UNIVERSITY OF CALIFORNIA,
IRVINE

Numerical, spectral, and group properties of random butterfly matrices

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematics

by

John Peca-Medlin

Dissertation Committee:
Michael Cranston, Chair
Thomas Trogdon
Roman Vershynin

2021

DEDICATION

To Arlo and Corbyn

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ACKNOWLEDGMENTS

I would like to thank my advisors, Tom Trogdon and Mike Cranston. Their patience and support has enabled me to more successfully traverse the many hills and occasional potholes on this mathematical landscape. In particular, I would like to thank Tom for introducing me to butterfly matrices and for being a continued resource, guide, and collaborator throughout this project.

This work was partially supported by the National Science Foundation through the grant NSF DMS-1916492 (AGEP-GRS Supplement), and by the UCI Math Department through the Community, Outreach, and Mentoring Program Fellowship.

Most importantly, I want to give a special thanks to my wife, Megan, for allowing me to pursue this academic endeavor and for helping provide me the sustenance – of food, coffee and belief – to make it all the way to the end. And last, I want to thank both of my sons, Arlo and Corbyn, for showing me how to never stop.

VITA

John Peca-Medlin

EDUCATION

Doctor of Philosophy in Mathematics **2021**
University California, Irvine *Irvine, CA*

Master of Science in Mathematics **2017**
University California, Irvine *Irvine, CA*

Bachelor of Arts in Mathematics **2006**
University of Chicago *Chicago, IL*

ABSTRACT OF THE DISSERTATION

Numerical, spectral, and group properties of random butterfly matrices

By

John Peca-Medlin

Doctor of Philosophy in Mathematics

University of California, Irvine, 2021

Michael Cranston, Chair

The recursive structure of butterfly matrices has been exploited to accelerate common methods in computational linear algebra. This was first developed by D. Stott Parker [31]. Recently, the machine learning community has taken particular interest in these applications. Butterfly structures can now be found integrated into architectures for software used in learning fast solvers for large linear systems and in image recognition, covering tasks such as early cancer identification or smart vehicle navigation [1, 6, 27]. These new advances have enabled less powerful computing systems, such as in mobile devices or portable smart devices, to effectively utilize computationally heavy tools that were previously unavailable. Although empirical evidence supports the use of butterfly matrices in these newer technologies, the literature on the mathematical theory that justified these results is lacking. Building on research started in [37], I will give a fuller picture of the numerical, spectral, and group properties of particular ensembles of random butterfly matrices. This document will provide a stronger mathematical foundation to further support the approaches already found in practice and can inform future applications not yet explored.

Chapter 1

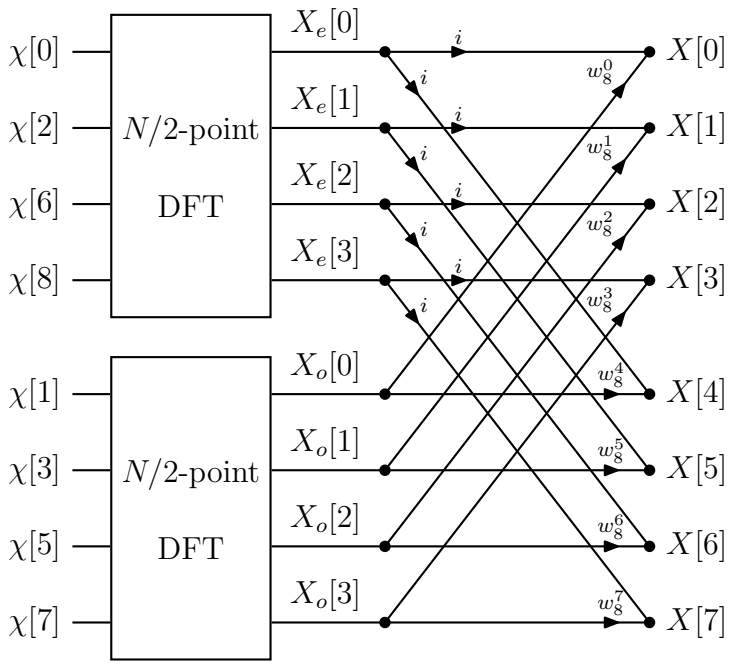
Introduction

> In the light of the moon a little egg lay on a leaf.
>
> Eric Carle

Random butterfly matrices were first introduced by D. Stott Parker in 1995 as a matrix realization of a recursively constructed randomization algorithm related to the Fast Fourier Transform (FFT). Butterfly matrices get their name from the associated data flow "butterfly" diagram used in the radix-2 Cooley-Tukey FFT (see Figure 1.1, which is adapted from [29]). The FFT has many desirable attributes and has become a modern staple in a wide swath of disciplines, including engineering, mathematics and music. The FFT has been in wide use since the mid-1960s, when Cooley and Tukey published their eponymous algorithm that allowed efficient computation of the Fourier transform, enabling its use in signal processing and image compression [3]. The IEEE magazine *Computing in Science & Engineering* has cited the FFT as one of the 10 most important algorithms of the 20th century [12].

Parker's focus on butterfly matrices was on their application in removing the need for pivoting in Gaussian elimination. We say a matrix is **block degenerate** if some upper left subblock is

Figure 1.1: Data flow radix-2 butterfly diagram for a DFT for $N = 8$, adapted from [29]



singular, in which case an attempt at running Gaussian elimination without pivoting would halt at an attempt to divide by 0. Parker showed the following application of butterfly matrices can ensure any nonsingular matrix can be transformed into a block nondegenerate matrix.

Theorem 1.1 ([31])**.** *If A is a nonsingular square matrix and U, V are random butterfly matrices, then U^*AV is block nondegenerate with probability 1.*

For an order N matrix A and a vector $\mathbf{x} \in \mathbb{R}^N$, the matrix-vector product $A\mathbf{x}$ can be computed using $O(N^2)$ arithmetic operations. Using the block structure of an order N butterfly matrix Ω, one can carry out the matrix-vector multiplication $\Omega\mathbf{x}$ using $O(N \log_2 N)$ steps. Hence, we can carry out a matrix-matrix multiplication in $O(N^2 \log_2 N)$ steps. So when solving the linear system

$$A\mathbf{x} = \mathbf{b} \tag{1.1}$$

using Gaussian elimination, which takes $O(N^3)$ steps, we can instead solve the equivalent system

$$(UAV^*)(V\mathbf{x}) = U\mathbf{b} \tag{1.2}$$

now using Gaussian elimination *without pivoting*, for U, V random butterfly matrices. The above discussion shows that only $O(N^2 \log_2 N)$ operations are needed to get from (1.1) to (1.2), and this does not impact the leading-order complexity of Gaussian elimination.

The additional scans needed to use Gaussian elimination with partial pivoting, compared to Gaussian elimination without pivoting, also do not impact the leading complexity (unlike for complete pivoting). However, the costs of moving large amounts of data using pivoting can be substantial on many high performance machines, as the data movement can interrupt the data flow. In [2], Baboulin, Li and Rouet showed that pivoting comprised approximately 20 percent of the total process time in an implementation of Gaussian elimination with full pivoting. Pivoting and the communication overhead to coordinate data movements also is a hindrance for parallel architectures and block algorithms [31]. Removing the need of pivoting would clear up this potential bottleneck to enable faster computations.

The use of butterfly architectures have recently spiked in the machine learning and image processing communities, with a particular rise in Convolution Neural Network (CNN) architectures [1, 6, 27]. However, applications are ahead of the theoretical understanding of the properties of random butterfly matrices. Building on top of [31, 37], the purpose of this document is to fill in some of these missing pieces.

1.1 Outline

The remainder of Chapter 1 will introduce tools and results in algebra, linear algebra, numerical analysis, probability and random matrix theory that will be used throughout this document. In particular, later chapters will rely heavily on the properties of the Kronecker product of matrices (Section 1.4.4) and an application of the Subgroup algorithm (Section 1.6.4). A sufficient background for Gaussian elimination is provided, which includes a formal proof regarding the uniqueness of the factors in the LU factorization using Gaussian elimination with partial pivoting (see Theorem 1.10).

Additional preliminary material is established to support two independent results found in Appendices A and B. The first project (Appendix A) relates to the divisibility of random integers, which can be computed in a simple closed form using the characteristic function (see Proposition A.1). This can be used to give alternative proofs for standard results of limiting distributions of particular random variables, such as the binomial distribution.

The second project (Appendix B) provides a novel result showing spacings between ocean waves can be accurately modeled using random matrices. Similar random matrix statistics have been found to model other mathematical objects as well as physical systems, such as transportation systems. Our project provides a sequence of new filtering techniques for ocean wave data, using both frequency and time domain. We then establish standard statistical significance of the model fit for the normalized spacings between peaks of successive ocean waves to the Wigner surmise, which approximately models the spacings between eigenvalues of Gaussian Unitary Ensemble random matrices. The remaining chapters focus on butterfly matrices.

Chapter 2 gives a thorough background for butterfly matrices, including group and topological properties, as well as introducing new general butterfly models of arbitrary order. A particular focus of this document falls on butterfly matrices formed using Kronecker products.

These structures enable closed-formed computations for particular statistics or factorizations that are usually difficult or can only be approximated throughout the randomized linear algebra literature. Additionally, new connections between butterfly matrices and Hadamard matrices are established, including a new method to generate Hadamard matrices using butterfly matrices. Chapter 3 introduces the main definitions for random butterfly matrices, which includes an overview of classification of the general Haar-butterfly models, which are random butterfly matrices formed using Kronecker products. Chapter 4 gives an overview of some spectral properties of butterfly matrices. This chapter begins a particular focus on the Haar-butterfly models, whose Kronecker product structure enables closed-form matrix factorizations, such as the eigenvalue and LU decompositions.

Chapter 5 focuses on numerical properties of butterfly matrices. This chapter explores the application of random butterfly matrices to remove the need for pivoting when using Gaussian elimination to solve a linear system. Parker showed randomizing a linear system on the left and right (two-sided randomization) by random butterfly matrices almost surely enables a linear system to have an LU factorization (see Theorem 1.1). Some implementations of random butterfly models already in practice have used a simplified one-sided randomization architecture. We present a result that shows one-sided randomization does not ensure an LU factorization exists when using Haar-butterfly models. Section 5.2 gives an overview on sampling Haar orthogonal matrices before introducing a butterfly QR algorithm, which enables Haar orthogonal matrices sampling using butterfly matrices.

Section 5.3 provides a novel result giving the full distribution of the growth factors of Haar-butterfly matrices using no pivoting, partial pivoting, or rook pivoting. This is a significant step forward for analysis of growth factors of random matrices, which has been limited to first moment estimates in the existing literature. Additional results relating to complete pivoting growth factors are also introduced, which connect to an open problem relating to Hadamard matrices.

1.2 Notation

Throughout the majority of the text, for an integer n we will write $N = 2^n$ so that $n = \log_2 N$. Write $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$, respectively, for the integers, rationals, real numbers and complex numbers. Let $\mathbb{F}$ denote one of these sets during this section, although most of the text will use $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. Write $\mathbb{1}_A$ for the indicator function of the set A and $\delta_{ij} = \delta(i, j)$ for the Kronecker delta function. Let $[n] = \{1, 2, \ldots, n\}$.

The power set of a set X will be denoted $2^X = \{A : A \subset X\}$. The **commutator** of two elements in a ring is $[x, y] = xy - yx$, and we say x, y commute if $[x, y] = \mathbf{0}$. For $z = a + bi \in \mathbb{C}$, we write $\bar{z} = a - bi$ for the complex conjugate, and $\bar{\mathbf{x}}$ for the vector where we take the complex conjugate of each component of $\mathbf{x}$. Similarly, treating an $n \times m$ matrix as a vector in $\mathbb{C}^{nm}$, we can define $\overline{A}$ for $A \in \mathbb{C}^{n\times m}$. For matrix notation, we write A^T to denote the transpose of A so that $(A^T)_{ij} = A_{ji}$, and A^* to denote the conjugate transpose $A^* = \overline{A^T}$. For $f : R \to S$ a function, we will occasionally shorthand the induced map from $R^n \to S^n$ such that f is applied at each component again as $f : R^n \to S^n$ (e.g., $f(\mathbf{x}) = (f(x_1), \ldots, f(x_n))$ for $\mathbf{x} = (x_1, \ldots, x_n) \in R^n$). If $X, Y \subset Z$ for Z a multiplicatively closed set, then $XY = \{xy \in Z : x \in X, y \in Y\}$.

Let $\mathbf{e}_i$ denote the standard basis elements of $\mathbb{F}^n$ with j^{th} component δ_{ij}. Let $\mathbf{E}_{ij} = \mathbf{e}_i\mathbf{e}_j^T$ be the standard basis elements of $\mathbb{F}^{n\times m}$, the collection of $n \times m$ matrices with entries in $\mathbb{F}$. Sometimes I will also write $M_n(\mathbb{F})$ or $M_{m,n}(\mathbb{F})$ to denote $\mathbb{F}^{n\times n}$ and $\mathbb{F}^{n\times m}$, respectively, while $\mathbb{F}$ may be suppressed if it does not needed to be emphasized. For $A \in \mathbb{F}^{n\times m}$, write A_{ij} to denote the element in row i and column j (where $A_{ij} = \mathbf{e}_i^T A \mathbf{e}_j$). Note $(\mathbf{E}_{k\ell})_{ij} = \delta_{ik}\delta_{j\ell}$. By definition, we have

$$\mathbf{E}_{ij}\mathbf{E}_{k\ell} = \delta_{jk}\mathbf{E}_{i\ell}. \tag{1.3}$$

Write $\mathbf{I}$ for identity matrix and $\mathbf{0}$ for the zero matrix or vector, where the dimensions are either made explicit by using appropriate subscripts (e.g., $\mathbf{I}_2$ is the order 2 identity matrix) or implicit from context. Let $\mathbf{1}_n = \sum_i \mathbf{e}_i$ be the vector of all ones. We will write $f(n) = O(g(n))$ if there exists a constant $C > 0$ such that $f(n) \leq Cg(n)$ for n sufficiently large.

Let $S^{n-1} = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 = 1\}$, with $S^1 = \{(\cos\theta, \sin\theta) : \theta \in [0, 2\pi]\}$. Let $\mathbb{T}^n = (S^1)^n$ denote the higher dimensional tori, with $\mathbb{T}^2 = S^1 \times S^1$ the standard torus.

Let $\mathrm{O}(n)$ denote the set of order n (real) orthogonal matrices O such that $O^{-1} = O^T$ and $\mathrm{U}(n)$ the set of order n (complex) unitary matrices U with $U^{-1} = U^*$. Let $\mathrm{SO}(n) \subset \mathrm{O}(n)$ and $\mathrm{SU}(n) \subset \mathrm{U}(n)$ denote the special orthogonal and special unitary subgroups, which are defined by the additional criterion that such matrices have unit determinant.

Let $|A|$ denote the matrix such that $|A|_{ij} = |A_{ij}|$, i.e., apply the absolute value entrywise throughout the matrix. For $b \in \mathbb{R}$ and $n > 0$, let $b \pmod n$ denote the (unique) number $d \in [0, n)$ such that $d \equiv b \pmod n$.

1.3 Algebra

This section gives an overview of tools and definitions from algebra that will be used later in the text.

1.3.1 Divisibility

We say for $d, n \in \mathbb{Z}$ that d **divides** n, written $d \mid n$, if there exists $c \in \mathbb{Z}$ such that $n = dc$. Write the **greatest common divisor** of n and m as

$$\gcd(n, m) = \min\{|d| : d \mid n, d \mid m\} = \min_{a,b \in \mathbb{Z}} |an + bm|. \tag{1.4}$$

Recall Euclid's algorithm can be used to compute $\gcd(n, m)$ as follows: with $n \geq m$, and write $n = qm + r$ for integers q, r with $0 \leq r \leq m$; if $r = 0$, return m, otherwise set $(n, m) = (m, r)$ and restart. In particular, working backward through this algorithm, one would then be able to write d as a linear combination of n and m. Let the **least common multiple** of n and m be $\operatorname{lcm}(n, m) = \min\{M \in \mathbb{Z} : n, m \mid M; M \geq 1\}$. Recall $nm = \gcd(n, m) \cdot \operatorname{lcd}(n, m)$. If $\gcd(n, m) = 1$, then n and m are **relatively prime** or **coprime**. $p \in \mathbb{Z}$ is a **prime number** if $d \mid p$ implies $d = \pm 1$ or $d = \pm p$. Let $\mathcal{P} = \{2, 3, 5, \ldots\}$ denote the set of all positive prime numbers. The following result shows that $\mathbb{Z}$ is a unique factorization domain such that every integer can be written as a unique product of powers of prime numbers:

Theorem 1.2 (Fundamental Theorem of Arithmetic)**.** *If n is a positive integer, then there exists a unique sequence of nonnegative integers* $\mathbf{e}$ *such that $n = \prod_{p \in \mathcal{P}} p^{e_p}$.*

A nonzero integer is **composite** if it is not prime.

1.3.2 Group theory

A nonempty set G with an associative binary operator (which we will denote by multiplication) is a **group** if there exists $e \in G$ such that $ge = eg = g$ for all $g \in G$, and for every $g \in G$ there exists $g^{-1} \in G$ such that $gg^{-1} = g^{-1}g = e$. I will often write $e = 1$ when convenient. G is **abelian** if the group action is commutative, i.e., $ab = ba$ for all $a, b \in G$. Let C_n denote the **cyclic group of order** n such that $C_n = \{x^k : x^n = 1, k \in \mathbb{Z}\}$. $g_1, g_2 \in G$ are **conjugate** if there exist $h \in G$ such that $g_1 = hg_2h^{-1}$. Conjugacy constitutes an equivalence relationship on G, and hence G is a disjoint union of its conjugacy classes. If G is abelian, then every element constitutes a distinct conjugacy class. We say $g \in G$ has **order** k if k is the minimal positive integer such that $g^k = 1$. Moreover, if g has order k, then $g^m = 1$ if and only if $k \mid m$.

$H \subset G$ is a **subgroup** of G if H is closed under the the group action of G and inverses. A

straightforward verification confirms

Proposition 1.1 (Subgroup criterion)**.** *Let G be a group and $\varnothing \neq H \subset G$. Then H is a subgroup of G if and only if $xy^{-1} \in H$ for all $x, y \in H$.*

N is **normal** in G if N is a subgroup of G and for all $x \in N, y \in G$, $yxy^{-1} \in N$. The quotient space of **(left) cosets** of a subgroup H, denoted G/H, is the collection of sets $\{xH : x \in G\}$ where $xH = \{xh : h \in H\}$. Let $[G : H] = |G/H|$ denote the **index** of H in G. If N is a normal subgroup, then G/N is a group with group action $(xN)(yN) = xyN$. If $[G : H] = 2$, then H is necessarily normal in G since the left and right nontrivial coset of H must align so that $xH = Hx$ for $x \notin H$. A function $f : G \to H$ for groups G and H is a **group homomorphism** if $f(ab) = f(a)f(b)$. Furthermore, if f is bijective, then f is a **group isomorphism**, and we write $G \cong H$ to denote G and H are isomorphic groups. Let $\ker f = \{g \in G : f(g) = 1\}$ be the **kernel** of a group homomorphism f. Then $\ker f$ is a normal subgroup of G while $f(G)$ is a subgroup of H.

Theorem 1.3 (Fundamental Homomorphism Theorem)**.** *If $f : G \to H$ is a group homomorphism, then $G/\ker f \cong f(G)$.*

Proof. This follows naturally through restriction from the commutative diagram

$$\begin{array}{ccc} G & \xrightarrow{\pi} & G/\ker f \\ & \searrow^{f} & \downarrow{\scriptstyle \widetilde{f}} \\ & & H \end{array}$$

for $\pi : G \to G/\ker f$ the natural quotient map and $\widetilde{f}$ the induced map from this diagram. □

In particular, if f is a surjective group homomorphism, then $G/\ker f \cong H$. For example, $\det : \mathrm{O}(N) \to \{\pm 1\}$ is a group homomorphism with $\ker \det = \mathrm{SO}(N)$.

If N is normal in G, then HN is a subgroup of G: for any $hn, h'n' \in HN$ we have $(hn)(h'n')^{-1} = (hh'^{-1})(h'(nn'^{-1})h'^{-1}) \in HN$. Furthermore, if $H \cap N = \{1\}$, then $HN =: H \ltimes N$ is a **semidirect product** of H and N. Write $H \ltimes_\varphi N$ for $\varphi(n)$ an inner group homomorphism on H such that $(h_1n_1)(h_2n_2) = (h_1\varphi(n_1)(h_2))(n_1n_2)$. For example, the dihedral group of $2n$ elements can be defined as

$$D_{2n} = C_2 \ltimes C_n = \langle ab^k : a^2 = b^n = 1, aba = b^{-1} \rangle, \tag{1.5}$$

where then $\varphi(a)(b) = b^{-1}$. If $G = HK$ where $H \cap K = \{1\}$ and both H and K are normal in G, then $G = H \times K$ is the **(direct) product** of H and K; equivalently, if $G = H \ltimes_\varphi K$ where $\varphi(k) : H \to H$ is the identity map for all k, then $G = H \times K$.

Let V be a **vector space** over k, which is an abelian group under addition while addition satisfies the distribution property with respect to scalar multiplication by elements in k. Say V has dimension n if V is spanned by a set of n linearly independent vectors in V. Let $\mathrm{GL}(V)$ denote the **general linear group**, which consists of the invertible maps on the vector space V with composition as the group action. Note $\mathrm{GL}(\mathbb{C}) = \mathbb{C}\backslash\{0\}$, which has **torsion part** (i.e., the elements of finite order) the **roots of unity**. If $\rho : G \to \mathrm{GL}(V)$ where V is dimension n, then we say ρ is an n-dimensional **representation** of G. A subspace $W \subset V$ is G**-invariant** if $\rho(g)W \subset W$ for all $g \in G$. In this case, the restriction of $\widetilde{\rho} : G \to W$ such that $\widetilde{\rho}(g) = \rho(g)|_W$ is a **subrepresentation** of ρ; this term is also used for the subspace W itself. The trivial subrepresentations of ρ are the associated subrepresentations for $W = V$ or $W = \{0\}$. If the only subrepresentations of ρ are trivial, then ρ is **irreducible**. Note 1-dimensional representations are always irreducible. Every finite-dimensional representation of G can be written as a direct sum of irreducible subrepresentations, which will be called its **factors**. $\chi_\rho = \mathrm{Tr} \circ \rho$ denotes the **character** of ρ. Moreover, the characters of ρ can be used to uniquely determine the irreducible factors of ρ. Every finite-dimensional representation of G is uniquely determined by $\chi_\rho(g)$ for $g \in G$. If $g \in G$ has order k, then $\rho(g)^k = \rho(g^k) =$

$\rho(1) = \mathbf{I}$. Hence, if λ is an eigenvalue of $\rho(g)$ then $\lambda^k = 1$ (see Section 1.4) so that λ is a k^{th} **root of unity**, and hence $\lambda = e^{2\pi ji/k}$ for some $j \in \mathbb{Z}$. In particular, note if $|G| = n$ is finite, then any 1-dimensional representation of G must satisfy $\rho(G) \subset \{e^{i\pi k/n} : k \in [n]\}$.

1.3.3 Symmetric group and permutation matrices

Let S_n denote the **symmetric group** of n elements, which consists of the set of bijections on $[n]$ with composition as the group action. Note $|S_n| = n! = n(n-1)\cdots 3 \cdot 2 \cdot 1$. Elements of S_n will be called **permutations**. Using cycle notation, a k**-cycle** permutation can be written of the form $\sigma = (a_1\ a_2\ \cdots\ a_k)$, where $\sigma(a_i) = a_{i+1}$ for $i < k$, $\sigma(a_k) = a_1$ and $\sigma(b) = b$ for $b \notin \{a_1, \ldots, a_k\}$. The **transpositions** are the permutations of the form $(i\ j)$ for $i \neq j$. Recall S_n can be generated by the transpositions $(1\ j)$ for $j = 2, \ldots, n$, and every permutation has a unique cycle decomposition as the product of distinct cycles. This last decomposition can be used to define a set map from S_n to the partitions of n objects based on the weakly increasing cycle lengths in this decomposition, which define the permutation's **cycle type**. The conjugacy classes of S_n are the permutations of the same cycle type, which follows directly from the fact

$$\tau(a_1\ a_2\ \cdots\ a_k)\tau^{-1} = (\tau(a_1)\ \tau(a_2)\ \cdots\ \tau(a_k)) \tag{1.6}$$

for $\tau \in S_n$. We can define the group homomorphism $\operatorname{sgn} : S_n \to \{\pm 1\}$ by $\operatorname{sgn}(\tau) = -1$ for any transposition τ and $\operatorname{sgn}(\sigma_1\sigma_2) = \operatorname{sgn}(\sigma_1)\operatorname{sgn}(\sigma_2)$ for any $\sigma_1, \sigma_2 \in S_N$. Then $A_n = \ker(\operatorname{sgn})$ is the **alternating group** of n elements.

Let $\mathcal{P}_n$ denote the order n **permutation matrices**, which is the left regular representation of the action of S_n on $\{\mathbf{e}_i : i \in [n]\}$, i.e, $P_\sigma \in \mathcal{P}_n$ for $\sigma \in S_n$ satisfies $P_\sigma \mathbf{e}_i = \mathbf{e}_{\sigma(i)}$. For $D \in \mathcal{D}_n \cap \{\pm 1\}^{n\times n}$ and $P \in \mathcal{P}_n$, then DP is a **signed permutation matrix**.

Note $P_\sigma^T P_\sigma = \mathbf{I}$, so that $P_\sigma^T = P_\sigma^{-1}$, establishing $\mathcal{P}_n \subset \mathrm{O}(n)$. Note also $\det(P_\sigma) = \mathrm{sgn}(\sigma)$, which then shows the **alternating permutation matrices** $\mathcal{A}_n$ satisfy $\mathcal{A}_n = \mathcal{P}_n \cap \mathrm{SO}(n)$. Also, note

$$(P_\sigma^T A P_\sigma)_{ij} = (P_\sigma \mathbf{e}_i)^T A (P_\sigma \mathbf{e}_j) = \mathbf{e}_{\sigma(i)}^T A \mathbf{e}_{\sigma(j)} = A_{\sigma(i),\sigma(j)}, \tag{1.7}$$

which is equivalent to (1.6).

Furthermore, note $\rho : S_n \to P_n$ such that $\rho(\sigma) = P_\sigma$ is an n-dimensional representation of S_n. Note ρ is not irreducible since $P_\sigma \mathbf{1}_n = \mathbf{1}_n$ for all $\sigma \in S_n$, so that $W_1 = \mathrm{span}(\mathbf{1}_n)$ and $W_2 = W_1^\perp$ are both subrepresentations of ρ. In fact, W_1 and W_2 are both irreducible: W_1 is trivially irreducible since it is 1-dimensional. To see W_2 is irreducible, suppose $W \subset W_2$ is S_n-invariant. If $W \neq \{0\}$, for $\mathbf{v} = \sum_i v_i \mathbf{e}_i \in W$ with $\mathbf{v} \neq \mathbf{0}$, then

$$P_{(1\ k)}\mathbf{v} - \mathbf{v} = (v_k \mathbf{e}_1 + v_1 \mathbf{e}_k) - (v_1 \mathbf{e}_1 + v_k \mathbf{e}_k) = (v_k - v_1)(\mathbf{e}_1 - \mathbf{e}_k) \tag{1.8}$$

for all k. Since $\mathbf{v} \perp \mathbf{1}_n$ and $\mathbf{v} \neq \mathbf{0}$, then necessarily $v_j \neq v_1$ for some j and hence $\mathbf{e}_1 - \mathbf{e}_j \in W$ by (1.8). It follows then $P_{(i\ j)}(\mathbf{e}_1 - \mathbf{e}_j) = \mathbf{e}_1 - \mathbf{e}_i \in W$ for all $i \geq 2$ so that W has dimension at least $n - 1 = \dim(W_2)$ and hence $W = W_2$.

Note the character $\chi(\sigma)$ gives the number of fixed points of $\sigma \in S_n$ as well as the sum of the k^{th} powers of the eigenvalues of σ. If σ is an n-cycle, then

$$\chi(\sigma^k) = \mathrm{Tr}(P_\sigma^k) = \sum_{j=1}^{n} e^{2\pi j k/n} = \begin{cases} n & \text{if } n \mid k \\ 0 & \text{if } n \nmid k \end{cases} \tag{1.9}$$

since $\sigma^k(i) = i$ for some i if and only if $\sigma^k(i) = i$ for all i if and only if $\sigma^k = 1$ if and only if $n \mid k$. An application of (1.9) to the divisibility of random integers is explored further in Appendix A.

1.3.4 Topological groups

We call $(X, \mathcal{T})$ a **topological space** if $\varnothing, X \in \mathcal{T} \subset 2^X$ and T is closed under finite intersections and arbitrary unions, and we say a set O is open in X if $O \in \mathcal{T}$. A topological space K is **compact** if any open covering admits a finite subcover. A topological space X is **connected** if it cannot be written as the disjoint union of two open sets. The product of a collection of compact spaces is compact by Tychenoff's theorem. Finite products of compact spaces can be verified to satisfy the open subcover criterion through induction. For $(Y, \mathcal{T}')$ another topological space, we say $f : X \to Y$ is **continuous** if $f^{-1}(O) \in \mathcal{T}$ whenever $O \in \mathcal{T}'$. A bijective continuous map between two topological spaces is called a **homeomorphism**, and we write $X \cong Y$ if X and Y are **homeomorphic**. We will call G a **topological group** if G is a topological space that is a group under multiplication such that multiplication and inverses are continuous maps on G. A space is **Polish** if it is metrizable, separable, and complete. These will include all of the spaces of interest in the current document. If not stated explicitly, one can assume a topological space means a Polish topological space.

An important example that we be revisited frequently: $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\} = \{e^{i\theta} : \theta \in [0, 2\pi)\}$ for the unit circle in $\mathbb{C}$, which is a compact topological group under the group action of multiplication.

1.4 Linear algebra

1.4.1 Background

For $\mathbf{x} \in \mathbb{C}^n$, define the p-norm $\|\cdot\|_p$ as

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \tag{1.10}$$

for $|x_i| = (x_i\overline{x_i})^{1/2}$. In particular, for $\langle \mathbf{u}, \mathbf{v}\rangle = \mathbf{u}^*\mathbf{v}$ the **inner product** or **dot product** of $\mathbf{u}, \mathbf{v} \in \mathbb{C}^n$, we have $\|\mathbf{x}\|_2^2 = \langle \mathbf{x}, \mathbf{x}\rangle$. Write $\mathbf{x} \perp \mathbf{y}$ if $\langle \mathbf{x}, \mathbf{y}\rangle = 0$. Note further if $U \in \mathrm{U}(n)$, then $\|U\mathbf{v}\|_2 = \|\mathbf{v}\|_2$.

I will write $\det A$ for the **determinant** of $A \in M_n$. The determinant is a polynomial of the entries of A, which can be realized using the form

$$\det A = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} A_{i,\sigma(i)}. \tag{1.11}$$

Note $\operatorname{sgn}(\sigma^{-1}) = \operatorname{sgn}(\sigma)$ since $1 = \operatorname{sgn}(1) = \operatorname{sgn}(\sigma\sigma^{-1}) = \operatorname{sgn}(\sigma)\operatorname{sgn}(\sigma^{-1})$. It follows $\det A = \det A^T$ since

$$\begin{aligned}\det A^T &= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} (A^T)_{i,\sigma(i)} = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} A_{\sigma(i),i} = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma^{-1}) \prod_{i=1}^{n} A_{i,\sigma^{-1}(i)} \\ &= \det A.\end{aligned}$$

It follows $\det A^* = \overline{\det A}$.

We will write $\operatorname{rank} A$ for the **rank** of A, which is the dimension of the column space of A. Recall $\operatorname{rank} A = \operatorname{rank} A^T = \operatorname{rank} A^*$, so it follows $\operatorname{rank} A$ is also the dimension of the row space of A and hence $\operatorname{rank} A \leq \min(m, n)$ for $A \in M_{m,n}$. A matrix $A \in M_{m,n}$ is **full rank** if $\operatorname{rank} A = \min(m, n)$. For $A \in \mathbb{C}^{n\times n}$, we will write $\operatorname{Tr}(A)$ for the **trace**, which is the sum of the diagonal entries of A. Note $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$ whenever $A, B^T \in \mathbb{C}^{n\times m}$. It follows if $B = CAC^{-1}$, then $\operatorname{Tr} B = \operatorname{Tr}(CAC^{-1}) = \operatorname{Tr}((AC^{-1})C) = \operatorname{Tr} A$.

We will say λ is an **eigenvalue** for a square matrix A if there exists a nonzero $\mathbf{v}$ such that $A\mathbf{v} = \lambda\mathbf{v}$, with such a $\mathbf{v}$ then being called an associated **eigenvector**. Equivalently, λ is an eigenvalue of $A \in M_n$ if λ is a root of the **characteristic polynomial** of A,

$$\sigma_A(x) := \det(x\mathbf{I} - A), \tag{1.12}$$

which is a degree n monic polynomial in $\mathbb{C}[x]$. Note if $A\mathbf{v} = \lambda\mathbf{v}$ then $A^k\mathbf{v} = \lambda^k\mathbf{v}$ for any nonnegative integer k. Moreover, if A is nonsingular, then this also holds for $k \in \mathbb{Z}$, so that λ^k is an eigenvalue of A^k whenever λ is an eigenvalue. Similarly, $c\lambda$ is an eigenvalue of cA if λ is an eigenvalue of A. By the Fundamental Theorem of Algebra, $\sigma_A(x)$ has at least one root since $\mathbb{C}$ is algebraically closed, and hence every square matrix has at least one eigenvalue. An **eigenspace** for the eigenvalue λ of A, denoted $V_A(\lambda)$, is the span of the eigenvectors associated with λ, where we assume the convention $V_A(\lambda) = \{\mathbf{0}\}$ for when this space is empty. Recall if λ is an eigenvalue for $U \in \mathrm{U}(N)$ then $|\lambda| = 1$ since if $\mathbf{v}$ is an associated unit eigenvector for λ then

$$1 = \|\mathbf{v}\|_2 = \|U\mathbf{v}\|_2 = \|\lambda\mathbf{v}\|_2 = |\lambda|\|\mathbf{v}\|_2 = |\lambda|.$$

Recall A and A^T have the same eigenvalues since

$$\sigma_A(x) = \det(x\mathbf{I} - A) = \det((x\mathbf{I} - A)^T) = \det(x\mathbf{I} - A^T) = \sigma_{A^T}(x).$$

It follows A and A^* have eigenvalues that are conjugate to one another: if λ is an eigenvalue of A, then it is also an eigenvalue of A^T, say with eigenvector $\mathbf{v}$; it follows then $\overline{\mathbf{v}}$ is an eigenvector of A^* for eigenvalue $\overline{\lambda}$ since $A^*\overline{\mathbf{v}} = \overline{A^T\mathbf{v}} = \overline{\lambda\mathbf{v}} = \overline{\lambda}\overline{\mathbf{v}}$. Moreover, if $A = BDB^{-1}$ for D diagonal, then necessarily D_{jj} is an eigenvalue of A with associated eigenvector $B\mathbf{e}_j$. We call such a matrix **diagonalizable**. If A is diagonalizable, then $\operatorname{Tr} A$ is the sum of its eigenvalues.

A matrix $A \in M_n$ is **nonsingular** if A^{-1} exists, and **singular** otherwise. Recall A is singular if and only if $\det A = 0$ if and only if 0 is an eigenvalue for A if and only if A is not full rank.

Recall AB and BA have the same eigenvalues for $A, B \in M_n$: let λ be an eigenvalue for AB. If $\lambda = 0$, then $\det AB = (\det A)(\det B) = \det BA = 0$, so that $\lambda = 0$ is also an eigenvalue for BA. If $\lambda \neq 0$, let $\mathbf{v}$ be an eigenvector for AB, such that $AB\mathbf{v} = \lambda\mathbf{v}$, and in particular,

$B\mathbf{v} \neq 0$ (since $\lambda\mathbf{v} \neq 0$). It follows then $B\mathbf{v}$ is an eigenvector for BA for eigenvalue λ since $BA(B\mathbf{v}) = B(AB\mathbf{v}) = B(\lambda\mathbf{v}) = \lambda B\mathbf{v}$.

A matrix is **symmetric** if $A^T = A$ and **Hermitian** if $A^* = A$, while a matrix is **skew-symmetric** or **skew-Hermitian** if $A^T = -A$ or $A^* = -A$. Let $\mathrm{Sym}_N, \mathrm{Herm}_N, \mathrm{Skew}_N^{\mathrm{S}}, \mathrm{Skew}_N^{\mathrm{H}}$ denote, respectively, the symmetric, Hermitian, skew-symmetric and skew-Hermitian order N matrices. For $R \subset \mathbb{C}$ closed under complex conjugation, for any $A \in M_N(R)$ we have $AA^T, A + A^T \in \mathrm{Sym}, A - A^T \in \mathrm{Skew}^{\mathrm{S}}, AA^*, A + A^* \in \mathrm{Herm}$ and $A - A^* \in \mathrm{Skew}^{\mathrm{H}}$, while also $[A, A^T] \in \mathrm{Skew}^{\mathrm{S}}$ and $[A, A^*] \in \mathrm{Skew}^{\mathrm{H}}$. Recall

$$M_N(R) = \mathrm{Sym}(R) \oplus \mathrm{Skew}^{\mathrm{S}}(R) = \mathrm{Herm}(R) \oplus \mathrm{Skew}^{\mathrm{H}}(R), \tag{1.13}$$

where we note

$$A = \frac{1}{2}(A + A^T) + \frac{1}{2}(A - A^T) = \frac{1}{2}(A + A^*) + \frac{1}{2}(A - A^*).$$

Also, recall a Hermitian matrix has real eigenvalues since if $\mathbf{v}$ is an unit eigenvector for Hermitian A with eigenvalue λ, then

$$\lambda = \lambda\langle\mathbf{v}, \mathbf{v}\rangle = \langle\mathbf{v}, \lambda\mathbf{v}\rangle = \langle\mathbf{v}, A\mathbf{v}\rangle = \langle A^*\mathbf{v}, \mathbf{v}\rangle = \langle A\mathbf{v}, \mathbf{v}\rangle = \langle\lambda\mathbf{v}, \mathbf{v}\rangle = \bar{\lambda}\langle\mathbf{v}, \mathbf{v}\rangle = \bar{\lambda}.$$

A matrix $A \in \mathrm{Herm}_N$ is **positive definite (semi-definite)** if all of its eigenvalues are positive (nonnegative), or equivalently if for all $\mathbf{v} \neq 0$ we have $\mathbf{v}^* A\mathbf{v} = \langle\mathbf{v}, A\mathbf{v}\rangle > 0$ ($\mathbf{v}^* A\mathbf{v} = \langle\mathbf{v}, A\mathbf{v}\rangle \geq 0$). $A \in \mathrm{Herm}_N$ is **negative definite (semi-definite)** if $-A$ is positive definite (semi-definite). A matrix A is positive semi-definite if and only if there exists a positive semi-definite matrix B such that $B^2 = A$. For such a matrix B corresponding to A, we we use the notation $B = \sqrt{A}$. If A is positive definite, then $\sqrt{A}$ is unique.

For any matrix A, A^*A is positive semi-definite since for any $\mathbf{v}$ we have

$$\mathbf{v}^*A^*A\mathbf{v} = \langle A\mathbf{v}, A\mathbf{v}\rangle = \|A\mathbf{v}\|_2^2 \geq 0,$$

and $\operatorname{rank} A = \operatorname{rank} A^*A$.

The **singular values** of $A \in \mathbb{C}^{N\times M}$ are the (nonnegative) eigenvalues of $\sqrt{A^*A}$. It follows a matrix is singular if and only if 0 is a singular value of A. The **singular value decomposition (SVD)** of A is $A = U\Sigma V^*$ for $U \in \mathrm{U}(N)$ and $V \in \mathrm{U}(M)$, with Σ being a diagonal $N \times M$ matrix consisting of the singular values of A. Recall if A, A' are nonsingular, then $\operatorname{rank} B = \operatorname{rank} AB = \operatorname{rank} BA'$, so that $\operatorname{rank} A = \operatorname{rank} \Sigma$ is then the number of positive singular values of A.

Since the singular values of a matrix $A \in \mathbb{C}^{N\times M}$ for $N \leq M$ are nonnegative, then we can order them from smallest to largest, say $\sigma_1 = \max \operatorname{diag}(\Sigma)$ and $\sigma_N = \min \operatorname{diag}(\Sigma)$ for $A = U\Sigma V^*$ the SVD decomposition of A. Recall the **(induced) operator norm** of A is

$$\|A\| = \max_{\|\mathbf{v}\|=1} \|A\mathbf{v}\| = \max_{\mathbf{v}} \frac{\|A\mathbf{v}\|}{\|\mathbf{v}\|}. \tag{1.14}$$

In finite dimensions, all norms are equivalent.

Let $A \in \mathbb{C}^{N\times M}$. Let

$$\|A\|_F = (\operatorname{Tr}(AA^*))^{1/2} = \left(\sum_{i,j} |A_{i,j}|^2\right)^{1/2} \tag{1.15}$$

denote the **Frobenius norm**. Note if $A = U\Sigma V^*$, then $AA^* = U\Sigma^2 U^*$ so that

$$\|A\|_F^2 = \operatorname{Tr}(AA^*) = \operatorname{Tr}(\Sigma^2) = \sum_{k=1}^{n} \sigma_k^2 \tag{1.16}$$

for σ_k the singular values of A. Let $\|\cdot\|_{\max}$ be the elementwise **max norm** of a matrix

defined by

$$\|A\|_{\max} = \max_{i,j} |A_{ij}|. \tag{1.17}$$

Recall the max norm is not submultiplicative, which is apparent from

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ & 1 \end{bmatrix}. \tag{1.18}$$

Let $\|\cdot\|_\infty$ denote the induced ℓ_∞ matrix norm, which satisfies the max row sum property

$$\|A\|_\infty = \max_i \sum_{j=1}^{N} |A_{ij}|. \tag{1.19}$$

Note $\|\cdot\|_{\max}$, $\|\cdot\|_\infty$ and $\|\cdot\|_F$ are invariant under row or column permutations or unit multiples while, considering only these three norms, only $\|\cdot\|_F$ is invariant under general orthogonal or unitary transformations: it enough to note for

$$B(\theta) = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \tag{1.20}$$

the (counterclockwise) **rotation matrix**, $\|B(\theta)\|_\infty = |\cos\theta| + |\sin\theta| \neq 1 = \|\mathbf{I}_2\|_\infty$ and $\|B(\theta)\|_{\max} = \max(|\cos\theta|, |\sin\theta|) \neq 1 = \|\mathbf{I}_2\|_{\max}$ when $\theta \notin \frac{\pi}{2} + \mathbb{Z}$.

