1} + \beta (L_k^{-1})_{k,1} p_k \quad (4.4.31)$$

On the third line, we factor out the last column for the matrix  $P_k$ . Next, we wish to derive the recurrence between  $p_{k+1}$  and  $p_k$ . Which is:

$$P_k = Q_k U_k^{-1} \quad (4.4.32)$$

$$P_k U_k = Q_k \quad (4.4.33)$$

$$\implies \beta_{k-1} p_{k-1} + u_k p_k = q_k \quad (4.4.34)$$

$$u_k p_k = q_k - \beta_{k-1} p_{k-1} \quad (4.4.35)$$

$$p_k = u_k^{-1} (q_k - \beta_{k-1} p_{k-1}) \quad (4.4.36)$$

We made use of the fact that the matrix  $U_k$  is unit upper bidiagonal. Next, we seek for the recurrences of the parameters  $u_k, l_k$ . Let's consider the recurrence using the block structure of the matrices:

$$T_k = L_k U_k \quad (4.4.37)$$

$$T_{k+1} = \begin{bmatrix} T_k & \beta_k \xi_k \\ \beta_k \xi_k^T & \alpha_{k+1} \end{bmatrix} = \begin{bmatrix} L_k & \mathbf{0} \\ l_k \xi_k^{-1} & 1 \end{bmatrix} \begin{bmatrix} U_k & \eta_k \xi_k \\ \mathbf{0} & u_{k+1} \end{bmatrix} \quad (4.4.38)$$

$$= \begin{bmatrix} L_k U_k & \eta_k L_k \xi_k \\ l_k \xi_k^T U_k & \eta_k l_k \xi_k^T \xi_k + u_{k+1} \alpha_k \end{bmatrix} \quad (4.4.39)$$

$$= \begin{bmatrix} T_k & \eta_k (L_k)_{:,k} \\ l_k (U_k)_{k,:} & \eta_k l_k + u_{k+1} \end{bmatrix} \quad (4.4.40)$$

$$= \begin{bmatrix} T_k & \eta_k \\ l_k u_k & \eta_k l_k + u_{k+1} \end{bmatrix} \quad (4.4.41)$$

Note that I changed the upper diagonal of matrix  $U$  at the top to be  $\eta_k$  instead of  $\beta_k$  so we have a chance to convince ourselves that  $\beta_k$  for the upper diagonal of  $T_k$  are indeed the same as the  $\eta_k$  for the upper diagonal of matrix  $U_k$ . From the results above,  $\eta_k = \beta_k$  as expected, and  $l_k = \beta_k/u_k$ ,  $u_{k+1} = \alpha_{k+1} - \beta_k l_k$ , and hence, to sum up the recurrence relation we have:

$$\begin{cases} u_{k+1} &= \alpha_{k+1} - \beta_k^2/u_k \\ l_k &= \beta_k/u_k \end{cases} \quad (4.4.42)$$

The base case is  $u_1 = \alpha_1$ . The recurrence of the parameter  $u_k$  is immediately useful for figuring out the recurrence for  $x_k$ , And to figure out the recurrence relations of  $(L_k^{-1})_{k,1}$ , we consider the following fact:

$$L_k^{-1} L_k = I \quad (4.4.43)$$

$$\begin{bmatrix} L_k^{-1} & \mathbf{0} \\ s_k^T & d_{k+1} \end{bmatrix} \begin{bmatrix} L_k & \mathbf{0} \\ l_k \xi_k^T & 1 \end{bmatrix} = I \quad (4.4.44)$$

$$\begin{bmatrix} I & \mathbf{0} \\ s_k^T L_k + d_{k+1} l_k \xi_k^T & d_{k+1} \end{bmatrix} = I \quad (4.4.45)$$

It equals to the identity matrix therefore  $d_{k+1} = 1$ , and it has to be that the lower diagonal sub vector in the results has to be zero. For the bi-lower unit diagonal matrix  $L_k$ , we cannot predict the structure, most of the time it's likely to be dense and unit lower triangular. We are interested in look for the first element of the vector  $s_k^T$ , the equality will assert:

$$s_k^T L_k + d_{k+1} l_k \xi_k^T = \mathbf{0} \quad (4.4.46)$$

$$L_k^T s_k + d_{k+1} l_k \xi_k = \mathbf{0} \quad (4.4.47)$$

$$s_k + L^{-T} d_{k+1} l_k \xi_k = \mathbf{0} \quad (4.4.48)$$

$$(s_k)_1 + d_{k+1} l_k ((L_k^{-1}) \xi_k)_1 = 0 \quad (4.4.49)$$

$$(s_k)_1 + d_{k+1} l_k (L_k^{-1})_{k,1} = 0 \quad (4.4.50)$$

$$\implies (s_k)_1 = -l_k (L_k^{-1})_{k,1} \quad (4.4.51)$$

$$(s_k)_1 = (L_{k+1}^{-1})_{k+1,1} \quad \text{by definition} \quad (4.4.52)$$

$$\implies (L_{k+1}^{-1})_{k+1,1} = -l_k (L_k^{-1})_{k,1} \quad (4.4.53)$$

To assert the fact that  $L_k^{-1}L_k$  is identity, we carefully consider the last row excluding the bottom right element vector:  $s^T L_k + d_{k+1} l_k \xi_k^T$  which will have to be a zero vector. From the first line to the second line, we took the transpose of the vector. From the second to the third line we multiplied both side by inverse of  $L_k^T$ . And from the third to the forth line, we took out only the first element in the vector. Observe that the first element of the vector  $L_k^{-1} \xi_k$  is just  $(L_k^{-1})_{k,1}$ .

Therefore the recurrence for the step size into the direction of the conjugate vector requires us to use the newest element  $l_k$  from  $L_{k+1}$  and the previous step size in the direction of the conjugate vector  $p_k$ . The short recurrence allows us to build up another algorithm that is just as efficient as CG algorithm but running Lanczos Algorithm at the same time.