Using $\|\cdot\|_2$, we have $\|A\|_2 = \sigma_1$: let $\mathbf{v}$ be a unit eigenvector for $\sqrt{A^*A}$ for σ_1, then

$$\sigma_1 = \|\sigma_1 \mathbf{v}\|_2 = \|\sqrt{A^*A}\mathbf{v}\|_2 = \langle \sqrt{A^*A}\mathbf{v}, \sqrt{A^*A}\mathbf{v}\rangle^{1/2} = \langle \mathbf{v}, A^*A\mathbf{v}\rangle^{1/2} = \langle A\mathbf{v}, A\mathbf{v}\rangle^{1/2} = \|A\mathbf{v}\|_2,$$

which shows $\sigma_1 \leq \|A_2\|$. Equality follows from the fact any unit vector $\mathbf{v} \in \mathbb{C}^M$ is a linear combination of the columns of $V \in \mathrm{U}(M)$, with coefficients with unit 2-norm: this follows

since if $V\mathbf{u} = \mathbf{v}$ then $\|\mathbf{u}\|_2 = \|V\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1$ so that $\mathbf{v} = \sum_i u_i \mathbf{v}_i$ and hence

$$\|A\mathbf{v}\|_2 = \|\sum_i u_i \sqrt{AA^*}\mathbf{v}_i\|_2 = \|\sum_i \sigma_i u_i \mathbf{v}_i\|_2 = \|\sum_i \sigma_i u_i \mathbf{e}_i\|_2 = (\sum_i \sigma_i^2 u_i^2)^{1/2} \leq \sigma_1,$$

so that also $\|A\|_2 \leq \sigma_1$ and hence $\|A\|_2 = \sigma_1$. Similarly, if $A \in \mathbb{C}^{n\times n}$ is nonsingular (viz., $\sigma_n > 0$), then $\|A^{-1}\|_2 = \dfrac{1}{\sigma_n}$.

We define the **condition number** of A for A nonsingular as

$$\kappa(A) = \|A\|\|A^{-1}\|. \tag{1.21}$$

Additionally write $\kappa_p(A)$ if $\|\cdot\| = \|\cdot\|_p$, which will be referred to as the p-condition number. In particular, note

$$\kappa_2(A) = \frac{\sigma_1}{\sigma_n} \tag{1.22}$$

The condition number of a matrix A controls how much the output of a vector multiplied by A using floating-point arithmetic can change with respect to a small change in the input vector. The 2-condition number is then minimized whenever $\sigma_1 = \sigma_n > 0$. When A is square, then this occurs only when A is a scalar multiple of a unitary matrix: then $\Sigma = \sigma\mathbf{I}$, and $A = U\Sigma V^* = \sigma UV^*$, where $UV^* \in \mathrm{U}(n)$. In particular, unitary matrices have minimal 2-condition number.

Let

$$\mathcal{D}_n = \mathrm{span}\{\mathbf{E}_{ii} : i = 1, \ldots, n\} \subset \mathbb{C}^{n\times n} \tag{1.23}$$

be the set of diagonal matrices of order n and

$$\mathcal{L}_n(k) = \mathrm{span}\{\mathbf{E}_{ij} : i + k \geq j\} \subset \mathbb{C}^{n\times n} \tag{1.24}$$

be the set of lower triangular square matrices of order n with only nonzero entries found k diagonals below the main diagonal, where we use the convention $L_n(k) = \{\mathbf{0}\}$ when $k \geq n$. Furthermore, let $\mathcal{L}_n := \mathcal{L}_n(0)$ be the set of lower triangular square matrices of order n. Note $\mathcal{L}_n(k) \supset \mathcal{L}_n(\ell)$ for $k \leq \ell$. Clearly

$$\mathcal{L}_n = \mathcal{D}_n \oplus \mathcal{L}_n(1). \tag{1.25}$$

Moreover, simple computations show

$$\mathcal{D}_n^2 = \mathcal{D}_n, \tag{1.26}$$

$$\mathcal{L}_n(k) + \mathcal{L}_n(\ell) = \mathcal{L}_n(\min(k, \ell)), \text{ and} \tag{1.27}$$

$$\mathcal{D}_n\mathcal{L}_n(k) = \mathcal{L}_n(k)\mathcal{D}_n = \mathcal{L}_n(k). \tag{1.28}$$

Also, we have

Lemma 1.1.

$$\mathcal{L}_n(k)\mathcal{L}_n(\ell) = \mathcal{L}_n(k + \ell) \tag{1.29}$$

Proof. Suppose $\mathbf{E}_{i_1 j_1} \in \mathcal{L}_n(k)$ and $\mathbf{E}_{i_2 j_2} \in \mathcal{L}_n(\ell)$. We have $\mathbf{E}_{i_1 j_1}\mathbf{E}_{i_2 j_2} = \delta_{j_1 i_2}\mathbf{E}_{i_1 j_2} \neq \mathbf{0}$ if and only if $j_1 = i_2$ and $i_1 + k + \ell \geq j_1 + \ell = i_2 + \ell \geq j_2$. The result follows. □

Letting $k = \ell = 0$ and a simple induction argument, we have

Corollary 1.1. *Lower triangular square matrices are closed under multiplication, i.e.,* $\mathcal{L}_n^2 = \mathcal{L}_n$.

Corollary 1.2. *Strictly lower triangular matrices are nilpotent. In particular,* $\mathcal{L}_n(1)^n = \{\mathbf{0}\}$.

Moreover, an important property of lower triangular matrices is the following result:

Lemma 1.2. *$\mathcal{L}_n$ is closed under inverses when they exist.*

Proof. Suppose $L \in \mathcal{L}_n$ is nonsingular. Let $D \in \mathcal{D}_n$ and $N \in \mathcal{L}_n(1)$ such that $L = D - N$, so that D is also nonsingular (since $\det D = \det L \neq 0$). First suppose $D = \mathbf{I}$. By Corollary 1.2 we have $N^n = \mathbf{0}$, so that

$$\mathbf{I} = \mathbf{I} - N^n = (\mathbf{I} - N)(\mathbf{I} + N + \cdots + N^{n-1}) = L(\mathbf{I} + N + \cdots + N^{n-1})$$

and hence

$$L^{-1} = \mathbf{I} + N + \cdots + N^{n-1} \in \mathcal{L}_n$$

since $\mathbf{I}, N^k \in \mathcal{L}_n$ for each k. Now for general D, we have $L = D(\mathbf{I} - D^{-1}N) =: DL_0$. By the prior case, $L_0^{-1} \in \mathcal{L}_n$ so that $L^{-1} = L_0^{-1}D^{-1} \in \mathcal{L}_n$, using also Corollary 1.1. □

Let

$$\mathcal{U}_n(k) = \operatorname{span}\{\mathbf{E}_{ij} : j + k \geq i\} \subset \mathbb{C}^n \tag{1.30}$$

and $\mathcal{U}_n := \mathcal{U}_n(0)$, the set of upper triangular square matrices of order n. By using transposition, along with the fact $A \in \mathcal{L}_n(k)$ if and only if $A^T \in \mathcal{U}_n(k)$, we have immediately

$$\mathcal{U}_n = \mathcal{D}_n \oplus \mathcal{U}_n(1) \tag{1.31}$$

$$\mathcal{U}_n(k) + \mathcal{U}_n(\ell) = \mathcal{U}_n(\min(k, \ell)), \text{ and} \tag{1.32}$$

$$\mathcal{D}_n\mathcal{U}_n(k) = \mathcal{U}_n(k)\mathcal{D}_n = \mathcal{U}_n(k). \tag{1.33}$$

Moreover,

Corollary 1.3.

$$\mathcal{U}_n(k)\mathcal{U}_n(\ell) = \mathcal{U}_n(k + \ell) \tag{1.34}$$

Corollary 1.4. *Upper triangular square matrices are closed under multiplication, i.e.,* $\mathcal{U}_n^2 = \mathcal{U}_n$.

Corollary 1.5. *Strictly upper triangular matrices are nilpotent. In particular,* $\mathcal{U}_n(1)^n = \{\mathbf{0}\}$.

Corollary 1.6. $\mathcal{U}_n$ *is closed under inverses when they exist.*

1.4.2 Matrix factorization

This section will outline several matrix factorizations that have useful applications in computational linear algebra. The general application of interest for these factorizations will be in solving a linear system $A\mathbf{x} = \mathbf{b}$. If A is nonsingular, then this has the unique solution $\mathbf{x} = A^{-1}\mathbf{b}$. Even for moderately sized A, it can be very expensive to calculate A^{-1} directly. If $A = BC$, then one can solve $A\mathbf{x} = \mathbf{b}$ by first solving $B\mathbf{y} = \mathbf{b}$ and then $C\mathbf{x} = \mathbf{y}$. If B and C have certain desirable properties (e.g., unitary or triangular), then each separate step can be significantly simpler in a computational sense. Moreover, finding such a factorization once then enables one to solve *any* linear system using the same A and a different input $\mathbf{b}$.

For any matrix $A \in \mathbb{C}^{n\times n}$, there exists $U \in \mathrm{U}(n)$ and an upper triangular matrix T such that $A = UTU^*$. This is called the **Schur decomposition** of A.

Theorem 1.4. *Every* $A \in \mathbb{C}^{n\times n}$ *has a Schur decomposition.*

Proof. The existence of such a decomposition can be established by induction on n along with the fact A has at least one eigenvalue λ: this would allow us to construct $U_1 \in \mathrm{U}(n)$ with the first columns being an orthonormal basis of $V(\lambda)$ and the remaining columns an orthonormal basis of $V(\lambda)^\perp$ so that

$$A = U_1 \begin{bmatrix} \lambda\mathbf{I} & A_1 \\ 0 & A_2 \end{bmatrix} U_1^*,$$

where the main result follows from using induction on A_2 to find $V \in \mathrm{U}$ such that V^*A_2V is upper triangular, so that then $U_2 = \mathbf{I} \oplus V \in \mathrm{U}(n)$ and $U = U_1U_2$ satisfies U^*AU is upper triangular. $\square$

This shows every square matrix is **(unitarily) triangularizable**. We say a set of matrices is **simultaneously triangularizable** if each matrix can be triangularized by the same unitary matrix. Note for $A = UTU^*$ the Schur decomposition of A, we have T_{11} is an eigenvalue for A with associated eigenvector $\mathbf{u}_1$, the first column of U, since

$$A\mathbf{u}_1 = UT\mathbf{e}_1 = T_{11}\mathbf{u}_1.$$

This can be used to establish the following result:

Lemma 1.3. *For $A, B \in \mathbb{C}^{n\times n}$, if $[A, B] = \mathbf{0}$ then A, B are simultaneously triangularizable.*

Proof. Note first commuting matrices preserve eigenspaces: if λ is an eigenvalue of A with eigenspace $V_A(\lambda)$, then for $\mathbf{v} \in V_A(\lambda)$, we have

$$A(B\mathbf{v}) = B(A\mathbf{v}) = B(\lambda\mathbf{v}) = \lambda B\mathbf{v},$$

which shows $BV_A(\lambda) \subset V_A(\lambda)$, i.e., $V_A(\lambda)$ is B-stable.

Now using induction on n (where the result is trivial for $n = 1$), first note that A and B commuting implies there is a non-trivial subspace of minimal dimension that is invariant under the actions of A and B: this just follows from the well-ordering principle along with the result $\mathbb{C}^n$ is A- and B-invariant.

Next, note any nontrivial A-invariant subspace must contain an eigenvector of A: Suppose $\mathcal{S}$ is A-invariant and has dimension $k > 0$, with basis $\mathbf{s}_1, \ldots, \mathbf{s}_k$ and let $S = \begin{bmatrix} \mathbf{s}_1 & \cdots & \mathbf{s}_k \end{bmatrix} \in \mathbb{C}^{n\times k}$. Then $AS = SC$ for some $C \in \mathbb{C}^{k\times k}$ since $A\mathcal{S} \subset \mathcal{S}$. Moreover, since C has at least

one eigenvalue/eigenvector pair, say $(\lambda, \mathbf{v})$, then $S\mathbf{v} \neq 0$ (since S has full-rank), and we see $AS\mathbf{v} = SC\mathbf{v} = \lambda S\mathbf{v}$, showing $S\mathbf{v} \in \mathcal{S}$ is an eigenvector of A.

Now we claim $\mathcal{S}$, a minimal non-trivial subspace that is both A- and B-invariant, consists of shared eigenvectors for A and B. Suppose first $\mathcal{S}$ contains some vector $\mathbf{v} \neq 0$ that is not an eigenvector for A. By above, $\mathcal{S}$ does contain an eigenvector of A for an eigenvalue λ, so that $\mathcal{S}' = \mathcal{S} \cap V_A(\lambda) \neq \varnothing$ and $\mathcal{S}' \neq \mathcal{S}$ since $\mathbf{v} \notin \mathcal{S}'$. Since $BV_A(\lambda) \subset V_A(\lambda)$ (and obviously $AV_A(\lambda) \subset V_A(\lambda)$) and $B\mathcal{S} \subset \mathcal{S}$, then $B\mathcal{S}' \subset \mathcal{S}'$, which shows $\mathcal{S}'$ is both A- and B-invariant, which contradicts the minimality of $\mathcal{S}$. It follows then $\mathcal{S}$ consists precisely of eigenvectors of both A and B, and so $\mathcal{S}$ has an orthonormal basis of simultaneous eigenvectors of A and B.

It follows now we can form $U = U_1 \oplus U_2 \in \mathrm{U}(n)$ for U_1 an orthonormal basis of $\mathcal{S}$ and U_2 an orthonormal basis of $\mathcal{S}^\perp$ such that

$$A = U \begin{bmatrix} D_1 & A_1 \\ 0 & A_2 \end{bmatrix} U^* \quad \text{and} \quad B = U \begin{bmatrix} D_2 & B_1 \\ 0 & B_2 \end{bmatrix} U^*$$

for D_1, D_2 diagonal matrices. Since $[A, B] = \mathbf{0}$ then necessarily $[A_2, B_2] = \mathbf{0}$. Applying the inductive hypothesis now yields $V \in \mathrm{U}$ such that V^*A_2V and V^*B_2V are triangular, so that $U(I \oplus V) = U_1 \oplus U_2V \in \mathrm{U}(n)$ simultaneously triangularlizes A and B. □

A simple expansion of this argument yields that a finite set of mutually commuting matrices are simultaneously triangularizable.

A matrix A is **(unitarily) diagonalizable** if there exists $U \in \mathrm{U}$ and a diagonal matrix D such that $A = UDU^*$. In particular, the columns of U are eigenvectors of A with the corresponding diagonal element of $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ being the associated eigenvalue:

$$AU\mathbf{e}_i = UDU^*(U\mathbf{e}_i) = UD\mathbf{e}_i = \lambda_i U\mathbf{e}_i.$$

A set of matrices in $\mathbb{C}^{n\times n}$ is **simultaneously diagonalizable** if each matrix is diagonalizable with respect to the same $U \in \mathrm{U}(n)$. A matrix A is **normal** if $[A, A^*] = \mathbf{0}$, with diagonal matrices and $\mathrm{U}(n)$ being examples of normal matrices (e.g., $UU^* = \mathbf{I} = U^*U$). Recall the following results, for which we will include proofs for completeness.

Lemma 1.4. *An upper triangular normal matrix is diagonal.*

Proof. Suppose T is upper triangular and normal, then $TT^* = T^*T$ and in particular

$$\|T\mathbf{e}_i\|_2^2 = \langle T\mathbf{e}_i, T\mathbf{e}_i\rangle = \langle \mathbf{e}_i, T^*T\mathbf{e}_i\rangle = \langle \mathbf{e}_i, TT^*\mathbf{e}_i\rangle = \|T^*\mathbf{e}_i\|_2^2$$

for all i. This shows the norm of each row of T equals the norm of each corresponding column. Since T has (at most) one non-zero entry in its first column (viz., $T\mathbf{e}_1 = T_{11}\mathbf{e}_1$) occurring in the diagonal, then the same is true of its first row, so that $T_{1,j} = 0$ if $j \neq 1$. Continuing inductively, it follows the only non-zero entries of T can exist in its diagonal. □

Lemma 1.5. *Let $A, B \in \mathbb{C}^{n\times n}$ be normal. A, B are simultaneously diagonalizable if and only if $[A, B] = \mathbf{0}$.*

Proof. Suppose A, B are simultaneously diagonalizable, so there exists $U \in \mathrm{U}(n)$ and diagonal D_1, D_2 such that $A = UD_1U^*$ and $B = UD_2U^*$. Since diagonal matrices commute, we have

$$[A, B] = U[D_1, D_2]U^* = \mathbf{0},$$

showing A and B commute.

Now suppose $[A, B] = \mathbf{0}$. By Lemma 1.3, we have A and B are simultaneously triangularizable, so that there exists $U \in \mathrm{U}(n)$ and upper triangular T_1 and T_2 such that $A = UT_1U^*$ and $B = UT_2U^*$. Since A and B are normal, then

$$[T_1, T_1^*] = U^*[A, A^*]U = \mathbf{0} = U^*[B, B^*]U = [T_2, T_2^*],$$

which yields T_1 and T_2 are both normal upper triangular matrices, and hence diagonal by Lemma 1.4. □

Again, this can be easily expanded to conclude a finite set of mutually commuting normal matrices are simultaneously diagonalizable.

Theorem 1.5 (Spectral Theorem)**.** *A square matrix is normal if and only if it is diagonalizable.*

Proof. If A is diagonalizable, then there exists $U \in \mathrm{U}(n)$ and diagonal D such that $A = UDU^*$. Now $[D, D^*] = \mathbf{0}$ since diagonal matrices are normal, so

$$[A, A^*] = [UDU^*, UD^*U^*] = U[D, D^*]U^* = \mathbf{0}.$$

Now suppose A is normal, so that $[A, A^*] = 0$. Since A has a Schur decomposition with $A = UTU^*$ for $U \in \mathrm{U}$ and T upper triangular, we have

$$[T, T^*] = U^*[UTU^*, UT^*U^*]U = U^*[A, A^*]U = \mathbf{0}.$$

Lemma 1.4 yields T is diagonal since it is an upper triangular normal matrix. □

One immediate application of this is:

Theorem 1.6. *Every matrix has an SVD decomposition.*

Proof. Let A be an order $n \times m$ matrix and let $M = A^*A$, which is positive semi-definite. Since M is normal of order n, then there exist $V \in \mathrm{U}(n)$ and diagonal D such that $M = VDV^*$ by Theorem 1.5. Since M is positive semidefinite, then D is a diagonal matrix of nonnegative numbers. By relabeling (i.e., replacing D with $P_\sigma D P_\sigma^T$ and V with VP_σ^T for

some $\sigma \in S_n$) we can assume $D = D_1 \oplus \mathbf{0}$ where D_1 has positive diagonal entries, and write $V = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$ such that $V_1^* M V_1 = D_1$ and $V_2^* M V_2 = \mathbf{0}$. Now define $U_1 = AV_1 D_1^{-1/2}$, where D_1^k denotes the diagonal matrix with associated entries of D_1 raised to the k^{th} power, so that $U_1 D_1^{\frac{1}{2}} V_1^* = A - (AV_2)V_2^* = A$ (since $V_2^* M V_2 = (AV_2)^* AV_2 = \mathbf{0}$). Since $U_1^* U_1 = D_1^{-\frac{1}{2}} V_1^* A^* A V_1 D_1^{-\frac{1}{2}} = D_1^{-\frac{1}{2}} D_1 D_1^{-\frac{1}{2}} = \mathbf{I}$, then the columns of U_1 are orthonormal and so can be extended to form $U = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \in \mathrm{U}(m)$. Now let $\Sigma = \begin{bmatrix} D^{\frac{1}{2}} \oplus \mathbf{0} \\ \mathbf{0} \end{bmatrix}$ be order $n \times m$. By construction, it follows $A = U\Sigma V^*$. □

We say A has an LU factorization if there exist a lower unit triangular matrix L and an upper triangular matrix U such that $A = LU$. Here I am adopting the Doolittle definition that requires L have unit diagonal rather than the Crout definition where the unit diagonal is required on the U factor. By considering A^T, we can move between these definitions. A standard result shows an LU factorization is unique for nonsingular A when it exists:

Theorem 1.7 ([16])**.** *There exists a unique LU factorization of $A \in \mathbb{R}^{n\times n}$ if and only if $A_{:k,:k}$ is nonsingular for all $k = 1, \ldots, n-1$. If $A_{:k,:k}$ is singular for some $k < n$, an LU factorization may exist but it is not unique.*

Proof. Suppose $A_{:k,:k}$ is nonsingular for all $k = 1, \ldots, n-1$. We can prove the first statement using induction on n. The result is trivial for $n = 1$. If the result holds for $n-1$, then writing $A \in \mathbb{R}^{n\times n}$ as

$$A = \begin{bmatrix} \widetilde{A} & \mathbf{b} \\ \mathbf{c}^T & A_{nn} \end{bmatrix} \tag{1.35}$$

note $\widetilde{A} = A_{:n-1,:n-1}$ also satisfies $\widetilde{A}_{:k,:k}$ is nonsingular for all $k = 1, \ldots, n-2$. Hence, $\widetilde{A} = \widetilde{L}\widetilde{U}$ is a unique LU factorization by the inductive hypothesis, where we note $\widetilde{L} = L_{:n-1,:n-1}$ and

$\widetilde{U} = U_{:n-1,:n-1}$ are both nonsingular since also $0 \neq \det \widetilde{A} = \det \widetilde{L} \det \widetilde{U}$. It follows

$$A = \begin{bmatrix} \widetilde{L} & \mathbf{0} \\ \mathbf{c}^T \widetilde{U}^{-1} & 1 \end{bmatrix} \begin{bmatrix} \widetilde{U} & \widetilde{L}^{-1}\mathbf{b} \\ \mathbf{0} & A_{nn} - \mathbf{c}^T \widetilde{A}^{-1}\mathbf{b} \end{bmatrix} =: LU. \tag{1.36}$$

This establishes the existence of an LU factorization. For the uniqueness result: if also $A = L_0 U_0$, then $(L_0^{-1}L)U = U_0 \in \mathcal{U}_n$. Since L and L_0 are unit lower triangular, then so is $L_0^{-1}L$, so that $L_0^{-1}L = \mathbf{I} + N$ for some $N \in \mathcal{L}_n(1)$. Since then $U_0 = U + NU \in \mathcal{U}_n$, then $0 = (U_0)_{ij} = (NU)_{ij} = (N^T\mathbf{e}_i)^T(U\mathbf{e}_j)$ for $i > j$. This establishes $N^T\mathbf{e}_i \perp U\mathbf{e}_j$ for all $j = 1, \ldots, i-1$. Since $U_{:n-1,:n-1}$ is nonsingular, then

$$N^T\mathbf{e}_i \in \operatorname{span}(U\mathbf{e}_j : j = 1, \ldots, i-1)^\perp = \operatorname{span}(\mathbf{e}_j : j = 1, \ldots, i-1)^\perp = \operatorname{span}(\mathbf{e}_j : j \geq i).$$

Since $N^T \in \mathcal{U}_n(1)$ then $N^T\mathbf{e}_i = \mathbf{0}$ for all i and hence $N = \mathbf{0}$. It follows $L_0^{-1}L = \mathbf{I}$ so that $L_0 = L$ and $U_0 = U$.

Conversely, suppose A has a unique LU factorization with $A = LU$. We will again use induction on n to show $\det A_{:k,:k} \neq 0$ for $k = 1, \ldots, n-1$. The result is trivial for $n = 1$ so assume the result holds for $n-1$. Note then for $\widetilde{A} = A_{:n-1,:n-1}$, $\widetilde{L} = L_{:n-1,:n-1}$, and $\widetilde{U} = U_{:n-1,:n-1}$ then necessarily $\widetilde{A} = \widetilde{L}\widetilde{U}$. Moreover, this must be unique since $\det \widetilde{A}_{:k,:k} \neq 0$ for all k. For reference, note

$$A = \begin{bmatrix} \widetilde{L} & \mathbf{0} \\ \mathbf{v}^T & 1 \end{bmatrix} \begin{bmatrix} \widetilde{U} & \mathbf{u} \\ \mathbf{0} & u \end{bmatrix} = \begin{bmatrix} \widetilde{L}\widetilde{U} & \widetilde{L}\mathbf{u} \\ \mathbf{v}^T\widetilde{U} & \mathbf{v}^T\mathbf{u} + u \end{bmatrix}. \tag{1.37}$$

It follows then $\widetilde{U}_{jj} = U_{jj} \neq 0$ for all $j = 1, \ldots, n-2$ by the inductive hypothesis. If

$U_{n-1,n-1} = 0$, then $\mathbf{e}_n^T \widetilde{U} = \mathbf{0}$, and hence for any $\alpha \in \mathbb{R}$

$$A = \begin{bmatrix} \widetilde{L} & \mathbf{0} \\ \mathbf{v}^T + \alpha \mathbf{e}_n^T & 1 \end{bmatrix} \begin{bmatrix} \widetilde{U} & \mathbf{u} \\ \mathbf{0} & u - \alpha \mathbf{e}_n^T \mathbf{u} \end{bmatrix} \tag{1.38}$$

using (1.37), which contradicts the uniqueness of the LU decomposition of A. Hence, necessarily $U_{n-1,n-1} \neq 0$ as well so that $\det A_{:k,:k} = \det U_{:k,:k} = \prod_{j=1}^{k} U_{jj} \neq 0$ for all $k = 1, \ldots, n-1$. □

If $A \in \mathbb{C}^{n \times n}$, then there exist $Q \in \mathrm{U}(n)$ and $R \in \mathcal{U}_n$ where $R_{jj} > 0$ for all j such that $A = QR$. If $A \in \mathbb{R}^{n \times n}$, then $Q \in \mathrm{O}(n)$ and R is real. This is called the QR decomposition of A. Every square complex matrix has a QR decomposition. Common methods to compute this decomposition, which is unique whenever A is nonsingular, include the **Gram-Schmidt orthogonalization**, **Householder reflections**, or **Givens rotations**. I will give an overview of Gram-Schmidt here and postpone a discussion about Householder reflections and Givens rotations until Section 5.2.

If $\mathcal{V} = \{\mathbf{v}_i : i \in [n]\}$ is a collection of linearly independent vectors, then the Gram-Schmidt process can be used to generate an orthonormal basis of $\mathrm{span}(\mathcal{V})$. For simplicity, I will only consider the square case such that $\mathbf{v}_i \in \mathbb{C}^n$ for each i and so $\mathcal{V}$ constitutes the columns of a nonsingular order n matrix. It follows as:

Algorithm 1 Gram-Schmidt Process

procedure GramSchmidt($\mathcal{V}$)
  $\mathcal{B} = \varnothing$
  $n = \mathrm{size}(\mathcal{V})$
  for $i = 1 : n$ **do**
    $\mathbf{q}_i = \mathbf{v}_i - \sum_{j=1}^{i-1} \langle \mathbf{q}_j, \mathbf{v}_i \rangle \mathbf{q}_j$
    $\mathbf{q}_i = \dfrac{\mathbf{q}_i}{\|\mathbf{q}_i\|_2}$
    $\mathcal{B} = \mathcal{B} \cup \{\mathbf{q}_i\}$
    $\mathcal{V} = \mathcal{V} \backslash \{\mathbf{v}_i\}$
  return $\mathcal{B}$

By construction, if $\mathbf{q}_i, \mathbf{q}_j \in \mathcal{B} = \textsc{GramSchmidt}(\mathcal{V})$, then $\langle \mathbf{q}_i, \mathbf{q}_j \rangle = \delta_{ij}$, so that $\mathcal{B}$ is an orthonormal basis of $\text{span}(\mathcal{V})$. Moreover, note

$$\langle \mathbf{v}_i, \mathbf{q}_i \rangle = \langle C_i \mathbf{q}_i + \sum_{j=1}^{i-1} \langle \mathbf{q}_j, \mathbf{v}_i \rangle \mathbf{q}_j, \mathbf{q}_i \rangle = C_i \langle \mathbf{q}_i, \mathbf{q}_i \rangle + \sum_{j=1}^{i-1} \langle \mathbf{q}_j, \mathbf{v}_i \rangle \langle \mathbf{q}_j, \mathbf{q}_i \rangle = C_i$$

for

$$C_i^2 = \|\mathbf{v}_i\|_2^2 - \sum_{j=1}^{i-1} |\langle \mathbf{q}_j, \mathbf{v}_i \rangle|^2 \neq 0. \tag{1.39}$$

This suffices to establish:

Theorem 1.8. *If $A \in \mathbb{C}^{n \times n}$ is nonsingular, then A has a unique QR decomposition.*

Proof. For $\mathbf{a}_i = A\mathbf{e}_i$, let $\mathcal{V} = \{\mathbf{a}_i : i \in [n]\}$. For $\mathbf{q}_i \in \textsc{GramSchmidt}(\mathcal{V})$, let $\widetilde{Q} \in \mathrm{U}(n)$ such that $\widetilde{Q}\mathbf{e}_i = \mathbf{q}_i$. Let $\widetilde{R} \in \mathcal{U}_n$ such that $\widetilde{R}_{ij} = \langle \mathbf{q}_i, \mathbf{a}_j \rangle$, where we note $\widetilde{R}_{ij} = 0$ for $i > j$ by construction. Let $\widetilde{D} \in \mathcal{D}_n$ such that $\widetilde{D}_{jj} = \frac{\overline{R}_{jj}}{\|R_{jj}\|_2}$, where we note $R_{jj} \neq 0$ by (1.39). It follows then $\widetilde{D} \in \mathrm{U}(n)$, and so for $Q = \widetilde{Q}\widetilde{D}^*$ and $R = \widetilde{D}\widetilde{R}$, we have $A = QR$ is a QR decomposition of A, where we note R has positive diagonal by construction with $R_{jj} = \|\widetilde{R}_{jj}\|_2 > 0$. This establishes existence.

For uniqueness, suppose also $A = \widehat{Q}\widehat{R}$. Since R and $\widehat{R}$ have positive diagonals, then both are nonsingular, so we have $Q^*\widehat{Q} = R\widehat{R}^{-1} \in \mathcal{U}_n \cap \mathrm{U}(n)$. In particular, then $R\widehat{R}^{-1}$ is an upper triangular normal matrix and so it must be diagonal by Lemma 1.4; say $D = R\widehat{R}^{-1}$. Since $R, \widehat{R}$ have positive diagonal entries, then so does D, and since $D \in \mathrm{U}(n)$ then necessarily $D = \mathbf{I}$. It follows then $Q = \widehat{Q}$ and $R = \widehat{R}$. □

For any positive semi-definite A, there exists a lower-triangular matrix L such that $A = LL^*$; this is called the **Cholesky decomposition** of A, which is not unique if there are zeroes on the diagonal. We do have

Theorem 1.9. *If A is positive definite, then A has a unique Cholesky decomposition*

Proof. Since A is positive definite, then there exists $B = \sqrt{A}$ such that $A = BB^*$ (let $B = U\sqrt{D}V$ for $A = UDV$ the unique SVD factorization, where D is a diagonal matrix of positive singular values of A and $\sqrt{D}$ is the corresponding diagonal matrix where the square root function is applied to the diagonal entries of D). Let $B^* = QR$ be the QR decomposition of B^*, where $Q \in \mathrm{U}$ and R upper triangular with positive diagonal entries. Then $A = BB^* = (QR)^*QR = (R^*Q^*)(QR) = R^*R = LL^*$ for $L = R^*$ lower triangular with positive diagonal. This shows existence.

For uniqueness, assume $A = LL^* = L_0L_0^*$ for L_0 with positive diagonal. In particular, then L and L_0 are nonsingular, so that $\mathbf{I} = L^{-1}(L_0L_0^*)(L^*)^{-1} = (L^{-1}L_0)(L^{-1}L_0)^*$. Hence, $(L^{-1}L_0)^{-1} = (L^{-1}L_0)^* \in \mathcal{L}_n \cap \mathcal{U}_n = \mathcal{D}_n$. Write $L^{-1}L_0 = D$ where necessarily D has positive diagonal entries since both L and L_0 do. It follows $\mathbf{I} = DD^* = D^2$ such that $(D_{jj})^2 = 1$ for all j. Since $D_{jj} > 0$, then $D = \mathbf{I}$ so that $L = L_0$. □

1.4.3 Gaussian elimination

Gaussian elimination (**GE**) remains the most prominent approach to solving linear systems $A\mathbf{x} = \mathbf{b}$. A standard description involves using elementary row operators to zero out below the main diagonal, moving one column at a time from left to right, which is outlined in Algorithm 2.

GE without pivoting (**GENP**) results in the factorization $A = LU$ for L lower triangular with positive unit diagonal with

$$L_{ij} = \frac{A_{ij}^{(j)}}{A_{jj}^{(j)}} \tag{1.40}$$

Algorithm 2 Gaussian Elimination

```
1: procedure GaussianElimination(A)
2:     n = size(A)
3:     L = I_n
4:     U = A
5:     for i = 1 : n − 1 do
6:         L(i + 1 : n, i) = U(i + 1 : n, i)/U(i, i)
7:         for j = i + 1 : n do
8:             U(j, :) = U(j, :) − L(j, i)U(j, :)
9:     return [L, U]
```

for $i > j$ where $A^{(k)}$ be the matrix with zeros below the first $k-1$ diagonal entries that results before the k^{th} step in GENP, and $U = A^{(N)}$ is upper triangular with $U_{jj} = A^{(j)}_{jj}$. Such a factorization is the associated LU **factorization** of A. A standard result shows GENP can be carried out in $\frac{2}{3}N^3 + O(N^2)$ flops (see Section 1.5.1).

The **pivot** using GE is the leading entry in the remaining order $n-k+1$ untriangularized lower block of $A^{(k)}$. If zero is encountered in the pivot then the algorithm would terminate to avoid dividing by zero. Such a matrix would not have an LU factorization using GENP. **Pivoting**, which involves a sequence of row or column changes to change the pivot, is sometimes necessary to enable an LU factorization to be possible for a nonsingular matrix A. This would result in a factorization $PAQ = LU$ for P and Q permutation matrices, which will be called the LU factorization for the associated pivoting strategy. Although pivoting adds computational costs to GE, pivoting is often employed for numerical stability in the computed solution. Selecting a pivoting strategy for GE often involves weighing accuracy against computation time.

The most common pivoting strategy is GE with partial pivoting (**GEPP**). GEPP involves an additional scan at each intermediate step to find the entry of max norm in $A^{(k)}$ below the diagonal within the column and then a row swap when this value exceeds the norm of the initial pivot, which yields a factorization $PA = LU$ for a permutation matrix P. GE with complete pivoting (**GECP**) involves a scan through the lower-untriangularized

remaining block of $A^{(k)}$ for the max norm entry, followed by row and column permutations to move the max norm entry to the pivot. This results in a $PAQ = LU$ factorization for permutation matrices P, Q. For this document, we will assume a pivot search chooses the pivot with minimal taxi cab distance with respect to the row and column indices to the main leading diagonal entry in the remaining subblock, prioritizing minimal row index distance over column index in the case of a tie.

GE with rook pivoting (**GERP**) involves iteratively scanning within each associated row and column to find the max norm entry to find the candidate pivot. This is followed then by the associated row and column swaps to move the resulting entry to the pivot. The name of the pivoting scheme is derived from the limitation on the pivot scans to paths a rook could make on a chess board. Note the added complexity for pivoting using GERP is bounded below by twice the complexity of pivoting using GEPP to the full complexity of pivoting using GECP. See [32] for further discussion regarding GERP. We will not explore additional numerical experiments for GERP beyond highlighting additional connections to our chosen models. Note for this document we will assume GERP always sequences column scans before row scans at each intermediate step.

Remark 1.1. *I will write $A^{(k)}$ to denote the matrix with zeros below the first $k-1$ diagonal entries using GE with a specified pivot scheme. By default, I will use $A^{(k)}$ to denote the intermediate GENP step. Let $L^{(k)}$ denote the sequence of lower triangular row operations such that $L^{(k)}A = A^{(k)}$. If ambiguous, I will specify the pivoting strategy in consideration in the superscript (e.g., $A^{(k,PP)}, L^{(k,CP)}$), where then $A^{(k)} = A^{(k,NP)}$.*

The total operational costs of these pivoting schemes differs only in the approaches to find each pivot. If we consider a comparison between two elements during a pivot scan as a flop, the additional scans from these pivoting schemes add $O(N^2)$ and $O(N^3)$ flops, respectively, for GEPP and GECP, while GERP ranges from twice the GEPP complexity to the full GECP complexity. In [32], computations with iid models show that on average GERP is

$O(N^2)$ as well, only accounting for 3 times more scans than GEPP.

Of note, GEPP can be carried out without affecting the leading order of complexity for GE. In practice, however, pivoting can lead to significant computational overhead due to memory storage. In [2], numerical experiments running GEPP on a order 10,000 random matrix using a hybrid CPU/GPU environment resulted in pivoting accounting for 20 percent of the total computation time.

Using GENP, Theorem 1.7 gives sufficient criteria for the existence and uniqueness of an LU factorization. However, uniqueness of LU factorization using different pivoting strategies is not invariant under row or column permutations: if $B = PAQ$ for permutation matrices P, Q, then the GE factorizations $P'AQ' = L'U'$ and $P''BQ'' = L''U''$ do not necessarily have $L' = L''$, $U' = U''$, $P' = P''P^T$ or $Q' = Q^TQ''$. Non-uniqueness results from a "tie" encountered during a pivot search. The following example illustrates this point:

Example 1.1. *Let*

$$A = \begin{bmatrix} 2 & 1 \\ 2 & 2 \end{bmatrix} \quad \textit{and} \quad B = P_{(1\ 2)}A = \begin{bmatrix} 2 & 2 \\ 2 & 1 \end{bmatrix}.$$

Using GEPP, no pivoting would be required for either A or B, so that A and B have distinct LU factors, with

$$A = \begin{bmatrix} 1 & \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ & 1 \end{bmatrix} \quad \textit{and} \quad B = \begin{bmatrix} 1 & \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 2 \\ & -1 \end{bmatrix}.$$

Recall that the L factor from GEPP satisfies:

$$|L_{ij}| \leq 1 \quad \text{for any } i > j. \tag{1.41}$$

A tie would occur using GEPP only when $|L_{ij}| = 1$ for some $i > j$. When no ties are encountered at any intermediate stage, GEPP results in unique P, L, and U factors, leading to:

Theorem 1.10. *Let A be a nonsingular square matrix. Then the factors in the GEPP factorization $PA = LU$ are invariant under row permutations on A iff $|L_{ij}| < 1$ for all $i > j$.*

Proof. By relabeling $B = PA$, we can assume $P = \mathbf{I}$. Suppose first $|L_{ij}| = 1$ for some $i > j$. Let j' and then i' be minimal such that this occurs. Since then $|B^{(j')}_{j'j'}| = |B^{(j')}_{i'j'}|$, adding the permutation $(i'\ j')$ would yield a different P_σ factor with $\sigma(j') = i' \neq j'$, so that $P_\sigma \neq \mathbf{I}$. Now suppose $|L_{ij}| < 1$ for all $i > j$. Suppose $P_\sigma B = L'U'$ is another GEPP factorization of B for some $\sigma \in S_n$. Suppose σ is a nontrivial permutation. Let i be the first non-fixed point of σ, and note then $i < \min(\sigma(i), \sigma^{-1}(i))$. It follows

$$|(L')_{\sigma(i),i}| = \left|\frac{(P_\sigma B)^{(i)}_{\sigma(i),i}}{(P_\sigma B)^{(i)}_{ii}}\right| = \left|\frac{B^{(i)}_{ii}}{B^{(i)}_{\sigma^{-1}(i),i}}\right| = \frac{1}{|L_{\sigma^{-1}(i),i}|} > 1.$$

This contradicts (1.41). It follows σ must be the trivial permutation so that $P = \mathbf{I}$. The uniqueness of the L and U factors follows from Theorem 1.7. □

1.4.4 Direct sum of matrices and Kronecker product

Let $A \oplus B \in \mathbb{R}^{(n_1+n_2)\times(m_1+m_2)}$ denote the **direct sum** of A and B, which is the block diagonal matrix with blocks $A \in \mathbb{R}^{n_1\times m_1}$ and $B \in \mathbb{R}^{n_2\times m_2}$, i.e.,

$$A \oplus B = \begin{bmatrix} A & \\ & B \end{bmatrix}. \tag{1.42}$$

Note $(A \oplus B)^* = A^* \oplus B^*$, $(A \oplus B)^T = A^T \oplus B^T$ and $\operatorname{Tr}(A \oplus B) = \operatorname{Tr}(A) + \operatorname{Tr}(B)$. We can then write $\bigoplus_{j=1}^{n} d_j$ for the diagonal matrix $D \in \mathbb{R}^{n \times n}$ such that $D_{jj} = d_j$. If A and B are square matrices, then $\det(A \oplus B) = \det(A)\det(B)$, so $(A \oplus B)^{-1} = A^{-1} \oplus B^{-1}$ if either side exists. Moreover, if A and B are both unitary or orthogonal, then so is $A \oplus B$.

Let $A \otimes B \in \mathbb{R}^{n_1 n_2 \times m_1 m_2}$ denote the **Kronecker product** of $A \in \mathbb{R}^{n_1 \times m_1}$ and $B \in \mathbb{R}^{n_2 \times m_2}$, given by

$$A \otimes B = \begin{bmatrix} A_{11}B & \cdots & A_{1,m_1}B \\ \vdots & \vdots & \vdots \\ A_{n_1,1}B & \cdots & A_{n_1,m_1}B \end{bmatrix}. \tag{1.43}$$

Note $(A \otimes B)^* = A^* \otimes B^*$, $(A \otimes B)^T = A^T \otimes B^T$ and $\operatorname{Tr}(A \otimes B) = \operatorname{Tr}(A)\operatorname{Tr}(B)$. Recall $A \otimes B = P(B \otimes A)Q$ for perfect shuffle permutation matrices P, Q. See [8] for an overview of properties and structures of perfect shuffles. If A and B are both square then $Q = P^T$, so that $A \otimes B$ and $B \otimes A$ are conjugate. Also, recall the **mixed-product property** of the Kronecker product: if the products AC and BD can be computed, then

$$(A \otimes B)(C \otimes D) = AC \otimes BD. \tag{1.44}$$

A useful mnemonic for the mixed-product property is "the product of the Kronecker products is the Kronecker product of the products".

If $n_1 = m_1 = n$ and $n_2 = m_2 = m$, then

$$\begin{aligned} \det(A \otimes B) &= \det((A \otimes \mathbf{I}_m)(\mathbf{I}_n \otimes B)) = \det(A \otimes \mathbf{I}_m)\det(\mathbf{I}_n \otimes B) \\ &= \det(\mathbf{I}_m \otimes A)\det(\mathbf{I}_n \otimes B) = \det(\bigoplus^{m} A)\det(\bigoplus^{n} B) \\ &= \det(A)^m \det(B)^n, \end{aligned}$$

and so $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ when either side exists. Moreover, we see if A, B are both unitary or orthogonal, then so is $A \otimes B$.

Note the mixed-product property and the bilinearity of the Kronecker product yields if $A\mathbf{u} = \lambda\mathbf{u}$ and $B\mathbf{v} = \mu\mathbf{v}$, then

$$(A \otimes B)(\mathbf{u} \otimes \mathbf{v}) = A\mathbf{u} \otimes B\mathbf{v} = (\lambda\mathbf{u}) \otimes (\mu\mathbf{v}) = (\lambda\mu)(\mathbf{u} \otimes \mathbf{v}). \tag{1.45}$$

The Kronecker product in particular allows simplified matrix norm calculations:

Lemma 1.6. *If $\|\cdot\|$ is an induced matrix norm or $\|\cdot\|_{\max}$, then $\|A \otimes B\| = \|A\|\|B\|$ for $A \in \mathbb{R}^{n_1 \times m_1}$ and $B \in \mathbb{R}^{n_2 \times m_2}$.*

Proof. The result for $\|\cdot\| = \|\cdot\|_{\max}$ follows immediately from the definition of $\|\cdot\|_{\max}$. If $\|\cdot\|$ is an induced matrix norm, then the multiplicative property is established in [26, Theorem 8]. □

Additionally, the Kronecker product allows for straightforward matrix factorizations determined directly from the corresponding factorizations of each factor. Of particular utility for these factorizations is the fact that several classes of matrices of interest are closed under $\otimes$. For example, as seen above, orthogonal and unitary matrices are closed under $\otimes$. Similarly, direct computations would verify lower triangular, diagonal, and upper triangular matrices are closed under $\otimes$. Combining this with the mixed-product property, this yields certain factorizations can be computed by factoring each Kronecker product factor separately. This is illustrated in the following result, which will be used repeatedly in Chapter 5:

Lemma 1.7. *For $j = 1, 2, \ldots, n$, if $P_j A_j Q_j = L_j U_j$ for permutation matrices P_j, Q_j and unit lower triangular L_j and upper triangular U_j and $A = \bigotimes_{j=1}^{n} A_j$, then $PAQ = LU$ for*

permutation matrices $P = \bigotimes_{j=1}^{n} P_j$, $Q = \bigotimes_{j=1}^{n} Q_j$ *and unit lower triangular* $L = \bigotimes_{j=1}^{n} L_j$ *and upper triangular* $U = \bigotimes_{j=1}^{n} U_j$.

Proof. This follows directly from the mixed-product property (1.44) and induction on n. □

1.5 Numerical analysis

Note my particular focus within numerical analysis will be on rounding error analysis, so I will limit an overview of relevant ideas here. For further background, even on the following topics, see [16, 35].

1.5.1 Complexity

Floating-point arithmetic is a method to provide an approximate representation of real numbers that enables machines to carry out common computational processes. This remains to be the most commonly used method by computers for representing real numbers. Alternative methods include fixed-point representations and logarithmic number systems. Properties of these methods as well as other alternative methods, including comparisons to floating-point arithmetic, can be further explored in standard texts (e.g., [16, 22]).

Standard implementations of floating-point arithmetic results in working with a fixed number of significant digits. This is in contrast to exact arithmetic, which does not have a finite cut-off and can handle exactly irrational numbers. The **machine epsilon** for a fixed computing system, denoted by $\epsilon = \epsilon_{\text{machine}}$, is defined by

$$\epsilon = \min\{x > 0 : fl(1 + x) \neq 1\}, \tag{1.46}$$

where $fl(\cdot)$ is used to denote rounding to a fixed significand.

Complexity of algorithms or processes are often measured in terms of the counts of floating-point arithmetic operations or **flops**, which will include additions, subtractions, multiplications, divisions and sometimes other common or black box computations (e.g., square roots, comparisons between two numbers, samples of a random number). Using standard floating arithmetic, we have

$$fl(x \text{ op } y) = (x \text{ op } y)(1+\delta) \tag{1.47}$$

for $|\delta| \leq \epsilon$, where op is a stand-in for a flop. I will revisit this shortly after the discussion about relative error.

As a short illustration of some of these ideas, I will include a few straightforward applications of complexity computations for common computational processes.

Example 1.2 (Matrix-vector multiplication)**.** *Let $A \in \mathbb{R}^{n\times m}$ and $\mathbf{x} \in \mathbb{R}^m$. Each component of $A\mathbf{x}$ consists of a dot product of a row from A and $\mathbf{x}$, which then involves m multiplications and $m-1$ additions, so $2m-1$ flops overall. In total, then $A\mathbf{x}$ takes $n(2m-1)$ total flops to compute.*

Example 1.3 (Matrix-matrix multiplication)**.** *Let $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{m\times k}$. Then AB takes $n(2m-1)k$ total flops.*

Example 1.4. *We can analyze the complexity of Algorithm 1, again applied only to the square case: for step i in the **for** loop, each dot product takes n multiplications and $n-1$ additions and so $2n-1$ flops. So $\langle \mathbf{q}_j, \mathbf{v}_i\rangle \mathbf{q}_j$ takes n more multiplications, meaning $3n-1$ flops. So then $\mathbf{v}_i - \langle \mathbf{q}_j, \mathbf{v}_i\rangle \mathbf{q}_j$ then takes n subtractions on top of the prior step, so there are $4n-1$ flops and hence $\mathbf{v}_i - \sum_{j=1}^{i-1} \langle \mathbf{q}_j, \mathbf{v}_i\rangle \mathbf{q}_j$ takes $(i-1)(4n-1)$ flops. Then vector normalization takes a dot product, a square root, and n divisions, so $3n$ additional flops. Hence, step i takes $(i-1)(4n-1)+3n = (4n-1)i + 1 - n$ flops. So running Algorithm 1*

then takes

$$\sum_{i=1}^{n}((4n-1)i+1-n) = (4n-1)\sum_{i=1}^{n} i + n(1-n) = \frac{1}{2}(4n-1)n(n+1) + n(1-n)$$
$$= 2n^3 + \frac{1}{2}n(n+1) = 2n^3 + O(n^2)$$

total flops.

1.5.2 Stability

For a procedure with exact solution $\mathbf{x}$ and computed solution $\hat{\mathbf{x}}$, the **relative error** is

$$\frac{\|\mathbf{x}-\hat{\mathbf{x}}\|}{\|\mathbf{x}\|}. \tag{1.48}$$

Example 1.5. *We can explore the relative error of a flop. For scalars x and y, let $\mathbf{x} = (x \text{ op } y)$ and $\hat{\mathbf{x}} = fl(\mathbf{x})$ where* op *is a stand-in for a flop (e.g., addition or multiplication). By* (1.47)*, then*

$$\frac{\|\mathbf{x}-\hat{\mathbf{x}}\|}{\|\mathbf{x}\|} = \frac{\|\mathbf{x}(1-(1+\delta))\|}{\|\mathbf{x}\|} = |\delta| \leq \epsilon. \tag{1.49}$$

For a procedure $y = f(x)$, the **forward error** of the computed output $\hat{\mathbf{y}}$ is the relative error of $\hat{\mathbf{y}}$ relative to $\mathbf{y}$. The **backward error** of $\hat{\mathbf{y}}$ is

$$\min\left\{\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} : \hat{\mathbf{y}} = f(\mathbf{x}+\delta\mathbf{x})\right\}. \tag{1.50}$$

In particular, if the backward error of $\hat{\mathbf{y}}$ is sufficiently small relative to ϵ (e.g., $O(\epsilon)$), then $\hat{\mathbf{y}}$ is the exact answer to a slightly perturbed input.

If a procedure always produces a small backward error, then the procedure is called **back-**

ward stable. A lot of the framework used in backward error analysis was established by Wilkinson in his analysis of Gaussian elimination with partial pivoting [39]. Section 5.3 will look deeper into this particular result.

1.6 Probability background

1.6.1 Background

We say $(\Omega, \mathcal{A}, \mathbb{P})$ is a **probability space** if $\Omega \in \mathcal{A} \subset 2^\Omega$ is a σ-algebra and $\mathbb{P} : \mathcal{A} \to [0,1]$ is a measure such that $\mathbb{P}(\Omega) = 1$. Elements of $\mathcal{A}$ are called **events**. Two events A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. A map $X : \Omega \to \mathbb{F}$ is a **random variable** if X is measurable with respect to $\mathcal{A}$ and $\mathcal{B}_\mathbb{F}$, the Borel set of $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, which is the σ-algebra generated by the open sets in $\mathbb{F}$; i.e., $f^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}_\mathbb{F}$. For X a random variable, we define the expectation,

$$\mathbb{E}X = \int X \ \mathrm{d}\,\mathbb{P} = \int X(\omega) \ \mathrm{d}\,\mathbb{P}(\omega). \tag{1.51}$$

If $\mathbb{E}|X|^2 < \infty$, we define the **variance** of X to be $\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 - \mathbb{E}X^2 - (\mathbb{E}X)^2$. We can define the induced measure $\mathbb{P}_X$ by $\mathbb{P}_X(A) = \mathbb{P}(X \in A)$. If $\mathbb{P}_X = \mathbb{P}_Y$, then we say X and Y are **equal in distribution**, which is denoted $X \sim Y$. If X is a real random variable, then the (cumulative) **distribution function** of X, $F_X(t) = \mathbb{P}(X \leq t)$, completely determines the distribution of X. Let $\mathrm{supp}(X) = \{\omega \in \Omega : |X(\omega)| > 0\}$ denote the **support** of X. If $\mathrm{supp}(X)$ is discrete, then X is a discrete random variable, and we can define the **probability mass function** $\mathbb{P}(X = t)$ for $t \in \mathrm{supp}(X)$. If X is a continuous real random variable, then we can define the **density function** $f_X(t) = \frac{d}{dt}F_X(t)$ for $t \in \mathrm{supp}(X)$. We define the **characteristic function** of X by $\varphi_X(t) = \mathbb{E}e^{itX}$, which completely determines the distribution of X. We say X and Y are **independent** if $[X \in A]$ and

$[Y \in B]$ are independent for all $A, B \in \mathcal{A}$. Equivalently, X and Y are independent if and only if $\mathbb{E}f(X)g(Y) = (\mathbb{E}f(X))(\mathbb{E}g(Y))$ for all continuous bounded f, g. Induction can extend this definition to include a finite collection of independent random variables. We say a family of random variables $\{X_i : i \in \mathcal{I}\}$ is independent if any finite subcollection of events using distinct X_i is independent, and the family is **iid** (independent and identically distributed) if additionally $X_i \sim X_1$ for each i.

Suppose $\{X_i : i = 1, 2, \ldots\}$ is a family of real random variables. We say X_n **converges in r-mean** if

$$\lim_{n\to\infty} \mathbb{E}|X_n - Y|^r = 0 \tag{1.52}$$

for some random variable Y defined on the same probability space as X_i. (Such a probability space is guaranteed to exist by Carathéodory's Extension Theorem.) We say X_n **converges in distribution** to a random variable Y if for all t in the continuity set of F_Y we have

$$\lim_{n\to\infty} F_{X_n}(t) = F_Y(t). \tag{1.53}$$

We say X_n **convergences in probability** if for all $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|X_n - Y| > \varepsilon) = 0 \tag{1.54}$$

for a random variable Y. We say X_i **converges almost surely** to Y if $\mathbb{P}(X_n \to Y) = 1$. Almost sure convergence and convergence in r-mean both imply convergence in probability, which implies convergence in distribution, while the reverse implications in general do not hold.

If P, Q are two probability measures on a finite group G, then we can define the **convolution**

of P and Q, written $P * Q$ as

$$P * Q(s) = \sum_{t \in G} P(st^{-1})Q(t). \tag{1.55}$$

If G is a compact group, then for measurable $A \subset G$,

$$P * Q(A) = \int \mathbb{1}_A(s) \ \mathrm{d}\, P(st^{-1}) \, \mathrm{d}\, Q(t). \tag{1.56}$$

1.6.2 Common distributions

Below are several common distributions that will be studied in the following text:

Normal distribution

If $X \sim N(\mu, \sigma^2)$ then X has density

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \tag{1.57}$$

while $\mathbb{E}X = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Let $Z \sim N(0,1)$ denote a standard normal random variable, where we note if $X \sim N(\mu, \sigma^2)$ then $X \sim \sigma Z + \mu$. We write $X \sim N_{\mathbb{C}}(0,1)$ if $X = X_1 + X_2 i$ for X_i iid $N(0, \frac{1}{2})$.

Uniform distribution

If $\mathcal{A}$ is a finite set, then we say $X \sim \mathrm{Uniform}(A)$ if for $a \in A$ we have mass function

$$\mathbb{P}(X = a) = \frac{1}{|A|}. \tag{1.58}$$

If $X \sim \text{Uniform}(a, b)$ then X has support on (a, b) where

$$F_X(t) = \frac{t-a}{b-a} \tag{1.59}$$

for $t \in (a, b)$, and so has density

$$f_X(t) = \frac{1}{b-a} \mathbb{1}_{(a,b)}(t). \tag{1.60}$$

The following result will be used a few times throughout the document:

Lemma 1.8. *If $X \sim \text{Uniform}(0, 1)$ and Y is independent of X, then $(X+Y) \pmod 1 \sim X$.*

Proof. Since $X + Y \mid Y \sim \text{Uniform}(Y, Y + 1)$ and so $(X + Y) \pmod 1 \mid Y \sim \text{Uniform}(0, 1)$, then

$$\mathbb{P}((X + Y) \pmod 1 \leq t) = \mathbb{E}\mathbb{P}((X + Y) \pmod 1 \leq t \mid Y) = \mathbb{E}\mathbb{P}(X \leq t) = \mathbb{P}(X \leq t).$$

□

For example, if $\theta \sim \text{Uniform}([0, 2\pi))$, then since $\cos x$ and $\sin x$ are periodic of period 2π, then $\sin\theta = \cos(\frac{\pi}{2} - \theta) = \cos(\theta - \frac{\pi}{2}) \sim \cos\theta$.

Arcsine distribution

If $Y \sim \text{Arcsine}(0, 1)$ then for $t \in [0, 1]$

$$F_Y(t) = \frac{2}{\pi} \arcsin\sqrt{t} \tag{1.61}$$

with density

$$f_Y(t) = \frac{1}{\pi} \frac{1}{\sqrt{t(1-t)}} \mathbb{1}_{(0,1)}(t). \tag{1.62}$$

The Arcsine distribution can be generated using uniform random variables:

Lemma 1.9. *If* $\theta \sim \text{Uniform}([0, 2\pi))$ *then* $\sin^2(2\theta) \sim \text{Arcsine}(0, 1)$.

Proof. Let $\varphi \sim \text{Uniform}([0, \pi))$. Note first $2\theta \pmod{\pi} \sim \theta \pmod{\pi} \sim \varphi$, so since $|\sin(x)|$ has period π then $|\sin(2\theta)| \sim \sin\varphi$. Now note for $t \in [0, 1]$

$$\mathbb{P}(|\sin 2\theta| \leq t) = \mathbb{P}(\sin\varphi \leq t) = \mathbb{P}(\varphi \in [0, \arcsin t] \cup [\pi - \arcsin t, \pi]) = \frac{2}{\pi} \arcsin t.$$

Hence, $\mathbb{P}(\sin^2(2\theta) \leq t) = \mathbb{P}(|\sin(2\theta)| \leq \sqrt{t}) = \frac{2}{\pi} \arcsin\sqrt{t}$. □

Cauchy distribution

For $X \sim \text{Cauchy}(1)$, then

$$F_X(t) = \frac{1}{\pi} \arctan t + \frac{1}{2} \tag{1.63}$$

with density

$$f_X(t) = \frac{2}{\pi} \frac{1}{1+t^2}. \tag{1.64}$$

Note $X \sim -X$ (since $\arctan x$ is an odd function), and so

$$\mathbb{P}(|X| \leq t) = 2\mathbb{P}(X \leq t) - 1 = \frac{2}{\pi} \arctan t. \tag{1.65}$$

In particular, $\mathbb{P}(|X| \leq 1) = \frac{1}{2}$.

If $X \sim \text{Cauchy}(1)$, then X has no absolute moments of order $k \geq 1$: by the Cauchy inequality, it suffices to show X does not have a finite absolute first moment (since $\mathcal{L}^1(\mathbb{P}) \supset \mathcal{L}^k(\mathbb{P})$ for $k \geq 1$). Note if $|x| \geq 1$ then $1 + \frac{1}{x^2} \leq 2$ so that $|x| f_X(x) \geq \frac{1}{\pi} \frac{1}{|x|}$. It follows

$$\begin{aligned}\mathbb{E}|X| &= \int_{-\infty}^{\infty} |x| f_X(x) \ \mathrm{d}\, x = 2 \int_0^{\infty} |x| f_X(x) \ \mathrm{d}\, x \geq 2 \int_1^{\infty} |x| f_X(x) \ \mathrm{d}\, x \\ &\geq \frac{2}{\pi} \int_1^{\infty} \frac{1}{|x|} \ \mathrm{d}\, x = \infty.\end{aligned}$$

Similarly, one can show $\mathbb{E}|X|^k < \infty$ for $k < 1$ and hence $\mathbb{E} \ln |X| < \infty$.

Cauchy random variables can also be generated using uniform random variables, as follows:

Lemma 1.10. *If* $\theta \sim \text{Uniform}([0, 2\pi))$, *then* $\tan\theta, \cot\theta \sim \text{Cauchy}(1)$.

Proof. Recall first $\tan(\pi(Y - \frac{1}{2})) \sim \text{Cauchy}(1)$ for $Y \sim \text{Uniform}(0, 1)$ since for $t \in (0, 1)$ we have

$$\mathbb{P}(\tan(\pi(Y - \frac{1}{2})) \leq t) = \mathbb{P}(Y \leq \frac{1}{\pi} \arctan t + \frac{1}{2}) = \frac{1}{\pi} \arctan t + \frac{1}{2} = \mathbb{P}(X \leq t).$$

Note $\pi(Y - \frac{1}{2}) \sim \text{Uniform}(-\frac{\pi}{2}, \frac{\pi}{2})$ and $\pi(Y - \frac{1}{2}) \pmod{\pi} \sim \text{Uniform}(0, \pi)$ while $\theta \pmod{\pi} \sim \text{Uniform}(0, \pi)$. Using the periodicity of $\tan x$, we have

$$\tan\theta = \tan(\theta \pmod{\pi}) \sim \tan(\pi(Y - \frac{1}{2}) \pmod{\pi}) = \tan(\pi(Y - \frac{1}{2})) \sim \text{Cauchy}(1).$$

Hence, we have also $\cot\theta \sim \cot(\theta - \frac{\pi}{2}) = -\tan\theta \sim \text{Cauchy}(1)$ by Lemma 1.8. □

Lemma 1.11. *If* $\theta \sim \text{Uniform}([0, 2\pi))$ *and* $X \sim \text{Cauchy}(1)$, *then*

$$\min(|\tan\theta|, |\cot\theta|) \sim |X| \mid |X| \leq 1.$$

Proof. Using Lemma 1.10, then for $t \in [0,1]$, we have

$$\begin{aligned}
&\mathbb{P}(\min(|\tan\theta|, |\cot t|) \leq t) \\
&\quad = 1 - \mathbb{P}(\min(|X|, \frac{1}{|X|}) \geq t) = 1 - \mathbb{P}(|X| \geq t, \frac{1}{|X|} \geq t) \\
&\quad = 1 - \mathbb{P}(t \leq |X| \leq \frac{1}{t}) = 1 + \mathbb{P}(|X| \leq t) - \mathbb{P}(|X| \leq \frac{1}{t}) \\
&\quad = 1 + \frac{2}{\pi}\left(\arctan t - \arctan\frac{1}{t}\right) = 1 + \frac{2}{\pi}\left(2\arctan t - \frac{\pi}{2}\right) \\
&\quad = \frac{4}{\pi}\arctan t = \frac{\mathbb{P}(|X| \leq t)}{\mathbb{P}(|X| \leq 1)} \\
&\quad = \mathbb{P}(|X| \leq t \mid |X| \leq 1)
\end{aligned}$$

using also the fact $\arctan t$ and $\arctan \frac{1}{t}$ comprise complementary angles when $t > 0$. □

Haar measure

If G is a compact and separable topological group, then there exists a (left and right) invariant (inner and outer regular) Radon measure μ, called the **Haar measure** on G, such that $\mu(G) = 1$ and $\mu(A) = \mu(gA) = \mu(Ag)$ for any measurable $A \subset G$. The following is a classical result first due to Weil:

Theorem 1.11 ([38])**.** *Let G be a locally compact Hausdorff group. Then there exists a left (right) Haar probability measure on G.*

Write $X \sim \text{Haar}(G)$ if $\mathbb{P}_X = \mu$. This allows one to sample uniformly from G. If G is finite, then $\mu(g) = \frac{1}{|G|}$, so that $\text{Haar}(G) = \text{Uniform}(G)$.

Note if G is also compact, then it is **unimodular**, so that a left Haar measure and right Haar measure necessarily coincide. If $\varphi : G_1 \to G_2$ is a group isomorphism of compact Polish groups and μ_1 is the Haar measure on G_1, then the **push forward** measure $\mu_2 = \mu_1 \circ \varphi^{-1}$ is the Haar measure on G_2.

If H is a closed normal group of a locally compact Hausdorff group G, then the quotient map $\pi : G \to G/H$ induces the Haar measure on G/H through the push forward measure

$$\mu_{G/H} = \frac{1}{C}\mu \circ \pi^{-1}, \tag{1.66}$$

where C is a normalizing constant so that $\mu_{G/H}$ is a probability measure: since H is normal, then we can write $\pi^{-1}(A) = AH = HA$ for $A \subset G/H$, where we note then

$$gH \cdot A = gH \cdot AH = gH \cdot HA = g \cdot HA = g \cdot \pi^{-1}(A)$$

so that

$$\mu_{G/H}(gH \cdot A) = \frac{1}{C}\mu(g\pi^{-1}(A)) = \frac{1}{C}\mu(\pi^{-1}(A)) = \mu_{G/H}(A);$$

we further note this measure is Radon since its compact sets are of the form KH for K compact in G.

For example, the Haar measure on $\mathbb{T}$ can be defined as the push forward of the Haar measure on $[0, 2\pi)$ using $\theta \mapsto e^{i\theta}$, so that it has density $\frac{1}{2\pi i}\frac{\mathrm{d}\, z}{z}$.

1.6.3 Universality

For complex systems, computing exact solutions for common questions can be difficult – if not impossible – as the number of variables increases. However, a particular phenomenon arises in some instances when an increase in the number of variables leads to higher structure and predictability in the system, and enables asymptotic analysis of particular statistics. This phenomenon is called a **universality principle**. Universality principles often focus on the limiting distribution for a family of random objects, which can have weak restrictions on their initial structures (e.g, iid with finite first moment while being agnostic of the particular initial distribution). This then leads to a series of other closely tied shared statistics for

objects in the same universality class.

The most famous universality result, which is perhaps the most widely used result from mathematics, is the central limit theorem.

Theorem 1.12 (Central limit theorem (CLT)). *Let* $\{X_i : i = 1, 2, \ldots\}$ *be a family of iid random variables with finite second moments such that* $\mathbb{E}X_1 = \mu$ *and* $\operatorname{Var} X_1 = \sigma^2$. *For* $S_n = \sum_{i=1}^n X_i$, $\dfrac{S_n - n\mu}{\sqrt{n}\sigma}$ *converges to* $Z \sim N(0, 1)$ *in distribution.*

Note in the CLT, the limiting distribution has nothing to do with the actual distributions of the X_i other than the fact they have a finite second moment.

Another famous universality result is:

Theorem 1.13 (Strong Law of Large Numbers ((S)LLN)). *Let* $\{X_i : i = 1, 2, \ldots\}$ *be a sequence of random variables with finite first moment such that* $\mathbb{E}X_1 = \mu$. *Then* $\frac{1}{n}S_n$ *converges almost surely to* μ.

The Weak Law of Large Numbers (WLNN) is a corollary involving this convergence holding in probability, although this can be proved directly (and much more simply) without using the SLNN. Other universality results will be explored further in Section 1.7.

1.6.4 Subgroup algorithm

Suppose G is a finite group and let $1 = H_0 \subset H_1 \subset H_2 \subset \cdots \subset H_k = G$ be a chain of subgroups inside G. Any element $g \in G$ can be written as $g = h_1 h_2 \cdots h_k$ for $h_1 \in H_1$ and $h_k \in H_k/H_{k-1}$. This yields a surjective map $f : \prod_{j=1}^k H_k/H_{k-1} \to G$. If one can sample H_k/H_{k-1} uniformly from each coset, then this map would produce a uniform element of G. This is the idea behind the Subgroup algorithm, which was introduced by Diaconis and Shahshahani in [9].

Theorem 1.14 (Subgroup algorithm, [9]). *Let G be a compact Polish topological group group and H a closed subgroup. For $\pi : G \to G/H$ the quotient map, let $\varphi : G/H \to G$ be measurable such that $\pi \circ \varphi$ is the identity map on G/H (using the Axiom of choice). Let $T : G \to G/H \times H$ be defined by*

$$T(g) = (\pi(g), (\varphi\pi(g))^{-1}g). \tag{1.67}$$

T is bijective and bimeasurable with inverse $T^{-1}(x, h) = \varphi(x)h$. Let $dP_G, dP_H, dP_{G/H}$ be invariant probability measures on $G, H, G/H$, respectively. For $\widetilde{\varphi} : G/H \to G$ a measure preserving transformation of G/H, let $\mathrm{d}\,\widetilde{P}_{G/H}$ denote the image of $\mathrm{d}\,P_{G/H}$ under $\widetilde{\varphi}$. Then

$$\mathrm{d}\,P_G = \mathrm{d}\,\widetilde{P}_{G/H} * \mathrm{d}\,P_H. \tag{1.68}$$

Note (1.68) yields

$$\int_G f(g)\ \mathrm{d}\,P_G(g) = \int_{G/H}\int_H f(gh)\ \mathrm{d}\,P_H(h)\,\mathrm{d}\,P_{G/H}(gH) \tag{1.69}$$

for any integrable f.

If one has a way of uniformly sampling from a subgroup and the quotient space of its cosets, then one can uniformly sample from the group itself.

Example 1.6. *Note $[\mathrm{O}(n) : \mathrm{SO}(n)] = 2$ and $\mathbf{I}_{n-1} \oplus -1, P_{(1\ 2)} \in \mathrm{O}(n)\backslash\mathrm{SO}(n)$. It follows then if $X \sim \mathrm{Haar}(\mathrm{SO}(n))$ and $Y \sim \mathrm{Bernoulli}(\frac{1}{2})$ then $X(\mathbf{I}_{n-1} \oplus (-1)^Y) \sim \mathrm{Haar}(\mathrm{O}(n))$ and $P^Y_{(1\ 2)}X \sim \mathrm{Haar}(\mathrm{O}(n))$.*

A straightforward implication of Theorem 1.14 is the following, which I will make excessive use of in Chapter 3:

Corollary 1.7. *Let $G = \prod_{j=1}^n G_j$ for G_j a compact Polish topological group. Then G is*

a compact Polish topological group. Moreover, if $X_j \sim \text{Haar}(G_j)$*, then* $X = \prod_{j=1}^n X_j \sim \text{Haar}(G)$*.*

Proof. First, note G is a compact Polish space since it is the product of compact Polish spaces. Moreover, since multiplication on G and inverses are performed componentwise, on which each operation is continuous, then these operations are continuous on G so that G is also a topological group. It follows G has a unique Haar probability measure.

For the last statement we will use induction on n. The result is trivial for $n = 1$. Now consider $G = G_1 \times G_2$, and let μ_i denote the Haar probability measure on G_i and μ the Haar probability measure on G. Let $H = G_1 \times 1$, which is a normal closed subgroup of G that is naturally isomorphic with G_1. We see further $G/H \cong G_2$ via the commutative diagram of short exact sequences:

$$\begin{array}{ccccccccc} 1 & \longrightarrow & G_1 & \longrightarrow & G & \xrightarrow{\pi_2} & G_2 & \longrightarrow & 1 \\ & & \downarrow\cong & & \downarrow\cong & & \downarrow\cong & & \\ 1 & \longrightarrow & H & \longrightarrow & G & \xrightarrow{\pi} & G/H & \longrightarrow & 1 \end{array}$$

where π_2 is the projection map onto the second coordinate, which is an open map and is hence continuous. Since each of the associated maps are continuous, then the induced group isomorphisms are homeomorphisms. For $\varphi_1 : G_1 \to H$ and $\varphi_2 : G_2 \to G/H$ the isomorphisms $x \mapsto (x, 1)$ and $y \mapsto (1, y)H$, then the push forward measures $\widetilde{\mu}_i = \mu_i \circ \varphi_i^{-1}$ are the Haar measures for H and G/H. By the Subgroup Algorithm (Theorem 1.14), we have $\mu = \widetilde{\mu}_2 * \widetilde{\mu}_1$. Using (1.69), for measurable $A \subset G$, we see

$$\begin{aligned} \mu(A) &= \int_G \mathbb{1}_A(x, y) \ \mathrm{d}\,\mu(x, y) \\ &= \int_{G/H} \int_H \mathbb{1}_A((x, y)(h, 1)) \ \mathrm{d}\,\widetilde{\mu}_1(h, 1)\, \mathrm{d}\,\widetilde{\mu}_2((x, y)H) \\ &= \int_{G/H} \int_H \mathbb{1}_A(xh, y) \ \mathrm{d}\,\mu_1(\varphi_1^{-1}(h, 1))\, \mathrm{d}\,\mu_2(\varphi_2^{-1}((x, y)H) \end{aligned}$$

$$
\begin{aligned}
&= \int_{G_2} \int_{G_1} \mathbb{1}_A(xh, y) \ \mathrm{d}\,\mu_1(h)\,\mathrm{d}\,\mu_2(y) \\
&= \int_{G_2} \left(\int_{G_1} \mathbb{1}_{x(A_y)}(h) \ \mathrm{d}\,\mu_1(h) \right) \mathrm{d}\,\mu_2(y) \\
&= \int_{G_2} \mu_1(x(A_y))\,\mathrm{d}\,\mu_2(y) \\
&= \int_{G_2} \mu_1(A_y)\,\mathrm{d}\,\mu_2(y) \\
&= (\mu_1 \times \mu_2)(A),
\end{aligned}
$$

where $A_y = \{x \in G_1 : (x, y) \in A\}$ is the sector of A; we used the invariance of μ_1 in the penultimate line and Fubini's theorem for the last line. It follows $\mu = \mu_1 \times \mu_2$. For the general case, where we assume the result holds for $n-1$, we can reduce to the $n = 2$ case by writing $G = G_1' \times G_n$ for $G_1' = \prod_{j=1}^{n-1} G_j$, which then has Haar measure $\prod_{j=1}^{n-1} \mu_j$ by the inductive hypothesis. □

Corollary 1.8. *If $G = \bigotimes_{j=1}^n G_j$ for G_j a compact topological subgroup of $\mathrm{U}(m_j)$ for $m_j \geq 2$. Then G is a compact topological subgroup of $\mathrm{U}(N)$ for $N = \prod_{j=1}^n m_j$. Moreover, if $X_j \sim \mathrm{Haar}(G_j)$, then $X = \bigotimes_{j=1}^n X_j \sim \mathrm{Haar}(G)$.*

Proof. Note $\varphi : \prod_{j=1}^n G_j \mapsto G$ given by $\varphi(\prod_{j=1}^n X_j) = \bigotimes_{j=1}^n X_j$ is a surjective continuous group homomorphism using the mixed-product property. Note this map is not necessarily injective since $\mathbf{I}_N = (-\mathbf{I}_{m_1}) \otimes (-\mathbf{I}_{N/m_1})$, while the kernel is a compact subgroup of $\prod_{j=1}^n G_j$. We do have, though, G is isomorphic to $(\prod_{j=1}^n G_j)/K$ for compact $K = \ker \varphi$. By Corollary 1.7, $\mu = \prod_{j=1}^n \mu_j$ is the Haar probability measure on $\prod_{j=1}^n G_j$ for μ_j the Haar probability measure on G_j. The result then follows by (1.66) since $\mu_G = \frac{1}{C}\mu \circ \varphi^{-1}$ is the Haar measure on G for C a normalizing constant. □

1.7 Random Matrix Theory

Random Matrix Theory (RMT) is the study of matrices whose entries are random variables. Of particular interest is the study of the spectral and numerical properties of these matrices. A focus of a lot of RMT research relates to determining universality properties for different families of random matrices. In RMT, famous universality results similarly follow limiting distributions determined by particular Gaussian ensembles.

1.7.1 Common distributions

We say X is sampled from the **Ginibre ensemble**, written $X \sim \mathrm{Gin}(n,m)$ or $X \sim \mathrm{Gin}(n)$ if $n = m$, if X is an order $n \times m$ random matrix with X_{ij} iid $N(0,1)$, and write $X \sim \mathrm{Gin}_{\mathbb{C}}(n,m)$ or $\mathrm{Gin}_{\mathbb{C}}(n)$ if $n = m$ when X_{ij} are iid $N_{\mathbb{C}}(0,1)$. Note if $X \sim \mathrm{Gin}(n,m)$ and $U \in \mathrm{O}(n)$, then $UX \sim \mathrm{Gin}(n,m)$: First, suppose $X \sim \mathrm{Gin}(n,1)$. Since the components are iid $N(0,1)$, then their joint density is multiplicative in terms of the density of $X_i \sim N(0,1)$. That is, for measurable $A \subset \mathbb{R}^n$

$$\mathbb{P}(X \in A) = \frac{1}{(2\pi)^{n/2}} \int_A e^{-\frac{1}{2}\|\mathbf{x}\|_2^2} \prod_{j=1}^{n} \mathrm{d}\, x_j. \tag{1.70}$$

Since $\|U\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ for any $U \in \mathrm{O}(n)$, it follows $\mathbb{P}(UX \in A) = \mathbb{P}(X \in A)$ for any $U \in \mathrm{O}(n)$. It follows if $X \sim \mathrm{Gin}(n,m)$ then $UX \sim \mathrm{Gin}(n,m)$ for any $U \in \mathrm{O}(n)$. Similarly, $UX \sim \mathrm{Gin}_{\mathbb{C}}(n,m)$ if $X \sim \mathrm{Gin}_{\mathbb{C}}(n,m)$ and $U \in \mathrm{U}(n)$. Also, if $X \sim \mathrm{Gin}(n,m)$, then $X^T \sim \mathrm{Gin}(m,n)$. It follows

Lemma 1.12. *If $X \sim \mathrm{Gin}(n)$ and $U \in \mathrm{O}(n)$, then $UXU^T \sim \mathrm{Gin}(n)$.*

Similarly, $UXU^* \sim \mathrm{Gin}_{\mathbb{C}}(n)$ if $U \in \mathrm{U}(n)$ and $X \sim \mathrm{Gin}_{\mathbb{C}}(n)$.

We say X is sampled from the **Gaussian orthogonal ensemble**, written $X \sim \mathrm{GOE}(n)$,

if $X = \frac{1}{\sqrt{2}}(G + G^T)$ for $G \sim \text{Gin}(n)$. Note if $X \sim \text{GOE}(n)$, then $X_{ii} \sim N(0, 2)$ and $X_{ij} \sim N(0, 1)$ for $i > j$. Similarly, we say X is sampled from the **Gaussian unitary ensemble**, written $X \sim \text{GUE}(n)$ if $X = \frac{1}{\sqrt{2}}(G + G^*)$ for $G \sim \text{Gin}_{\mathbb{C}}(n)$. By the invariance of $\text{Gin}(n)$ by conjugation of orthogonal matrices by Lemma 1.12, it follows $UXU^T \sim \text{GOE}(n)$ if $X \sim \text{GOE}(n)$ and $U \in \text{O}(n)$. Similarly, $UXU^* \sim \text{GUE}(n)$ if $X \sim \text{GUE}(n)$ and $U \in \text{U}(n)$.

We say X is a **(real) Wigner matrix** if X is (symmetric) Hermitian and its off-diagonal entries are iid with mean 0 and variance 1, whose diagonal is iid and independent of its off-diagonal entries and has a finite first moment. For example, $\text{GOE}(n)$ and $\text{GUE}(n)$ are real and complex Wigner ensembles.

Since $\text{SO}(n), \text{O}(n), \text{SU}(n)$ and $\text{U}(n)$ are compact Polish topological spaces (in fact, they are Lie groups), then one can define the Haar measure on each group. A natural question then is how to actually sample a matrix from this distribution. Using the invariance of $\text{Gin}(n)$ under left multiplication by $\text{O}(n)$, Stewart provided the following construction:

Theorem 1.15 ([33])**.** *Let $G \sim \text{Gin}(n)$ and $G = QR$ is the QR factorization of G, where R has positive diagonal. Then $Q \sim \text{Haar}(\text{O}(n))$.*

This result will be explored more extensively in Section 5.2.

1.7.2 Universality in RMT

A particular measure of interest associated with a given matrix is the **empirical spectral distribution (ESD)**, defined by

$$\mu_A = \frac{1}{n} \sum_{j=1}^{n} \delta_{\lambda_j(A)} \tag{1.71}$$

for $\lambda_j(A)$ denoting the eigenvalues of an order n Hermitian matrix A. Note if A is a random Hermitian matrix, then μ_A is a random measure on $\mathbb{R}$, while if A is a random unitary matrix, then μ_A is a random measure on $\mathbb{T}$. Using the Riesz representation theorem, we can define the **density of states** of μ_A as

$$\int f \ \mathrm{d}\,\mathbb{E}\mu_A := \mathbb{E}\int f \ \mathrm{d}\,\mu_A \tag{1.72}$$

for $f \in C_c(\mathbb{R})$, the continuous functions on $\mathbb{R}$ with compact support. Another type of convergence of particular interest in RMT is the following: For A_n a sequence of random matrices, we say μ_{A_n} **converges in expectation** to a probability measure $\mathbb{P}$ if $\mathbb{E}\mu_{A_n}$ **converges in the vague topology** to $\mathbb{P}$, that is,

$$\lim_{n\to\infty} \mathbb{E}\int f \ \mathrm{d}\,\mu_{A_n} = \int f \ \mathrm{d}\,\mathbb{P} \tag{1.73}$$

for all $f \in C_c(\mathbb{R})$. This is a weaker notion of convergences since convergence in probability (and so also convergence in expectation or convergence in r-mean) implies convergence in expectation. Note if a sequence of random matrices with increasing order have ESDs that converge in expectation to a fixed probability measure, then taking the average of a large number of independent samples would yield a picture close to the limiting distribution by the LLN. If the ESDs converge in probability or almost surely to $\mathbb{P}$, then for a large enough order matrix, one sample would be sufficient to generate an approximate picture of the limiting distribution.

One of the most famous results in RMT is the following, due to Wigner. Define the **semicircular law** on $\mathbb{R}$ by the density

$$\mu_{\mathrm{sc}}(\mathrm{d}\,x) = \frac{1}{2\pi}\sqrt{4-x^2}\mathbb{1}_{[-2,2]}(x)\,\mathrm{d}\,x. \tag{1.74}$$

Theorem 1.16 (Semicircular law)**.** *If X_n is an order n Wigner matrix. Then $\mu_{\frac{1}{\sqrt{n}}X_n}$ con-*

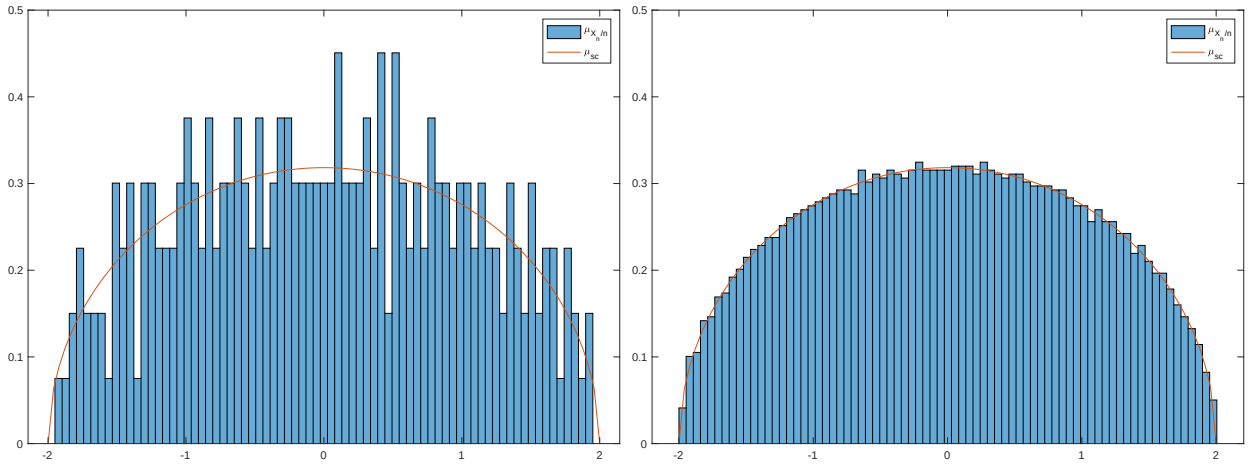


Figure 1.2: μ_{sc} versus $\mu_{\frac{1}{n}X_n}$ for $X_n \sim \mathrm{GOE}(n)$ for $n = 256$ and $n = 4096$

verges almost surely (and hence in probability and in expectation) to μ_{sc} *as* $n \to \infty$.

Figure 1.2 show one sample of $X \sim \mathrm{GOE}(256)$ is sufficient to get a decent approximation of μ_{sc}, while one gets a very close approximation with $X \sim \mathrm{GOE}(4096)$.