**Remark 4.4.1.** The derivation might seems excessive for the discussion, but it's part of the analysis of the Conjugate Gradient. It derived the conjugate gradient without using the fact that  $A$  is symmetric positive definite providing potential to new algorithm that can solve symmetric indefinite system directly. The Lanczos Algorithm for linear system (we refer to the method derived in the above section) is a special case of FOM (**CITATION NEEDED**) when the matrix  $A$  is symmetric. The above algorithm is just FOM with a short term recurrences for its parameters, and it's based on the fact that  $A$  is symmetric. It's implied from the above derivation that under exact arithmetic, CG can be applied to symmetric indefinite system, if we have the luck where  $T_j$  is non-singular for all  $j \leq k$ . Recall that when we derived the CG algorithm, we convert it into solving the system:  $P_k^T r_0 = P_k^T A P_k w$ , and when the matrix  $A$  is indefinite, we can still solve the system and get the saddle points for the indefinite error norm. However there are 2 problems solving Symmetric Indefinite using the CG is that. The first problem is the bad numerical accuracy of using the recurrence to perform LU decomposition on  $T_K$  without pivoting. Another problem is when  $T_k$  is singular during one iteration of the Lanczos Algorithm. For example,  $T_j$  will become singular when  $A$  is a symmetric tridiagonal matrix with zeros as its diagonals and all ones on it's subdiagonals. However, these problems can be solved by considering something other than LU without pivoting. In fact, there are works of Paige, Y. Saad extending the idea and derived algorithms that can solve a symmetric indefinite system without using the norm equation. **CITATION NEEDED**.

## 5 Effects of Floating Point Arithmetic

In this section, we highlights the practical concerns of the algorithm and showcase it with numerical experiments with analysis, in hope to get deeper insights about the behaviors of Lanczos and Conjugate Gradient under floating point arithmetic.

### 5.1 Partial Orthogonalization and Full Orthogonalization

In this section, we use some of the results and insight obtained from the derivation of the CG algorithm to develop a theoretical fix for the loss of precision under floating point arithmetic for the CG algorithm. The floating errors inside of the CG algorithm is manifested as a loss of orthogonality and loss of conjugacy for the vectors  $r_k, p_k$ . To fix this, we simply use the CDM algorithm's projector to re-orthogonalize the conjugate vectors and the residual

vectors so the A-Orthogonality and Orthogonality are both preserved for the  $r_k, p_k$  vectors. Such idea is not new and it's stated in the original paper by Hestenes and Stiefel back in 1952; here we present the second more computationally expensive idea in the paper by Hestenes and Stiefel, but using the quantities we derived in the first section. Firstly recall the proof for [proposition 3.9](#). Next, we inductively consider the case where the newest residual vector is involving some round off error  $\bar{r}_{k+1}$  and it breaks the orthogonality conditions  $\bar{r}_{k+1} \perp r_j \forall 0 \leq j \leq k$ :

$$(AP_k)^T \bar{r}_k = \begin{bmatrix} \langle p_0, A\bar{r}_k \rangle \\ \langle p_1, A\bar{r}_k \rangle \\ \vdots \\ \langle p_{k-1}, A\bar{r}_k \rangle \end{bmatrix} \quad (5.1.1)$$

$$= a_{k-1}^{-1} \langle r_k, (r_{k-1} - r_k) \rangle \xi_k + \sum_{j=0}^{k-1} \langle p_j, A\bar{r}_k \rangle \xi_j \quad (5.1.2)$$

$$p_k := \bar{r}_k + b_k p_k - \frac{\langle \bar{r}_k, r_{k-1} \rangle}{\langle r_{k-1}, r_{k-1} \rangle} p_k - \sum_{j=0}^{k-1} \frac{\langle p_j, A\bar{r}_k \rangle}{\langle p_k, Ap_k \rangle} p_j \quad (5.1.3)$$

$$r_k := \bar{r}_k - \sum_{j=0}^{k-1} \langle \hat{r}_j, \bar{r}_k \rangle \hat{r}_j \quad (5.1.4)$$

Here, we generate the conjugate vectors  $p_k$  correctly by faithfully reproducing the term  $(AP_k)^T \bar{r}_k$ , and then we update the residual  $\bar{r}_k$  into  $r_k$  by orthogonalizing it against all previous residual vectors. Doing this require the expensive storage of previous vectors  $p_k$ . One can use alternative formulas to A-orthogonalize  $p_k$  and re-orthogonalize  $r_k$ . In addition, we have the options for partially orthogonalizing the  $r_k, p_k$  vectors for less memory usage.

## 5.2 Relative Errors of CG Under Floating Points

We investigate the relative energy norm of the error for fully orthogonalized CG, partially orthogonalized CG, and CG without orthogonalizations numerically.

### 5.2.1 Experiments

Here we use full reorthogonalization to emulate the effects of exact arithmetic. the step required for convergence is less than or equal to the number of unique eigenvalues for the symmetric definite matrix, this is established in part (2.2) of the discussion. However in practice, this is not always the case. Similar experiments are conducted in Greenbaum's works (**CITATION NEEDED**). In this section, we replicate the same set of experiments using modern Julia to showcase the extra steps required for the CG algorithm to converge. For testing the convergence of the algorithm we use a 16 digits floats to exaggerate the floating point round-off error. The spectrum of the matrix we are using are taken to be the same from Greenbaum's works (**CITATION NEEDED**):

$$\lambda_{\min} + \left( \frac{j}{N-1} \right) (\lambda_{\max} - \lambda_{\min}) \rho^{N-j+1} \quad \forall 1 \leq j \leq N-1, 0 \leq \rho \leq 1 \quad (5.2.1)$$

If the value of  $\rho$  is close to zero, then the eigenvalues are clustered around the origin, if it's close to 1, then the eigenvalues are tend to be more evenly distributed around the interval  $[\lambda_{\min}, \lambda_{\max}]$ . The Chebyshev bound is no longer a tight bound because the distribution of