Of particular interest are the spacings between successive eigenvalues from a Hermitian ensemble, with a distinction made between edge statistics and bulk statistics of the spectral picture. Focusing on the bulk statistics for $X \sim \mathrm{GUE}(n)$, Wigner showed that the successive spacings normalized to 1 (approximately) follow the Wigner surmise ($\beta = 2$) distribution,

$$\mu_{\mathrm{WS}}(\mathrm{d}\, x) \approx \frac{32}{\pi^2} x^2 e^{-\frac{4}{\pi}x^2} \ \mathrm{d}\, x. \tag{1.75}$$

(1.75) is exact when $n = 2$. A similar universality result shows this holds for X sampled from a Wigner ensemble.

The origin of RMT less than a century ago arose from a need for statistical modeling of physical systems. Eugene Wigner introduced a lot of results in early RMT through the lens of nuclear physics. Computations relating to the level spacings between the nuclei of

heavy atoms led to his conjecture that became the Wigner surmise. Other results that have shown empirical Wigner surmise spacing statistics include prime number gaps, the spacings between parallel parked cars, gaps in arrival times of subway trains in New York City or buses in Cuernavaca, and even the spacings between rods in chicken eyes [20, 21, 24, 25, 28]. Appendix B outlines a novel result showing the spacings between ocean waves satisfy the Wigner surmise.

Chapter 2

Butterfly matrices

> I cannot separate the aesthetic pleasure of seeing a butterfly and the scientific pleasure of knowing what it is.
>
> Vladimir Nabokov

I will first introduce a structure for butterfly matrices of order $N = 2^n$. This is the model originally studied in [31, 37]. I will then introduce a general structure of butterfly models for $N = m^n$, along with a butterfly model that uses the prime factorization of N close in relationship to the Cooley-Tukey FFT construction. This chapter will provide a thorough foundation for the remainder of this document.

2.1 Order $N = 2^n$ butterfly matrices

By default this section will assume $N = 2^n$. Note the definitions of the generalized rotation matrices and butterfly factors work for any even sized matrices, but the focus will often return to this strict assumption on N.

2.1.1 Rotation matrices

Definition 2.1. *A **generalized rotation matrix** is an order N matrix for N even of the form*

$$\begin{bmatrix} C & S \\ -S & C \end{bmatrix} \tag{2.1}$$

*where C, S are commuting, symmetric $N/2$ order real matrices such that $C^2 + S^2 = \mathbf{I}$. The **scalar rotation matrices** and **diagonal rotation matrices** are the corresponding generalized rotation matrices formed using, respectively, scalar or diagonal matrices C, S; these are denoted by $\mathcal{R}(N)$ and $\mathcal{R}_d(N)$.*

Note the generalized rotation matrices of order 2 are precisely the (clockwise) rotational matrices, $SO(2)$.

Since $[C, S] = \mathbf{0}$ and C, S are symmetric, then C, S are simultaneously diagonalizable, that is, there exists a $Q \in \mathrm{U}(N)$ such that

$$C = Q\Lambda_1 Q^* \quad \text{and} \quad S = Q\Lambda_2 Q^*. \tag{2.2}$$

Moreover, we necessarily have

$$\Lambda_1 = \bigoplus_{j=1}^{N/2} \cos(\theta_j) \quad \text{and} \quad \Lambda_2 = \bigoplus_{j=1}^{N/2} \sin(\theta_j), \tag{2.3}$$

since the corresponding real eigenvalues (since C, S are symmetric) need to satisfy the Pythagorean identity $\lambda_C^2 + \lambda_S^2 = 1$, meaning (λ_C, λ_S) lies on the unit circle, and hence $(\lambda_C, \lambda_S) = (\cos(\theta), \sin(\theta))$ for some θ.

A lot can be established about generalized rotation matrices from their definitions, such as:

Proposition 2.1. *The generalized rotation matrices belong to* $\mathrm{SO}(N)$*. Moreover, the generalized rotation matrices generated by simultaneously diagonalizable* C, S *matrices (viz., they all mutually commute with one another) form an abelian subgroup of* $\mathrm{SO}(N)$*.*

Additionally, $\mathcal{R}(N)$ *forms an abelian subgroup of* $\mathrm{SO}(N)$*, and* $\mathcal{R}(N) \cong \mathbb{T}$ *for all* $n \geq 1$*, both as a group and topologically.*

Remark 2.1. *It is useful to note that*

$$\mathcal{R}(N) = \mathrm{SO}(2) \otimes \mathbf{I}_{N/2}. \tag{2.4}$$

Recall there exists a perfect shuffle Q *such that* $A \otimes B = Q(A \otimes B)Q^T$*. Hence, we can find a permutation matrix* Q *such that for* $R_N(\theta) = R(\theta) \otimes \mathbf{I}_{N/2} \in \mathcal{R}(N)$ *for* $R(\theta) \in \mathcal{R}(2) = \mathrm{SO}(2)$*, we have*

$$Q(R(\theta) \otimes \mathbf{I}_{N/2})Q^T = \mathbf{I}_{N/2} \otimes R(\theta) = \bigoplus^{N/2} R(\theta). \tag{2.5}$$

If $R(\boldsymbol{\theta}) \in \mathcal{R}_d(N)$*, then using the same* Q *we have*

$$QR(\boldsymbol{\theta})Q^T = \bigoplus_{j=1}^{N/2} R(\theta_{N/2-j+1}) \tag{2.6}$$

where $R(\theta_j) \in \mathcal{R}(2) = \mathrm{SO}(2)$.

Proof of Proposition 2.1. Note the identity matrix can be realized as a generalized rotation matrix formed using $(C, S) = (\mathbf{I}, 0)$. Now we check

$$\begin{bmatrix} C & S \\ -S & C \end{bmatrix} \begin{bmatrix} C & S \\ -S & C \end{bmatrix}^T = \begin{bmatrix} C & S \\ -S & C \end{bmatrix} \begin{bmatrix} C & -S \\ S & C \end{bmatrix} = \begin{bmatrix} C^2 + S^2 & [S, C] \\ [C, S] & C^2 + S^2 \end{bmatrix} = \mathbf{I}. \tag{2.7}$$

Using Schur's complement formula and the fact C, S commute, we have

$$\det \begin{bmatrix} C & S \\ -S & C \end{bmatrix} = \det(C^2 + S^2) = \det(\mathbf{I}) = 1$$

if C is nonsingular (a slightly modified argument is needed if C is singular[1]). These collectively show the generalized rotation matrices belong to $\mathrm{SO}(N)$ and are closed under complements (since the inverse of the generalized rotation matrix formed by (C, S) is the generalized rotation matrix formed by $(C, -S)$).

Next, we see these are closed under multiplication, and multiplication is commutative: we first compute

$$\begin{bmatrix} C_1 & S_1 \\ -S_1 & C_1 \end{bmatrix} \begin{bmatrix} C_2 & S_2 \\ -S_2 & C_2 \end{bmatrix} = \begin{bmatrix} C_1C_2 - S_1S_2 & C_1S_2 + S_1C_2 \\ -(S_1C_2 + C_1S_2) & -S_1S_2 + C_1C_2 \end{bmatrix}. \tag{2.8}$$

Since $[C_1, C_2] = [S_1, S_2] = [C_1, S_2] = [S_1, C_2] = \mathbf{0}$, then it is clear these matrices commute with one another (we can freely interchange the indices above), and also $(C_1C_2 - S_1S_2)^T =$

[1] If C is singular, use the decompositions $C = Q\Lambda_1 Q^*, S = Q\Lambda_2 Q^*$ from (2.3) to reduce to the diagonal C and S case, followed then by using the corresponding row transpositions to switch any zeros in C with the ± 1 in S followed then by a sign change for the row/column moved from $-S$, which will preserve the determinant. The resulting nonsingular C' and S' still satisfy the conditions of C, S, and so form a generalized rotation matrix with the same determinant as the matrix formed by C, S.

$C_2C_1 - S_2S_1 = C_1C_2 - S_1S_2$, $(C_1S_2 + S_1C_2)^T = S_2C_1 + C_2S_1 = C_1S_2 + S_1C_2$, $[C_1C_2 - S_1S_2, C_1S_2 + S_1C_2] = \mathbf{0}$ (since each is a composition of commuting matrices) and

$$(C_1C_2 - S_1S_2)^2 + (C_1S_2 + S_1C_2)^2 = (C_1^2 + S_1^2)(C_2^2 + S_2^2) = \mathbf{I}.$$

In particular, we have the diagonal rotation matrices form a subgroup of $\mathrm{SO}(N)$, with a smaller subgroup formed by the scalar rotation matrices, $\mathcal{R}(N)$, which follows directly from Remark 2.1 and the mixed-product property.

To see $\mathcal{R}(N) \cong \mathbb{T}$, it suffices to show the result for $N = 2$. This follows directly from the relationship $R_N(\theta) \cong \mathrm{SO}(2)$ via the map $B(\theta) \otimes \mathbf{I}_{N/2} \mapsto B(\theta)$. Moreover, this is a bijective map, which is a group homomorphism by the mixed-product property and a homeomorphism since it is equivalently a projection onto the middle 2×2 block of $R_N(\theta)$ and is hence an open map. Using both (2.7) and (2.8), we have

$$R_2(\theta)^{-1} = R_2(-\theta) \quad \text{and} \quad R_2(\theta)R_2(\varphi) = R_2(\theta + \varphi), \tag{2.9}$$

which then completes the task at hand.

Next, since multiplication in $\mathcal{R}(2) = \mathrm{SO}(2)$ is equivalent to addition of angles modulo 2π, then $\mathcal{R}(2) = \mathrm{SO}(2) \cong \mathbb{T}$. In particular, the map $B(\theta) \mapsto e^{i\theta}$ is then a group homomorphism, which is clearly an isomorphism. Moreover, this map is continuous: it is the composition of the projection map sending $B(\theta)$ onto its first column in $\mathbb{R}^2$ followed by the isometry $\mathbb{R}^2 \to \mathbb{C}$. Again, since projections are open maps, then this map is also a homeomorphism. It follows $\mathcal{R}(N)$ and $\mathbb{T}$ are isomorphic and homeomorphic. □

2.1.2 Order $N = 2^n$ butterfly matrices

Definition 2.2. *A **butterfly matrix**, denoted collectively as* $\mathrm{B}(N)$*, is an iteratively defined matrix of order* $N = 2^n$*, where we start with* $\{1\}$ *if* $N = 1$*, of the following form:*

$$\begin{bmatrix} CA_1 & SA_2 \\ -SA_1 & CA_2 \end{bmatrix} = \begin{bmatrix} C & S \\ -S & C \end{bmatrix} \begin{bmatrix} A_1 & \\ & A_2 \end{bmatrix}, \tag{2.10}$$

where $A_1, A_2 \in \mathrm{B}(N/2)$*, and* C, S *form a generalized rotation matrix. A butterfly matrix is **simple** if* $A_1 = A_2$ *at each iterative step; otherwise, such a butterfly matrix is **nonsimple**.*

Note a butterfly matrix is the product of a generalized rotation matrix and a block diagonal butterfly matrix. A simple walkthrough of how to build a butterfly matrix is established in the next two examples.

Example 2.1. *The order 2 butterfly matrices (which are necessarily simple) are comprised precisely by the (clockwise) rotation matrices,* $\mathrm{SO}(2)$*,*

$$A = B(\theta) = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}.$$

Example 2.2. *Using* $A = B(\theta)$ *from the prior example and now* $(C_2, S_2) = (cos\varphi, \sin\varphi)\mathbf{I}_2$*, we can form the order 4 simple butterfly matrix*

$$\begin{bmatrix} C_2A & S_2A \\ -S_2A & C_2A \end{bmatrix} = \left[\begin{array}{cc|cc} \cos\varphi\cos\theta & \cos\varphi\sin\theta & \sin\varphi\cos\theta & \sin\varphi\sin\theta \\ -\cos\varphi\sin\theta & \cos\varphi\cos\theta & -\sin\varphi\sin\theta & \sin\varphi\cos\theta \\ \hline -\sin\varphi\cos\theta & -\sin\varphi\sin\theta & \cos\varphi\cos\theta & \cos\varphi\sin\theta \\ \sin\varphi\sin\theta & -\sin\varphi\cos\theta & -\cos\varphi\sin\theta & \cos\varphi\cos\theta \end{array}\right].$$

Example 2.3. *For some concrete order 4 examples, here are two simple butterfly matrices:*

$$\left[\begin{array}{rr|rr} -0.65 & -0.15 & -0.72 & -0.17 \\ 0.15 & -0.65 & 0.17 & -0.72 \\ \hline 0.72 & 0.17 & -0.65 & -0.15 \\ -0.17 & 0.72 & 0.15 & -0.65 \end{array}\right] \quad \textit{and} \quad \left[\begin{array}{rr|rr} 0.32 & -0.19 & 0.80 & -0.47 \\ 0.19 & 0.32 & 0.47 & 0.80 \\ \hline -0.80 & 0.47 & 0.32 & -0.19 \\ -0.47 & -0.80 & 0.19 & 0.32 \end{array}\right];$$

and two nonsimple butterfly matrix:

$$\left[\begin{array}{rr|rr} -0.51 & 0.86 & 0.06 & 0.05 \\ -0.86 & -0.51 & -0.05 & 0.06 \\ \hline 0.04 & -0.07 & 0.80 & 0.59 \\ 0.07 & 0.04 & -0.59 & 0.80 \end{array}\right] \quad \textit{and} \quad \left[\begin{array}{rr|rr} -0.22 & 0.97 & -0.13 & 0.03 \\ -0.97 & -0.22 & -0.03 & -0.13 \\ \hline 0.03 & -0.13 & -0.97 & 0.19 \\ 0.13 & 0.03 & -0.19 & -0.97 \end{array}\right].$$

.

Since the generalized rotation matrices are the main building blocks of butterfly matrices, we immediate get the following result:

Proposition 2.2. $\mathrm{B}(N) \subset \mathrm{SO}(N)$.

Proof. Using induction on n, the result is trivial for $n = 0$. Now assume $\mathrm{B}(N/2) \subset \mathrm{SO}(N/2)$. We have then the block diagonal matrices with blocks from $\mathrm{B}(N/2)$ are also in $\mathrm{SO}(N)$, and so $\mathrm{B}(N) \subset \mathrm{SO}(N)$ since this consists precisely of products of two special orthogonal matrices, using also Proposition 2.1. □

Remark 2.2. *Note the formation of a butterfly matrix $B = RD$ for R a generalized rotation matrix and D a block diagonal butterfly matrix is unique only when $n = 1$ (there is only one possible block diagonal butterfly matrix of order 2, viz.,* $\mathbf{I}$*), since for $n \geq 2$, $-R, -D \in \mathrm{B}(N)$ and so*

$$B = RD = (-R)(-D).$$

This shows the map $\mathcal{R}(N) \times \bigoplus^{2} \mathrm{B}(N/2) \to \mathrm{B}(N)$ *is a 2:1 map when* $N \geq 4$.

Definition 2.3. *The* ***diagonal butterfly matrices*** *and* ***scalar butterfly matrices*** *are the butterfly matrices formed iteratively using diagonal and scalar rotation matrices, respectively. Let* $\mathrm{B}_s(N)$ *denote the simple scalar butterfly matrices. For* $B \in \mathrm{B}_s(N)$*, we will write* $B = B(A, \theta)$ *for* $A \in B_s(N/2)$ *if* B *is of the form*

$$\begin{bmatrix} \cos\theta A & \sin\theta A \\ -\sin\theta A & \cos\theta A \end{bmatrix} = \begin{bmatrix} \cos\theta \mathbf{I} & \sin\theta \mathbf{I} \\ -\sin\theta \mathbf{I} & \cos\theta \mathbf{I} \end{bmatrix} \begin{bmatrix} A & \\ & A \end{bmatrix} = \begin{bmatrix} A & \\ & A \end{bmatrix} \begin{bmatrix} \cos\theta \mathbf{I} & \sin\theta \mathbf{I} \\ -\sin\theta \mathbf{I} & \cos\theta \mathbf{I} \end{bmatrix}. \tag{2.11}$$

Unless otherwise stated, $\mathrm{B}(N)$ *denotes the nonsimple scalar butterfly matrices.*

Remark 2.3. *For* $\mathrm{B}_s(N)$ *then* (2.10) *can be written as*

$$B(\boldsymbol{\theta}) = \bigotimes_{j=1}^{n} B(\theta_{n-j+1}) = B(\theta_n) \otimes \cdots \otimes B(\theta_1) \tag{2.12}$$

for $B(\theta_j) \in \mathrm{SO}(2)$*. This form will be used frequently throughout this document. Additionally,* $\mathrm{B}(N)$ *(the nonsimple scalar butterfly matrices) are of the form*

$$(B(\theta) \otimes \mathbf{I}_{N/2})(B(\boldsymbol{\theta}) \oplus B(\boldsymbol{\varphi})) \tag{2.13}$$

for $B(\theta) \in \mathrm{B}(2)$ *and* $B(\boldsymbol{\theta}), B(\boldsymbol{\varphi}) \in \mathrm{B}(N/2)$.

2.1.3 Matrix-vector multiplication

Parker's original interest in using butterfly matrices for reducing complexity for common computations relied on the fact products using these matrices can be carried out very efficiently. This is outlined explicitly in this section. Algorithms 3 and 4 show how one compute

the product of by a butterfly matrix with a matrix $\mathbf{V} \in \mathbb{R}^{N\times M}$. Both algorithms show how to implement this multiplication by first splitting the input matrix into two blocks of equal size (Step 2), implementing the multiplication on each subblock (Steps 3 and 4 in Algorithm 3; Step 3 in Algorithm 4), and then stitching the outputs together to form the final output product (Step 5 in Algorithm 3; Step 4 in Algorithm 4).

Note one does not need to store the matrix itself to compute the product, but one would need storage of the input parameters to generate the matrix if one needs to undo this multiplication. See [2, 31] for discussions relating to data storage. I will not explore data storage further in this document.

Algorithm 3 Butterfly matrix-vector multiplication

1: **procedure** BMult($\mathbf{V}$)
2: $\begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix} = \mathbf{V}$ for $\mathbf{V}_1, \mathbf{V}_2 \in \mathbb{R}^{N/2\times M}$
3: $\mathbf{V}_1 = \text{BMult}(\mathbf{V}_1)$
4: $\mathbf{V}_2 = \text{BMult}(\mathbf{V}_2)$
5: $\mathbf{V} = \begin{bmatrix} C\mathbf{V}_1 + S\mathbf{V}_2\\ -S\mathbf{V}_1 + C\mathbf{V}_2\end{bmatrix}$
6: **return** $\mathbf{V}$

Algorithm 4 Simple butterfly matrix-vector multiplication

1: **procedure** SBMult($\mathbf{V}$)
2: $\begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix} = \mathbf{V}$ for $\mathbf{V}_1, \mathbf{V}_2 \in \mathbb{R}^{N/2\times M}$
3: $\begin{bmatrix}\mathbf{V}_1 & \mathbf{V}_2\end{bmatrix} = \text{SBMult}(\begin{bmatrix}\mathbf{V}_1 & \mathbf{V}_2\end{bmatrix})$
4: $\mathbf{V} = \begin{bmatrix} C\mathbf{V}_1 + S\mathbf{V}_2\\ -S\mathbf{V}_1 + C\mathbf{V}_2\end{bmatrix}$
5: **return** $\mathbf{V}$

We can determine the complexity in Algorithm 3: considering only the case when C and S are diagonal, then for $\mathbf{V}$ a vector (i.e., $M = 1$) we have $C\mathbf{V}_1 + S\mathbf{V}_2$ takes $2 \cdot N/2$ multiplications and $N/2$ additions for $3N/2$ total flops, so that $\mathbf{V} = \begin{bmatrix} C\mathbf{V}_1 + S\mathbf{V}_2\\ -S\mathbf{V}_1 + C\mathbf{V}_2\end{bmatrix}$ takes $3N$ flops. Each recursive step in the middle then accounts for $3 \cdot N/2$ flops plus two more smaller recursive

steps. Hence, since there are $n = \log_2 N$ total recursive steps, we have

$$3N + 2(3N/2 + 2(3N/4 + \cdots)) = 3N + 3N + 3N + \cdots = 3Nn$$

total flops. (Equivalently, one can see complexity satisfies the recurrence $\alpha_n = 3N + 2\alpha_{n-1}$, so that $\alpha_n = 3Nn$ can be verified through induction.) The calculation is identical for Algorithm 4. Hence, both of these methods are substantially faster than the (dense) matrix-vector multiplication, which takes $N(2N-1) = 2N^2 + O(N)$ flops (see Example 1.2).

Note to attain this $O(Nn)$ complexity order, it is necessary for C, S to be matrices such that matrix-vector multiplication takes $O(N)$ flops. Above we considered only the scalar and diagonal cases, which satisfy this property. For simplicity, we will focus on these models.

2.1.4 Butterfly block factors

Focusing on a scalar butterfly matrix B, we see inductively from (2.10) that B can be written of the form

$$B = D_0 D_1 \cdots D_{n-1} \tag{2.14}$$

where D_j is a block diagonal matrix with 2^j blocks of order 2^{n-j} and each block is a scalar rotation matrix in $\mathcal{R}(2^{n-j})$. If $B \in \mathrm{B}_s(n)$ then D_j would then have identical blocks. These can be used to better establish the topological picture of butterfly matrices.

Definition 2.4. *For N and j such that $2j \mid N$, the **butterfly block factors**, denoted by $\mathcal{D}^j(N) = \bigoplus^j \mathcal{R}(N/j)$, are the set of order N block diagonal matrices of the form*

$$\bigoplus_{\ell=1}^{j} A_\ell, \tag{2.15}$$

*such that $A_k \in \mathcal{R}(N/j)$ for each k. The **simple butterfly block factors**, denoted by $\mathcal{D}_s^j(N)$, comprise the subset of $\mathcal{D}^j(N)$ such that each block is identical.*

*The **(simple) diagonal butterfly block factors**, denoted by $\mathcal{D}_d^j(N)$ ($\mathcal{D}_{ds}^j(N)$), are comprised of the block diagonal matrices such that $A_k \in \mathcal{R}_d(N/j)$ for each k (with $A_k = A_1$ for all k).*

Note the condition $2j \mid N$ ensures that N/j is an even integer, and hence $R_{N/j}$ is defined. In particular, $\mathcal{D}^1(N) = \mathcal{R}(N)$ when N is even. Also, $D_j \in \mathcal{D}^{2^j}(N)$ in (2.14).

Remark 2.4. *Building off of Remark 2.3, we see*

$$\mathcal{D}^j(N) = \bigoplus^{j} (\mathrm{SO}(2) \otimes \mathbf{I}_{N/2j}) \tag{2.16}$$

and

$$\mathcal{D}_s^j(N) = \mathbf{I}_j \otimes \mathrm{SO}(2) \otimes \mathbf{I}_{N/2j}. \tag{2.17}$$

Proposition 2.3. *$\mathcal{D}^j(N)$ and $\mathcal{D}_s^j(N)$ are compact abelian subgroups of $\mathrm{SO}(N)$, where for $B(\boldsymbol{\theta}), B(\boldsymbol{\varphi}) \in \mathcal{D}^j(N)$ ($\mathcal{D}_s^j(N)$) then $B(\boldsymbol{\theta})B(\boldsymbol{\varphi}) = B(\boldsymbol{\theta} + \boldsymbol{\varphi}) \in \mathcal{D}^j(N)$ ($\mathcal{D}_s^j(N)$). As groups and topologically $\mathcal{D}^j(N) \cong \mathbb{T}^j$ and $\mathcal{D}_s^j(N) \cong \mathbb{T}$. If $2j, 2k \mid N$ with $j \neq k$, then*

$$\mathcal{D}^j(N) \cap \mathcal{D}^k(N) = \left\{ \bigoplus_{\ell=1}^{\gcd(j,k)} A_\ell : A_\ell = \pm \mathbf{I}_{N/\gcd(j,k)} \right\}, \tag{2.18}$$

while

$$\mathcal{D}_s^j(N) \cap \mathcal{D}^k(N) = \mathcal{D}^j(N) \cap \mathcal{D}_s^k(N) = \mathcal{D}_s^j(N) \cap \mathcal{D}_s^k(N) = \{\pm \mathbf{I}_N\}. \tag{2.19}$$

Proof. The first statement follows directly from Proposition 2.1 (and Corollary 1.8).

To establish (2.18), suppose $A \in \mathcal{D}^j(N) \cap \mathcal{D}^k(N)$, and so there exist $\boldsymbol{\theta}, \boldsymbol{\varphi}$ such that

$$A = \bigoplus_{i=1}^{j} R_{N/j}(\theta_i) = \bigoplus_{i=1}^{k} R_{N/k}(\varphi_i).$$

By equating diagonals of both representations, we have each N/j interval of diagonal entries are equal as these are the blocks in the $\bigoplus^j \mathcal{R}(N/j)$ setting, and similarly each N/k interval of diagonal entries are equal. Considering then overlapping intervals, we have each

$$m := \operatorname{lcm}(N/j, N/k)$$

interval of diagonal entries must be equal.

Since $nm = \gcd(n,m) \cdot \operatorname{lcm}(n,m)$ and

$$\gcd(N/j, N/k) = \min_{a,b\in\mathbb{Z}} \left| a\frac{N}{j} + b\frac{N}{k} \right| = \frac{N}{jk} \min_{a,b\in\mathbb{Z}} |ak + bj| = \frac{N}{jk} \gcd(j,k) = \frac{N}{\operatorname{lcm}(j,k)},$$

then

$$m = \frac{N^2/jk}{\gcd(N/j, N/k)} = \frac{N^2/jk}{N/\operatorname{lcm}(j,k)} = \frac{N}{\gcd(j,k)} \tag{2.20}$$

and hence

$$\frac{N}{m} = \gcd(j,k). \tag{2.21}$$

We have

$$\cos\theta_1 = \cos\theta_2 = \cdots = \cos\theta_b = \cos\varphi_1 = \cdots = \cos\varphi_a$$

where

$$b = \frac{m}{N/j} = \frac{mj}{N} = \frac{j}{\gcd(j,k)} \quad \text{and} \quad a = \frac{k}{\gcd(j,k)}. \tag{2.22}$$

In general,

$$\cos\theta_{db+1} = \cos\theta_{db+s} = \cos\varphi_{da+1} = \cos\varphi_{da+t}, \tag{2.23}$$

where $s = 1, \ldots, b$ and $t = 1, \ldots, a$, and $d = 0, \ldots, \gcd(j,k) - 1$. (Note $j/b = \gcd(j,k) = k/a$.)

Fix d between $0, \ldots, \gcd(j,k) - 1$. Now consider the $(dm+1)^{th}$ column of A, which is

$$\cos\theta_{db+1}\mathbf{e}_{dm+1} - \sin\theta_{db+1}\mathbf{e}_{dm+N/2j+1} = \cos\varphi_{da+1}\mathbf{e}_{dm+1} - \sin\varphi_{da+1}\mathbf{e}_{dm+N/2k+1}. \tag{2.24}$$

Using the linear independence of the $\mathbf{e}_i$, we have $\sin\theta_{db+1} = 0$ so that $\pm 1 = \cos\theta_{db+1}$, and hence

$$0 = \sin\theta_{db+1} = \sin\theta_{db+s} = \sin\varphi_{da+1} = \sin\varphi_{da+t}$$

for $s = 1, \ldots, b$ and $t = 1, \ldots, a$ using now (2.23). In particular, the $m \times m$ block starting at the $dm+1$ diagonal entry is $\pm\mathbf{I}$. Since d was arbitrary, then we have (2.18).

(2.19) then follows directly from (2.18) along with the definition of $\mathcal{D}_s^j(N)$ and $\mathcal{D}_s^k(N)$, viz., that the blocks need be identical. □

Note if also $j \mid k$ so that $\gcd(j,k) = j$, then (2.18) yields

$$\mathcal{D}^j(N) \cap \mathcal{D}^k(N) = \left\{ \bigoplus_{\ell=1}^{j} A_\ell : A_\ell = \pm\mathbf{I}_{N/j} \right\}.$$

In particular, we can establish a straightforward relationship when considering $N = 2^n$.

Corollary 2.1. *If $N = 2^n$, then for $1 \le j_1 < j_2 < \cdots < j_k \le n$, then*

$$\bigcap_{i=1}^{k} \mathcal{D}^{2^{j_i}}(N) = \left\{ \bigoplus_{\ell=1}^{2^{j_1}} A_\ell : A_\ell = \pm \mathbf{I}_{2^{n-j_1}} \right\} \tag{2.25}$$

and

$$\bigcap_{i=1}^{k} \mathcal{D}_s^{2^{j_i}}(N) = \bigcap_{\ell \in \mathcal{S}} \mathcal{D}_s^{2^{j_\ell}}(N) \cap \bigcap_{i=1}^{k} \mathcal{D}^{2^{j_i}}(N) = \{\pm \mathbf{I}\} \tag{2.26}$$

for $\varnothing \neq \mathcal{S} \subset [k]$.

Note the right hand side in (2.26) considers the intersection in (2.25) with at least one $\mathcal{D}^{2^{j_i}}(N)$ replaced by $\mathcal{D}_s^{2^{j_i}}(N)$.

Proof. Using (2.18), we have

$$\mathcal{D}^{2^{j_1}}(N) \cap \mathcal{D}^{2^{j_i}}(N) = \left\{ \bigoplus_{\ell=1}^{2^{j_1}} A_\ell : A_\ell = \pm \mathbf{I}_{2^{n-j_1}} \right\}$$

for $i = 2, \ldots, k$. Hence, (2.25) follows after noting

$$\bigcap_{i=1}^{k} \mathcal{D}^{2^{j_i}}(N) = \bigcap_{i=2}^{k} \mathcal{D}^{2^{j_1}}(N) \cap \mathcal{D}^{2^{j_i}}(N).$$

(2.26) follows immediately from (2.25) by noting the blocks need be identical. □

Using a similar argument, updated appropriately in light of (2.1), one can also show:

Proposition 2.4. *$\mathcal{D}_d^j(N)$ and $\mathcal{D}_{ds}^j(N)$ are compact abelian subgroups of $\mathrm{SO}(N)$, where for $B(\boldsymbol{\theta}), B(\boldsymbol{\varphi}) \in \mathcal{D}_d^j(N)$ $(\mathcal{D}_{ds}^j(N))$ then $B(\boldsymbol{\theta})B(\boldsymbol{\varphi}) = B(\boldsymbol{\theta} + \boldsymbol{\varphi}) \in \mathcal{D}_d^j(N)$ $(\mathcal{D}_{ds}^j(N))$. As groups and topologically, $\mathcal{D}_d^j(N) \cong \mathbb{T}^{N/2}$ and $\mathcal{D}_{ds}^j(N) \cong \mathcal{T}^{N/2j}$.*

2.1.5 Topological properties of butterfly matrices

We then can outline the topological picture of the scalar butterfly matrices.

Remark 2.5. *I will now exclusively use* $\mathrm{B}(N)$ *and* $\mathrm{B}_s(N)$ *to denote the scalar nonsimple and simple butterfly matrices, respectively.*

Proposition 2.5. $\mathrm{B}(N)$ *and* $\mathrm{B}_s(N)$ *are compact spaces in* $\mathrm{SO}(N)$*, which are homeomorphic to quotients of higher dimensional tori* $\mathbb{T}^n$ *and* $\mathbb{T}^{N-1}$*, respectively.*

Proof. By Proposition 2.3, we have $\mathcal{D}^{2^j}(N)$ and $\mathcal{D}_s^{2^j}(N)$ are compact for each $j = 0, \dots, n-1$, so that $\prod_{j=0}^{n-1} \mathcal{D}^j(N)$ and $\prod_{j=0}^{n-1} \mathcal{D}_s^j(N)$ are each compact. The map

$$f_n : \mathrm{SO}(N)^n \to \mathrm{SO}(N)$$

given by $(D_0, D_1, \dots, D_{n-1}) \mapsto D_0 D_1 \cdots D_{n-1}$ is continuous (since $\mathrm{SO}(N)$ is a Lie group then matrix multiplication is continuous). By (2.14), we have

$$f_n\left(\prod_{j=0}^{n-1} \mathcal{D}^{2^j}(N)\right) = \mathrm{B}(N) \quad \text{and} \quad f_n\left(\prod_{j=0}^{n-1} \mathcal{D}_s^{2^j}(N)\right) = \mathrm{B}_s(n).$$

It follows $\mathrm{B}(N)$ and $\mathrm{B}_s(N)$ are each compact since they are the continuous images of compact spaces.

By Proposition 2.3, we have topologically

$$\prod_{j=0}^{n-1} \mathcal{D}^{2^j}(N) \cong \prod_{j=0}^{n-1} \mathbb{T}^{2^j} = \mathbb{T}^{\sum_{j=0}^{n-1} 2^j} = \mathbb{T}^{N-1}$$

and

$$\prod_{j=0}^{n-1} \mathcal{D}_s^{2^j}(N) \cong \prod_{j=0}^{n-1} \mathbb{T} = \mathbb{T}^n.$$

Since

$$g_n : \prod_{j=0}^{n-1} \mathcal{D}^{2^j}(N) \to \mathrm{B}(N) \quad \text{and} \quad h_n : \prod_{j=0}^{n-1} \mathcal{D}_s^{2^j}(N) \to \mathrm{B}_s(N)$$

given both by $(D_0, D_1, \ldots, D_{n-1}) \mapsto D_0 D_1 \cdots D_{n-1}$ are continuous surjective maps, then they are quotient maps. This establishes the last statement. □

These arguments can be updated appropriately to establish the following result for the diagonal butterfly matrices:

Proposition 2.6. *The diagonal simple and nonsimple butterfly matrices are compact spaces in* $\mathrm{SO}(N)$*, which are homeomorphic to quotients of higher dimensional tori* $\mathbb{T}^{N-1}$ *and* $\mathbb{T}^{\frac{1}{2}Nn}$*, respectively.*

2.1.6 Butterfly parameters

I will refer to the number of angles θ that are needed to generate a given butterfly matrix as the needed **butterfly parameters**. Using the iterative structure of butterfly matrices, we can enumerate the needed parameters to construct some particular classes of butterfly matrices. Note that this corresponds directly to the dimension of the butterfly matrices as manifolds.

Proposition 2.7. *The simple and nonsimple scalar butterfly matrices are constructed, respectively, using* $n = \log_2 N$ *and* $N - 1$ *parameters, while the simple and nonsimple diagonal butterfly matrices are constructed, respectively, using* $N - 1$ *and* $\frac{1}{2}Nn$ *parameters.*

Proof. Note the scalar cases follow from Proposition 2.5 and Proposition 2.6. Instead, I can independently verify these results using only (2.10).

Let α_n be the number of parameters needed for an order N such matrix, and note $\alpha_0 = 0$ for all cases above. Then we can verify each of the above counts by solving the corresponding

iterative equations following from (2.10) of the form

$$\alpha_n = 2^{a_n} + 2^{\delta}\alpha_{n-1} \tag{2.27}$$

where $\delta = 0$ in the simple case and $\delta = 1$ in the nonsimple case and $a_n = 0$ in the scalar case and $a_n = n - 1$ in the diagonal case, which is done with straightforward induction. □

Note the prior proposition shows we can write a scalar or diagonal butterfly matrix as $B(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in [0, 2\pi)^{\alpha_n}$ for α_n is the number of needed butterfly parameters. Note if $n \geq 1$ and $B(\boldsymbol{\theta})$ is a scalar or diagonal butterfly matrix, then using (2.27) we can write $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$ so that

$$B(\boldsymbol{\theta}) = R(\boldsymbol{\theta}_3) \begin{bmatrix} B(\boldsymbol{\theta}_1) & \\ & B(\boldsymbol{\theta}_2) \end{bmatrix}, \tag{2.28}$$

for $\boldsymbol{\theta}_1 = (\theta_1, \ldots, \theta_{\alpha_{n-1}}), \boldsymbol{\theta}_2 = (\theta_{\delta\alpha_{n-1}+1}, \ldots, \theta_{(\delta+1)\alpha_{n-1}})$, and $\boldsymbol{\theta}_3 = (\theta_{2^{\delta}\alpha_{n-1}+1}, \ldots, \theta_{\alpha_n})$ where $\delta = 0$ if $B(\boldsymbol{\theta})$ is simple and $\delta = 1$ if $B(\boldsymbol{\theta})$ is nonsimple, and $R(\boldsymbol{\varphi})$ is the order N diagonal or scalar rotational matrix formed by

$$(C, S)(\boldsymbol{\varphi}) = \bigoplus_{i=1}^{N/2} (\cos(\varphi_i), \sin(\varphi_i)),$$

where $\boldsymbol{\varphi} = \varphi \mathbf{1}_{N/2}$ in the scalar case.

2.1.7 Reverse butterfly matrices

Since one desirable attribute of a butterfly matrix is that we can efficiently multiply a vector by it (see Section 2.1.3), we would also like to be able to undo this multiplication. By Proposition 2.2, we know if $B = RD \in \mathrm{B}(N)$ where R is a generalized rotation matrix and

D a block diagonal butterfly matrix, then $B^{-1} = B^T = D^T R^T$. In general, $B^{-1} \notin \mathrm{B}(N)$. For this, we will introduce a new subset of special orthogonal matrices:

Definition 2.5. *A **reverse butterfly matrix**, collectively denoted by* $\mathrm{B}^R(N)$ *for* $N = 2^n$ *with* $\mathrm{B}^R(1) = \{1\}$*, is an iteratively defined matrix of order* N *of the following form:*

$$\begin{bmatrix} CA_1 & SA_1 \\ -SA_2 & CA_2 \end{bmatrix} = \begin{bmatrix} A_1 & \\ & A_1 \end{bmatrix} \begin{bmatrix} C & S \\ -S & C \end{bmatrix}, \tag{2.29}$$

where $A_1, A_2 \in \mathrm{B}^R(N/2)$*, and* C, S *form a generalized rotation matrix. Analogously, define **simple**, **nonsimple**, **diagonal** and **scalar** reverse butterfly matrices.*

By (2.11), the prior conversation shows there is a one-to-one correspondence between $\mathrm{B}(N)$ and $\mathrm{B}^R(N)$ given by the map $B \mapsto B^{-1}$. This immediately yields the following results:

Corollary 2.2. $\mathrm{B}^R(N) \subset \mathrm{SO}(N)$*, while simple and nonsimple scalar (diagonal) reverse butterfly matrices are formed, respectively, using* n *and* $N-1$ *(*$N-1$ *and* $\frac{1}{2}Nn$*) parameters.*

We can also write a scalar or diagonal reverse butterfly matrix of the form

$$B^R(\boldsymbol{\theta}) \quad \text{for } \boldsymbol{\theta} \in [0, 2\pi)^{\alpha_n}$$

where α_n is the number of needed butterfly parameters to generate to same order scalar or diagonal butterfly matrix. For $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$ we then have

$$B^R(\boldsymbol{\theta}) = \begin{bmatrix} B^R(\boldsymbol{\theta}_1) & \\ & B^R(\boldsymbol{\theta}_2) \end{bmatrix} R(\boldsymbol{\theta}_3). \tag{2.30}$$

We can now establish the following:

Proposition 2.8. *If $B(\boldsymbol{\theta})$ is a scalar or diagonal butterfly matrix, then $B(\boldsymbol{\theta})^{-1}$ is the corresponding scalar or diagonal reverse butterfly matrix of the form $B^R(-\boldsymbol{\theta})$.*

Proof. We will use induction on n. The result is trivial for $n = 0$. Now assume the result holds for $\mathrm{B}(N/2)$. Using (2.9), (2.28), (2.30), and the inductive hypothesis, for $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_2)$ appropriately partitioned, we have

$$\begin{aligned} B(\boldsymbol{\theta})^{-1} &= \begin{bmatrix} B(\boldsymbol{\theta}_1)^{-1} & \\ & B(\boldsymbol{\theta}_2)^{-1} \end{bmatrix} R(\boldsymbol{\theta}_3)^{-1} \\ &= \begin{bmatrix} B^R(-\boldsymbol{\theta}_1) & \\ & B^R(-\boldsymbol{\theta}_2) \end{bmatrix} R(-\boldsymbol{\theta}_3) \\ &= B^R(-\boldsymbol{\theta}). \end{aligned}$$

□

2.1.8 Group properties of butterfly matrices

This section will explore particular group properties of scalar butterfly matrices. Appendix C will outline more properties relating the scalar butterfly matrices with respect to the larger group generated by $\mathrm{B}(N)$.

Which butterfly matrices can satisfy a group structure relies on the following result:

Lemma 2.1. *Let A_1, A_2 be order $N/2$ normal matrices. A generalized rotation matrix formed by C, S with nonsingular S commutes with $A_1 \oplus A_2$ if and only if A_1, A_2 both commute with C, S and $A_1 = A_2$.*

In particular, a scalar rotation matrix formed by $(C, S) = (\cos\theta, \sin\theta)\mathbf{I}$ commutes with a block diagonal matrix with block entries $A_1, A_2 \in \mathrm{B}(N/2)$ if and only if $S = \mathbf{0}$ or $S \neq \mathbf{0}$ and

$A_1 = A_2$.

Proof. We first compute the commutator of a generalized rotation matrix and a block diagonal matrix:

$$\left[\begin{bmatrix} C & S \\ -S & C \end{bmatrix}, \begin{bmatrix} A_1 & \\ & A_2 \end{bmatrix}\right] = \begin{bmatrix} [C, A_1] & SA_2 - A_1S \\ A_2S - SA_1 & [C, A_2] \end{bmatrix}. \tag{2.31}$$

If $A_1 = A_2$ and $[C, A_1] = [C, A_2] = [S, A_1] = [S, A_2] = \mathbf{0}$, then (2.31) evaluates to $\mathbf{0}$. Conversely, suppose (2.31) evaluates to $\mathbf{0}$. It follows $[C, A_1] = [C, A_2] = \mathbf{0}$. For $i = 1, 2$, since A_i is normal then A_i is diagonalizable, and so C and A_i are both simultaneously diagonalizable, as then are also S and A_i, so that also $[S, A_1] = [S, A_2] = \mathbf{0}$. It follows $\mathbf{0} = SA_1 - A_2S = S(A_1 - A_2)$. Since S is nonsingular, then $A_1 = A_2$. The scalar case follows immediately from this result, where we note now the scalar rotation matrix with $S = \mathbf{0}$ is necessarily $\pm\mathbf{I}$ and hence commutes with any matrix. □

Note the condition $[C, A_1] = [C, A_2] = \mathbf{0}$ is always satisfied by the scalar rotation matrices. It follows:

Proposition 2.9. $\mathrm{B}_s(N)$ *is a compact abelian subgroup of* $\mathrm{SO}(N)$.

Proof. We already established $\mathrm{B}_s(N)$ is a compact space in $\mathrm{SO}(N)$ in Proposition 2.5, so it remains to establish $\mathrm{B}_s(N)$ is an abelian group.

We will again use induction on n. The result for $n = 1$ follows from Proposition 2.1. Now suppose $\mathrm{B}_s(N/2)$ is an abelian subgroup of $\mathrm{SO}(N/2)$. Let $\mathcal{D}$ be the collection of block diagonal matrices whose two blocks are identically taken from $\mathrm{B}_s(N/2)$, and $\mathcal{R}(N)$ the abelian group of scalar rotation matrices of order N (using Proposition 2.1). By the inductive hypothesis, $\mathcal{D} \cong \mathrm{B}_s(N/2)$ is an abelian group, as is then $\mathcal{R}(N) \times \mathcal{D}$. Since elements from $\mathcal{R}(N)$ and $\mathcal{D}$ commute by Lemma 2.1, then $\mathcal{R}(N)\mathcal{D}$ is a subgroup of $\mathrm{SO}(N)$, and

$\mathrm{B}_s(N) = \mathcal{R}(N)\mathcal{D}$ by construction. Moreover, the map $\mathcal{R}(N) \times \mathcal{D} \to \mathrm{B}_s(N)$ given by $(R, D) \mapsto RD$ is a surjective group homomorphism, using also (2.10). It follows $\mathrm{B}_s(N)$ is isomorphic to a quotient of an abelian group and is hence abelian. $\square$

Alternatively, one can inductively show $h_n : \prod_{j=0}^{n-1} \mathcal{D}_s^{2^j}(N) \to \mathrm{B}_s(N)$ given by $(D_0, \dots, D_{n-1}) \mapsto D_0 \cdots D_{n-1}$ from Proposition 2.5 is a surjective group homomorphism, where $\prod_{j=0}^{n-1} \mathcal{D}_s^{2^j}(N) \cong \mathbb{T}^n$ is abelian, with kernel

$$\ker h_n = \{((-\mathbf{I})^{a_1}, (-\mathbf{I})^{a_2}, \dots, (-\mathbf{I})^{a_n}) : \mathbf{a} \in \{0,1\}^n, \sum_{i=1}^{n} a_i \in 2\mathbb{Z}\}, \tag{2.32}$$

using (2.26). In particular, we see $\ker h_n$ is abelian with (group) order 2, and $|\ker h_n| = N/2$ (the number of ways of choosing an even number of indices among the n options is $2^{n-1} = N/2$), so that $\ker h_n \cong (\mathbb{Z}/2\mathbb{Z})^{n-1}$. It follows

$$\mathrm{B}_s(N) \cong \mathbb{T}^n / K \tag{2.33}$$

for

$$K := \{((-1)^{a_1}, \dots, (-1)^{a_n}) \in \mathbb{T}^n : \mathbf{a} \in \{0,1\}^n, \sum_{i=1}^{n} a_i \in 2\mathbb{Z}\} \cong (\mathbb{Z}/2\mathbb{Z})^{n-1}.$$

The following is immediate from this proof:

Corollary 2.3. *The scalar rotation matrices are a compact subgroup of* $\mathrm{B}_s(N)$.

Also, we can extend Proposition 2.8 in the case of simple scalar butterfly matrices:

Corollary 2.4. *For* $B(\boldsymbol{\theta}), B(\boldsymbol{\varphi}) \in \mathrm{B}_s(N)$, $B(\boldsymbol{\theta})^{-1} = B(-\boldsymbol{\theta}) \in \mathrm{B}_s(N)$ *and* $B(\boldsymbol{\theta})B(\boldsymbol{\varphi}) = B(\boldsymbol{\theta} + \boldsymbol{\varphi}) \in \mathrm{B}_s(N)$.

Proof. It remains only to show $B(\boldsymbol{\theta})B(\boldsymbol{\varphi}) = B(\boldsymbol{\theta} + \boldsymbol{\varphi})$. The result for $n = 1$ follows from Proposition 2.1, so assume the result holds for $\mathrm{B}_s(N/2)$. Write $\boldsymbol{\theta} = (\boldsymbol{\theta}', \theta_n)$ and $\boldsymbol{\varphi} = (\boldsymbol{\varphi}', \varphi_n)$.

Using Lemma 2.1, Proposition 2.1, (2.28), (2.9), and the inductive hypothesis, we have

$$\begin{aligned}
B(\boldsymbol{\theta})B(\boldsymbol{\varphi}) &= R(\theta_n)\begin{bmatrix} B(\boldsymbol{\theta}') & \\ & B(\boldsymbol{\theta}') \end{bmatrix} R(\varphi_n)\begin{bmatrix} B(\boldsymbol{\varphi}') & \\ & B(\boldsymbol{\varphi}') \end{bmatrix} \\
&= R(\theta_n)R(\varphi_n)\begin{bmatrix} B(\boldsymbol{\theta}')B(\boldsymbol{\varphi}') & \\ & B(\boldsymbol{\theta}')B(\boldsymbol{\varphi}') \end{bmatrix} \\
&= R(\theta_n+\varphi_n)\begin{bmatrix} B(\boldsymbol{\theta}'+\boldsymbol{\varphi}') & \\ & B(\boldsymbol{\theta}'+\boldsymbol{\varphi}') \end{bmatrix} \\
&= B(\boldsymbol{\theta}+\boldsymbol{\varphi}).
\end{aligned}$$

□

In Section C, we show the nonsimple butterfly matrices are not closed under multiplication, that is, $\mathrm{B}(N)^2 \not\subset \mathrm{B}(N)$ (cf., Proposition C.1). This can potentially be exploited later to use butterfly matrices to approximate $\mathrm{SO}(N)$, and hence $\mathrm{O}(N)$ (see Section 5.2).

2.1.9 Connectedness

A natural question with butterfly matrices is far can a butterfly matrix change through perturbations of the input butterfly parameters. This can be answered explicitly by computing bounds for $\|B(\boldsymbol{\theta}) - B(\boldsymbol{\theta}+\boldsymbol{\varepsilon})\|_F$, which show the map $\boldsymbol{\theta} \mapsto B(\boldsymbol{\theta})$ is Lipshitz continuous. In particular, this shows $\mathrm{B}_s(N)$ and $\mathrm{B}(N)$ are connected. Only scalar butterfly matrices are considered in this section. Analogous results can be attained for diagonal butterfly matrices.

First, recall $\|\cdot\|_F$ is invariant under unitary transformations. Using also the triangle in-

equality, we have for $U_i, V_i \in \mathrm{U}(N)$ then

$$\|U_1V_1 - U_2V_2\|_F \leq \|U_1 - U_2\|_F + \|V_1 - V_2\|_F. \tag{2.34}$$

Moreover, directly by the definition of $\|\cdot\|_F$,

$$\|A \oplus B\|_F^2 = \|A\|_F^2 + \|B\|_F^2. \tag{2.35}$$

Next, we will explore the continuity of the butterfly factors.

Lemma 2.2. *Let $\varepsilon \in \mathbb{R}$. Then*

$$\|\mathbf{I}_k - R_k(\varepsilon)\|_F \leq \sqrt{k}|\varepsilon|. \tag{2.36}$$

Proof. Since k is necessarily even, we can consider $2k$. Recall

$$R_{2k}(\theta) = B(\theta) \otimes \mathbf{I}_k = Q^T(\mathbf{I}_k \otimes B(\theta))Q$$

for $B(\theta) \in \mathrm{SO}(2)$ and Q a perfect shuffle matrix. It follows

$$\begin{aligned}
\|\mathbf{I}_{2k} - R_{2k}(\varepsilon)\|_F^2 &= \|\mathbf{I}_{2k} - Q^T(\mathbf{I}_k \otimes B(\varepsilon))Q\|_F^2 = \|\mathbf{I}_k \otimes (\mathbf{I}_2 - B(\varepsilon))\|_F^2 \\
&= \|\bigoplus^k (\mathbf{I}_2 - B(\varepsilon))\|_F^2 = k\|\mathbf{I}_2 - B(\varepsilon)\|_F^2 \\
&= k \operatorname{Tr}\left((\mathbf{I}_2 - B(\varepsilon))(\mathbf{I}_2 - B(\varepsilon))^*\right) \\
&= k \operatorname{Tr}(2\mathbf{I}_2 - (B(\varepsilon) + B(-\varepsilon))) \\
&= k \operatorname{Tr}(2(1 - \cos\varepsilon)\mathbf{I}_2) = 4k(1 - \cos\varepsilon) \\
&\leq 2k\varepsilon^2
\end{aligned}$$

using the bilinearity of $\otimes$ and (2.35), while the last inequality follows directly from the

elementary bound

$$1 - \cos x \leq \frac{x^2}{2} \tag{2.37}$$

for all $x \in \mathbb{R}$. □

This leads to the result for simple scalar butterfly matrices:

Proposition 2.10. *Let $B(\boldsymbol{\theta}) \in \mathrm{B}_s(N)$ and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$. Then*

$$\|B(\boldsymbol{\theta}) - B(\boldsymbol{\theta} + \boldsymbol{\varepsilon})\|_F \leq \sqrt{N}\|\boldsymbol{\varepsilon}\|_1. \tag{2.38}$$

Proof. Recall $B(\boldsymbol{\theta} + \boldsymbol{\varepsilon}) = B(\boldsymbol{\theta})B(\boldsymbol{\varepsilon})$. Using the mixed-product property, we have

$$\begin{aligned}
\|B(\boldsymbol{\theta}) - B(\boldsymbol{\theta} + \boldsymbol{\varepsilon})\|_F &= \|B(\boldsymbol{\theta})(\mathbf{I}_N - B(\boldsymbol{\varepsilon}))\|_F \\
&= \|\mathbf{I}_N - \bigotimes_{j=1}^{n} B(\varepsilon_{n-j+1})\|_F \\
&= \|\mathbf{I}_N - \prod_{k=1}^{n} \bigotimes_{j=1}^{n} B(\varepsilon_{n-j+1})^{\delta_{k,j}}\|_F \\
&= \|\mathbf{I}_N^n - \prod_{k=1}^{n} \mathbf{I}_{N2^{-(n-k+1)}} \otimes R_{2^{n-k+1}}(\varepsilon_{n-k+1})\|_F \\
&\leq \sum_{k=1}^{n} \|\mathbf{I}_N - \mathbf{I}_{N2^{-k}} \otimes R_{2^k}(\varepsilon_k)\|_F \\
&= \sum_{k=1}^{n} \|\mathbf{I}_{N2^{-k}} \otimes (\mathbf{I}_{2^k} - R_{2^k}(\varepsilon_k))\|_F \\
&= \sum_{k=1}^{n} \| \bigoplus^{N2^{-k}} (\mathbf{I}_{2^k} - R_{2^k}(\varepsilon_k))\|_F \\
&= \sum_{k=1}^{n} \sqrt{N2^{-k}}\|\mathbf{I}_{2^k} - R_{2^k}(\varepsilon_k)\|_F \\
&\leq \sum_{k=1}^{n} \sqrt{N2^{-k}}\sqrt{2^k}|\varepsilon_k| \\
&= \sqrt{N}\|\boldsymbol{\varepsilon}\|_1
\end{aligned}$$

using (2.34) and Lemma 2.2, respectively, for the inequalities. □

The result for scalar butterfly matrices is established similarly, focusing first on the diagonal block butterfly factors:

Lemma 2.3. *Let $B(\boldsymbol{\theta}) \in \mathcal{D}^k(N)$ and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$. Then*

$$\|B(\boldsymbol{\theta}) - B(\boldsymbol{\theta} + \boldsymbol{\varepsilon})\|_F \leq \sqrt{\frac{N}{k}} \|\boldsymbol{\varepsilon}\|_2. \tag{2.39}$$

Proof. Similarly, $B(\boldsymbol{\theta} + \boldsymbol{\varepsilon}) = B(\boldsymbol{\theta})B(\boldsymbol{\varepsilon})$, and recall necessarily $k \mid N$. It follows

$$\begin{aligned}
\|B(\boldsymbol{\theta}) - B(\boldsymbol{\theta} + \boldsymbol{\varepsilon})\|_F &= \|B(\boldsymbol{\theta})(\mathbf{I}_N - B(\boldsymbol{\varepsilon}))\|_F \\
&= \|\mathbf{I}_N - \bigoplus_{j=1}^{k} R_{N/k}(\varepsilon_j)\|_F \\
&= \|\bigoplus_{j=1}^{k} (\mathbf{I}_{N/k} - R_{N/k}(\varepsilon_j))\|_F \\
&= \left(\sum_{j=1}^{k} \|\mathbf{I}_{N/k} - R_{N/k}(\varepsilon_j)\|_F^2 \right)^{1/2} \\
&\leq \left(\sum_{j=1}^{k} \frac{N}{k} \varepsilon_j^2 \right)^{1/2} \\
&= \sqrt{\frac{N}{k}} \|\varepsilon\|_2
\end{aligned}$$

using (2.35) and Lemma 2.2 for the inequality. □

Now we can establish the result for general scalar butterfly matrices.

Proposition 2.11. *Let $B(\boldsymbol{\theta}) \in \mathrm{B}(N)$ and $\boldsymbol{\varepsilon} \in \mathbb{R}^{N-1}$. Then*

$$\|B(\boldsymbol{\theta}) - B(\boldsymbol{\theta} + \boldsymbol{\varepsilon})\|_F \leq \sqrt{N-1} \|\boldsymbol{\varepsilon}\|_2 \tag{2.40}$$

Proof. Let $\boldsymbol{\theta}_j \in \mathbb{R}^{N2^{-j}}$ such that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ and

$$B(\boldsymbol{\theta}) = B_n(\boldsymbol{\theta}_n) \cdots B_1(\boldsymbol{\theta}_1) = \prod_{j=1}^{n} B_{n-j+1}(\boldsymbol{\theta}_{n-j+1}) \tag{2.41}$$

for $B_j(\boldsymbol{\theta}_j) \in \mathcal{D}^{2^j}(N)$. Similarly decompose $\boldsymbol{\varepsilon}$ so that

$$\begin{aligned}
\|B(\boldsymbol{\theta}) - B(\boldsymbol{\theta} + \boldsymbol{\varepsilon})\|_F &= \|\prod_{j=1}^{n} B_{n-j+1}(\boldsymbol{\theta}_{n-j+1}) - \prod_{j=1}^{n} B_{n-j+1}(\boldsymbol{\theta}_{n-j+1} + \boldsymbol{\varepsilon}_{n-j+1})\|_F \\
&\leq \sum_{j=1}^{n} \|B_j(\boldsymbol{\theta}_j) - B_j(\boldsymbol{\theta}_j + \boldsymbol{\varepsilon}_j)\|_F \\
&\leq \sum_{j=1}^{n} \sqrt{N2^{-j}} \|\boldsymbol{\varepsilon}_j\|_2 \\
&\leq \left(\sum_{j=1}^{n} N2^{-j}\right)^{1/2} \left(\sum_{j=1}^{n} \|\boldsymbol{\varepsilon}_j\|_2^2\right)^{1/2} \\
&= \sqrt{N-1} \|\boldsymbol{\varepsilon}\|_2.
\end{aligned}$$

using (2.34), Lemma 2.3, and the Cauchy-Schwarz inequality, respectively, for the above inequalities. □

2.2 Order $N = m^n$ butterfly matrices

Now I will introduce a new butterfly structure of order $N = m^n$ for any integer $m \geq 2$, which is a generalization of the scalar butterfly matrices.

The most general structure I will consider will be the following:

Definition 2.6. *For G a class of groups such that $G(m)$ is a compact subgroup of $\mathrm{GL}(\mathbb{C}^m)$, a m-**butterfly** G **matrix**, collectively denoted by $\mathrm{B}(m, n, G)$, with $\mathrm{B}(m, 0, G) = G(1)$, is an*

iteratively defined matrix of order $N = m^n$ of the following form:

$$(A \otimes \mathbf{I}_{N/m}) \bigoplus_{j=1}^{m} B_j, \tag{2.42}$$

*where $B_j \in \mathrm{B}(m, n-1, G)$ and $A \in G(m)$. The **simple m-butterfly G matrices**, denoted by $\mathrm{B}_s(m, n, G)$, are formed such that $B_j = B \in \mathrm{B}_s(m, n-1, G)$ for all j; such matrices are of the form*

$$\bigotimes_{j=1}^{n} B_j \tag{2.43}$$

for $B_j \in G(m)$.

Definition 2.7. *Let $\mathrm{B}(m, n) = \mathrm{B}(m, n, \mathrm{SO})$ and $\mathrm{B}_s(m, n) = \mathrm{B}_s(m, n, \mathrm{SO})$ denote the **m-butterfly matrices** and the **simple m-butterfly matrices**.*

Note for $N = 2^n$ then

$$\mathrm{B}(N) = \mathrm{B}(2, n) = \mathrm{B}(2, n, \mathrm{SO}) \quad \text{and} \quad \mathrm{B}_s(N) = \mathrm{B}_s(2, n) = \mathrm{B}_s(2, n, \mathrm{SO}). \tag{2.44}$$

So far we have only considered butterfly matrices initiated using $G = \mathrm{SO}$, but a lot of the arguments will go through similarly, with some (notable) exceptions using different G.

Proposition 2.12. *Let $N = m^n$ and let $G = \mathrm{SO}, \mathrm{O}, \mathrm{SU}$, or U. Then $\mathrm{B}(m, n, G)$ is a compact subset of $G(N)$, which is homeomorphic to a quotient of $G(m)^{\frac{N-1}{m-1}}$. Moreover, $\mathrm{B}_s(m, n, G)$ is a compact topological group that is isomorphic and homeomorphic to a quotient of $G(m)^n$, and is abelian if and only if $m = 2$ and $G = \mathrm{SO}$.*

Remark 2.6. *I am abstaining from considering also the symplectic models in this document. Similar arguments should be able to derive similar results for these models if one is interested.*

Proof of Proposition 2.12. This follows directly from induction using results in Section 1.4.4 and Definition 2.6: Let

$$B = (A \otimes \mathbf{I}_{N/m}) \bigoplus_{j=1}^{m} B_j. \tag{2.45}$$

We have B is orthogonal or unitary, respectively, if A, B_j are all orthogonal or unitary. Moreover,

$$\begin{aligned} \det B &= \det\left((A \otimes \mathbf{I}_{N/m})\left(\bigoplus_{j=1}^{m} B_j\right)\right) = \det(A \otimes \mathbf{I}_{N/m}) \det\left(\bigoplus_{j=1}^{m} B_j\right) \\ &= \det(A)^{N/m} \prod_{j=1}^{m} \det(B_j) \end{aligned}$$

yields B has unit determinant when all of A, B_j have unit determinant. If $A = \bigotimes_{j=1}^{n} A_j, B = \bigotimes_{j=1}^{n} B_j \in \mathrm{B}_s(m, n, G)$, then $A_j B_j^{-1} \in G$ for all j and hence $AB^{-1} = \bigotimes_{j=1}^{n} A_j B_j^{-1} \in \mathrm{B}_s(m, n, G)$, which shows $\mathrm{B}_s(m, n, G)$ is a subgroup of $\mathrm{SO}(N), \mathrm{O}(N), \mathrm{SU}(N)$ and $\mathrm{U}(N)$ whenever G is, respectively, $\mathrm{SO}, \mathrm{O}, \mathrm{SU}, \mathrm{U}$.

To see $\mathrm{B}(m, n, G)$ is compact, we use induction on n. The result is trivial for $n = 1$. If the result holds for $n - 1$, then $G(m) \times \mathrm{B}(m, n-1, G)^m$ is compact by the inductive hypothesis. Hence, the multiplication map $f : G(m) \times \mathrm{B}(m, n-1, G)^m \to \mathrm{U}(N)$ with $f(A, B_1, \ldots, B_m) = (A \otimes \mathbf{I}_{N/m}) \bigoplus_{j=1}^{m} B_j$, which is continuous since $\mathrm{U}(N)$ is a Lie group, has a compact image $f(G(m) \times \mathrm{B}(m, n-1, G)^m) = \mathrm{B}(m, n, G)$. It follows then $\mathrm{B}(m, n, G)$ is a quotient of

$$G(m) \times \mathrm{B}(m, n-1, G)^m \cong G(m)^{1+m\left(\frac{N/m-1}{m-1}\right)} = G(m)^{\frac{N-1}{m-1}}. \tag{2.46}$$

Using the same argument with $\mathrm{B}_s(m, n, G)$ yields this is a compact space homeomophic to a quotient of $G \times \mathrm{B}(m, n-1, G) \cong G^n$ (again using induction to justify the last equivalence). Moreover, the multiplication map then is also a group homomorphism (by the mixed-product

property), so that $\mathrm{B}_s(m,n,G)$ is also isomorphic as a group to a quotient of G^n.

The first statement regarding $\mathrm{B}_s(m,n,G)$ follows directly from Corollary 1.8.

For the last statement, we have $\mathrm{B}_s(2,n,\mathrm{SO})$ is abelian by Proposition 2.9. To see this is the only possible abelian butterfly model, it is enough to note $\mathrm{O}(2)$, $\mathrm{SU}(2)$ and $\mathrm{SO}(3)$ are nonabelian: sufficiency follows since $\mathrm{O}(2)$ embeds into $\mathrm{O}(N) \subset \mathrm{U}(N)$ and $\mathrm{SU}(2)$ embeds into $\mathrm{SU}(N) \subset \mathrm{U}(N)$ (via $A \mapsto A \oplus \mathbf{I}_{m^n-2}$) for $N \geq 2$, while $\mathrm{SO}(3)$ embeds into $\mathrm{SO}(N)$ (via $A \mapsto A \oplus \mathbf{I}_{N-3}$) for $N \geq 3$. Hence, we have each of these embed into a subgroup of $\mathrm{B}_s(m,n,G)$ by considering, appropriately, $(\mathrm{O}(2) \oplus \mathbf{I}_{m-2}) \otimes \mathbf{I}_{N/m}$, $(\mathrm{SU}(2) \oplus \mathbf{I}_{m-2}) \otimes \mathbf{I}_{N/m}$ and $(\mathrm{SO}(3) \oplus \mathbf{I}_{m-3}) \otimes \mathbf{I}_{N/m}$.

To see $\mathrm{O}(2)$ is nonabelian, we note

$$\begin{bmatrix} 1 & \\ & -1 \end{bmatrix}^T \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & \\ & -1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}. \tag{2.47}$$

To see $\mathrm{SU}(2)$ is nonabelian, we note

$$\begin{bmatrix} i & \\ & -i \end{bmatrix}^* \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} i & \\ & -i \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}. \tag{2.48}$$

To see $\mathrm{SO}(3)$ is nonabelian, we note

$$\begin{bmatrix} -1 & & \\ & 1 & \\ & & -1 \end{bmatrix}^T \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & & \\ & 1 & 1 \\ & -1 & 1 \end{bmatrix} \begin{bmatrix} -1 & & \\ & 1 & \\ & & -1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & & \\ & 1 & -1 \\ & 1 & 1 \end{bmatrix}. \tag{2.49}$$

□

Remark 2.7. *For $N = m^n$, recall $\mathrm{O}(m)$ and $\mathrm{SO}(m)$ are $\binom{m}{2}$-manifolds while $\mathrm{U}(m)$ and*

$\mathrm{SU}(m)$ *are* (m^2-1)*-manifolds. Proposition 2.12 shows* $\mathrm{B}(m,n,\mathrm{SO})$ *and* $\mathrm{B}(m,n,\mathrm{O})$ *are quotients of* $\frac{N-1}{m-1}\binom{m}{2}$*-manifolds while* $\mathrm{B}(m,n,\mathrm{SU})$ *and* $\mathrm{B}(m,n,\mathrm{U})$ *are quotients of* $\frac{N-1}{m-1}(m^2-1)$*-manifolds. Similarly,* $\mathrm{B}_s(m,n,\mathrm{SO})$ *and* $\mathrm{B}_s(m,n,\mathrm{O})$ *are quotients of* $n\binom{m}{2}$*-manifolds while* $\mathrm{B}_s(m,n,\mathrm{SU})$ *and* $\mathrm{B}_s(m,n,\mathrm{U})$ *are quotients of* $n(m^2-1)$*-manifolds. In the case of* $G=\mathrm{SO}$ *and* $m=2$*, then the quotients maps have finite kernels and so* $\mathrm{B}(n)$ *and* $\mathrm{B}_s(n)$ *are themselves* $N-1$*- and* n*-manifolds (by Proposition 2.7).*

2.3 Order $N=\prod_{j=1}^{k} p_j^{e_j}$ butterfly matrices

Now we will explore butterfly matrices of arbitrary orders. First we will consider a butterfly matrix formed only using m-butterfly matrices, such as the 2-butterfly matrices. We will then explore a general structure depending directly on the prime factorization of N.

Another choice for G in Definition 2.6 can include the following:

Definition 2.8. *Let* p *be a prime number. The* p***-nary butterfly matrices*** *of order* N*, denoted by* $\mathrm{B}(p,N)$*, are of the form*

$$\mathrm{B}(p,N)=\bigoplus_{j=0}^{\lfloor \log_p N\rfloor}\ \bigoplus^{a_{\lfloor \log_p N\rfloor -j+1}} \mathrm{B}_s(p,\lfloor \log_p N\rfloor -j+1) \tag{2.50}$$

where $N=\sum_{j\geq 0} a_j p^j$ *for* $a_j\in\{0,1,\ldots,p-1\}$ *for all* j*. The* ***binary butterfly matrices*** *and* ***ternary butterfly matrices*** *are, respectively,* $\mathrm{B}(2,N)$ *and* $\mathrm{B}(3,N)$*.*

Remark 2.8. *The* a_j *give the* p*-nary representation of* N*.*

Remark 2.9. *I am not using a subscript to denote the use of the simple butterfly structures since this is the only model that satisfies a group structure.*

A straightforward calculation verifies:

Corollary 2.5. $\mathrm{B}(p,N)$ *is a compact topological subgroup of* $\mathrm{SO}(N)$ *isomorphic with a quotient of* $\mathrm{SO}(p)^{\sum_{j=1}^{\lfloor \log_p N \rfloor} j a_j}$ *for* $N = \sum_{j\geq 0} a_j p^j$*, and* $\mathrm{B}(p,N)$ *is abelian if and only if* $p = 2$*.*

Remark 2.10. *Using Remark 2.7,* $\mathrm{B}(p,N)$ *is a quotient of a* $\left(\sum_{j=1}^{\lfloor \log_p N \rfloor} j a_j\right)\binom{p}{2}$*-manifold.*

Remark 2.11. *There are certain limitations in using this simpler structure for a general order butterfly matrix. If* $N = 2k+1$ *is odd, then* $B\mathbf{e}_N = B^T\mathbf{e}_N = \mathbf{e}_N$ *for* $B \in \mathrm{B}(2,N)$*. So if* $B, B' \in \mathrm{B}(2,N)$*, then* $B(\mathbf{I}_{2k} \oplus \theta)B'^T = \mathbf{I}_{2k} \oplus \theta$*. If one desired to randomize a vector using a butterfly transformation to spread out the weights among its components, then one is out of luck using only binary butterfly matrices if most of the weight is in the last component.*

We will now explore a different structure that is more robust in its randomization properties. This butterfly structure is an exact analogue of the Cooley-Tukey FFT algorithm that allows an FFT decomposition of a composite integer.

I will first give the most general structure I will consider here:

Definition 2.9. *Let* G *denote a class of groups such that* $G(m)$ *is a compact subgroup of* $\mathrm{GL}(\mathbb{C}^m)$ *for any positive integer* m*. Let* $N = \prod_{j=1}^{k} p_j^{e_j}$ *be the prime factorization of* N *such that* $p_j < p_{j+1}$ *for each* j*. An order* N ***butterfly*** G ***matrix****, collectively denoted by* $\mathrm{B}(N,G)$*, with* $\mathrm{B}(1,G) = G(1)$*, is the matrix of the form*

$$\bigotimes_{j=1}^{k} B_j \tag{2.51}$$

where $B_j \in \mathrm{B}(p_j, e_j, G)$*. The* ***simple butterfly*** G ***matrices****, denoted by* $\mathrm{B}_s(N,G)$*, are formed using* $B_j \in \mathrm{B}_s(p_j, e_j, G)$ *for all* j*.*

Definition 2.10. *Let* $\mathrm{B}(N) = \mathrm{B}(N, \mathrm{SO})$ *and* $\mathrm{B}_s(N) = \mathrm{B}_s(N, \mathrm{SO})$ *denote the* ***butterfly matrices*** *and the* ***simple butterfly matrices****.*

Remark 2.12. *Note Definition 2.10 is consistent with the scalar butterfly matrices definition in Definition 2.3 when* $N = 2^n$*.*

Note the ordering in (2.51) then enforces the smallest prime factors to constitute the top Kronecker factors. If one desired, one could instead choose the descending order of the prime factors for the Kronecker ordering, or an ordering of the form

$$\bigotimes_{j=1}^{n} B_{\sigma(j)} \tag{2.52}$$

for any $\sigma \in S_k$, again when $N = \prod_{j=1}^{k} p_j^{e_j}$. Any fixed ordering then can be transformed into another order through a sequence of perfect shuffle transformations.

Example 2.4. *Using examples* $B_1 \in \mathrm{SO}(2)$ *and* $B_2 \in \mathrm{SO}(3)$ *with*

$$B_1 = \begin{bmatrix} 0.68 & 0.73 \\ -0.73 & 0.68 \end{bmatrix}$$

and

$$B_2 = \begin{bmatrix} -0.22 & 0.90 & 0.38 \\ -0.96 & -0.14 & -0.23 \\ -0.15 & -0.42 & 0.90 \end{bmatrix}$$

together produce the order 6 butterfly matrices

$$B_1 \otimes B_2 = \left[\begin{array}{ccc|ccc} -0.15 & 0.61 & 0.26 & -0.16 & 0.66 & 0.28 \\ -0.66 & -0.10 & -0.15 & -0.71 & -0.10 & -0.17 \\ -0.10 & -0.29 & 0.61 & -0.11 & -0.31 & 0.66 \\ \hline 0.16 & -0.66 & -0.28 & -0.15 & 0.61 & 0.26 \\ 0.71 & 0.10 & 0.16 & -0.66 & -0.10 & -0.15 \\ 0.11 & 0.31 & -0.66 & -0.10 & -0.29 & 0.61 \end{array}\right]$$

and

$$B_2 \otimes B_1 = \left[\begin{array}{cc|cc|cc} -0.15 & -0.16 & 0.61 & 0.66 & 0.26 & 0.28 \\ 0.16 & -0.15 & -0.66 & 0.61 & -0.28 & 0.26 \\ \hline -0.66 & -0.71 & -0.10 & -0.10 & -0.15 & -0.16 \\ 0.71 & -0.66 & 0.10 & -0.10 & 0.17 & -0.15 \\ \hline -0.10 & -0.11 & -0.29 & -0.31 & 0.61 & 0.66 \\ 0.11 & -0.10 & 0.31 & -0.29 & -0.66 & 0.61 \end{array}\right].$$

For $Q = P_{(2\ 4\ 5\ 3)}$, *we have* $Q^T(B_1 \otimes B_2)Q = B_2 \otimes B_1$.

Moreover, repeating the previous arguments yield

Proposition 2.13. *Let* $G = \mathrm{SO}, \mathrm{O}, \mathrm{SU}$ *or* U. *Then* $\mathrm{B}(N, G)$ *is a compact subset of* $G(N)$. *For* $N = \prod_{j=1}^{k} p_j^{e_j}$ *for* $p_j < p_{j+1}$, *then* $\mathrm{B}(N, G)$ *is homeomorphic to a quotient of*

$$\prod_{j=1}^{k} G(p_j)^{\frac{p_j^{e_j}-1}{p_j-1}}. \tag{2.53}$$

Moreover, $\mathrm{B}_s(N, G)$ *is a compact topological group that is isomorphic and homeomorphic to a quotient of*

$$\prod_{j=1}^{k} G(p_j)^{e_j}, \tag{2.54}$$

and is abelian if and only if $N = 2^n$ *and* $G = \mathrm{SO}$.

2.4 Other butterfly matrix models

Other considerations one can explore could involve fixing the **depth** of a butterfly matrix in terms of limiting how many butterfly factors are used. This can be done either by cutting off the factors to the left or right, which would have equivalent impact on randomization

applications. As noted above, $\mathcal{R}(N), \mathcal{D}^j(N)$ and $\mathcal{D}_s^j(N)$ can be defined as long as N is even and $2j \mid N$, which would lead to similar constructions of butterfly matrices that would result from limiting the depth on the right: that is, instead of starting with $\mathrm{B}(1) = \{1\}$ we can start with $\mathrm{B}^{(m)}(1) = \{\mathbf{I}_m\}$ and then iteratively build butterfly matrices $\mathrm{B}^{(m)}(2^n)$ of order $2^n m$. The results in Section 2.1 would follow through nearly identically for these models.

[2] looked specifically at limiting depths of similar butterfly models, indicating even desired numerical accuracy calculations sufficed from their experiments with very small depth, such as 2 or 3 levels. Limiting depth, especially in the one-sided preconditioning model, can be shown to have undesirable results (see Section 5.3). However, the payoff for the reduction in complexity by using fewer parameters is a worthwhile balance to explore further.

2.5 Hadamard matrices

This chapter will end with a brief discussion on Hadamard matrices. Some connections to butterfly matrices and potential applications are outlined below.

2.5.1 History and background

Definition 2.11. *$H \in \{\pm 1\}^{n \times n}$ is a **Hadamard matrix** if $HH^T = n\mathbf{I}_n$.*

Equivalently, H is a Hadamard matrix if its rows and columns of only ± 1 are orthogonal.

In 1893[2], Hadamard investigated the question of how large the $\det A$ could be for $A \in$

[2]The same year as the World's Columbian Exposition where HH Holmes silently wrecked havoc on the streets of Chicago.

$[-1,1]^{n\times n}$. He presented the bound

$$|\det(A)| \leq n^{n/2}. \tag{2.55}$$

Hadamard observed that Sylvester had provided a recursive construction for certain order 2^n matrices that satisfied this upper bound, that were constructed as

$$H_0 = \bigotimes^{n} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}. \tag{2.56}$$

A simple observation shows if H is a Hadamard matrix, then

$$(\det H)^2 = \det HH^T = \det(n\mathbf{I}_n) = n^n \tag{2.57}$$

so that H satisfies the upper bound in (2.55).

Using Sylvester's construction, with these being called **Sylvester Hadamard matrices** for different order 2 Hadamard matrices in each intermediate Kronecker factor, a Hadamard matrix can be constructed for every power of 2. Note a similar construction using the form $(H_1 \otimes \mathbf{I}_{N/2})(H_2 \oplus H_3)$ for H_1 an order 2 Hadamard matrix and H_2, H_3 each order $N/2$ Hadamard matrices is called the **Walsh-Hadamard construction**. The existence of Hadamard matrices for a given order remains the most famous open question relating to Hadamard matrices. Other than $n = 1$ and $n = 2$, a simple consideration of the first few rows of a Hadamard matrix shows that the order of a Hadamard matrix must be divisible by 4. This leads to the famous open problem:

Conjecture 1. *A Hadamard matrix of order $4n$ exists for every positive integer n.*

Once a Hadamard matrix is shown to exist of order k, then one can construct through Kronecker products Hadamard matrices of order $k2^n$ for any positive integer, and similarly

of order km using an order m Hadamard matrix. Currently, the lowest integer divisible by 4 that a Hadamard matrix has not been verified to exist is of order $668 = 4 \cdot 167$ [11]. If a Hadamard matrix exists of order n, then multiplying on the left and right by permutation matrices generate distinct Hadamard matrices. This shows there are *at least* $(n!)^2 = O(2^{n \log n})$ Hadamard matrices of order n. It is an interesting question in itself to look into how many Hadamard matrices exist of a fixed order, and whether this matches the same scaling as the lower bound.

Conjecture 2. *There are $O(2^{n \log n})$ distinct Hadamard matrices of order n if a Hadamard matrix exists of order n.*

In a recent result, Ferber, Jain and Zhao present an upper bound for the number of Hadamard matrices of $2^{\frac{1}{2}(1-c)n^2}$ for n sufficiently large and an absolute constant $c > 0$ [14]. This is still far from establishing Conjecture 2

2.5.2 Butterfly Hadamard matrices

One can define an equivalence class on Hadamard matrices by relating matrices that are transformations of one another through a sequence of signed permutation matrices or transposition. I will say two Hadamard matrices are in the same **class** if they are equivalent using this relationship. It can be shown there is only one class of Hadamard matrices (the Sylvester class) for orders up to 12, but it stops there: there are 4 distinct classes of order 16 Hadamard matrices. Hadamard matrices remain a continued focus of research. Butterfly matrices have the potential application as being a continuous generalization of the Sylvester Hadamard matrices, as is illustrated here:

Proposition 2.14. *Let $N = 2^n$ and $\boldsymbol{\theta} \in (\frac{\pi}{4} + \frac{\pi}{2}\mathbb{Z})^n$. For $B(\boldsymbol{\theta}) \in \mathrm{B}_s(N)$, then $\sqrt{N}B(\boldsymbol{\theta})$ is a Hadamard matrix in the Sylvester class.*

Proof. Let $H = \sqrt{N}B(\boldsymbol{\theta})$. Since $|\cos(\theta_j)| = |\sin(\theta_j)| = \frac{1}{\sqrt{2}}$ for all j, then $H_{ij} = \sqrt{N}B(\boldsymbol{\theta})_{ij} = \pm\sqrt{N}2^{-n/2} = \pm 1$ so that $H \in \{\pm 1\}^{N\times N}$. Checking also

$$HH^T = NB(\boldsymbol{\theta})B(\boldsymbol{\theta})^T = N\mathbf{I}_N \tag{2.58}$$

verifies H is a Hadamard matrix.

To check H is in the Sylvester Hadamard class, note first for $n = 1$ we see

$$\sqrt{2}B\left(\frac{\pi}{4}\right) = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & \\ & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$\sqrt{2}B\left(\frac{3\pi}{4}\right) = \begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} -1 & \\ & 1 \end{bmatrix}$$

$$\sqrt{2}B\left(\frac{5\pi}{4}\right) = \begin{bmatrix} -1 & -1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} -1 & \\ & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$\sqrt{2}B\left(\frac{7\pi}{4}\right) = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & \\ & -1 \end{bmatrix}$$

Let

$$P(\theta) = \begin{cases} \begin{bmatrix} 1 & \\ & -1 \end{bmatrix} & \text{if } \theta = \frac{\pi}{4} \\ -\mathbf{I}_2 & \text{if } \theta = \frac{5\pi}{4} \\ \mathbf{I}_2 & \text{otherwise} \end{cases}$$

and

$$Q(\theta) = \begin{cases} \begin{bmatrix} -1 & \\ & 1 \end{bmatrix} & \text{if } \theta = \frac{3\pi}{4} \\ \begin{bmatrix} 1 & \\ & -1 \end{bmatrix} & \text{if } \theta = \frac{7\pi}{4} \\ \mathbf{I}_2 & \text{otherwise.} \end{cases}$$

It follows then for $P = \bigotimes_{j=1}^{n} P(\theta_{n-j+1})$ and $Q = \bigotimes_{j=1}^{n} Q(\theta_{n-j+1})$, then $PHQ = H_0$ using the mixed-product property. □

This gives one way to generate a Sylvester Hadamard matrix using the butterfly matrices, with $4^n = N^2$ possible butterfly Hadamard matrices of order N. Rather than fixing given angles, one can generate a Hadamard matrix using (almost) any butterfly matrix combined with the sgn function, which when applied to a matrix acts componentwise so that $(\operatorname{sgn}(A))_{ij} = \operatorname{sgn}(A_{ij}) = \frac{A_{ij}}{|A_{ij}|}$, where we define $\operatorname{sgn}(0) := 0$. If $A \in (\mathbb{R}\backslash\{0\})^{n\times m}$ then $\operatorname{sgn}(A) \in \{\pm 1\}^{n\times m}$.