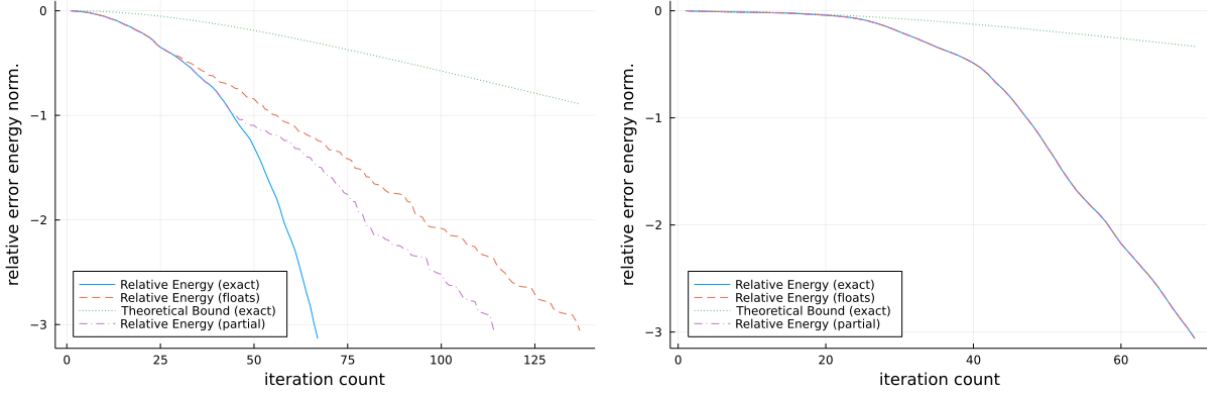


Figure 5.2.1: The relative energy norm error for different methods. Blue solid line: The exact conjugate gradient convergence. Purple dotdashed: The conjugate gradient that is partially orthogonalized with previous 8 residual and conjugate vectors. orange dashed: the Original conjugate gradient. Green dot: The tighter theoretical upper bound derived by Chebyshev (4.3.20). Left:  $\rho = 0.9$ ; right:  $\rho = 1$ , the matrix is  $64 \times 64$

the eigenvalues are not perfectly uniform. The partially orthogonalized methods diverges from the exact error after more steps of iterations compare to the relative error without any orthogonalizations. These are seemed in (fig 5.2.1 left). In contrast on the right, then the eigenvalues are uniformly distributed, the convergence of all 3 types of methods aligns with the exact convergence in (fig 5.2.1 right).

**Remark 5.2.1.** The convergence is disappointing under floating point arithmetic and the promised efficiency of the algorithm is not there anymore if the matrix is not necessarily ill-conditioned. Just from (fig 5.2.1) it seems like outlier eigenvalues provide fast convergence under full orthogonalization, but not for floating point.

## 5.3 Paige's Floating Point Analysis

In this section we present the backwards analysis of the algorithm in a more thorough manner and exam its consequences. When floating point arithmetic is used, the eigenvalues of the tridiagonal matrices might introduce ghost eigenvectors for the Lanczos Iterations, and using the equivalence of the Lanczos iterations and CG, we can capture a posteriori bound on how much the error is exactly during the iterations of CG.

### 5.3.1 Bounding the Relative Residuals

Recall from proposition [Proposition 4.6](#) that the residual of the CG can be expressed in terms of the Lanczos vectors. However, the Lanczos iterations under floating doesn't produces perfectly orthogonal Lanczos Vectors (The lost of orthogonality is experimented and visualized in the next section),  $\tilde{Q}_k$  is not quiet orthogonal, which would also means that  $\tilde{Q}_k^H A \tilde{Q}_k \approx T_k$  where  $T_k$  is the results from the Lanczos iterations will not equal to  $\tilde{Q}_k^T A \tilde{Q}_k$ . However the Lanczos iterations will still solve for  $y_k$  using the expression  $y_k = \beta T_k^{-1} \xi_1$ , the algorithm still thought  $T_k$  produced by itself is perfectly tridiagonal. As a result the algorithm

never quiet find the optimal under the conjugate basis. At first glance, tiny round-off errors in the Lanczos vectors is very problematic.

Surprisingly the recurrence formula for Lanczos still holds to some extend, in addition we can leverage the fact that  $y_k$  is solved exactly and assume:  $y_k = \beta T^{-1} \xi_1$  is at least, exact. Then it left us with fewer type of floating errors to keep track of. Next, we proceed to look for the residual of the CG algorithm by assuming that the Lanczos recurrences:

$$AQ_k = Q_{k+1} \begin{bmatrix} T_k \\ \beta_k \xi_k^T \end{bmatrix} + F_k \quad (5.3.1)$$

Reader please reflect on the fact that the  $Q_k$  which is not orthogonal, and we are fixing the recurrences with  $F_k$ , a matrix representing the floating error to correct it so that the equality holds true.  $\|F_k\|$  is small and it will be stated later.

$$r_k = r_0 - AQ_k y_k \quad (5.3.2)$$

$$r_k = r_0 - \left( Q_{k+1} \begin{bmatrix} T_k \\ \beta_k \xi_k^T \end{bmatrix} + F_k \right) y_k \quad (5.3.3)$$

$$r_k = \underbrace{\left( r_0 - Q_{k+1} \begin{bmatrix} T_k \\ \beta_k \xi_k^T \end{bmatrix} y_k \right)}_{= -\beta_k \xi_k^T y_k q_{k+1}} + F_k \beta T_k^{-1} \xi_1 \quad (5.3.4)$$

$$\implies \frac{\|r_{k+1}\|}{\|r_0\|} \leq \beta_k \|\xi_k^T T_k^{-1} \xi_1 q_{k+1}\| + \|F_k T_k^{-1} \xi_1\| \quad (5.3.5)$$

$$\frac{\|r_{k+1}\|}{\|r_0\|} \leq \beta_k |\xi_k^T T_k^{-1} \xi_1| + \|F_k\| \|T_k^{-1} \xi_1\| \quad (5.3.6)$$

We make use of [Proposition 4.6](#) and obtain a similar expression because it didn't make use of the fact that  $Q_k$  is orthogonal. This time, we take  $F_k$  into account. The residual is now bounded by the sum of scalar  $\xi_k^T T_k^{-1} \xi_1$  and the floating point error matrix  $F_k$  produced by the Lanczos iterations.

**Remark 5.3.1.** We can bound the first term that made up the upper bound for the residual of CG using previous convergence results of CG under exact arithmetic; recall [CG convergence rate \(theorem 3\)](#). It can be applied here for the first term in (5.3.6):  $\beta_k |\xi_k^T T_k^{-1} \xi_1|$ .