Proposition 2.15. *Let $N = 2^n$ and $\boldsymbol{\theta} \in [0, 2\pi)^n \backslash (\frac{\pi}{2}\mathbb{Z})^n$. For $B(\boldsymbol{\theta}) \in \mathrm{B}_s(N)$ then $\operatorname{sgn}(B(\boldsymbol{\theta}))$ is a Hadamard matrix in the Sylvester class.*

Proof. It suffices to note sgn is constant on fixed sectors, where $\operatorname{sgn}((A_{ij}(0,\infty))^{N\times N}) = A$ for $A \in \{\pm 1\}^{N\times N}$ and so is continuous on each sector. The result then follows by Propositions 2.10 and 2.16. □

Note $\operatorname{sgn}(DA) = \operatorname{sgn}(D)\operatorname{sgn}(A)$ if D is a diagonal matrix. It follows then

$$\operatorname{sgn}(A \otimes B) = \operatorname{sgn}(A) \otimes \operatorname{sgn}(B). \tag{2.59}$$

and

$$\operatorname{sgn}\big((A \otimes \mathbf{I}_{N/2})(B \oplus C)\big) = (\operatorname{sgn}(A) \otimes \mathbf{I}_{N/2})(\operatorname{sgn}(B) \oplus \operatorname{sgn}(C)). \tag{2.60}$$

for $A \in \mathbb{R}^{2\times 2}$ and $B, C \in \mathbb{R}^{N/2}$. In particular, for C, S diagonal, then

$$\operatorname{sgn}\left(\begin{bmatrix} CA_1 & SA_2 \\ -SA_1 & CA_2 \end{bmatrix}\right) = \begin{bmatrix} \operatorname{sgn}(C)\operatorname{sgn}(A_1) & \operatorname{sgn}(S)\operatorname{sgn}(A_2) \\ -\operatorname{sgn}(S)\operatorname{sgn}(A_1) & \operatorname{sgn}(C)\operatorname{sgn}(A_2) \end{bmatrix}. \tag{2.61}$$

For $B(\boldsymbol{\theta}) \in \mathrm{B}_s(N)$, from (2.59) we have

$$\operatorname{sgn}(B(\boldsymbol{\theta})) = \sqrt{N}B(\boldsymbol{\theta}') \quad \text{where} \quad \theta'_j = \frac{\pi}{4}\left(2\left\lfloor \frac{2\theta_j}{\pi} \right\rfloor + 1\right). \tag{2.62}$$

This gives an alternative proof of Proposition 2.15. Additionally, from (2.61) we have $\operatorname{sgn}(B)$ is a Walsh-Hadamard matrix for also B a nonsimple scalar butterfly matrix, simple diagonal butterfly matrix and nonsimple diagonal butterfly matrix assuming $\theta_j \notin \frac{\pi}{2}\mathbb{Z}$ for any j. The nonsimple scalar butterfly matrix case follows immediately from (2.61). The diagonal case is less direct:

Proposition 2.16. *If $B(\boldsymbol{\theta})$ is a scalar or diagonal order N butterfly matrix with $\theta_j \notin \frac{\pi}{2}\mathbb{Z}$ for all j, then* $\operatorname{sgn}(B)$ *is a Hadamard matrix.*

Proof. It suffices to consider only the general case of $B = B(\boldsymbol{\theta})$ a nonsimple diagonal butterfly matrix. Since $\theta_j \notin \frac{\pi}{2}\mathbb{Z}$, then $\operatorname{sgn}(B) \in \{\pm 1\}^{N\times N}$. To see $\operatorname{sgn}(B)\operatorname{sgn}(B)^T = N\mathbf{I}_N$, we will use induction on n. The result is immediate for $n = 1$ while the $n = 2$ case follows from Proposition 2.15. Assume the result holds for $n - 1$. From (2.61), for $(C, S)(\boldsymbol{\theta}) =$

$\bigoplus_{j=1}^{N/2}(\cos\theta_j, \sin\theta_j)$ and A_1, A_2 order $N/2$ diagonal butterfly matrices, we have

$$\operatorname{sgn}(B) = \begin{bmatrix} \operatorname{sgn}(C) & \operatorname{sgn}(S) \\ -\operatorname{sgn}(S) & \operatorname{sgn}(C) \end{bmatrix} \begin{bmatrix} \operatorname{sgn}(A_1) & \\ & \operatorname{sgn}(A_2) \end{bmatrix} \tag{2.63}$$

By the inductive hypothesis, $\operatorname{sgn}(A_i)\operatorname{sgn}(A_i)^T = \frac{N}{2}\mathbf{I}_{N/2}$ so that

$$(\operatorname{sgn}(A_1) \oplus \operatorname{sgn}(A_2))(\operatorname{sgn}(A_1) \oplus \operatorname{sgn}(A_2))^T = \frac{N}{2}\mathbf{I}_N. \tag{2.64}$$

Since C and S are nonsingular diagonal matrices, then $\operatorname{sgn}(C)^2 = \operatorname{sgn}(S)^2 = \mathbf{I}_{N/2}$ and $[\operatorname{sgn}(C), \operatorname{sgn}(S)] = \mathbf{0}$. It follows

$$\begin{bmatrix} \operatorname{sgn}(C) & \operatorname{sgn}(S) \\ -\operatorname{sgn}(S) & \operatorname{sgn}(C) \end{bmatrix} \begin{bmatrix} \operatorname{sgn}(C) & \operatorname{sgn}(S) \\ -\operatorname{sgn}(S) & \operatorname{sgn}(C) \end{bmatrix}^T = 2\mathbf{I}_N. \tag{2.65}$$

Together, this yields $\operatorname{sgn}(B)\operatorname{sgn}(B)^T = N\mathbf{I}_N$. □

2.5.3 Constructions of Hadamard matrices

Note the Sylvester and Walsh-Hadamard constructions only work for $N = 2^n$ or $N = mk$ where a Hadamard matrix is know to exist of order m and k. Alternative constructions of Hadamard matrices include the Paley construction, which uses tools from finite fields, and the Williamson construction, which uses the sum of four squares [30, 41]. Applications of these only work for very particular orders and do not cover all potential multiple of 4 orders.

A very rudimentary method to (try to) construct a Hadamard matrix is the rejection method. For $4 \mid n$, one can sample a random matrix in $\{\pm 1\}^{n\times n}$ and keep the resulting matrix $\widehat{H}$ if $\widehat{H}\widehat{H}^T = n\mathbf{I}_n$. This approach is not viable using any reasonable measurement of the term: If

$\widehat{H} \in \text{Haar}(\{\pm 1\}^{n\times n}$, then

$$p_n = \mathbb{P}(\widehat{H}\mathbf{e}_1 \perp \widehat{H}\mathbf{e}_2) = \frac{\binom{n}{n/2}}{2^n} \tag{2.66}$$

as this is the probability a simple symmetric random walk is at 0 at step n (noting $4 \mid n$ yields p_n is nonzero, while if $2 \nmid n$ then $p_n = 0$). Hence, the probability $\widehat{H}$ has orthogonal columns is bounded above by p_n^{n-1}, which is the probability that the first column of $\widehat{H}$ is orthogonal to the remaining columns. Using Stirling's approximation formula,

$$p_n \sim \sqrt{\frac{2}{\pi n}} \tag{2.67}$$

so that

$$p = \mathbb{P}(\widehat{H}\widehat{H}^T = n\mathbf{I}_n) \leq p_n^{n-1} \lesssim n^{\frac{1}{2}(1-n)}. \tag{2.68}$$

For example, $n^{\frac{1}{2}(1-n)} \approx 8.651 \cdot 10^{-943}$ for $n = 668$ (the smallest order that has the potential of having a Hadamard matrix) where we note $p_{668} \approx 0.03086$. Hence, (using the Geometric(p) distribution) the expected number of trials needed to produce one Hadamard matrix using this method is

$$\frac{1}{p} \gtrsim n^{\frac{1}{2}(n-1)}. \tag{2.69}$$

For $n = 668$, this means (approximately) at least $1.156 \cdot 10^{942}$ trials would be needed on average for a first success. For comparison, there are an estimated 10^{82} particles in the observable universe. Considering there have been about $4.355 \cdot 10^{17}$ seconds since the Big Bang, this method is not reasonable to say the least, especially considering the lower bound in (2.69) is far from sharp.

Perhaps one can increase the chance of producing a Hadamard matrix using this crude

rejection method by instead choosing $\widehat{H} = \operatorname{sgn}(B)$ for $B \in \mathrm{B}(N)$ for $N = \prod_{j=1}^{k} p_j^{e_j}$. For $N = 2^n$, this will always produce a Hadamard matrix by Proposition 2.16. Unfortunately this does not work for any other p: Let $n = 2^k m$ for m odd. $\operatorname{sgn}(\mathrm{SO}(m))$ cannot have the first two columns orthogonal since a simple random walk cannot be at 0 at an odd step. It follows then $\frac{1}{\sqrt{n}} \operatorname{sgn}(B_1 \otimes B_2) = \frac{1}{\sqrt{m}} \operatorname{sgn}(B_1) \otimes 2^{-k/2} \operatorname{sgn}(B_2)$ for $B_1 \in \mathrm{SO}(m)$ and $B_2 \in \mathrm{B}(2^k)$ cannot be orthogonal since $B_1 B_1^T \neq m\mathbf{I}_m$. One can still increase the probability the first two columns are orthogonal by using $\operatorname{sgn}(H)$ for $H \sim \mathrm{SO}(n)$ for $4 \mid n$.

Example 2.5. *Using an experiment with* 10^6 *samples of* $\widehat{H} = \operatorname{sgn}(H)$ *for* $H \sim \mathrm{Haar}(\mathrm{O}(n))$ *for* $n = 12$*, the sample proportion* $\hat{p} = 0.305073$ *of trials such that the first two columns are orthogonal is larger than* $p_{12} = 0.2256$*. However, no Hadamard matrices were found in this experiment.*

This method does not seem very promising for any practical applications in the current state.

Additional applications of butterfly matrices to questions of interest involving Hadamard matrices will be further explored in Section 5.3.

Chapter 3

Random butterfly matrices

> A butterfly can flap its wings in Peking and in Central Park you get rain instead of sunshine.
>
> Ian Malcolm

The previous chapter considered deterministic butterfly matrices and their properties. Now we will consider an ensemble of orthogonal random matrices, using the same structures already established. This chapter focuses mostly on definitions and preliminary results relating to these random butterfly models, while Chapters 4 and 5 will look more into specific spectral and numerical properties and applications of these models.

3.1 Order $N = 2^n$ random butterfly matrices

We will again start first by considering $N = 2^n$, where we first start with the general definition for the matrices along with the corresponding butterfly factors.

Definition 3.1. *A **random butterfly matrix** is a butterfly matrix whose generating generalized rotation matrices are random matrices such that*

$$\Sigma := \{(C_j, S_j)\}_{j \geq 1} \tag{3.1}$$

is an independent sequence of pairs of matrices, where each pair (C_j, S_j) generates a random generalized rotation matrix of order 2^{j-1}. We will denote the set of random butterfly matrices of order N using the sequence of pairs of random matrices Σ collectively as $\mathrm{B}(N, \Sigma)$.

*We will use $\mathcal{D}^j(N, \Sigma)$ ($\mathcal{D}^j_s(N, \Sigma)$) to denote the corresponding **(simple) random butterfly block factor ensemble** and $\mathcal{R}(N, \Sigma)$ for the corresponding **random rotation ensemble** generated by Σ. Similarly, we will denote the corresponding **diagonal butterfly block factor ensembles** by $\mathcal{D}^j_d(N, \Sigma)$ and $\mathcal{D}^j_{ds}(N, \Sigma)$ and the **random diagonal rotation ensemble** by $\mathcal{R}_d(N, \Sigma)$.*

Note a random butterfly matrix $B \sim \mathrm{B}(N, \Sigma)$ is generated by a random generalized rotation matrix $R = R(C_n, S_n)$ and two iid random butterfly matrices $A_1, A_2 \sim \mathrm{B}(N/2, \Sigma)$.

Define the independent sequence of pairs of random matrices

$$\Sigma_S := \{(\cos\theta_j, \sin\theta_j)\mathbf{I}_{2^j} : \theta_j \sim \mathrm{Uniform}([0, 2\pi)) \text{ iid}, j \geq 1\}. \tag{3.2}$$

By Proposition 2.1, $\mathcal{R}(N) \cong \mathbb{T}$. Note also $\mathbb{T} \cong [0, 2\pi)/\sim$, the additive quotient space modulo 2π, using the map $\theta \to e^{i\theta}$. Since the uniform and Haar measures on $[0, 2\pi)/\sim$ coincide, then the resulting push-forward measure on $\mathcal{R}(N)$ induced by composing these isomorphisms is the Haar measure on $\mathcal{R}(N)$. This establishes the following result:

Proposition 3.1. *If $\theta \sim \mathrm{Uniform}([0, 2\pi))$, then $B(\theta) \otimes \mathbf{I}_{N/2} \sim \mathrm{Haar}(\mathcal{R}(N))$.*

Random ensembles formed using Σ_S will be particularly emphasized throughout the remain-

der of this document.

First, we can study $\mathcal{D}^j(N, \Sigma_S)$ and $\mathcal{D}_s^j(N, \Sigma_S)$. A straightforward argument now shows:

Proposition 3.2. *If* $2j \mid N$, $\mathcal{D}^j(N, \Sigma_S) \sim \mathrm{Haar}(\mathcal{D}^j(N))$ *and* $\mathcal{D}_s^j(N, \Sigma_S) \sim \mathrm{Haar}(\mathcal{D}_s^j(N))$.

Proof. This follows directly from Corollary 1.7 and Proposition 2.3. □

Definition 3.2. *We will refer to* $\mathrm{B}(N, \Sigma_S)$ *as the* ***random butterfly matrices*** *(dropping the* scalar *descriptor, and sometimes also dropping the* random *descriptor when it is unnecessary), and* $\mathrm{B}_s(N, \Sigma_S)$ *as the* ***Haar-butterfly matrices****.*

Note, the naming for $\mathrm{B}_s(N, \Sigma_S)$ is suggestive, particularly for this result:

Theorem 3.1 ([37])**.** $\mathrm{B}_s(N, \Sigma_S) \sim \mathrm{Haar}(\mathrm{B}_s(N))$

Proof. We will continually make use of the form

$$\mathrm{B}_s(N) = \bigotimes_{j=1}^{n} \mathrm{SO}(2). \tag{3.3}$$

Hence, the desired result follows from an application of Corollary 1.8 to $\mathrm{B}_s(N)$. □

The conclusions also follow from Corollary 1.7 and Proposition 3.2 with respect to the multiplication map $\prod_{j=1}^{n} \mathcal{D}_s^{2^j}(N) \to \mathrm{B}_s(N)$.

Remark 3.1 (Left invariance)**.** *The original proof presented in [37] used the following useful criteria, which verifies the left (and hence right since compact groups are unimodular) invariance of* $\mathrm{B}_s(N, \Sigma_S)$*: Let* $B(\boldsymbol{\theta}) \in \mathrm{B}_s(N)$ *and* $B(\boldsymbol{\varphi}) \sim \mathrm{B}_s(N, \Sigma_S)$*. Since* $\varphi_i \sim \mathrm{Uniform}([0, 2\pi))$ *then* $\theta_i + \varphi_i \pmod{2\pi} \sim \mathrm{Uniform}([0, 2\pi))$ *iid for all* $i = 1, \ldots, n$ *by Lemma 1.8. Hence,* $B(\boldsymbol{\theta})B(\boldsymbol{\varphi}) = B(\boldsymbol{\theta} + \boldsymbol{\varphi}) \sim \mathrm{B}_s(N, \Sigma_S)$*, using also Corollary 2.4.*

This can also be used to give an alternative proof of Theorem 3.1 in terms of the induced measure resulting from the multiplication map on $\mathrm{SO}(N)$.

We will also occasionally consider the **random diagonal butterfly matrices**, $\mathrm{B}(N, \Sigma_D)$ and $\mathrm{B}_s(N, \Sigma_D)$ where

$$\Sigma_D = \{\bigoplus_{k=1}^{2^j}(\cos\theta_j, \sin\theta_j) : \theta_j \sim \mathrm{Uniform}([0, 2\pi)) \text{ iid}, j \geq 1\}. \tag{3.4}$$

A similar argument would yield the diagonal butterfly diagonal factors (i.e., $\bigoplus^j Q\mathcal{R}(N/j)Q^T$ for Q a perfect shuffle as in (2.4) and (2.5)) sampled using Σ_D would be equal in distribution to the Haar diagonal butterfly diagonal factors. I will not explore this particular result further.

3.2 Random butterfly Hadamard matrices

One method of producing random Hadamard matrices of order $N = 2^n$ is to start with one Hadamard matrix, and then multiply each row by independent **Rademacher random variables** (these take values ± 1 each with probability $\frac{1}{2}$). Using Proposition 2.16, we can sample a random Hadamard matrix of order N by $H = \mathrm{sgn}(B)$ for $B \sim \mathrm{B}(N, \Sigma)$ or $B \sim \mathrm{B}_s(N, \Sigma)$ for $\Sigma = \Sigma_S$ or $\Sigma = \Sigma_D$. Since $\mathbb{P}(\bigcup_{i,j}[B_{ij} = 0]) = 0$ for $B \sim \mathrm{B}(N, \Sigma)$ or $B \sim \mathrm{B}_s(N, \Sigma)$ for $\Sigma = \Sigma_S$ or $\Sigma = \Sigma_D$, we then have $\mathbb{P}(\mathrm{sgn}(B) \in \{\pm 1\}^{N \times N}) = 1$. Hence, by Proposition 2.16 then $\mathrm{sgn}(B)$ is almost surely a Hadamard matrix. So we can now introduce:

Definition 3.3. *For $H = \mathrm{sgn}(B)$, then H is a **Haar-butterfly Hadamard matrix** if $B \sim \mathrm{B}_s(N, \Sigma_S)$, H is a **random butterfly Hadamard matrix** if $B \sim \mathrm{B}(N, \Sigma_S)$, H is a **random simple diagonal butterfly Hadamard matrix** if $B \sim \mathrm{B}_s(N, \Sigma_D)$ and H is a **random diagonal butterfly Hadamard matrix** if $B \sim B(N, \Sigma_D)$.*

If $B \sim \mathrm{B}_s(N, \Sigma_S)$, then $H = \mathrm{sgn}(B)$ would then be uniformly sampled from the $4^n = N^2$ distinct Haar-butterfly Hadamard matrices. Using $B \sim B(N, \Sigma_D)$, then $\mathrm{sgn}(B)$ would be uniformly sampled from the *at most* $4^{\frac{1}{2}nN} = N^N$ diagonal butterfly Hadamard matrices (using Proposition 2.7). Future work can determine the exact number of Hadamard matrices attainable by the butterfly Hadamard models.

3.3 Order $N = m^n$ random butterfly matrices

Considering then the general case, we can similarly define random m-butterfly G matrices:

Definition 3.4. *For G a class of groups such that $G(m)$ is a compact subgroup of $\mathrm{GL}(\mathbb{C}^m)$, a **random m-butterfly G matrix**, denoted $\mathrm{B}(m, n, \mathrm{Haar}(G))$, is of the form $\mathrm{Haar}(G(m))$ if $n = 1$ and*

$$(A \otimes \mathbf{I}_{N/m}) \bigoplus_{j=1}^{m} B_j, \tag{3.5}$$

*where $B_j \sim \mathrm{B}(m, n-1, \mathrm{Haar}(G))$ are iid and $A \sim \mathrm{Haar}(G(m))$ independent of the B_j if $n \geq 2$. The **random simple m-butterfly G matrices**, denoted by $\mathrm{B}_s(m, n, \mathrm{Haar}(G))$, are of the form*

$$\bigotimes_{j=1}^{m} B_j \tag{3.6}$$

*for $B_j \sim \mathrm{Haar}(G(m))$ iid. The **random m-butterfly matrices** and **Haar m-butterfly matrices**, denoted by $\mathrm{B}(m, n, \mathrm{Haar})$ and $\mathrm{B}_s(m, n, \mathrm{Haar})$, respectively, are the corresponding m-butterfly SO matrices.*

Similarly to Theorem 3.1, we have:

Proposition 3.3. $\mathrm{B}_s(m, n, \mathrm{Haar}(G)) \sim \mathrm{Haar}(\mathrm{B}_s(m, n, G))$

Proof. Use Corollary 1.8 and proposition 2.12. □

3.4 Order $N = \prod_{j=1}^{k} p_j^{e_j}$ random butterfly matrices

Continuing on with the structure of Chapter 2, we will explore random butterfly matrices of general orders. We will start with the p-nary structure before visiting the Cooley-Tukey analogous butterfly models.

Definition 3.5. *Let p be a prime number. The **Haar p-nary butterfly matrices** of order N, denoted by $\mathrm{B}(p, N, \mathrm{Haar})$, are of the form*

$$\mathrm{B}(p, N, \mathrm{Haar}) = \bigotimes_{j=0}^{\lfloor \log_p N \rfloor} \bigoplus^{a_{\lfloor \log_p N \rfloor - j + 1}} \mathrm{B}_s(p, \lfloor \log_p N \rfloor - j + 1, \mathrm{Haar}) \tag{3.7}$$

*where $N = \sum_{j \geq 0} a_j p^j$ for $a_j \in \{0, 1, \ldots, p-1\}$ for all j. The **Haar binary butterfly matrices** and **Haar ternary butterfly matrices** are, respectively, $\mathrm{B}(2, N, \mathrm{Haar})$ and $\mathrm{B}(3, N, \mathrm{Haar})$.*

Again, using a similar argument, we have

Proposition 3.4. $\mathrm{B}(p, N, \mathrm{Haar}) \sim \mathrm{Haar}(\mathrm{B}(p, N))$

Proof. Now use Corollaries 1.7, 1.8 and 2.5 and Proposition 3.3. □

We can next introduce the general random butterfly matrices, akin to the Cooley-Tukey structure:

Definition 3.6. *Let G denote a class of groups such that $G(m)$ is a compact subgroup of $\mathrm{GL}(\mathbb{C}^m)$ for any positive m. Let $N = \prod_{j=1}^{k} p_j^{e_j}$ be the prime factorization of N such that*

$p_j < p_{j+1}$ for each j. An order N ***random butterfly*** *G* ***matrix****, collectively denoted by* $\mathrm{B}(N, \mathrm{Haar}(G))$*, is a matrix of the form*

$$\bigotimes_{j=1}^{k} B_j \tag{3.8}$$

where $B_j \sim \mathrm{B}(p_j, e_j, \mathrm{Haar}(G))$ are mutually independent for each j. The ***Haar-butterfly*** *$G* ***matrices****, denoted by $\mathrm{B}_s(N, \mathrm{Haar}(G))$, are formed using independent $B_j \sim \mathrm{B}_s(p_j, e_j, \mathrm{Haar}(G))$ for each j. Let $\mathrm{B}(N, \mathrm{Haar}) = \mathrm{B}(N, \mathrm{Haar}(\mathrm{SO}))$ and $\mathrm{B}_s(N, \mathrm{Haar}) = \mathrm{B}_s(N, \mathrm{Haar}(\mathrm{SO}))$ denote the* ***random butterfly matrices*** *and the* ***Haar-butterfly matrices****.*

Remark 3.2. *Again, for $N = 2^n$, we can note the Definitions 3.1 and 3.6 are consistent for $\mathrm{B}(N, \Sigma_S) = \mathrm{B}(N, \mathrm{Haar})$ and $\mathrm{B}_s(N, \Sigma_S) = \mathrm{B}_s(N, \mathrm{Haar})$.*

A similar argument yields:

Proposition 3.5. $\mathrm{B}_s(N, \mathrm{Haar}(G)) \sim \mathrm{Haar}(\mathrm{B}_s(N, G))$.

Proof. Use Corollary 1.8 and Proposition 3.3. □

Chapter 4

Spectral properties of butterfly matrices

> Float like a butterfly, sting like a bee.
> The hands can't hit what the eyes
> can't see.
>
> Muhammad Ali

Typical questions in random matrix theory (RMT) focus on computing or estimating the spectra of random matrices. These approaches often use the empirical spectral distribution (ESD),

$$\mu_A = \frac{1}{n}\sum_{j=1}^{n} \delta_{\lambda_j(A)} \tag{4.1}$$

for $\lambda_j(A)$ the eigenvalues of A for $j = 1, \ldots, n$ when A is normal. In Section 1.7, famous results in RMT involve identifying universality principles for large ensembles of random matrices in terms of the limiting distributions for the ESDs (e.g., Theorem 1.16). It is often

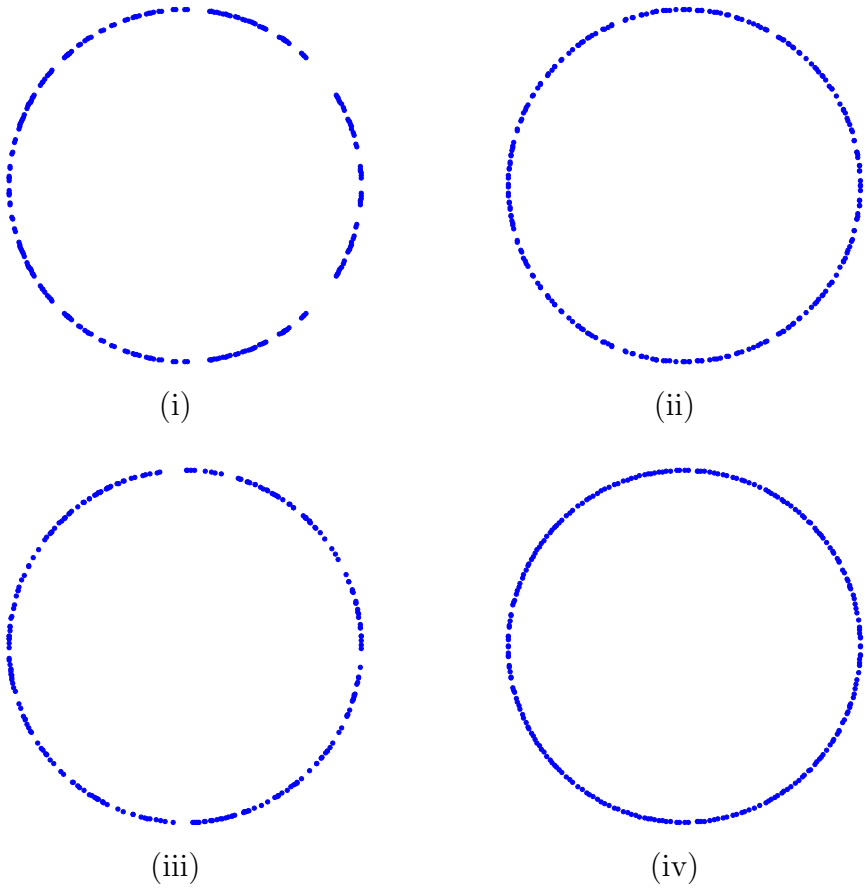


Figure 4.1: (i) $\mathrm{B}_s(2^8, \Sigma_S)$ eigenvalues, (ii) $\mathrm{B}(2^8, \Sigma_S)$ eigenvalues, (iii) 256 sampled Uniform($\mathbb{T}$) points, and (iv) eigenvalues for a Haar($\mathrm{O}(2^8)$) matrix. All plots are restricted to $\mathbb{T}$ in the complex plane.

difficult or not reasonable to give a closed form for the exact eigenvalues and eigenvectors for a random matrix. The limiting distributions of the ESDs can give a good approximation of the spectral picture for a large random matrix, but are not necessarily a good fit for smaller order matrices.

This chapter will give an overview of the spectral picture of random butterfly matrices. Figure 4.1 provides a comparison for the spectral pictures for certain order $2^8 = 256$ random butterfly matrices versus the uniform distribution on $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$ (showing 256 independently sampled uniform points) and the spectral picture a random Haar($\mathrm{O}(2^8)$) matrix. Of note, some random butterfly matrices, including the Haar-butterfly matrices,

have the rare property that their eigenvalues and eigenvectors can be computed explicitly.

4.1 Butterfly factors

We can first give the full spectral picture for both the general rotation and block diagonal butterfly factors. For generality, we only need to assume N is even for this section.

4.1.1 Scalar butterfly factors

First, we will consider deterministic scalar rotation matrices. For j such that $2j \mid N$, recall $\mathcal{D}^j(N) = \bigoplus^j \mathcal{R}(N/j)$ is an abelian subgroup of $\mathrm{SO}(N)$ by Proposition 2.3, as is $\mathcal{D}^j_s(N) = \mathbf{I}_j \otimes \mathcal{R}(N/j)$, which is a subgroup of $\mathcal{D}^j(N)$. Since normal matrices commute if and only if they are simultaneously diagonalizable (cf. Lemma 1.5), then it follows $\mathcal{D}^j(N)$ act on identical eigenspaces, that is, all such matrices have the same eigenvectors. In particular, we see the following result first for simple scalar rotation matrices:

Lemma 4.1. *If $A = R(\theta) \in \mathcal{R}(N)$ for N even, then A has eigenvalues $e^{\pm i\theta}$ with eigenvectors $\mathbf{v} = \mathbf{e}_j \pm i\mathbf{e}_{N/2+j}$ for $j = 1, \ldots, N/2$.*

Proof. Compute

$$\begin{aligned} A\mathbf{v} &= A(\mathbf{e}_j \pm i\mathbf{e}_{N/2+j}) = A\mathbf{e}_j \pm iA\mathbf{e}_{N/2+j} \\ &= (\cos\theta\mathbf{e}_j - \sin\theta\mathbf{e}_{N/2+j}) \pm i(\sin\theta\mathbf{e}_j + \cos\theta\mathbf{e}_{N/2+j}) \\ &= (\cos\theta \pm i\sin\theta)(\mathbf{e}_j \pm i\mathbf{e}_{N/2+j}) \\ &= e^{\pm i\theta}\mathbf{v}. \end{aligned}$$

□

Example 4.1. *The rotation matrix $B(\theta) \in \mathcal{R}(2) = \mathrm{SO}(2)$ has eigenvalues $e^{\pm i\theta}$ with eigenvectors*

$$\mathbf{e}_1 \pm i\mathbf{e}_2 = \begin{bmatrix} 1 \\ \pm i \end{bmatrix}.$$

An alternative proof to Lemma 4.1 uses the Kronecker product structure of $\mathcal{R}(N) = \mathrm{SO}(2) \otimes \mathbf{I}_{N/2}$. In particular,

$$R(\theta) = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} = Q\Lambda(\theta)Q^* \quad \text{for} \quad \Lambda(\theta) = e^{i\theta} \oplus e^{-i\theta} \quad \text{and} \quad Q = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix} \tag{4.2}$$

yields then

$$R_N(\theta) = R(\theta) \otimes \mathbf{I}_{N/2} = Q_N \Lambda_N(\theta) Q_N^* \tag{4.3}$$

for

$$\Lambda_N(\theta) = \Lambda(\theta) \otimes \mathbf{I}_{N/2} = \begin{bmatrix} e^{i\theta}\mathbf{I}_{N/2} & \\ & e^{-i\theta}\mathbf{I}_{N/2} \end{bmatrix} \quad \text{and} \quad Q_N = Q \otimes \mathbf{I}_{N/2} = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{I}_{N/2} & \mathbf{I}_{N/2} \\ i\mathbf{I}_{N/2} & -i\mathbf{I}_{N/2} \end{bmatrix} \tag{4.4}$$

using the mixed-product property.

A simple application yields the general result for the butterfly block factors $\mathcal{D}^j(N)$ for arbitrary j such that $2j \mid N$:

Corollary 4.1. *If $B(\boldsymbol{\theta}) \in \mathcal{D}^j(N)$ for $2j \mid N$, then $B(\boldsymbol{\theta})$ has eigenvalues $e^{\pm i\theta_k}$ with eigenvectors $\mathbf{e}_{(k-1)N/j+\ell} \pm i\mathbf{e}_{(k-1)N/j+N/2j+\ell}$ for $\ell = 1, \ldots, N/2j$, $k = 1, \ldots, j$.*

Proof. Apply Lemma 4.1 to the k^{th} block of $B(\boldsymbol{\theta})$. □

Alternatively, and building off of (4.3) we have

$$\mathcal{D}_s^j(N) = \mathbf{I}_j \otimes \mathcal{R}(N/j) = \mathbf{I}_j \otimes \mathrm{SO}(2) \otimes \mathbf{I}_{N/2j}. \tag{4.5}$$

For $B(\theta) \in \mathcal{D}_s^j(N)$, then

$$B(\theta) = Q_j^N \Lambda_{j,s}^N(\theta)(Q_j^N)^* \tag{4.6}$$

for

$$\Lambda_{j,s}^N(\theta) = \mathbf{I}_j \otimes \Lambda_{N/j}(\theta) = \mathbf{I}_j \otimes \Lambda(\theta) \otimes \mathbf{I}_{N/2j} \quad \text{and} \quad Q_j^N = \mathbf{I}_j \otimes Q_{N/j} = \mathbf{I}_j \otimes Q \otimes \mathbf{I}_{N/2j}. \tag{4.7}$$

Since $\mathcal{D}_s^j(N)$ is a subgroup of $\mathcal{D}^j(N)$, which is abelian, then Q_j^N is always consists of the eigenvectors for $B(\boldsymbol{\theta}) \in \mathcal{D}^j(N)$. In particular, we have

$$B(\boldsymbol{\theta}) = Q_j^N \Lambda_j^N(\boldsymbol{\theta})(Q_j^N)^* \tag{4.8}$$

for

$$\Lambda_j^N(\boldsymbol{\theta}) = \bigoplus_{k=1}^{j} \Lambda_{N/j}(\theta_k). \tag{4.9}$$

Note $\Lambda_{j,s}^N(\theta) = \Lambda_j^N(\theta \mathbf{1}_j)$.

This outlines the complete spectral picture for scalar butterfly factors in the deterministic case. Interestingly, this also almost completely takes care of the random case. Since $\mathcal{R}(N)$, $\mathcal{D}^j(N)$ and $\mathcal{D}_s^j(N)$ are abelian, then each is simultaneously diagonalizable, as evident in (4.3) and (4.8). It follows then *any random distribution* put on these models will always have *deterministic* eigenvectors.

In particular, if $B(\theta) \sim \mathrm{Haar}(\mathcal{R}(N))$, then (4.3) holds where $\theta \sim \mathrm{Uniform}([0, 2\pi))$. If

$B(\theta) \sim \text{Haar}(\mathcal{D}_s^j(N))$, then (4.6) holds where $\theta \sim \text{Uniform}([0, 2\pi))$. If $B(\boldsymbol{\theta}) \sim \text{Haar}(\mathcal{D}^j(N))$, then (4.8) holds where $\theta_j \sim \text{Uniform}([0, 2\pi))$ are iid for each j.

We can then combine these results into the following proposition:

Proposition 4.1. *Let $B(\boldsymbol{\theta}) \sim \mathcal{D}^j(N, \Sigma)$ for $\boldsymbol{\theta} \in [0, 2\pi)^j$. Then*

$$B(\boldsymbol{\theta}) = Q_j^N \Lambda_j^N(\boldsymbol{\theta})(Q_j^N)^* \tag{4.10}$$

for

$$\Lambda_j^N(\boldsymbol{\theta}) = \bigoplus_{k=1}^{j} \Lambda(\theta_k) \otimes \mathbf{I}_{N/2j} \quad \textit{and} \quad Q_j^N = \mathbf{I}_j \otimes Q \otimes \mathbf{I}_{N/2j}, \tag{4.11}$$

where

$$\Lambda(\theta) = \begin{bmatrix} e^{i\theta} & \\ & e^{-i\theta} \end{bmatrix} \quad \textit{and} \quad Q = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix}. \tag{4.12}$$

Moreover, Q_j^N is deterministic and independent of $\boldsymbol{\theta}$.

If $B(\theta) \sim \mathcal{D}_s^j(N, \Sigma)$, then

$$B(\theta) = Q_j^N \Lambda_{j,s}^N(\theta)(Q_j^N)^* \tag{4.13}$$

where

$$\Lambda_{j,s}^N(\theta) = \mathbf{I}_j \otimes \Lambda(\theta) \otimes \mathbf{I}_{N/2j}. \tag{4.14}$$

If $B(\theta) \sim \mathcal{R}(N, \Sigma)$, then

$$B(\theta) = (Q\Lambda(\theta)Q^*) \otimes \mathbf{I}_{N/2} = (Q \otimes \mathbf{I}_{N/2})(\Lambda(\theta) \otimes \mathbf{I}_{N/2})(Q \otimes \mathbf{I}_{N/2})^*. \tag{4.15}$$

Moreover, for each case above, the corresponding eigenvectors are deterministic.

4.1.2 Diagonal butterfly factors

Now we can consider the diagonal butterfly factors. A lot of the work has been done already by make appropriate connections to the scalar butterfly factors. Recall if U_N is the perfect shuffle such that

$$U_N(A \otimes \mathbf{I}_{N/2})U_N^T = \mathbf{I}_{N/2} \otimes A = \bigoplus^{N/2} A \tag{4.16}$$

for $A \in \mathbb{R}^{2\times 2}$, then for $R(\boldsymbol{\theta}) \in \mathcal{R}_d(N)$, we have

$$U_N^T R(\boldsymbol{\theta}) U_N = B(\boldsymbol{\theta}) \in \mathcal{D}^{N/2}(N). \tag{4.17}$$

In particular, $\mathcal{R}_d(N) \cong \mathcal{D}^{N/2}(N)$. Hence, for $R(\boldsymbol{\theta}) \in \mathcal{R}_d(N)$, we have

$$R(\boldsymbol{\theta}) = (U_N Q_{N/2}^N)\Lambda_{N/2}^N(\boldsymbol{\theta})(U_N Q_{N/2}^N)^* \tag{4.18}$$

using Proposition 4.1.

For $2j \mid N$, we can write

$$\mathcal{D}_{ds}^j(N) = \mathcal{R}_d(N/j) \otimes \mathbf{I}_j = (U_{N/j}\mathcal{D}^{N/2j}(N/j)U_{N/j}^T) \otimes \mathbf{I}_j. \tag{4.19}$$

By Proposition 4.1, it follows immediately for $B(\boldsymbol{\theta}) \in \mathcal{D}_{ds}^j(N)$ then

$$B(\boldsymbol{\theta}) = Q_{j,d}^N \left(\Lambda_{N/2j}^{N/j}(\boldsymbol{\theta}) \otimes \mathbf{I}_j\right)(Q_{j,d}^N)^* = Q_{j,d}^N \left(\bigoplus_{k=1}^{N/2j} \Lambda(\theta_k) \otimes \mathbf{I}_j\right)(Q_{j,d}^N)^* \tag{4.20}$$

where

$$Q_{j,d}^N = U_{N/j} Q_{N/2j}^{N/j} \otimes \mathbf{I}_j = (U_{N/j}(\mathbf{I}_{N/2j} \otimes Q)) \otimes \mathbf{I}_j. \tag{4.21}$$

Again, since $\mathcal{D}_{ds}^j(N)$ is a subgroup of the abelian group $\mathcal{D}_d^j(N)$, then the eigenvectors coincide for both groups. For $B(\boldsymbol{\theta}) \in \mathcal{D}^j(N)(d) = \bigoplus^j \mathcal{R}_d(N/j)$, write $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_j)$ for $\boldsymbol{\theta}_\ell \in [0, 2\pi)^{N/2j}$ for each ℓ. Hence, we have

$$\begin{aligned} B(\boldsymbol{\theta}) &= Q_{j,d}^N \left(\bigoplus_{\ell=1}^{j} \Lambda_{N/2j}^{N/j}(\boldsymbol{\theta}_\ell) \right) (Q_{j,d}^N)^* = Q_{j,d}^N \left(\bigoplus_{\ell=1}^{j} \bigoplus_{k=1}^{N/2j} \Lambda(\theta_{(\ell-1)N/2j+k}) \right) (Q_{j,d}^N)^* \\ &= Q_{j,d}^N \left(\bigoplus_{m=1}^{N/2} \Lambda(\theta_m) \right) (Q_{j,d}^N)^* \end{aligned} \tag{4.22}$$

Note the diagonal factor is independent of j.

Now considering the case for random instead of deterministic diagonal butterfly factors, again the picture is already mostly painted. Since $\mathcal{R}_d(N), \mathcal{D}_d^j(N)$ and $\mathcal{D}_{ds}^j(N)$ are abelian for each j, then the eigenvectors are necessarily the same as in the deterministic case. Similarly, as with the scalar case, any distribution put on these models is confined to the eigenvalues themselves. In particular, the Haar probability measure then necessarily coincides with the uniform measure on each input angle through the induced push-forward measure from the map $\boldsymbol{\theta} \mapsto B(\boldsymbol{\theta})$ as in (4.22).

I will summarize these results in the following proposition:

Proposition 4.2. *Let $B(\boldsymbol{\theta}) \sim \mathcal{D}_d^j(N, \Sigma)$ for $\boldsymbol{\theta} \in [0, 2\pi)^N$ Then*

$$B(\boldsymbol{\theta}) = Q_{j,d}^N \Lambda_d^N(\boldsymbol{\theta}) (Q_{j,d}^N)^* \tag{4.23}$$

for

$$\Lambda_d^N(\boldsymbol{\theta}) = \bigoplus_{m=1}^{N/2} \Lambda(\theta_m) \quad \textit{and} \quad Q_{j,d}^N = (U_{N/j}(\mathbf{I}_{N/2j} \otimes Q)) \otimes \mathbf{I}_j \tag{4.24}$$

where Q and $\Lambda(\theta)$ are defined by (4.12) *and U_m is the perfect shuffle permutation such that $U_m(A \otimes B)U_m^T = B \otimes A$ for A and B square matrices of orders 2 and $m/2$.*

If $B(\boldsymbol{\theta}) \sim \mathcal{D}_{ds}^j(N, \Sigma)$ for $\theta \in [0, 2\pi)^{N/2j}$, then

$$B(\boldsymbol{\theta}) = Q_{j,d}^N \left(\bigoplus_{k=1}^{N/2j} \Lambda(\theta_k) \otimes \mathbf{I}_j \right) (Q_{j,d}^N)^*. \tag{4.25}$$

If $B(\boldsymbol{\theta}) \sim \mathcal{R}_d(N, \Sigma)$ for $\boldsymbol{\theta} \in [0, 2\pi)^{N/2}$, then

$$B(\boldsymbol{\theta}) = (U_N(\mathbf{I}_{N/2} \otimes Q)) \left(\bigoplus_{k=1}^{N/2} \Lambda(\theta_k) \right) (U_N(\mathbf{I}_{N/2} \otimes Q))^*. \tag{4.26}$$

Moreover, for each case above, the corresponding eigenvectors are deterministic.

4.2 Order $N = 2^n$ butterfly matrices

Proposition 2.12 showed $\mathrm{B}_s(m, n, G)$ is a compact topological group when $G = \mathrm{SO}, \mathrm{O}, \mathrm{SU}$, or U, and it is abelian if and only if $m = 2$ and $G = \mathrm{SO}$. This shows then that $\mathrm{B}_s(2^n) = \mathrm{B}_s(2, n, \mathrm{SO})$ has a similar decomposition akin to the butterfly factors in that the abelian structure mandates the group is simultaneously diagonalizable and so share the same eigenvectors, which carries over (as with the butterfly factors) to the Haar-butterfly case.

4.2.1 Haar-butterfly matrices

Next, will we focus on Haar-butterfly matrices. These results will build on top of the corresponding results for the butterfly factors. In particular, the Kronecker product structure enables direct spectral factorizations of Haar butterfly matrices.

The computational advantage of this structure is illustrated in the following result:

Lemma 4.2. *Let* $\mathbf{u}$ *be an eigenvector for* A *with associated eigenvalue* λ*, and let*

$$B = \begin{bmatrix} \cos\theta A & \sin\theta A \\ -\sin\theta A & \cos\theta A \end{bmatrix} = B(\theta) \otimes A.$$

Then B *has eigenvectors* $\mathbf{v} = \begin{bmatrix} \mathbf{u} \\ \pm i\mathbf{u} \end{bmatrix}$ *with associated eigenvalues* $\lambda e^{\pm i\theta}$.

Proof. Compute

$$B\mathbf{v} = \begin{bmatrix} \cos\theta A\mathbf{u} \pm i\sin\theta A\mathbf{u} \\ -\sin\theta A\mathbf{u} \pm i\cos\theta A\mathbf{u} \end{bmatrix} = \lambda(\cos\theta \pm i\sin\theta)\mathbf{v} = \lambda e^{\pm i\theta}\mathbf{v}.$$

□

This can be used to show that for any $B \sim \mathrm{B}_s(N, \Sigma_S)$, the eigenvalues of B are stochastic and uniformly distributed on $\mathbb{T}$ while the eigenvectors are *deterministic*.

First, I will introduce some notation and an auxiliary function:

Definition 4.1. *For* $\mathbf{y} \in \mathbb{Z}^n$ *and* $a \in \mathbb{C}$*, write* $a^{\mathbf{y}} := [a^{y_1}, a^{y_2}, \ldots, a^{y_n}]^T \in \mathbb{C}^n$.

Definition 4.2. *Define* $f_n : \{-1, 1\}^n \to \mathbb{Z}^N$*, where we define* $\{-1, 1\}^0 := \{0\}$*, iteratively*

as follows: set $f_0(0) = 0$ *and for* $\mathbf{x} = (\mathbf{x}', x_n) \in \{-1, 1\}^n$*, then*

$$f_n(\mathbf{x}) = f_n(\mathbf{x}', x_n) = \begin{bmatrix} f_{n-1}(\mathbf{x}') \\ f_{n-1}(\mathbf{x}') + x_n \end{bmatrix}. \tag{4.27}$$

Example 4.2. *We see*

$$f_1(x_1) = [0, x_1]^T, \qquad f_2(x_1, x_2) = [0, x_1, x_2, x_1 + x_2]^T, \text{ and}$$
$$f_3(x_1, x_2, x_3) = [0, x_1, x_2, x_1 + x_2, x_3, x_1 + x_3, x_2 + x_3, x_1 + x_2 + x_3]^T.$$

Proposition 4.3. *For every* $\mathbf{x} \in \{-1, 1\}^n$ *with* $n \geq 1$ *and every* $B = B(\theta) \in \mathrm{B}_s(N)$*,* $i^{f_n(\mathbf{x})}$ *is an eigenvector for* B *with associated eigenvalue* $e^{i\boldsymbol{\theta}\cdot\mathbf{x}}$*. In particular,*

$$B(\boldsymbol{\theta}) = \left(\bigotimes^n Q\right)\left(\bigotimes_{j=1}^n \Lambda(\theta_{n-j+1})\right)\left(\bigotimes^n Q\right)^* \tag{4.28}$$

for Q *and* $\Lambda(\theta)$ *are defined by* (4.12)*.*

If $B \sim \mathrm{B}_s(N, \Sigma_S)$*, then the eigenvalues and eigenvectors are exactly as in the* $\mathrm{B}_s(N)$ *case, where each eigenvalue is uniformly distributed on* $\mathbb{T}$ *(the one-point correlation of* B*) and the associated eigenvector is deterministic.*

Proof. First, I will provide a direct argument using induction on n. The result follows from Lemma 4.1 for $n = 1$. Now assume the result holds for $\mathrm{B}_s(N/2)$. Fix $\mathbf{x} \in \{-1, 1\}^n$. Write $\boldsymbol{\theta} = (\boldsymbol{\theta}', \theta_n)$ and $\mathbf{x} = (\mathbf{x}', x_n)$ for $\theta' \in ([0, 2\pi))^{n-1}, \mathbf{x}' \in \{-1, 1\}^{n-1}$. By the inductive hypothesis, $\lambda = e^{i\theta'\cdot\mathbf{x}'}$ is an eigenvalue of $B(\boldsymbol{\theta}') \in \mathrm{B}_s(n-1)$ with associated eigenvector $\mathbf{u} = i^{f_{n-1}(\mathbf{x}')}$. By Lemma 4.2, $B(\boldsymbol{\theta})$ then has eigenvalue

$$\lambda e^{i\theta_n x_n} = e^{i\boldsymbol{\theta}'\cdot\mathbf{x}' + i\theta_n x_n} = e^{i\boldsymbol{\theta}\cdot\mathbf{x}}$$

with eigenvector

$$\mathbf{v} = \begin{bmatrix} \mathbf{u} \\ x_n i\mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{u} \\ i^{x_n}\mathbf{u} \end{bmatrix} = \begin{bmatrix} i^{f_{n-1}(\mathbf{x}')} \\ i^{f_{n-1}(\mathbf{x}')+x_n} \end{bmatrix} = i^{f_n(\mathbf{x})},$$

where we note $\pm i = i^{\pm 1}$.

When $B(\boldsymbol{\theta}) \sim \mathrm{B}_s(N, \Sigma_S)$, it only remains to show the eigenvalues are each uniform on $\mathbb{T}$, but this follows directly from Lemma 1.8 since $\theta_n, 2\pi - \theta_n \sim \mathrm{Uniform}([0, 2\pi))$, so that $\boldsymbol{\theta} \cdot \mathbf{x} \pmod{2\pi} \sim \mathrm{Uniform}([0, 2\pi))$.

Alternatively, and perhaps more naturally in light of the Kronecker product form of $\mathrm{B}_s(N)$: Let $B(\boldsymbol{\theta}) \in \mathrm{B}_s(N)$. Since

$$B(\theta) = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} = Q\Lambda(\theta)Q^* \tag{4.29}$$

then

$$\begin{aligned} B(\boldsymbol{\theta}) &= \bigotimes_{j=1}^{n} B(\theta_{n-j+1}) = \bigotimes_{j=1}^{n} Q\Lambda(\theta_{n-j+1})Q^* \\ &= \left(\bigotimes^{n} Q\right)\left(\bigotimes_{j=1}^{n} \Lambda(\theta_{n-j+1})\right)\left(\bigotimes^{n} Q\right)^* \end{aligned}$$

using the mixed-product property. □

Remark 4.1. *Since $\mathcal{D}_s^j(N)$ is a subgroup of $\mathrm{B}_s(N)$, which is abelian, then the eigenvectors of $\mathrm{B}_s(N)$ are also eigenvectors of $\mathcal{D}_s^j(N)$. This can be used to give a direct relationship between Proposition 4.1 and Proposition 4.3.*

For instance, the mixed-product property and (4.13) *show how the spectral decompositions factor, respectively: Let $R(\theta) \in \mathrm{SO}(2)$ denote the corresponding standard rotation matrix.*

Note $B_j(\theta) = \mathbf{I}_j \otimes R(\theta) \otimes \mathbf{I}_{N/2j} \in \mathcal{D}_s^j(N)$ *has factorization* $B_j(\theta) = Q_j^N \Lambda_{j,s}^N(\theta)(Q_j^N)^*$ *for* $Q_j^N = \mathbf{I}_j \otimes Q \otimes \mathbf{I}_{N/j}$ *and* $\Lambda_{j,s}^N(\theta) = \mathbf{I}_j \otimes \Lambda(\theta) \otimes \mathbf{I}_{N/j}$ *by Proposition 4.1. Hence,*

$$B(\boldsymbol{\theta}) = \prod_{j=1}^{n} B_{n-j+1}(\theta_{n-j+1}) \in \mathrm{B}_s(N)$$

for $B_k(\theta) \in \mathcal{D}_s^j(N)$ *for each* k*, and so*

$$\bigotimes_{j=1}^{n} Q = \prod_{j=1}^{n} Q_j^N \quad \textit{and} \quad \bigotimes_{j=1}^{n} \Lambda(\theta_{n-j+1}) = \prod_{j=1}^{n} \Lambda_{j,s}^N(\theta_{n-j+1}).$$

Example 4.3. *A straightforward application of Proposition 4.3 yields*

$$B(\theta,\varphi) = \left[\begin{array}{cc|cc} \cos\varphi\cos\theta & \cos\varphi\sin\theta & \sin\varphi\cos\theta & \sin\varphi\sin\theta \\ -\cos\varphi\sin\theta & \cos\varphi\cos\theta & -\sin\varphi\sin\theta & \sin\varphi\cos\theta \\ \hline -\sin\varphi\cos\theta & -\sin\varphi\sin\theta & \cos\varphi\cos\theta & \cos\varphi\sin\theta \\ \sin\varphi\sin\theta & -\sin\varphi\cos\theta & -\cos\varphi\sin\theta & \cos\varphi\cos\theta \end{array}\right]$$

has eigenvalues $e^{i(\theta+\varphi)}, e^{i(\theta-\varphi)}, e^{i(-\theta+\varphi)}, e^{i(-\theta-\varphi)}$ *with respective eigenvectors*

$$i^{f_2(1,1)} = \begin{bmatrix} 1 \\ i \\ i \\ -1 \end{bmatrix}, \; i^{f_2(1,-1)} = \begin{bmatrix} 1 \\ i \\ -i \\ 1 \end{bmatrix}, \; i^{f_2(-1,1)} = \begin{bmatrix} 1 \\ -i \\ i \\ 1 \end{bmatrix}, \; i^{f_2(-1,-1)} = \begin{bmatrix} 1 \\ -i \\ -i \\ -1 \end{bmatrix}.$$

If $B(\theta,\varphi) \sim \mathrm{B}_s(4,\Sigma_S)$*, then each eigenvalue is uniformly distributed on* $\mathbb{T}$ *(but* not *jointly, as they must be complex conjugates) and the associated eigenvector is deterministic.*

4.2.2 Nonsimple scalar butterfly matrices

Proposition 4.3 shows the eigenvalues and eigenvectors of any simple scalar butterfly matrix, either random or nonrandom, can be completely determined in a closed form. The case for nonsimple scalar butterfly matrices is a different matter altogether. The case for $n = 2$ is simpler than $n \geq 3$, as the following lemma will show. This also suggests how the higher order matrices have have progressively more computationally complex spectra pictures.

Lemma 4.3. (a) *Let* $\mathbf{u}$ *be an eigenvector for both* A_1 *and* A_2*, for respective associated eigenvalues of* λ_1, λ_2*, and let*

$$B = \begin{bmatrix} \cos\theta A_1 & \sin\theta A_2 \\ -\sin\theta A_1 & \cos\theta A_2 \end{bmatrix}.$$

If $\sin\theta = 0$ *or* $\lambda_1 = 0$*, then* $[\mathbf{u}, \mathbf{0}]^T$ *is an eigenvector for* B *with eigenvalue* $\lambda_1 \cos\theta$*. If* $\sin\theta = 0$ *or* $\lambda_2 = 0$*, then* $[\mathbf{0}, \mathbf{u}]^T$ *is an eigenvector for* B *with eigenvalue* $\lambda_2 \cos\theta$*. If* $\sin\theta, \lambda_1, \lambda_2$ *are all nonzero, then* B *has eigenvalue*

$$\lambda_1 \cos\theta + \alpha\lambda_2 \sin\theta$$

with eigenvector $\mathbf{v} = \begin{bmatrix} \mathbf{u} \\ \alpha\mathbf{u} \end{bmatrix}$*, where*

$$\alpha = \alpha(\pm) = \frac{1}{2}\left[\left(1 - \frac{\lambda_1}{\lambda_2}\right)\cot\theta \pm \sqrt{\left(1 - \frac{\lambda_1}{\lambda_2}\right)^2 \cot^2\theta - 4\frac{\lambda_1}{\lambda_2}}\right],$$

using the principle branch of the logarithm to define the square root.

(b) *When* $\sin\theta \neq 0$ *and* $\lambda_1, \lambda_2 \in \mathbb{T}$*, then the discriminant*

$$\Delta := \left(1 - \frac{\lambda_1}{\lambda_2}\right)^2 \cot^2\theta - 4\frac{\lambda_1}{\lambda_2} \tag{4.30}$$

is nonzero.

Remark 4.2. *When the discriminant is nonzero, the two choices of* α *lead to two distinct eigenvalues and eigenvectors of* B*, with potentially two additional eigenvalues and eigenvectors obtained by taking complex conjugates.*

Proof of Lemma 4.3. (a) Since

$$B\begin{bmatrix}\mathbf{u}\\ \mathbf{0}\end{bmatrix} = \begin{bmatrix}\lambda_1\cos\theta\mathbf{u}\\ -\lambda_1\sin\theta\mathbf{u}\end{bmatrix} \quad \text{and} \quad B\begin{bmatrix}\mathbf{0}\\ \mathbf{u}\end{bmatrix} = \begin{bmatrix}\lambda_2\sin\theta\mathbf{u}\\ \lambda_2\cos\theta\mathbf{u}\end{bmatrix},$$

we see $[\mathbf{u}, \mathbf{0}]^T$ is an eigenvector for B with eigenvalue $\lambda_1\cos\theta$ if and only if $\lambda_1\sin\theta = 0$, and $[\mathbf{0}, \mathbf{u}]^T$ is an eigenvector for B with eigenvalue $\lambda_2\cos\theta$ if and only if $\lambda_2\sin\theta = 0$.

Now assume $\lambda_1\lambda_2\sin\theta \neq 0$, and assume $\alpha \neq 0$. Now we compute

$$B\mathbf{v} = \begin{bmatrix}\cos\theta A_1\mathbf{u} + \alpha\sin\theta A_2\mathbf{u}\\ -\sin\theta A_1\mathbf{u} + \alpha\cos\theta A_2\mathbf{u}\end{bmatrix} = \begin{bmatrix}(\lambda_1\cos\theta + \alpha\lambda_2\sin\theta)\mathbf{u}\\ (-\alpha^{-1}\lambda_1\sin\theta + \lambda_2\cos\theta)\alpha\mathbf{u}\end{bmatrix}.$$

We then have $\mathbf{v}$ is an eigenvector for B for eigenvalue $\lambda_1\cos\theta + \alpha\lambda_2\sin\theta$ if and only if

$$\lambda_1\cos\theta + \alpha\lambda_2\sin\theta = -\alpha^{-1}\lambda_1\sin\theta + \lambda_2\cos\theta$$

if and only if

$$\alpha^2 - \left(1 - \frac{\lambda_1}{\lambda_2}\right)\cot\theta\alpha + \frac{\lambda_1}{\lambda_2} = 0. \tag{4.31}$$

The result then follows by using the quadratic formula to solve (4.31) for α to find $\alpha(+)$ and $\alpha(-)$.

(b) Suppose $\Delta = 0$. We would have $\lambda := \frac{\lambda_1}{\lambda_2} \in \mathbb{T}$ satisfies $4\lambda = (1-\lambda)^2 \cot^2\theta$, with also necessarily $\cos\theta \neq 0$, and hence

$$\lambda^2 - 2(1 + 2\tan^2\theta)\lambda + 1 = 0.$$

Again using the quadratic formula, it would follow then

$$\lambda = (1 + 2\tan^2\theta) \pm \sqrt{(1 + 2\tan^2\theta)^2 - 1}.$$

Since $1 + 2\tan^2\theta > 1$ and $1 + 2\tan^2\theta > \sqrt{(1 + 2\tan^2\theta)^2 - 1}$, then $\lambda > 0$, so that necessarily $1 = |\lambda| = \lambda$. But then

$$4 = 4\lambda = (1-\lambda)^2 \cot^2\theta = 0,$$

a contradiction. □

Note Lemma 4.3(a) gives an alternative proof of Lemma 4.2, where we let $A_1 = A_2$ so that $\lambda_1 = \lambda_2 := \lambda$ and hence $\alpha = \pm i$ when $\lambda \neq 0$; otherwise, $[\mathbf{u}, \mathbf{0}]^T$ and $[\mathbf{0}, \mathbf{u}]^T$ and hence $[\mathbf{u}, \pm i\mathbf{u}]^T$ is an eigenvector for $\lambda = 0$.

Since $\mathrm{B}(2) = \mathrm{B}_s(2) = \mathrm{SO}(2)$ all have the same eigenvectors by Proposition 4.3, then we have the eigenvalues and eigenvectors for any $B \in \mathrm{B}(2)$ can be determined explicitly by Lemma 4.3.

Example 4.4. *Suppose* $\sin\psi \neq 0$. *Define the nonsimple butterfly matrix*

$$B = B(\theta, \varphi, \psi) = \begin{bmatrix} \cos\psi A_\theta & \sin\psi A_\varphi \\ -\sin\psi A_\theta & \cos\psi A_\varphi \end{bmatrix}$$

$$= \left[\begin{array}{cc|cc} \cos\psi\cos\theta & \cos\psi\sin\theta & \sin\psi\cos\varphi & \sin\psi\sin\varphi \\ -\cos\psi\sin\theta & \cos\psi\cos\theta & -\sin\psi\sin\varphi & \sin\psi\cos\varphi \\ \hline -\sin\psi\cos\theta & -\sin\psi\sin\theta & \cos\psi\cos\varphi & \cos\psi\sin\varphi \\ \sin\psi\sin\theta & -\sin\psi\cos\theta & -\cos\psi\sin\varphi & \cos\psi\cos\varphi \end{array}\right].$$

Let $\mathbf{u} = \begin{bmatrix} 1 \\ i \end{bmatrix}$. *Then* B *has eigenvalue(s)*

$$e^{i\theta}\cos\psi + \alpha e^{i\varphi}\sin\psi$$

with respective eigenvector(s) $\mathbf{v} = \begin{bmatrix} \mathbf{u} \\ \alpha\mathbf{u} \end{bmatrix}$ *where*

$$\alpha = \alpha(\pm) = \frac{1}{2}\left[(1 - e^{i(\theta-\varphi)})\cot\psi \pm \sqrt{(1 - e^{i(\theta-\varphi)})^2\cot^2\psi - 4e^{i(\theta-\varphi)}}\right].$$

If $\theta = \varphi$, *so that* B *is simple, then* $\alpha(\pm) = \pm i$.

Since $\|\mathbf{u}\|_2^2 = 2$ *then* $\|\mathbf{v}\|_2^2 = 2(1 + |\alpha|^2)$

Note also necessarily when $\sin\psi \neq 0$ *the discriminant is nonzero by Lemma 4.3(b). Hence, it follows all eigenvalues and eigenvectors for* B *can be determined explicitly from the above construction, when considering also complex conjugates.*

Explicitly determining the eigenvalues and eigenvectors for $B \in \mathrm{B}(N)$ or $B \sim \mathrm{B}(N, \Sigma_S)$ when $n \geq 3$ is not as straightforward. In particular, we can no longer use Lemma 4.3 since B(4) does not preserve eigenspaces, which should be expected since multiplication on B(2) is not commutative (nor is $\mathrm{B}(N)$ a group). As such, it follows $\mathrm{B}(4, \Sigma_S)$ has both stochastic eigenvalues and eigenvectors, while only the eigenvalues for $\mathrm{B}(2, \Sigma_S)$ were stochastic. But the rigid structure of $\mathrm{B}(N)$ forces some rigidity to carry over to the structure of the eigenvectors.

A closed form may be harder to attain for $n \geq 3$, but repeated empirical results suggest:

Conjecture 3. *Let $B = B(\boldsymbol{\theta}) \in \mathrm{B}(N)$ with*

$$B = R_N(\theta_{N-1}) \begin{bmatrix} B(\boldsymbol{\theta}_1) & \\ & B(\boldsymbol{\theta}_2) \end{bmatrix}$$

for $B(\boldsymbol{\theta}_1), B(\boldsymbol{\theta}_2) \in \mathrm{B}(N/2)$. For $\mathbf{u} = \begin{bmatrix} 1 \\ i \end{bmatrix}$, then all eigenvectors of B are of the form or are (complex) conjugate to the form

$$\mathbf{v} = \begin{bmatrix} \alpha_1 \mathbf{u} \\ \alpha_2 \mathbf{u} \\ \vdots \\ \alpha_{N/2} \mathbf{u} \end{bmatrix} \tag{4.32}$$

for some $\alpha_i \neq 0$ when $\sin\theta_{N-1} \neq 0$, otherwise they are completely determined by the eigenvectors of $B(\boldsymbol{\theta}_1)$ and $B(\boldsymbol{\theta}_2)$.

Example 4.5. *Sampling $B \in \mathrm{B}_s(8, \Sigma_S)$, we have*

$$B = \left[\begin{array}{cc|cc|cc|cc} 0.2793 & -0.4539 & 0.3040 & 0.4761 & 0.4842 & -0.1295 & 0.1166 & -0.3633 \\ 0.4539 & 0.2793 & -0.4761 & 0.3040 & 0.1295 & 0.4842 & 0.3633 & 0.1166 \\ \hline 0.2960 & -0.4811 & -0.2868 & -0.4492 & -0.3686 & 0.0986 & 0.1532 & -0.4773 \\ 0.4811 & 0.2960 & 0.4492 & -0.2868 & -0.0986 & -0.3686 & 0.4773 & 0.1532 \\ \hline 0.2266 & -0.3682 & 0.2466 & 0.3862 & -0.5970 & 0.1597 & -0.1437 & 0.4479 \\ 0.3682 & 0.2266 & -0.3862 & 0.2466 & -0.1597 & -0.5970 & -0.4479 & -0.1437 \\ \hline 0.2401 & -0.3902 & -0.2326 & -0.3644 & 0.4544 & -0.1215 & -0.1888 & 0.5884 \\ 0.3902 & 0.2401 & 0.3644 & -0.2326 & 0.1215 & 0.4544 & -0.5884 & -0.1888 \end{array}\right]. \tag{4.33}$$

B has eigenvectors of the form in (4.32) *where* $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 \end{bmatrix}^T$ *are found in*

$$\left\{ \begin{bmatrix} 0.573i \\ 0.2931 + 0.1435i \\ -0.0045 + 0.2001i \\ 0.0461 + 0.1515i \end{bmatrix}, \begin{bmatrix} -0.0293 + 0.1454i \\ -0.0132 + 0.2131i \\ 0.2716 - 0.0963i \\ -0.5911i \end{bmatrix}, \begin{bmatrix} -0.1715 - 0.2108i \\ 0.2573 + 0.0993i \\ 0.5410 \\ -0.0458 + 0.2353i \end{bmatrix}, \begin{bmatrix} -0.2591 + 0.0931i \\ -0.5214i \\ 0.0979 + 0.2733i \\ -0.2123 - 0.1518i \end{bmatrix} \right\}.$$

4.3 Almost uniform eigenvalue distribution

A natural direction in RMT with respect to random butterfly matrices is to determine the limiting distribution of the butterfly matrices of increasing order. This section builds off directly of results found in [37].

First, we will establish some standard background necessary for the following steps.

Definition 4.3. *A random matrix M_n has a* ***uniform eigenvalue distribution*** *if $\mathbb{E}\mu_{M_n}$ is the uniform measure on $\mathbb{T}$.*

A useful criterion for this is:

Theorem 4.1 ([37])**.** *A random unitary matrix Q of order N has uniform eigenvalue distribution if and only if*

$$\mathbb{E}\frac{1}{N}\operatorname{Tr} Q^k = \delta_{k0} \tag{4.34}$$

for $k = 0, 1, \ldots$

Definition 4.4. *For a sequence of matrices, M_n, the eigenvalues of the sequence are said to*

be ***almost surely uniform*** *if for each* $f \in C(\mathbb{T})$ *then*

$$\lim_{n\to\infty} \int_{\mathbb{T}} f \ \mathrm{d}\,\mu_{M_n} = \frac{1}{2\pi i} \int_{\mathbb{T}} f(z) \ \frac{\mathrm{d}\,z}{z} \quad \textit{almost surely,} \tag{4.35}$$

where μ_{M_n} *is the empirical spectral distribution of* M_n*.*

Again, I will restate a useful classical criterion (also given in [37]), which follows almost directly from the Borel-Cantelli lemma:

Theorem 4.2. *Let* M_n *be a sequence of random matrices of order* $N = N(n)$ *that is strictly increasing such that* M_n *has uniform eigenvalue distribution for each* n*. Suppose if for each* $k = 1, 2, \ldots$ *then*

$$\mathbb{E}\left|\frac{1}{N}\operatorname{Tr} M_n^k\right|^2 \leq C_k n^{1-c_k} \tag{4.36}$$

for absolute positive constants C_k, c_k*. Then the eigenvalues of the sequence* M_n *are almost surely uniform.*

We can reproduce the result found in [37] regarding Haar-butterfly matrices of order 2^n using a different argument using the tools established earlier:

Theorem 4.3 ([37])**.** *Let* $B_n \sim \mathrm{B}_s(2^n, \Sigma_S)$*. Then the sequence* B_n *has almost uniform eigenvalues.*

Proof. For an alternative but equivalent proof from that found in [37], recall $\operatorname{Tr}(A \otimes B) = \operatorname{Tr}(A)\operatorname{Tr}(B)$. Hence, if $B(\boldsymbol{\theta}) \sim \mathrm{B}_s(N, \Sigma_S)$, then

$$\begin{aligned}\operatorname{Tr}(B(\boldsymbol{\theta})^k) &= \operatorname{Tr}\left(\bigotimes_{j=1}^{n} B(\theta_{n-j+1})^k\right) = \prod_{j=1}^{n} \operatorname{Tr}(B(\theta_{n-j+1})^k) = \prod_{j=1}^{n} 2\cos k\theta_j \\ &= N \prod_{j=1}^{N} \cos k\theta_j.\end{aligned}$$

Since $\cos k\theta_j \sim \cos\theta_1$ if $\theta_j \sim \text{Uniform}([0, 2\pi))$ are iid, it follows

$$\mathbb{E}\left|\frac{1}{N}\operatorname{Tr}(B(\boldsymbol{\theta})^k)\right|^2 = \mathbb{E}\left|\prod_{j=1}^{n}\cos k\theta_j\right|^2 = (\mathbb{E}\cos^2\theta_1)^n = 2^{-n}. \tag{4.37}$$

The result then follows from Theorem 4.2. □

Next, we can extend this result to include the more general case with $\mathrm{B}_s(m, n, G)$.

Theorem 4.4. *Let* $B_n \sim \text{Haar}(\mathrm{B}_s(m, n, G)) = \mathrm{B}_s(m, n, \text{Haar}(G))$ *for* $G = \mathrm{SO}, \mathrm{O}, \mathrm{SU}$ *or* U. *Then* B_n *has almost uniform eigenvalues.*

Proof. Note if $A \sim \text{Haar}(G(m))$ for G as above, then

$$\frac{1}{m}\operatorname{Tr}(A^k) = \frac{1}{m}\sum_{\ell=1}^{m}\lambda_\ell^k \tag{4.38}$$

for λ_ℓ the eigenvalues of A, each of which is on $\mathbb{T}$ as is λ_ℓ^k for each k. Since this is then a convex combination of m points on $\mathbb{T}$, with the points being almost surely distinct, then $\left|\frac{1}{m}\operatorname{Tr}(A^k)\right| < 1$ almost surely. We can say a lot more than this, though.

In [7] (which corrects similar results in [10]), we have if $A \sim \text{Haar}(\mathrm{U}(m))$, then $\mathbb{E}|\operatorname{Tr}(A^k)|^2 = \min(m, k)$ then $B = \bigotimes_j^n B_j$ for $B_j \sim \text{Haar}(\mathrm{U}(m))$ iid, we have

$$\mathbb{E}|\operatorname{Tr}(B^k)|^2 = \prod_{j=1}^{n}\mathbb{E}|\operatorname{Tr}(B_j^k)|^2 = \min(m, k)^n \leq N. \tag{4.39}$$

Similarly, $\mathbb{E}|\operatorname{Tr}(B^k)|^2 \leq N$ for $B \sim \text{Haar}(\mathrm{B}_s(m, n, G))$ for $G = \mathrm{SO}, \mathrm{O}$ and SU (e.g., see also [19]). It follows then for $B \sim \text{Haar}(\mathrm{B}_s(m, n, G))$ for $G = \mathrm{SO}, \mathrm{O}, \mathrm{SU}$ or U we have

$$\mathbb{E}\left|\frac{1}{N}\operatorname{Tr} B^k\right|^2 \leq \frac{1}{N} = m^{-n}. \tag{4.40}$$

The result then follows again from Theorem 4.2. □

Note, however, this argument does not carry over directly to $\mathrm{Haar}(\mathrm{B}_s(N))$ for general $N = \prod_{j=1}^{k} p_j^{e_j}$: the upper bound does not decay sufficiently fast enough with respect to the growth of the orders of each sequence term, as the order $\frac{1}{N}$ remains the same scaling with respect to the sequence order growth. This is unlike when using the exponentially growing subsequences in Theorem 4.4 that then led to exponential decay. Convergence in expectation does carry over, at least.

Corollary 4.2. *Let $B_n \sim \mathrm{Haar}(\mathrm{B}_s(n, G))$ for $G = \mathrm{SO}, \mathrm{O}, \mathrm{SU}$ or U. Then μ_{B_n} converges in expectation to the uniform measure on $\mathbb{T}$.*

A sparsification argument may be able to strengthen this result. Future work can explore the potential of this direction.

4.4 CLT for moments of the trace

Since $\mathrm{Tr}(A \otimes B) = \mathrm{Tr}(A)\,\mathrm{Tr}(B)$, for $B \sim \mathrm{Haar}(\mathrm{B}_s(m, n, G))$, then $\mathrm{Tr}(B^k) = \prod_{j=1}^{n} \mathrm{Tr}(B_j^k)$ for $B_j \sim \mathrm{Haar}(G(m))$ iid. Furthermore, note $B \sim B^{-1}$ so that $\mathrm{Tr}(B^{-k}) \sim \mathrm{Tr}(B^k)$ for $k \in \mathbb{Z}\backslash\{0\}$ and $B \sim \mathrm{Haar}(\mathrm{B}_s(m, n, G))$. Since $\mathbb{E} \log^2 |\,\mathrm{Tr}(B_1^k)| < \infty$ for each m and $G = \mathrm{SO}, \mathrm{O}, \mathrm{SU}, \mathrm{U}$, we can apply the Central Limit Theorem to yield:

Theorem 4.5. *Let $B \sim \mathrm{Haar}(\mathrm{B}_s(m, n, G))$ for $B = \bigotimes_{j=1}^{n} B_j$ where $B_j \sim \mathrm{Haar}(G(m))$ iid for $G = \mathrm{SO}, \mathrm{O}, \mathrm{SU}$ or U. For $k \in \mathbb{Z}\backslash\{0\}$, let $\mu_k = \mathbb{E} \log |\,\mathrm{Tr}(B_1^k)|$ and $\sigma_k^2 = \mathrm{Var}(\log |\,\mathrm{Tr}(B_1^k)|)$, then for any real t we have*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\log |\,\mathrm{Tr}(B^k)| - n\mu_k}{\sqrt{n}\sigma_k} \leq t \right) = \mathbb{P}(Z \leq t) \tag{4.41}$$

for $Z \sim N(0, 1)$.

Example 4.6. *Consider $\mathrm{B}_s(2^n, \Sigma_S)$. This is again expanding on a result in [37]. Since*

$\cos k\theta \sim \cos\theta$ *for* $\theta \sim \text{Uniform}([0, 2\pi))$ *and* $k = 1, 2, \ldots,$ *we have*

$$\mu_k = \mathbb{E}\log|\operatorname{Tr}(B(\theta)^k)| = \mathbb{E}\log|2\cos k\theta| = \mathbb{E}\ln|\cos\theta| + \log 2 = 0 \tag{4.42}$$

and

$$\begin{aligned}
\sigma_k^2 &= \mathbb{E}\log^2|\operatorname{Tr}(B(\theta)^k)| = \mathbb{E}\log^2|2\cos k\theta| = \mathbb{E}\log^2|2\cos\theta| \\
&= \log^2 2 + (2\log 2)\mathbb{E}\log|\cos\theta| + \mathbb{E}\log^2|\cos\theta| \\
&= \log^2 2 - 2\log^2 2 + \frac{2}{\pi}\left(\frac{\pi^3}{24} + \frac{\pi}{2}\log^2 2\right) \\
&= \frac{\pi^2}{12}.
\end{aligned}$$

As is true with CLTs, the universal bell curve shape does not emerge until n *is sufficiently large, as evidenced by Figure 4.2.*

Remark 4.3. *Theorem 4.5 can be used to get asymptotics on quantiles for* $\operatorname{Tr}(B^k)$*. I will postpone a further exploration until a similar result after Corollary 5.3.*

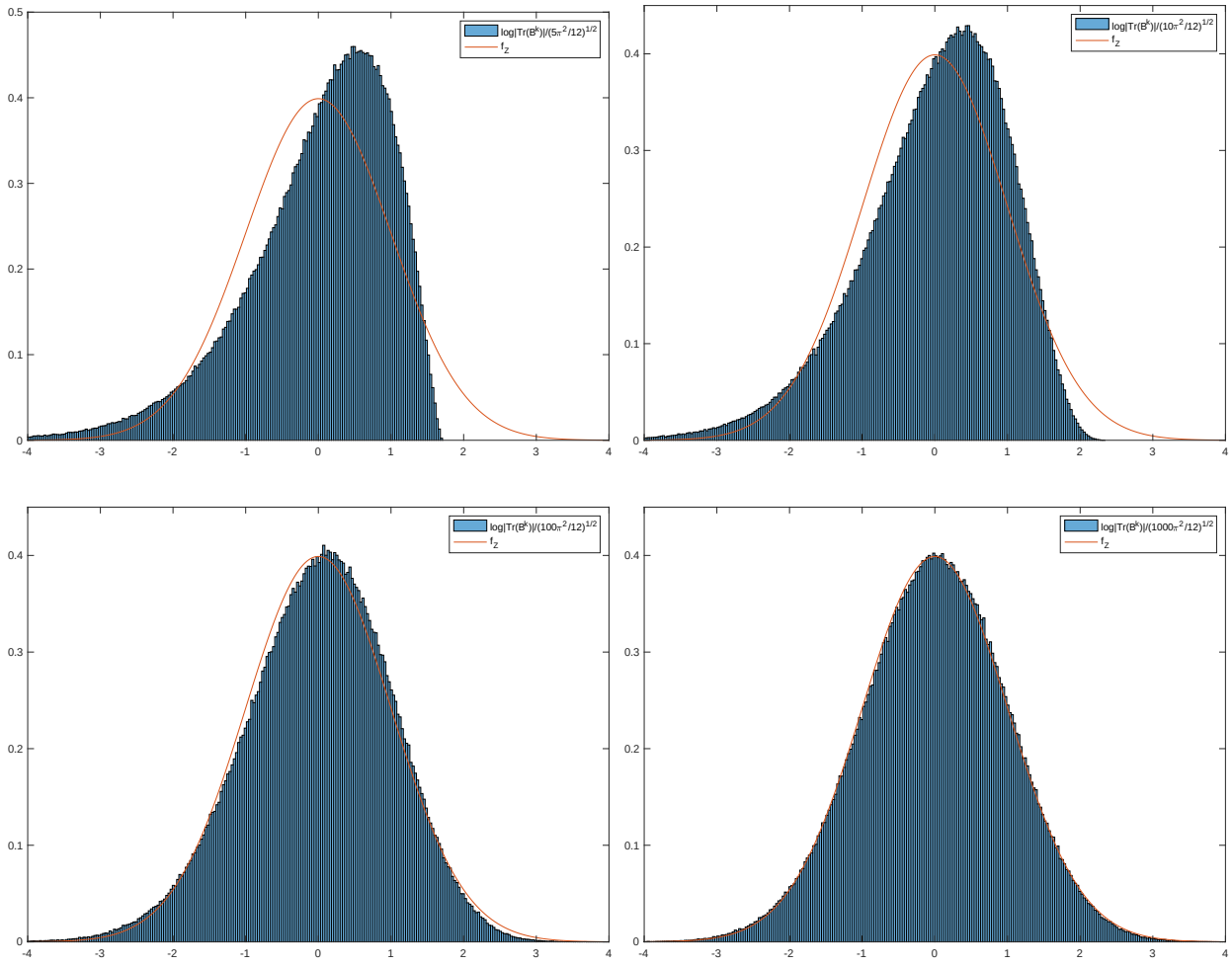


Figure 4.2: Histogram of $\frac{1}{\sqrt{n \cdot \frac{\pi^2}{12}}} \log |\operatorname{Tr}(B^k)|$ versus the standard normal density function f_Z for $Z \sim N(0,1)$ using 10^6 samples for $B \sim \mathrm{B}_s(2^n, \Sigma_S)$ for $n = 5, 10, 100, 1000$

Chapter 5

Numerical properties of butterfly matrices

> I know not if I was then a man
> dreaming I was a butterfly, or if I am
> now a butterfly dreaming I am a man.
>
> René Descartes

This chapter will focus on applications of butterfly matrices to reduce complexity in common applications in numerical linear algebra. Parker's original focus was on the application of butterfly matrices to transform a nonsingular matrix to a nondegenerate matrix so Gaussian elimination (GE) can be carried out without pivoting. A general overview of this application is presented here. The remainder of this chapter will focus almost exclusively on the $N = 2^n$ butterfly models. Section 5.2 will outline the hierarchy in randomness moving from $\mathrm{B}_s(N)$ to Haar(O(N)), which will include a method of sampling from Haar(O(N)) using butterfly matrices. A method to approximate Haar(O(N)) properties with butterfly matrices will also be introduced. Section 5.3 gives the full distribution of the growth factor of a Haar-butterfly

matrix. Previous results relating to the growth factors of (dense) random matrices was limited to estimates of the first moment of a random growth factor. The full distribution of the growth factor of a dense matrix is a significant step forward in the analysis of the numerical stability of methods in randomized linear algebra.

5.1 Nondegenerate transformation

In [31], Parker presents Theorem 1.1, which shows if A is nonsingular, then UAV^* is almost surely block nondegenerate for random butterfly matrices U, V. Again, this is a useful property if one wants to avoid pivoting, as big data movement and communication overhead commiserate with pivoting can impede parallel processing and block algorithms. Since matrix-matrix multiplication using butterfly matrices take $O(N^2 n)$ flops (see Section 2.1.3), multiplying a linear system on the left and right by a random butterfly matrix does not impact the leading order complexity of Gaussian elimination of $O(N^3)$. Additional motivations to remove pivoting are highlighted in [2, 31].

Parker's application specifically uses a *two-sided* butterfly transformation to form a nondegenerate linear system. One interesting and natural question to pose in light of Theorem 1.1 is whether one can gain enough benefit from using only a *one-sided* transformation. Through experiments, [2, 37] tested the impact of one-sided butterfly transformations. For a specific set of test cases, these experiments did both show benefits from the one-sided model. However, the benefit of these applications are limited in scope based on the actual test cases used and do not hold for all nonsingular matrices. Some potential applications can be shown to never achieve a block nondegenerate form through a one-sided butterfly transformation. For a given random butterfly matrix Ω, I can produce a nonsingular matrix A such that ΩA is nondegenerate. In fact, I can show a random permutation matrix P then has ΩP is nondegenerate with strictly positive probability.

Proposition 5.1. *Let $N = m^n$ and $B \sim \text{Haar}(\text{B}_s(m,n))$ for $n \geq 2$ and let $\sigma \in S_N$. Then a.s.*

$$\det\left(\begin{bmatrix}\mathbf{I}_2 & \mathbf{0}\end{bmatrix} BP_\sigma \begin{bmatrix}\mathbf{I}_2 \\ \mathbf{0}\end{bmatrix}\right) = 0 \tag{5.1}$$

if and only if $\sigma(1) \equiv \sigma(2) \pmod{m}$. Moreover, if $\sigma \sim \text{Uniform}(S_N)$ independent of B, then

$$\mathbb{P}\left(\det\left(\begin{bmatrix}\mathbf{I}_2 & \mathbf{0}\end{bmatrix} BP_\sigma \begin{bmatrix}\mathbf{I}_2 \\ \mathbf{0}\end{bmatrix}\right) = 0\right) = \frac{1}{m}\left(1 - \frac{m-1}{m^n - 1}\right). \tag{5.2}$$

Proof. Let $B = A \otimes C$ for $C \sim \text{Haar}(\text{SO}(m))$ and $A \sim \text{Haar}(\text{B}_s(m, n-1))$. Note

$$\begin{bmatrix}\mathbf{I}_2 & \mathbf{0}\end{bmatrix} = \begin{bmatrix}\mathbf{I}_2 & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{I}_m & \mathbf{0}\end{bmatrix} \tag{5.3}$$

so that

$$\begin{bmatrix}\mathbf{I}_2 & \mathbf{0}\end{bmatrix} B = \begin{bmatrix}\mathbf{I}_2 & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{I}_m & \mathbf{0}\end{bmatrix}(A \otimes C) = \begin{bmatrix}\mathbf{I}_2 & \mathbf{0}\end{bmatrix}\begin{bmatrix}A_{11}C & \cdots & A_{1m}C\end{bmatrix}$$

where we note $\mathbb{P}(A_{ij} = 0) = 0$ since $A \sim \text{Haar}(\text{B}_s(m,n))$. Moreover,

$$P_\sigma \begin{bmatrix}\mathbf{I}_2 \\ \mathbf{0}\end{bmatrix} = \begin{bmatrix}\mathbf{e}_{\sigma(1)} & \mathbf{e}_{\sigma(2)}\end{bmatrix}. \tag{5.4}$$

Let $\sigma(i) = q_i m + r_i$ for $q_i, r_i \in \mathbb{Z}$ with $0 \leq r_i < m$. Let

$$m_i = \delta_{0r_i} + r_i = \begin{cases} r_i & r_i \neq 0 \\ m & r_i = 0. \end{cases} \tag{5.5}$$

Then

$$\begin{bmatrix}\mathbf{I}_2 & \mathbf{0}\end{bmatrix} BP_\sigma \begin{bmatrix}\mathbf{I}_2 \\ \mathbf{0}\end{bmatrix} = \begin{bmatrix}\mathbf{I}_2 & \mathbf{0}\end{bmatrix} \begin{bmatrix}A_{1,1+q_1} C\mathbf{e}_{m_1} & A_{1,1+q_2} C\mathbf{e}_{m_2}\end{bmatrix}$$
$$= \begin{bmatrix}\mathbf{I}_2 & \mathbf{0}\end{bmatrix} C \begin{bmatrix}\mathbf{e}_{m_1} & \mathbf{e}_{m_2}\end{bmatrix} (A_{1,1+q_1} \oplus A_{1,1+q_2}).$$

If $\sigma(1) \equiv \sigma(2) \pmod m$ then $r_1 = r_2$, so that for $m_0 = m_1 = m_2$ we have

$$\begin{bmatrix}\mathbf{I}_2 & \mathbf{0}\end{bmatrix} BP_\sigma \begin{bmatrix}\mathbf{I}_2 \\ \mathbf{0}\end{bmatrix} = \begin{bmatrix}\mathbf{I}_2 & \mathbf{0}\end{bmatrix} CP_{(1\ m_0)} \begin{bmatrix}1 & 1 \\ 0 & 0\end{bmatrix} (A_{1,1+q_1} \oplus A_{1,1+q_2}). \tag{5.6}$$

has rank at most 1 and so is singular. If $\sigma(1) \not\equiv \sigma(2) \pmod m$, then $m_1 \neq m_2$ and

$$\begin{bmatrix}\mathbf{I}_2 & \mathbf{0}\end{bmatrix} BP_\sigma \begin{bmatrix}\mathbf{I}_2 \\ \mathbf{0}\end{bmatrix} = \begin{bmatrix}\mathbf{I}_2 & \mathbf{0}\end{bmatrix} CP_\tau (A_{1,1+q_1} \oplus A_{1,1+q_2}) \tag{5.7}$$

for $\tau = (2\ m_2)(1\ m_1) \in S_N$. Since $CP_\tau \in \mathrm{O}(N)$ has full rank then $(BP_\sigma)_{:2,:2}$ has rank equal to the rank of $(A_{1,1+q_1} \oplus A_{1,1+q_2})$, which is almost surely full rank.