This is true because if we were to perform an CG on the  $T_{k+1}$  produced by the finite precision algorithm with the initial Lanczos vector  $q_1$  being  $\xi_1^{(n)}$ , then its residual  $\bar{r}_k$  would be exact and it's given as  $-\beta_k T_k^{-1} \xi_k q_{k+1}$ , but with  $q_{k+1} = \xi_{k+1}^{(n)}$ , the  $k+1$  th standard basis vector in  $\mathbb{R}^n$ , and  $T_k = (T_{k+1})_{1:k, 1:k}$ . And to our excitement, we already have the bound for  $\bar{r}_{k+1}$  proven in [theorem 3](#).

### 5.3.2 Paige's Theorem and Backwards Analysis of CG

We now introduce a new theorem proposed by Paige (**CITATION NEEDED**). It gives a bound to the floating point errors for the Conjugate gradient by bounding the conditions number of  $T_k$  from Lanczos iterations. It's stated as follow:

**Theorem 4** (Paige's Theorem). The eigenvalues  $\theta_i^{(j)}, i = 1, \dots, j$  of the tridiagonal matrix  $T_j$  satisfies:

$$\lambda_1 - j^{5/2}\epsilon_2\|A\| \leq \theta_i^{(j)} \leq \lambda_n + j^{5/2}\epsilon_2\|A\| \quad (5.3.7)$$

$$\epsilon_2 := \sqrt{2} \max\{6\epsilon_0, \epsilon_1\} \quad (5.3.8)$$

Along with this theorem, the following quantities are also defined and cited in Greenbaum's work (**CITATION NEEDED**) from the work of Paige.

$$\epsilon_0 \equiv 2(n+4)\epsilon \quad (5.3.9)$$

$$\epsilon_1 \equiv 2(7+m\| |A| \|/\|A\|)\epsilon \quad (5.3.10)$$

$$\epsilon_0 < \frac{1}{12} \quad k(3\epsilon_0 + \epsilon_1) < 1 \quad (5.3.11)$$

$$\|F_k\| \leq \sqrt{k}(\epsilon_1)\|A\| \quad (5.3.12)$$

$$\|q_j^T q_j - 1\| \leq 2\epsilon_0 \quad (5.3.13)$$

$$\beta_j \leq \|A\|(1 + (2n+6)\epsilon + j(3\epsilon_0 + \epsilon_1)) \quad (5.3.14)$$

The quantity  $k$  is the current iterations number of the Lanczos Iterations,  $j \leq k$ .  $m$  is the maximum number of non-zero element in the matrix  $A$ . Using Paige's theorem, we can bound the conditions number for the matrix  $T_{k+1}$  produced by the finite precision Lanczos, which is given by:

$$\tilde{\kappa} = \frac{\lambda_n + (k+1)^{5/2}\epsilon_2\|A\|}{\lambda_1 - (k+1)^{5/2}\epsilon_2\|A\|} \quad (5.3.15)$$

Using (remark 5.3.1), we can make the following proposition

**Proposition 5.1.**

$$|\beta_k \xi_k^T T_k^{-1} \xi_1| \leq 2\sqrt{\tilde{\kappa}} \left( \frac{\sqrt{\tilde{\kappa}} - 1}{\sqrt{\tilde{\kappa}} + 1} \right)^k \quad (5.3.16)$$

Where  $\tilde{\kappa}$  is the bound of the condition number of the  $T_k$  matrix (from (5.3.15)), the tridiagonal matrix produced by the finite precision Lanczos.

*Proof.* Using the lemma 6.1.1 in appendix, we can derive the relations between the 2 norm of the relative residuals and the energy norm of the relative error:

$$\frac{\|Ae_k\|}{\|Ae_0\|} \leq \kappa(T_k) \frac{\|e_k\|_A}{\|e_0\|_A} \leq 2\sqrt{\kappa(T_k)} \left( \frac{\sqrt{\tilde{\kappa}} - 1}{\sqrt{\tilde{\kappa}} + 1} \right)^k \quad (5.3.17)$$

$$\frac{\|r_k\|}{\|r_0\|} = |\beta_k \xi_k^T T_k^{-1} \xi_1| \quad \text{by (remark 5.3.1)} \quad (5.3.18)$$

$$\implies |\beta_k \xi_k^T T_k^{-1} \xi_1| \leq 2\sqrt{\tilde{\kappa}} \left( \frac{\sqrt{\tilde{\kappa}} - 1}{\sqrt{\tilde{\kappa}} + 1} \right)^k \quad (5.3.19)$$



The third inequality is simply from [CG Convergence Rate \(theorem 3\)](#) when we assume that the eigenvalues are uniformly distributed convex hull of the spectrum of  $A$ . The first fraction is actually the relative error of the 2 norm of the residual because  $Ae_k = r_k$  by definition. Substituting the quantity  $\kappa(T_k)$ , the conditions number of the matrix  $T_k$ , which we figured out using Paige's theorem and denoted it as  $\tilde{\kappa}$ .  $\square$

Finally, if we assume that  $T_k^{-1}$  is actually invertible, which requires that the conditions for all the quantities:  $\epsilon_0, \epsilon_1$  holds true, and  $\lambda_1 - (k+1)^{5/2}\epsilon_2\|A\| > 0$ . Finally, we make can bound the relative residual of the CG algorithm by considering:

$$\frac{\|r_{k+1}\|}{\|r_0\|} \leq \beta_k \|\xi_k^T T_k^{-1} \xi_1 q_{k+1}\| + \|F_k T_k^{-1} \xi_1\| \quad (5.3.20)$$

$$\leq \beta_k |\xi_k^T T_k^{-1} \xi_1| \|q_{k+1}\| + \|F_k\| \|T_k^{-1} \xi_1\| \quad (5.3.21)$$

$$\leq 2\|q_{k+1}\| \sqrt{\tilde{\kappa}} \left( \frac{\sqrt{\tilde{\kappa}} - 1}{\sqrt{\tilde{\kappa}} + 1} \right)^k + \sqrt{k}(\epsilon_1) \|A\| \|T_k^{-1}\| \quad (5.3.22)$$

Now, observe that  $|q_j^T q_j - 1| \leq 2\epsilon_0$  from (5.3.13), which implies that  $\|q_{k+2}\|^2 \leq (1 + 3\epsilon_0)$  which is  $\|q_{k+1}\| \leq \sqrt{1 + 2\epsilon_0}$ . In pursuit of mathematical beauty, we look for alternative expression for the quantity  $\|A\| \|T_k^{-1}\|$  giving us:

$$\|A\| \|T_k^{-1}\| = \frac{\lambda_n}{\lambda_1 - k^{5/2}\epsilon_2\|A\|} \leq \tilde{\kappa} \quad (5.3.23)$$

$$\implies \frac{\|r_{k+1}\|}{\|r_0\|} \leq 2\sqrt{1 + 2\epsilon_0} \sqrt{\tilde{\kappa}} \left( \frac{\sqrt{\tilde{\kappa}} - 1}{\sqrt{\tilde{\kappa}} + 1} \right)^k + \sqrt{k}(\epsilon_1) \tilde{\kappa} \quad (5.3.24)$$

Which completes the proof for the upper bound on the convergence rate for the Conjugate Gradient Method under floating point arithmetic.

**Remark 5.3.2.** This proof showed that the Conjugate Gradient method for  $T_{k+1}$  that is non-singular then it will be backwards stable. As long as the value of  $\beta_k$  is non zero, and the eigenvalues of  $T_{k+1}$  is not containing zero, the conjugate gradient method will continue to converge in the future iterations. So in layman's terms, it doesn't matter if the round-off error accumulated, Conjugate Gradient will converge as long as the problem is not too insane, or  $A$  being too pathological to deal with.

Finally, I want to point out the fact that Paige's theorem is derived using forward error analysis on the Lanczos Iterations, which is the absolute worst case. For most cases in modern computing platforms, the summation process of vector dot products has much higher floating-accuracy compare to older computing platforms due to the use of parallelism, or floating point specific summation instructions, which reduces the relative sizes for the summands, hence reducing the total round-off error accumulations.

## 5.4 Ghost Eigenvalues and Losing Orthogonality

The name Ghost Eigenvalues refers to the phenomena where the Lanczos Algorithm seems to find eigenvalues that are extremely close to each other, when in fact, those extremely close

eigenvalues are a single eigenvalue of the original matrix  $A$ 's spectrum. More specifically, the eigenvalues in  $T_k$  seems to cluster closely around the true eigenvalues of the original matrix  $A$ , especially eigenvalues in the exterior of  $A$ 's spectrum. We know for a fact that the tridiagonal matrix produced via Lanczos can't have any repeated eigenvalues, this follows from (remark 3.4.2). This phenomena is in fact, due to the limited floating point accuracy of the Lanczos orthogonalization process.

#### 5.4.1 Ghost Eigenvalues Experiments

Here, we conducted numerical experiments and carefully reproduce the phenomena for a diagonal matrix  $A$  with diagonals given by the formula:  $\lambda_i = \left(-1 + \frac{2(i-1)}{(n-1)}\right)^3$  where  $A \in \mathbb{R}^{n \times n}$ . This matrix is particularly good for reproducing the phenomena. For this experiment, we set  $n = 64$  and we use Float64. We run the Lanczos iterations with  $q_1$  being the vector of all ones, we marked the smallest and largest 10 eigenvalues during the iterations and plotted out their trajectories from iteration 20 to 64. The results can be seemed in (figure 5.4.1 left). On figure 5.4.1 right, we made the plot for what would happen if the Lanczos iterations are free of numerical round off error. We didn't use exact arithmetic, instead, we simply reorthogonalized all the Lanczos vector  $q_k$  using all previous gotten Lanczos vectors to emulate the effects, which is just an Arnoldi iterations. Please bare in mind that there are eigenvalues in the middle interior part of the spectrum, they are just not plotted in the figure.

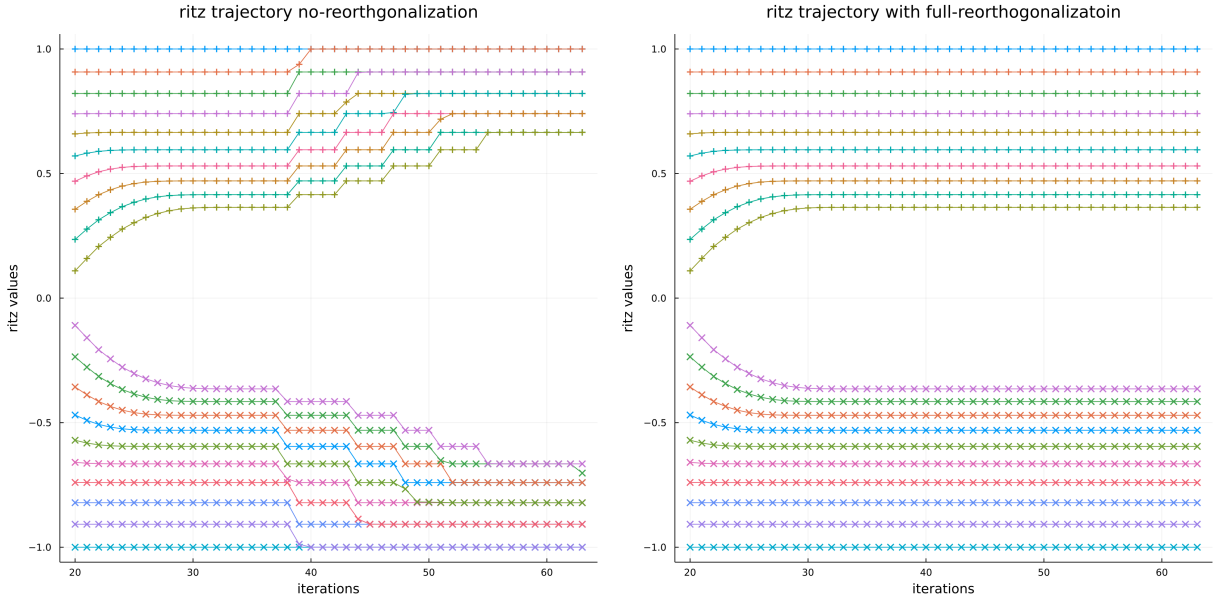


Figure 5.4.1: The highest and lowest 10 eigenvalues of the matrix  $T_k$  during the Lanczos Iterations are being tracked by their relative order. During each iterations, the first, the second, the third, ... etc eigenvalues of  $T_k$  are linked together by a line in a different color. Left is Lanczos with numerical round off error, right is Lanczos iterations that fully reorthogonalize  $q_k$  for each iteration which emulates the behaviors under exact arithmetic.

Recall from the Cauchy Interlace Theorem, the eigenvalues of the tridiagonal matrix  $T_{k+1}$  has to be in between each eigenvalues of  $T_k$  except for the first and the last eigenvalue of  $T_{k+1}$ . This implies that, the  $\theta_i^{(k)}$ , the  $i$  th eigenvalues during the  $k$  th iteration will move

monotonically upwards or downwards during the Lanczos iterations. The ghost eigenvalues on the figure appears when some of the interior eigenvalues suddenly switched to another eigenvalue's trajectory that is on the exterior of the spectrum. It appears as the matrix  $T_k$  has repeated eigenvalues but in fact they are not and they are just very close.

However, judging the eigenvalues of the matrix  $T_k$  alone will distinguish us whether 2 very close eigenvalues correspond to 2 different eigenvalues of  $A$  or it's due to the floating point round off error, even if their trajectories seems to suggest it. We can't tell it because if I keep the matrix  $T_k$  generated from an Lanczos iterations with huge amount of round off error and all it  $A$ , which has eigenvalues that are extremely close to each other and their path shifts over each other on the trajectory like in [fig 5.4.1](#). Now consider performing another Lanczos iterations  $A := T_k$  it but with the initial vector  $\xi_1$ , then we will exactly reproduce  $A$  itself because it's tridiagonal. But in this case, the eigenvalues of  $T_k$  is exact after termination of Lanczos iteration. In this case, all eigenvalues are actually presented in the original matrix  $A$ , which is just  $T_k$ , itself.

In fact, the ghost eigenvalues here are produced by floating point errors because firstly we know what the actual eigenvalues of  $A$  is, we made  $A$ . To make sense of it better intuitively, we observe conduct experiments to show that the lost of orthogonality of  $Q_k$  happens together with ghost eigenvalues on spectrum of  $T_k$ . If  $Q_k$  matrix is perfectly orthogonal, then there is no ghost eigenvalues, regardless of how the trajectories of the eigenvalues of  $T_k$  looks like. In fact, a corresponding plot of  $Q_k^H Q_k$  are plotted in ([fig 5.4.1 left](#)) for demonstrating lost of orthogonality for the same diagonal matrix  $A$  proposed earlier. We plotted the heat map of the matrix  $Q_k^H Q_k$  directly as well.

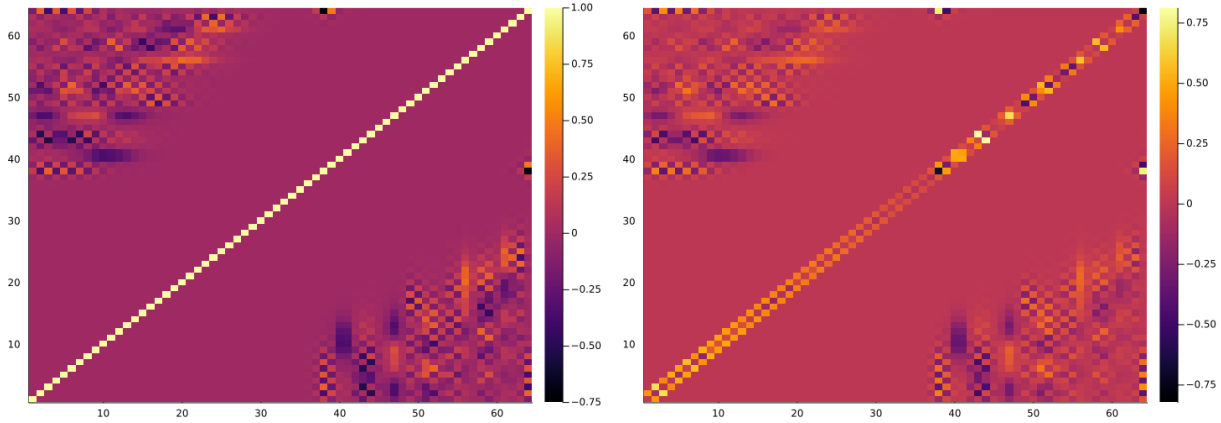


Figure 5.4.2: left: The heatmap of the plot of  $\log_2$  of the absolute values of the matrix  $Q_k^H Q_k$ . right: The plot of  $Q_k A Q_k$  from floating point Lanczos iterations

In addition to the lost orthogonality of the matrix  $Q_k$ , we also visualized the actual tridiagonal matrix reproduced by  $Q_k A Q_k$  which is plotted in ([fig 5.4.1 right](#)). Observe that most of the off tridiagonal entries are non-zero and relatively huge, which doesn't worry us too much because  $A$  itself has an extremely bad conditions number. What is important is the blob of non-zero entries on the top left and the bottom right of the plot. And this is a hint that the lost of orthogonality will create extra errors on the off tridiagonal parts of the matrix  $T_k$ .

**Remark 5.4.1.** As a final remark for these numerical experiments, I suggest an intuitive

way of understanding them. Which will be useful when we actually wish to analyze it rigorously. Simply put, the Lanczos Iterations might “forget” about the eigenvalues when it converged (Manifested as the usually stable trajectories of eigenvalues on the exterior of the spectrum for the matrix  $T_k$  in (fig 5.4.1)), and when it happens, the Lanczos vectors produced by the algorithm has lost its orthogonality correspondingly, which then causes the interior eigenvalues of  $T_k$  to shift, creating ghost eigenvalues for the matrix  $T_k$ .

#### 5.4.2 Losing Orthogonality on Ritz Vectors

To gain a better understanding, let's define the notion of Ritz values and Ritz vectors. For our discussion, the Ritz value  $\theta_i^{(k)}$  are the  $i$  th eigenvalues of the matrix  $T_k$  from the and the Ritz vectors are  $Q_k s_i^{(k)}$  where  $s$  is the  $i$  th eigenvector for the matrix  $T_k$ . Recall from [remark 3.4.2](#), the characteristic polynomial of  $T_k$  is the monic that minimizes the 2 norm error for the vector  $p_k(A)q_1$ ; intuitively, the Ritz values and Ritz vectors approximates eigenvalues and eigenvectors of matrix  $A$ . One would expect once the eigenvalue converged for  $T_k$ , say  $\theta_i^{(k)} = \lambda_i + \epsilon$  where  $\lambda_i$  is an eigenvalue of  $A$ , then the vector  $Q_k s_i^{(k)}$  approximate eigenvalue of  $A$  because:

$$AQ_k = Q_k T_k + q_{k+1} \beta_k \xi_k^T \quad (5.4.1)$$

$$AQ_k s_i^{(k)} = \theta_i^{(k)} Q_k s_i^{(k)} + q_{k+1} \beta_k \xi_k^T v \quad (5.4.2)$$

$$AQ_k s_i^{(k)} = \theta_i^{(k)} Q_k s_i^{(k)} + \beta_k q_{k+1} (s_i^{(k)})_k \quad (5.4.3)$$

$$AQ_k s_i^{(k)} = (\lambda_i + \epsilon) Q_k s_i^{(k)} + \beta_k q_{k+1} (s_i^{(k)})_k \quad (5.4.4)$$

$$AQ_k s_i^{(k)} = \lambda_i Q_k s_i^{(k)} + \epsilon Q_k s_i^{(k)} + \beta_k q_{k+1} (s_i^{(k)})_k \quad (5.4.5)$$

When  $\theta$  is close to  $\lambda_i$ , and the characteristic polynomial of  $T_k$  interpolates the eigenvalues of  $A$  better and better,  $\beta_k$  decreases (**NONTRIVIAL FACT**), the Ritz vector  $Q_k s_i^{(k)}$  approximates the eigenvector for  $A$ . Intuitively, as the Krylov subspace expanded, it contains more and more components of the Eigenvectors of  $A$ , and the Ritz vector will have more “room” to get closer to the eigenvector of  $A$ , by the approximation property of Lanczos.

However, under floating point arithmetic, once the Ritz vector  $s_i^{(k)}$  can be spanned by the Krylov subspace with small error, it will be the case that the projection of the latest Lanczos vector onto the Ritz vector begins to grow. On the plot showed in [figure 5.4.2](#), we projected the log absolute value of  $q_k^T Q_k s_i^{(k)}$  for  $i = 1, 2, 3$  for all  $k = 20, \dots, 64$ , and we used the same setup from the last section where we demonstrated ghost eigenvalues. For comparison, [figure 5.4.2](#) showed us the what happen in exact arithmetic. One very important observation to make from [figure 5.4.2](#) is that the peak of the blue curve, projection onto the largest Ritz value happens around iteration 3.8, and around that exact same iteration in [figure 5.4.1](#) is when the second largest eigenvalue of  $T_k$  decided to shifts over to the blue curve. Bear in mind that this is in a log plot, and without the log it looks like a sharp spike.

### 5.5 The Ritz Vector's View and Another Paige's Theorem

In this section, we prove another analysis of the Lanczos Iterations from Dammal's work(**Citation Needed**) where he introduced proof of another Paige's theorem that shows exactly how to

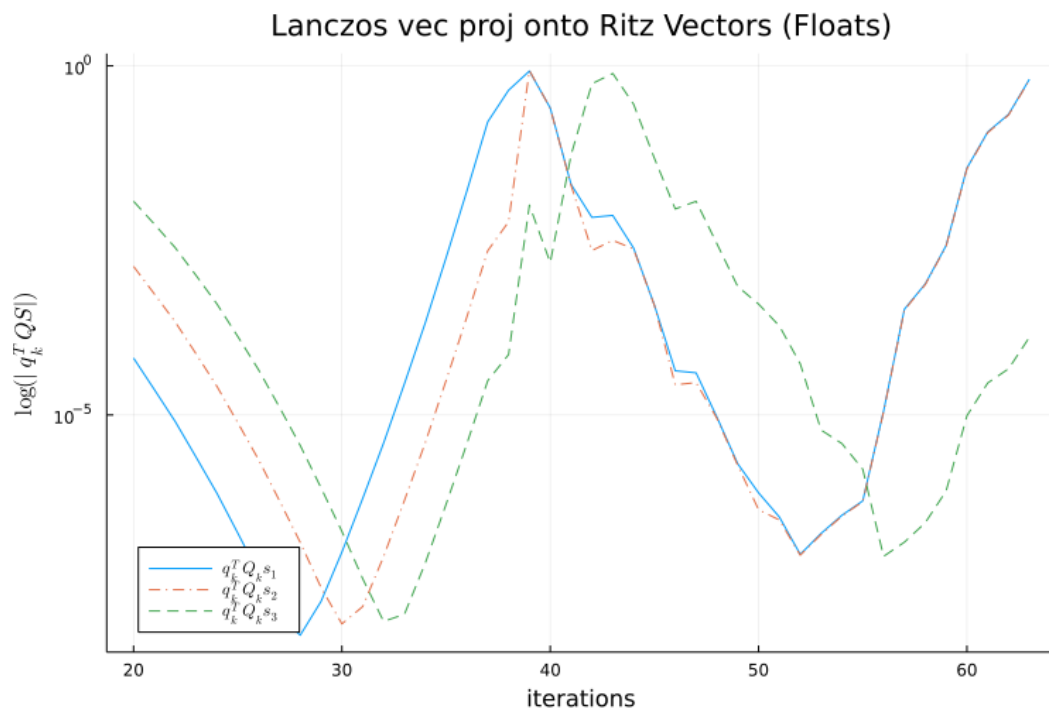


Figure 5.4.3: figure 4 caption

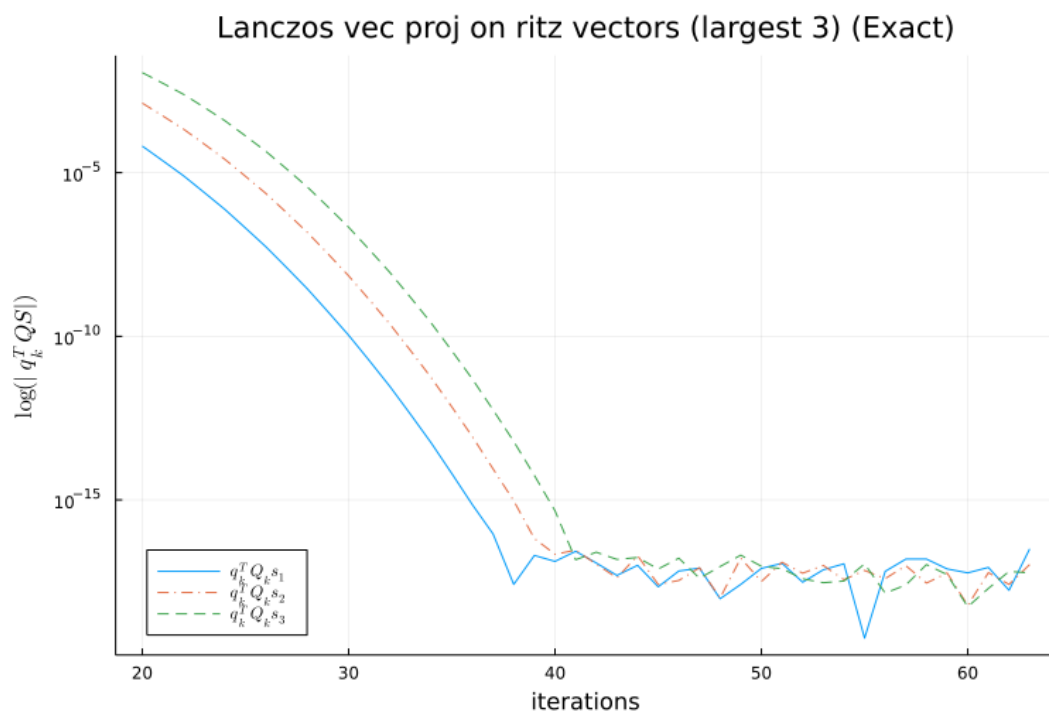


Figure 5.4.4: figure 5 caption

measure the lost of orthogonality of the Lanczos vectors while Lanczos algorithm is running. Here, we wish to prove it in much more details with more thoroughness.

The theorem highlights 2 important facts. The first is the lost of orthogonality of Lanczos Vector and Ghost eigenvalues appears at the same time and they are systematic. The second is that the projection of the Lanczos vector onto the converged Ritz vectors can be numerically attained, which tells us how much lost of orthogonality is occurring and which direction we need to re-orthogonalize so that the Lanczos vectors retains orthogonality. Here is the statement of the theorem:

**Theorem 5** (Another Paige's Theorem).

## 5.6 Forward Error Analysis

## 6 Appendix

### 6.1 Useful Lemmas

**Lemma 6.1.1** (Relative Energy Norm and Relative 2 Norm Conversions). Let  $A$  be a Positive Symmetric Positive Definite Matrix, then it can be said that:

$$\frac{\|Ax\|}{\|Ay\|} \leq \kappa(A) \frac{\|x\|_A}{\|y\|_A}$$

*Proof.* From the definition of included 2-norm of matrices, assuming that  $\lambda_1$  is the minimum eigenvalue of the matrix  $A$ , and  $\lambda_n$  the maximum, and the fact that matrix  $A$  has factorization  $A^{1/2}A^{1/2}$ :  $\square$

$$\lambda_1\|x\| \leq \|Ax\| \leq \lambda_n\|x\| \quad (6.1.1)$$

$$\sqrt{\lambda_1}\|x\| \leq \|A^{1/2}x\| \leq \sqrt{\lambda_n}\|x\| \quad (6.1.2)$$

$$\implies \sqrt{\lambda_1} \leq \frac{\|Ax\|}{\|A^{1/2}x\|} \leq \sqrt{\lambda_n} \quad (6.1.3)$$

Consider another vector  $y$ :

$$\sqrt{\lambda_1} \leq \frac{\|Ay\|}{\|A^{1/2}y\|} \leq \sqrt{\lambda_n} \quad (6.1.4)$$

Combining the two we have:

$$\sqrt{\lambda_1} \frac{\|Ax\|}{\|A^{1/2}x\|} \leq \sqrt{\lambda_n} \sqrt{\lambda_1} \quad (6.1.5)$$

$$\sqrt{\lambda_1} \sqrt{\lambda_n} \geq \sqrt{\lambda_n} \frac{\|Ay\|}{\|A^{1/2}y\|} \quad (6.1.6)$$

$$\implies \sqrt{\lambda_1} \frac{\|Ax\|}{\|A^{1/2}x\|} \leq \sqrt{\lambda_n} \frac{\|Ay\|}{\|A^{1/2}y\|} \quad (6.1.7)$$

$$\frac{\|Ax\|}{\|A^{1/2}x\|} \leq \sqrt{\kappa(A)} \frac{\|Ay\|}{\|A^{1/2}y\|} \quad (6.1.8)$$

$$\frac{\|Ax\|}{\|Ay\|} \leq \sqrt{\kappa(A)} \frac{\|A^{1/2}x\|}{\|A^{1/2}y\|} \quad (6.1.9)$$

$$\frac{\|Ax\|}{\|Ay\|} \leq \sqrt{\kappa(A)} \frac{\|x\|_A}{\|y\|_A} \quad (6.1.10)$$

### 6.2 Theorems, Propositions

**Proposition 6.1** (Krylov Subspace Grade Invariant Theorem). Once the subspace becomes linearly dependent, the subspace becomes invariant.

*Proof.*

$$K_k = [b \quad AB \quad \cdots \quad A^{k-1}b] \quad (6.2.1)$$

$$K_k \text{ Lin Dep} \implies A^{k-1}b = K_{k-1}c_k \quad (6.2.2)$$

$$\implies AK_k = K_k \underbrace{[e_2 \quad \cdots \quad e_k \quad c_k]}_{:=C_k} \quad (6.2.3)$$

$$\implies A^2K_k = AK_kC_k = K_kC_k^2 \quad (6.2.4)$$

$A^2K_k$  will span the same space as the range of the matrix  $K_k$ . □