Since $N = m^n$, then $[N]$ splits evenly among the residues modulo m. If $\sigma(1) \equiv \sigma(2) \pmod m$, then since $\sigma(2) \neq \sigma(1)$ there are $N/m - 1$ remaining choices of $j \equiv i \pmod 2$ for $\sigma(2)$ to take after $\sigma(1)$ is assigned, which have equal weight when $\sigma \sim \text{Uniform}\,(S_N)$. It follows

$$\mathbb{P}\left(\det\left(\begin{bmatrix}\mathbf{I}_2 & \mathbf{0}\end{bmatrix} BP_\sigma \begin{bmatrix}\mathbf{I}_2 \\ \mathbf{0}\end{bmatrix}\right) = 0\right) = \mathbb{P}(\sigma(1) \equiv \sigma(2) \pmod m)$$
$$= \frac{N(\frac{N}{m} - 1)(N-2)!}{N!} = \frac{\frac{N}{m} - 1}{N-1}$$
$$= \frac{1}{m}\left(1 - \frac{m-1}{m^n - 1}\right).$$

□

For example, if $B \sim \text{Haar}(\text{B}_s(m,n))$ and $\sigma \sim \text{Uniform}(S_N)$, then if $m = 2$ we have

$$\mathbb{P}\left(\det\left(\begin{bmatrix}\mathbf{I}_2 & \mathbf{0}\end{bmatrix} BP_\sigma \begin{bmatrix}\mathbf{I}_2 \\ \mathbf{0}\end{bmatrix}\right) = 0\right) = \frac{1}{2}\left(1 - \frac{1}{N-1}\right). \tag{5.8}$$

while if $n = 2$ then

$$\mathbb{P}\left(\det\left(\begin{bmatrix}\mathbf{I}_2 & \mathbf{0}\end{bmatrix} BP_\sigma \begin{bmatrix}\mathbf{I}_2 \\ \mathbf{0}\end{bmatrix}\right) = 0\right) = \frac{1}{m+1}. \tag{5.9}$$

Example 5.1. *For $n = 1$, then BP_σ is orthogonal for any $\sigma \in S_2$, and so is nonsingular. If $n = 2$, if $\sigma(1) = 1$ then $(BP_\sigma)_{:2,:2}$ is singular if and only if $\sigma(2) = 3$, for which only (2 3) and (2 3 4) satisfy this. If $\sigma(1) = 2$, then $(BP_\sigma)_{:2,:2}$ is singular if and only if $\sigma =$ (1 2 4) or $\sigma =$ (1 2 4 3). If $\sigma(1) = 3$, then $(BP_\sigma)_{:2,:2}$ is singular if and only if $\sigma =$ (1 3 2) or $\sigma =$ (1 3 4 2). Lastly, if $\sigma(1) = 4$, then $(BP_\sigma)_{:2,:2}$ is singular if and only if $\sigma =$ (1 4) or $\sigma =$ (1 4 3). Hence, the probability of choosing a permutation uniformly that results in $(BP_\sigma)_{:2,:2}$ being singular is*

$$\frac{4 \cdot 2}{4!} = \frac{1}{3} = \frac{1}{2} - \frac{1}{6}.$$

5.2 Haar orthogonal matrices

One common application of random Fourier transformations (RFTs) is to randomize an input vector to spread out the mass on each component to uniformize a vector before analysis. Without knowing beforehand the structure of an input vector, near localization of a vector can lead to undesirable numerical instabilities when using fixed-point arithmetic. To not limit the additional compounding and propagation of numerical errors, a desirable attribute

of a RFT is to have a small 2-condition number κ_2. Using real inputs, orthogonal transformations are desirable since they minimize κ_2. So a natural goal is to be able to sample an orthogonal transformation *uniformly*. This is accomplished by sampling a matrix using the Haar probability measure on $\mathrm{O}(N)$. This section will explore some existing methods to sample Haar orthogonal matrices and introduce a new sampling method using butterfly matrices.

5.2.1 Sampling Haar orthogonal matrices

As mentioned in Section 1.6.4, Stewart introduced a method to sample a Haar orthogonal matrix (see Theorem 1.15). Stewart's construction, which is still used widely for (exact) sampling of Haar orthogonal matrices uses **Householder reflections**.

For $\mathbf{v} \in \mathbb{R}^n$ is a vector, the projection map onto $\mathrm{span}(\mathbf{v})$ is given by $P_{\mathbf{v}} = \frac{\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}}$. If $\mathbf{u} \in \mathrm{span}(\mathbf{v})^\perp$, then $P_{\mathbf{v}}\mathbf{u} = \mathbf{0}$. A Householder reflection is the projection $H_{\mathbf{v}} = \mathbf{I}_n - 2P_{\mathbf{v}}$. Note $H_{\mathbf{v}}\mathbf{v} = -\mathbf{v}$ while if $\mathbf{u} \in \mathrm{span}(\mathbf{v})^\perp$, then $H_{\mathbf{v}}\mathbf{u} = \mathbf{u}$ so that $\det H_{\mathbf{v}} = -1$ (since 1 is an eigenvalue with multiplicity $n-1$ and -1 is an eigenvalue with multiplicity 1). Moreover, $H_{\mathbf{v}}$ is symmetric and since $H_{\mathbf{v}}^2\mathbf{v} = \mathbf{v}$ and $H_{\mathbf{v}}^2\mathbf{u} = \mathbf{u}$ for $\mathbf{u} \in \mathrm{span}(\mathbf{v})^\perp$, then $H_{\mathbf{v}}^2 = \mathbf{I}_n$, so that $H_{\mathbf{v}}^{-1} = H_{\mathbf{v}} = H_{\mathbf{v}}^T$. It follows $H_{\mathbf{v}}$ is orthogonal.

Householder reflections can be used to transform a matrix into an upper triangular matrix as follows: for $\mathbf{a}_i = A\mathbf{e}_i$, we have

$$H_{\|\mathbf{a}_1\|_2\mathbf{e}_1 - \mathbf{a}_1} A = \begin{bmatrix} \|\mathbf{a}_1\|_2 & * & \cdots & * \\ & * & \cdots & * \\ & \vdots & \vdots & \vdots \\ & * & \cdots & * \end{bmatrix} \tag{5.10}$$

One could then apply a Householder reflector to the bottom right order $n-1$ matrix $A^{(2)} =$

$H_1 A^{(1)}$. Note one does need to add a correction to the above construction to ensure the diagonal remains positive throughout. Hence, after n Householder reflector transformations, one can form an upper triangular matrix with positive diagonal R. Since each Householder reflector is orthogonal, then the transpose of the product of these Household reflectors is necessarily the Q factor in the QR factorization of A.

Now this can be combined with Stewart's result: if $A \sim \text{Gin}(n)$ and $A = QR$ is the QR factorization of A with R an upper triangular matrix with positive diagonal, then $Q \sim \text{Haar}(\text{O}(n))$. Hence, one can sample a Haar orthogonal matrix by using $G \sim \text{Gin}(n)$. Moreover, since $\text{Gin}(n)$ is invariant under orthogonal transformations, then $G_{2:,2:} \sim (H_{\mathbf{v}} G)_{2:,2:}$. Hence, one does not need to keep all of G but only needs to use Household reflectors using iid Gaussian vectors of length $n - i + 1$. Hence, one can form a product of a Haar orthogonal matrix and a vector using a sequence of Gaussian vectors, as outlined in Algorithm 5.

Algorithm 5 Haar orthogonal matrix-vector multiplication

1: **procedure** HOMULT($\mathbf{V}$)
2: $[n, m] = \text{size}(\mathbf{V})$
3: $\mathbf{U} = \mathbf{V}$
4: **for** $j = 1 : n$ **do**
5: $\mathbf{w} \sim \text{Gin}(n - j + 1, 1)$
6: $s = -\exp(-i \,\text{angle}(\mathbf{w}(1))$
7: $\mathbf{w} = s\mathbf{w}$
8: $\mathbf{w}(1) = \mathbf{w}(1) - \|\mathbf{2}\|_2$
9: $\mathbf{w} = \frac{\mathbf{w}}{\|\mathbf{w}\|_2}$
10: $\mathbf{U}_{j:,:} = \overline{s} \left(\mathbf{U}_{j:,:} - 2\mathbf{w}\mathbf{w}^T \mathbf{U}_{j:,:} \right)$
11: **return** $\mathbf{U}$

Applying Algorithm 5 to $\mathbf{I}_n$ then would produce $H \sim \text{Haar}(\text{O}(N))$.

Remark 5.1. *To sample* $U \sim \text{Haar}(\text{U}(n))$*, one would only need to update Step 5 to* $\mathbf{w} \sim \text{Gin}_{\mathbb{C}}(n - j + 1, 1)$ *in Algorithm 5.*

5.2.2 Givens rotations

Using Stewart's result, any algorithm to compute a QR factorization can be used to sample a Haar orthogonal matrix by applying the algorithm to $\mathrm{Gin}(n)$ and returning only the Q factor. One particular method is closely related to butterfly matrices:

Definition 5.1. *A **Givens rotation**, $G = G(\theta, i, j)$, is the special orthogonal matrix such*

$$R(\theta) = \begin{bmatrix} G_{ii} & G_{ij} \\ G_{ji} & G_{jj} \end{bmatrix} \in \mathcal{R}_2 = \mathrm{SO}(2),$$

and $G_{k\ell} = \delta_{k\ell}$ for any other k, ℓ.

Givens rotation matrices were introduced by Wallace Givens in the 1950s as an efficient means to introduce zeros in a matrix [15].

Example 5.2. *Let $\mathbf{x} = (r\cos\theta, r\sin\theta)$ for $r > 0$. Then*

$$G(\theta, 1, 2)\mathbf{x} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} r\cos\theta \\ r\sin\theta \end{bmatrix} = \begin{bmatrix} r \\ 0 \end{bmatrix}. \tag{5.11}$$

In particular, if $\mathbf{x} = (x, y) \in \mathbb{R}^2$ and $\mathbf{x} \neq \mathbf{0}$ and

$$\theta = \begin{cases} \arccos(\frac{x}{\|\mathbf{x}\|_2}) & \text{if } y \geq 0 \text{ and} \\ -\arccos(\frac{x}{\|\mathbf{x}\|_2}) & \text{if } y < 0, \end{cases}$$

then

$$R(\theta)\mathbf{x} = \begin{bmatrix} \|\mathbf{x}\| \\ 0 \end{bmatrix}. \tag{5.12}$$

It follows Givens rotations can be used to compute the QR decomposition of a nonsingular

matrix. Starting with the first column, use $N-1$ Givens rotations to progressively 0 out each element below the first entry. Now starting at the second entry on the second column, use $N-2$ Givens rotations to 0 out every below. Continuing on, after multiplication by $\binom{N}{2}$ total Givens rotations, we have necessarily an upper triangular matrix $\widetilde{R}$ with positive entries on its diagonal, with the possible exception of the last entry. Pulling out the sign of $\widetilde{R}_{nn}$ to form R such that $R_{nn} = |\widetilde{R}_{nn}|$ yields

$$A = \left(\prod_{i<j} G(\theta_{ij}, i, j) \right) (I_{N-1} \oplus \operatorname{sgn}(\widetilde{R}_{nn}))R = QR \tag{5.13}$$

for A nonsingular. Note the resulting Q is a product of (at most) $\binom{N}{2}$ Givens rotations, with the last term Q_{nn} completely determining the sign of $\det Q$, that is, $\det Q = Q_{nn}$. In particular, applying this to an element of $\mathrm{SO}(N)$ would have an R factor that is upper triangular with positive real entries, which must necessarily be the identity matrix. This yields:

Proposition 5.2. *The Givens rotations generate* $\mathrm{SO}(N)$.

In fact, every element of $\mathrm{SO}(N)$ is the product of at most $\binom{N}{2}$ Givens rotations.

5.2.3 Butterfly QR algorithm

Givens rotations are intimately related to butterfly matrices. This is established by noting $G = G(\theta, 1, 2) \in \mathrm{B}(N)$ for all $N \geq 1$, while also

$$P_\sigma G P_\sigma^T = G(\theta, \sigma(1), \sigma(2)), \tag{5.14}$$

which follows directly from (1.7) (and hence from (1.6)): apply the result

$$P_\sigma(\mathbf{e}_i \mathbf{e}_j^T)P_\sigma^T = (P_\sigma \mathbf{e}_i)(P_\sigma \mathbf{e}_j)^T = \mathbf{e}_{\sigma(i)}\mathbf{e}_{\sigma(j)}^T$$

to

$$G = \sum_{k=1}^{N} \mathbf{e}_k \mathbf{e}_k^T + (\cos\theta - 1)(\mathbf{e}_1\mathbf{e}_1^T + \mathbf{e}_2\mathbf{e}_2^T) + \sin\theta(\mathbf{e}_1\mathbf{e}_2^T - \mathbf{e}_2\mathbf{e}_1^T).$$

In particular, if $\sigma = (2\ j)(1\ i)$ for $i < j$, then $P_\sigma G P_\sigma^T = G(\theta, i, j)$. This then yields:

Theorem 5.1.

$$\langle P_\sigma \, \mathrm{B}(N) P_\sigma^T : \sigma = (2\ j)(1\ i), 1 \leq i < j \leq N \rangle = \mathrm{SO}(N).$$

Since $[\mathrm{O}(N) : \mathrm{SO}(N)] = 2$ and $P_{(1\ 2)} \in \mathcal{P}_N \backslash \mathcal{A}_N$, then we get also:

Corollary 5.1.

$$\langle \mathrm{B}(N), \mathcal{P}_N \rangle = \mathrm{O}(N).$$

In fact, we can provide an upper bound on the number of butterfly matrices needed to generate $M \in \mathrm{SO}(N)$. In particular, we can use $\mathcal{D}^{N/2}(N) \subset \mathrm{B}(N)$.

We want to carry out operations similar to how the Givens process generates the QR decomposition. For the first column, one butterfly matrix can 0 out the even indices, which followed then with $(2\ N/2+1)(4\ N/2+3)\cdots(N/2\ N-1)$ can assume the bottom half are zeroed out. In fact, the first n steps can be chosen such that $N/2$ zeros are introduced, and each successive step now needs to fix at least two columns. I will walk through some examples, where I keep track of which step I remove a given nonzero cell. I am not going to be exact with the permutation needed in these examples, but I want to outline the general Butterfly QR Algorithm.

Example 5.3. *Using a 4×4 matrix, which has $\binom{4}{2} = 6$ lower triangular entries, we see*

$$A = \begin{bmatrix} * & & & \\ * & * & & \\ * & * & * & \\ * & * & * & * \end{bmatrix}. \tag{5.15}$$

I will use the top two rows to eliminate the first entries from the last two rows. I can do this using $G(\theta_1, 1, 3)$ and $G(\theta_2, 2, 4)$ for θ_1 we first choose $B_1 \in \mathcal{D}^2(4)$ and $P_1 = P_{\sigma_1}$ for $\sigma_1 = (2\ 3)$ such that

$$P_1^T B_1 P_1 A = \begin{bmatrix} * & & & \\ * & * & & \\ [1] & * & * & \\ [1] & * & * & * \end{bmatrix}. \tag{5.16}$$

Next $B_2 \in \mathcal{D}^2(4)$ (and $\sigma_2 = 1$) such that

$$B_2 P_1^T B_1 P_1 A = \begin{bmatrix} * & & & \\ [2] & * & & \\ [1] & * & * & \\ [1] & [2] & * & * \end{bmatrix} \tag{5.17}$$

then $B_3 \in \mathcal{D}^2(4)$ and $\sigma_3 = (2\ 3)(1\ 2) = (1\ 3\ 2)$ such that

$$P_2^T B_3 P_3 B_2 P_1^T B_1 P_1 A = \begin{bmatrix} * & & & \\ [2] & * & & \\ [1] & [3] & * & \\ [1] & [2] & * & * \end{bmatrix} \tag{5.18}$$

Step	Zeros	Cumulative Zeros	Fixed Rows
1	2	2	
2	2	4	
3	1	5	1,4
4	1	6	1,2

Table 5.1: QR butterfly steps for $N = 4$

so that $B_4 = I \oplus B$ *and* $\sigma_4 = 1$ *with*

$$B_4 P_2^T B_3 P_2 B_2 P_1 B_1 A = \begin{bmatrix} * & & & \\ [2] & * & & \\ [1] & [3] & * & \\ [1] & [2] & [4] & * \end{bmatrix} \tag{5.19}$$

Table 5.1 provides an overview of the process through each step by keeping track of how many zeros were introduced along with which rows are fixed with respect to a Givens rotation.

Example 5.4. *Now consider the* 8×8 *case, which has* $\binom{8}{2} = 28$ *lower diagonal entries.*

$$\begin{bmatrix} * & & & & & & & \\ [3] & * & & & & & & \\ [2] & [5] & * & & & & & \\ [2] & [4] & [7] & * & & & & \\ [1] & [3] & [6] & [8] & * & & & \\ [1] & [3] & [5] & [7] & [9] & * & & \\ [1] & [2] & [4] & [6] & [8] & [10] & * & \\ [1] & [2] & [3] & [5] & [7] & [9] & [11] & * \end{bmatrix} \tag{5.20}$$

Table 5.2 similarly shows how many zeros are introduced in each step in the process for

Step	Zeros	Cumulative Zeros	Fixed Rows
1	4	4	
2	4	8	
3	4	12	
4	2	14	1,3,6,8
5	3	17	1,5
6	2	19	1,2,4,8
7	3	22	1,2
8	2	24	1,2,3,8
9	2	26	1,2,3,4
10	1	27	1,2,3,4,5,8
11	1	28	1,2,3,4,5,6

Table 5.2: QR butterfly steps for $N = 8$

$N = 8$.

Example 5.5. *And for the* 16×16 *case, with* $\binom{16}{2} = 120$ *lower diagonal terms, I will only*

include the final output:

$$\begin{bmatrix}
* & & & & & & & & & & & & & & & \\
[4] & * & & & & & & & & & & & & & & \\
[3] & [6] & * & & & & & & & & & & & & & \\
[3] & [5] & [8] & * & & & & & & & & & & & & \\
[2] & [5] & [7] & [10] & * & & & & & & & & & & & \\
[2] & [4] & [7] & [9] & [12] & * & & & & & & & & & & \\
[2] & [4] & [6] & [9] & [11] & [14] & * & & & & & & & & & \\
[2] & [4] & [6] & [8] & [10] & [13] & [16] & * & & & & & & & & \\
[1] & [3] & [5] & [8] & [10] & [12] & [15] & [18] & * & & & & & & & \\
[1] & [3] & [5] & [7] & [9] & [11] & [14] & [17] & [20] & * & & & & & & \\
[1] & [3] & [5] & [7] & [9] & [11] & [13] & [16] & [19] & [21] & * & & & & & \\
[1] & [3] & [4] & [6] & [8] & [10] & [12] & [15] & [18] & [20] & [22] & * & & & & \\
[1] & [2] & [4] & [6] & [8] & [10] & [12] & [14] & [17] & [19] & [21] & [23] & * & & & \\
[1] & [2] & [4] & [5] & [7] & [9] & [11] & [13] & [16] & [18] & [20] & [22] & [24] & * & & \\
[1] & [2] & [3] & [5] & [7] & [9] & [11] & [13] & [15] & [17] & [19] & [21] & [23] & [25] & * & \\
[1] & [2] & [3] & [4] & [6] & [8] & [10] & [12] & [14] & [16] & [18] & [20] & [22] & [24] & [26] & *
\end{bmatrix} \tag{5.21}$$

Table 5.3 again provides the accumulation of zeros for each step in the $N = 16$ case.

In particular, note the last line of each final output. Following this algorithm, we add a zero to the last row in the first n steps, and then the remaining steps we alternate in fixing the last row, so that there are

$$\alpha_n = n + 2 \cdot (N - (n + 1)) = 2(N - 1) - n \tag{5.22}$$

Step	Zeros	Cumulative Zeros	Fixed Rows
1	8	8	
2	8	16	
3	8	24	
4	8	32	
5	7	39	1,16
6	6	45	1,6,11,15
7	6	51	1,2,9,16
8	6	57	1,2,7,15
9	6	63	1,2,3,16
10	6	69	1,2,3,15
11	5	74	1,2,3,4,5,16
12	5	79	1,2,3,4,8,15
13	4	83	1,2,3,4,5,7,10,16
14	4	87	1,2,3,4,5,9,12,15
15	3	90	1,2,3,4,5,6,8,11,14,16
16	4	94	1,2,3,4,5,6,10,13
17	3	97	1,2,3,4,5,6,7,9,12,16
18	4	101	1,2,3,4,5,6,7,11
19	3	104	1,2,3,4,5,6,7,8,10,16
20	4	108	1,2,3,4,5,6,7,8
21	3	111	1,2,3,4,5,6,7,8,9,16
22	3	114	1,2,3,4,5,6,7,8,9,10
23	2	116	1,2,3,4,5,6,7,8,9,10,11,16
24	2	118	1,2,3,4,5,6,7,8,9,10,11,12
25	1	119	1,2,3,4,5,6,7,8,9,10,11,12,13,16
26	1	120	1,2,3,4,5,6,7,8,9,10,11,12,13,14

Table 5.3: QR butterfly steps for $N = 16$

total steps to complete the QR butterfly process. Hence, we need α_n block diagonal butterfly matrices to generate Haar(O(N)).

Note $\alpha_1 = 2(2-1)-1 = 1$, $\alpha_2 = 2(4-1)-2 = 4$ (compare to $\binom{4}{2} = 6$), $\alpha_3 = 2(8-1)-3 = 11$ (compare to $\binom{8}{2}$=28), $\alpha_4 = 2(16-1)-4 = 26$ (compare to $\binom{16}{2} = 120$). So $\alpha_5 = 2(32-1)-5 = 57$ (compare to $\binom{32}{2} = 496$) and $\alpha_6 = 2(64-1)-6 = 120$ (compare to $\binom{64}{2} = 2016$).

In the first n steps, as we eliminate precisely $N/2$ entries each time, this can be condensed into exactly one reverse diagonal butterfly matrix: by changing the affiliated permutations from the first n steps, we can let the large diagonal rotation matrix correspond to the zeroing

out of the bottom half of the components of the first column, then the next block diagonal component that zeros out the bottom of what remains in the first column and the bottom quarter of the second column, and so on. Hence, we need at most $2(N-n)-1$ butterfly matrices to generate a random element of $\mathrm{SO}(N)$ distributed according to the Haar measure.

Using this idea, we can outline Algorithm 6 to encapsulate this QR butterfly process:

Algorithm 6 Butterfly QR Algorithm

1: **procedure** ButterflyQR(A)
2: $n = \text{size}(A)$
3: $Q = \mathbf{I}_n$
4: $R = A$
5: **for** $j = 1 : n-1$ **do**
6: $\begin{bmatrix} Q_{j:,:} & R_{j:,:} \end{bmatrix} = P_\sigma \begin{bmatrix} Q_{j:,:} & R_{j:,:} \end{bmatrix}$ for $\sigma \in \mathcal{P}_{N-j+1}$ such that each leading column has all zero entries moved to the bottom, that is, $(P_\sigma R)_{ik} = 0$ and $(P_\sigma R)_{i-1,k-1} \neq 0$ implies $(P_\sigma R)_{\ell k} = 0$ for $\ell \geq i$
7: $\begin{bmatrix} Q_{j:,:} & R_{j:,:} \end{bmatrix} = (\bigoplus_{i=j}^{d} B(\boldsymbol{\theta}_i) \oplus \mathbf{I}_{j_i'}) \begin{bmatrix} Q_{j:,:} & R_{j:,:} \end{bmatrix}$ for $B(\boldsymbol{\theta}_i) \in \mathcal{D}_{j_i/2}^{j_i}$ that zeros out half of the nonzero terms below the diagonal of each leading column; that is, for $d = \max\{j : R_{N-1:N,:j-1} = \mathbf{0}\}$ and $n - j + 1 = \sum_{i=1}^{d} j_i + j_i'$ where $B(\boldsymbol{\theta}_i) R_{j:,i} = \sum_{k=0}^{\lfloor (n-j)/2 \rfloor} |r_k| \mathbf{e}_{j+2k}$
8: $Q = Q^T$
9: **return** $[Q, R]$

Let $\mathcal{D}^{\lfloor N/2 \rfloor}(N) = \mathcal{D}^{N/2}(N)$ if $2 \mid N$ and $\mathcal{D}^{\lfloor N/2 \rfloor}(N) = \mathcal{D}^{(N-1)/2}(N-1) \oplus 1$ if $2 \nmid N$. Combining the above results and observations leads to:

Theorem 5.2 (Butterfly QR Algorithm)**.**

$$\mathrm{SO}(N) = \prod_{j=1}^{2(N-1)-n} P_{\sigma_j} \mathcal{D}^{\lfloor N/2 \rfloor}(N) P_{\sigma_j^{-1}\tau_j} \tag{5.23}$$

for $\sigma_j, \tau_j \in S_N$.

Future work can explore how the Butterfly QR algorithm, which is a compression of the Givens QR algorithm, compares against the Householder reflection method to generate $\mathrm{Haar}(\mathrm{O}(N))$.

In particular, it would be interesting to explore how this algorithm can be used to approximate Haar orthogonal matrices by sampling $B(\boldsymbol{\theta}) \sim \text{Haar}(\mathcal{D}^{\lfloor N/2 \rfloor}(N))$, and how far away smaller products of butterfly factors are from Haar orthogonal.

5.2.4 Hierarchy of randomness

Now we can relate $\text{B}_s(N, \Sigma)$ to $\text{B}(N, \Sigma)$ to $\text{Haar}(\text{SO}(N))$ for $\Sigma = \Sigma_S$ and $\Sigma = \Sigma_D$. Note this hierarchy is natural in light of the fact the angle derived from applying a Givens rotation to a Gaussian vector is necessarily uniform. This can be established using only elementary calculus. Alternatively, this follows since for $G \sim \text{Ginibre}(2)$ then $G = QR$ for R upper triangular with positive diagonal, has $Q = R(\theta)(1 \oplus \text{sgn}(g_{22})) \sim \text{Haar}(\text{O}(2))$ so that $R(\theta) \sim \text{Haar}(\text{SO}(2))$ and hence $\theta \sim \text{Uniform}([0, 2\pi))$. This shows each Givens matrix transformation projects to the 2×2 case, such that necessarily the induced angle is $\text{Uniform}([0, 2\pi))$ at each stage. Hence, each angle that arises in the butterfly QR algorithm applied to a Ginibre matrix must also necessarily be uniform.

Using Section 2.1.6, we see $B \sim \text{B}_s(N, \Sigma_S)$ is sampled using n uniform angles, $B \sim \text{B}(N, \Sigma_S)$ is sampled using $N - 1$ uniform angles, $B \sim \text{B}_s(N, \Sigma_D)$ is sampled using $N - 1$ uniform angles, $\text{B}(N, \Sigma_D)$ is sampled using $\frac{1}{2}Nn$ uniform angles, and $\text{Haar}(\text{SO}(N))$ and $\text{Haar}(\text{O}(n))$ are sampled using $\binom{N}{2}$ uniform angles.

A simple consideration of the number of uniform angles needed to generate each model indicates the butterfly models are far from $\text{Haar}(\text{O}(N))$ by themselves. Appendix D explores one particular property relating to the signs of the components of each column that distinguishes butterfly matrices from Haar orthogonal matrices.

5.3 Growth factors of random butterfly matrices

This section will present a novel result that gives the full distribution of the growth factor of Haar-butterfly matrices. Standard results in the literature relating to growth factors of random matrices are limited to first moment estimates. The recursive structure of Haar-butterfly matrices enables us to go beyond this limited scope.

This section will be structured to explore the impact on the growth factor of preconditioning the linear system $A\mathbf{x} = \mathbf{b}$ using random butterfly matrices. Two basic models will be focused on in the following document.

First, we will consider the **naïve model** where $A = \mathbf{I}$. This is given its name since this linear system has the obvious solution $\mathbf{x} = \mathbf{b}$. In essence, this model will allow us to study how much we can mess things up using butterfly matrices. This will contain our significant results relating to the full distribution of the growth factors of Haar-butterfly matrices. The second model will be the **worst-case model** where $\rho(A) = 2^{N-1}$ maximizes the max-norm growth factor. This will enable us to study the potential dampening impacts on the growth factors using random butterfly matrices.

Other motivating goals that will be touched on in this section include exploring whether removing pivoting is a good idea to begin with. Parker's focus was on reducing computational complexity at the potential cost of numerical accuracy, so we want to explore how this cost-benefit analysis can play out with regard to precision.

A final motivating goal is to explore whether butterfly matrices could potentially be used to upgrade accuracy between pivoting strategies. This particular question can be revisited at a future point when more pivoting strategies are explored. Our initial focus limits our consideration to only GENP, GEPP, GERP and GECP.

5.3.1 Background

The growth factor of a matrix A is determined by the computed LU factorization $PAQ = LU$ using the corresponding pivoting scheme. We will focus on two particular growth factor definitions in this document. See [4] for an overview on other common definitions found in the literature, along with some explicit properties and relationships comparing different definitions.

The first growth factor we will visit is related to the max-norm of the associated factors encountered during GE, given by

$$\rho(A) := \frac{\|L\|_{\max} \cdot \max_k \|A^{(k)}\|_{\max}}{\|A\|_{\max}}. \tag{5.24}$$

This is the classical definition first used by Wilkinson in the 1960s in his error analysis on the backward stability of GEPP (note necessarily $\|L\|_{\max} = 1$ using GEPP) [39, 40]. Our experiments in the last section will focus on the growth factor derived from the ℓ_∞-induced matrix norm:

$$\rho_\infty(A) := \frac{\|L\|_\infty \|U\|_\infty}{\|A\|_\infty}. \tag{5.25}$$

The growth factor is an important component in controlling the relative error in a computed solution to a linear system using floating-point arithmetic. If $PAQ = LU$ is the computed LU factorization used to compute the approximate solution $\hat{\mathbf{x}}$ to the linear system $A\mathbf{x} = \mathbf{b}$ for nonsingular $A \in \mathbb{R}^{N\times N}$, then (cf. [39])

$$\frac{\|\mathbf{x} - \hat{\mathbf{x}}\|_\infty}{\|\mathbf{x}\|_\infty} \leq 4N^2 \epsilon \kappa_\infty(A) \rho(A) \tag{5.26}$$

for $\epsilon = \epsilon_{\text{machine}}$ and

$$\kappa_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty \tag{5.27}$$

is the ℓ_∞-induced condition number and (cf. [16, Section 9.7])

$$\frac{\|\mathbf{x} - \hat{\mathbf{x}}\|_\infty}{\|\mathbf{x}\|_\infty} \leq \gamma_{3N} \kappa_\infty(A) \rho_\infty(A) \tag{5.28}$$

where $\gamma_M = \dfrac{M\epsilon}{1 - M\epsilon}$.

This section will focus on the impact of the growth factor through orthogonal transformations as a means of controlling the relative error. Even through κ_∞ is not invariant under orthogonal transformations, the impact is relatively moderate. In particular, although the worst-case bounds inherent from the matrix norm inequalities $\|A\|_\infty \leq \sqrt{N}\|A\|_2$ yield $\kappa_\infty(A) \leq N\kappa_2(A) = N$, butterfly matrices have growth factors far from this worst case behavior. Even though we always have $\kappa_\infty(B) \leq N\kappa_2(B) = N$, in the next section we will see $\mathbb{E}\kappa_\infty(B) \approx N^{0.710719}$.

Previous results on growth factors of random matrices

Interest in growth factors of random matrices dates back to Wilkinson's original work establishing the backward stability of GEPP, which established maximal exponential growth factors of order 2^{N-1} that could lead to a loss of $N - 1$ bits of precision [39]. Early focus was on the worst-case behavior of growth factors, which led to very pessimistic views of precision using LU factorizations. In [36], Trefethen and Schreiber shifted the view away from the worst-case model to introduce average-case analysis of the stability of GE. They were interested why GEPP was successful in practice, with much higher precision than would be expected from the worst-case scenario. To accomplish this, they carried out experiments to

compute the growth factors of a variety of random matrices.

Through statistical arguments and numerical experiments, they showed that the average growth factor ρ using GEPP was no larger than $O(N)$ for various random matrices with iid entries. Limiting their experiments to matrices of order at most 2^{10}, they showed ρ using GEPP is approximately $N^{2/3}$ while ρ using GECP was approximately $N^{1/2}$ for ensembles with iid entries. They do conjecture further the growth factor should be asymptotically $O(N^{1/2})$ for both GEPP and GECP. Additionally, they observe in the iid case that only a few intermediate steps of GE were needed until the remaining entries exhibited approximately normal behavior. Hence, the Ginibre ensemble was a good stand-in for an approximately universal growth factor model for iid matrices.

In the same paper, Trefethern and Schreiber also experimented with Haar orthogonal matrices and observed the average growth factors were significantly larger than the iid models. This was not too surprising since orthogonal scaled Hadamard matrices have growth factors that are near the largest recorded using GECP [5, 13]. In [18], Higham, Higham and Pranesh establish $\rho(A) \gtrsim \frac{N}{4 \log N}$ for $A \sim \text{Haar}(\text{O}(N))$. Hence, they show asymptotically a *lower bound* growing at a higher rate than the iid ensemble growth factors.

One common note among these preceding works on random growth factors is analysis is limited to computing specific statistics or bounds relating to the growth factors, which only give a small glimpse at the distribution of random growth factors. Haar-butterfly matrices enable us to go beyond these prior limitations. We are able to pull back the curtain completely and see the full distribution of the growth factors of Haar-butterfly matrices.

5.3.2 Growth factors of Haar-butterfly matrices (Naïve model)

The naïve model enables us to study the growth factors of the random butterfly matrices directly. This section will focus explicitly on the growth factors of Haar-butterfly matrices, for which we are able to give the full distribution.

Computations using simple scalar butterfly or Haar-butterfly matrices are made significantly more tractable by combining (2.12) along with Lemmas 1.6 and 1.7. Section 5.3.4 will contain the proofs of the outstanding technical details. The resulting main statements regarding the full distribution of growth factors of Haar-butterfly matrices will be given here:

Theorem 5.3. (I) *Let $B = B(\boldsymbol{\theta}) \in \mathrm{B}_s(N)$. Then B has an LU factorization using GENP iff $\cos\theta_j \neq 0$ for all j. Moreover, using GENP then*

$$\rho(B) = \prod_{j=1}^{n}(1+\tan^2\theta_j) \tag{5.29}$$

$$\rho_\infty(B) = \prod_{j=1}^{n}(1+\max(|\tan\theta_j|, \tan^2\theta_j)), \tag{5.30}$$

with $1 \leq \rho(B) \leq \rho_\infty(B)$ while $1 = \rho(B) = \rho_\infty(B)$ iff $\cos\theta_j = 0$ for all j, $N \leq \rho(B) = \rho_\infty(B)$ iff $|\tan\theta_j| \geq 1$ for all j, and strict inequalities hold otherwise.

Using GEPP or GERP, then

$$\rho(B) = \prod_{j=1}^{n}(1+\min(\tan^2\theta_j, \cot^2\theta_j)) \tag{5.31}$$

$$\rho_\infty(B) = \prod_{j=1}^{n}(1+\min(|\tan\theta_j|, |\cot\theta_j|)), \tag{5.32}$$

with $1 \leq \rho(B) \leq \rho_\infty(B) \leq N$ with $1 = \rho(B) = \rho_\infty(B)$ iff $\cos\theta_j = 0$ or $\sin\theta_j = 0$ for all j, $\rho(B) = \rho_\infty(B) = N$ iff $|\tan\theta_j| = 1$ for all j, and strict inequalities hold otherwise.

(II) *Let $B \sim \mathrm{B}_s(N, \Sigma_S)$ and X_j iid* Cauchy(1)*. Then B has an LU factorization using GENP*

almost surely, with

$$\rho(B) \sim \prod_{j=1}^{n}(1+X_j^2)$$
$$\rho_\infty(B) \sim \prod_{j=1}^{n}(1+\max(|X_j|, X_j^2)),$$

with $1 \leq \rho(B) \leq \rho_\infty(B)$.

Using GEPP or GERP, B almost surely has unique factors with $PB = LU$ *and*

$$\rho(B) \sim \prod_{j=1}^{n}(1+X_j^2 \mid |X_j| \leq 1)$$
$$\rho_\infty(B) \sim \prod_{j=1}^{n}(1+|X_j| \mid |X_j| \leq 1)$$

with $1 \leq \rho(B) \leq \rho_\infty(B) \leq N$. *Moreover,* $\mathbb{P}(P = \mathbf{I}) = \frac{1}{N}$.

Using GENP, then $\rho(B)$ and $\rho_\infty(B)$ have no finite moments of any orders $k \geq 1$ since the absolute Cauchy has no finite moments of these sizes. However, since $\rho(B)$ and $\rho_\infty(B)$ are bounded when using pivoting, then we can calculate the average growth factors *exactly* rather than being restricted to first moment estimates, as has been the case in previous results with random growth factors:

Corollary 5.2. *Let* $B \sim \mathrm{B}_s(N, \Sigma_S)$. *Using GEPP or GERP, then*

$$\mathbb{E}\rho(B) = \left(\frac{4}{\pi}\right)^n = N^\alpha$$
$$\mathbb{E}\rho_\infty(B) = \left(1+\frac{\log 4}{\pi}\right)^n = N^\beta$$

for $\alpha = \log_2(\frac{4}{\pi}) \approx 0.34850387$ *and* $\beta = \log_2(1+\frac{\log 4}{\pi}) \approx 0.52734183$.

Remark 5.2. *We can now relate directly the growth factors of Haar-butterfly matrices to the growth factors of other random ensembles of matrices studied in [18, 36]. Using only GEPP,*

Theorem 5.3 and Corollary 5.2 show $\rho(B)$ is sublinear and $\mathbb{E}\rho(B) \approx N^{0.34850387}$. Of note, this is smaller than the first moment estimates of growth factors of iid ensembles using both GEPP and even GECP: Trefethen and Schreiber showed these were, respectively, about $N^{2/3}$ using GEPP and $N^{1/2}$ using GECP. This supports the underlying motivating question as to whether accuracy using different pivoting schemes can be upgraded by using butterfly preconditioning. Future work will consider the GECP factorization of simple butterfly matrices.

Section 5.3.4 will include the technical details needed to prove the above results relating to Haar-butterfly matrices. Many of the small technical steps are straightforward and follow from standard results relating the uniform distribution to the Cauchy and Arcsine distributions. Heavier technical machinery is used to establish the GENP, GEPP and GERP factorizations of $B \sim \mathrm{B}_s(N, \Sigma_S)$ satisfy

$$\max_k \|B^{(k)}\|_{\max} = \|U\|_{\max} \tag{5.33}$$

(see Proposition 5.6). The methods used in these steps currently do not generalize to other Haar-butterfly models initiated with $\mathrm{SO}(m)$ for $m > 2$.

Although $|X|$ has no finite moments of size $k \geq 1$ for $X \sim \text{Cauchy}(1)$, $\log(1+X^2)$ has finite moments of any order; in particular, $\mathbb{E}\log(1+X^2) = \log 4$ and $\mathbb{E}\log(1+X^2)^2 = \frac{\pi^2}{3} + (\log 4)^2$ so that $\operatorname{Var}\log(1+X^2) = \frac{\pi^2}{3}$. Since $\log\rho(B)$ (and $\log\rho_\infty(B)$) are then sums of iid terms, each of which is dominated by $1 + \log(1+X^2)$, then a simple consequence of the central limit theorem is:

Corollary 5.3. *Let $B \sim \mathrm{B}_s(N, \Sigma_S)$, $X \sim \text{Cauchy}(1)$ and $Z \sim N(0,1)$. Using GENP, GEPP, or GERP, then for any $t \in \mathbb{R}$*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\log\rho(B) - n\mu}{\sqrt{n}\sigma} \leq t\right) = \mathbb{P}(Z \leq t) \tag{5.34}$$

where $\mu = \mathbb{E}\log(1+X^2) = \log 4$, $\sigma^2 = \operatorname{Var}\log(1+X^2) = \frac{\pi^2}{3}$ *when using GENP and* $\mu = \mathbb{E}\log(1+X^2 \mid |X| \leq 1)$, $\sigma^2 = \operatorname{Var}\log(1+X^2 \mid |X| \leq 1)$ *when using GEPP or GERP.*

One application of this is that if a distribution is skewed then the average value is a skewed approximation of typical behavior one might encounter. The **median**, M_n, would perhaps be more desirable statistic to gauge behavior of $\rho(B)$. Since for n sufficiently large, then by Corollary 5.3 we have

$$\frac{1}{2} = \mathbb{P}(\rho(B) \leq M_n) = \mathbb{P}\left(\frac{\log \rho(B) - n\mu}{\sqrt{n}\sigma} \leq \frac{\log M_n - n\mu}{\sqrt{n}\sigma}\right) \approx \mathbb{P}\left(Z \leq \frac{\log M_n - n\mu}{\sqrt{n}\sigma}\right), \tag{5.35}$$

so that $\log M_n \approx n\mu$ and hence

$$M_n \approx N^{\mu/\log 2}. \tag{5.36}$$

Using GENP, then $\mu = \log 4 = 2\log 2$ and so the median $\rho(B)$ is asymptotically N^2, while the average is not even finite. Using pivoting, we have $\mathbb{E}(1+X^2 \mid |X| \leq 1) = \mu = \log 4 - \frac{4G}{\pi}$, using Catalan's constant

$$G = \sum_{n \geq 0} \frac{(-1)^n}{(2n+1)^2} \approx 0.91596559. \tag{5.37}$$

This yields M_n is asymptotic with $N^{2-\frac{4G}{\pi \log 2}} \approx N^{0.31746612}$, which is not too far off from $\mathbb{E}\rho(B) \approx N^{0.34850387}$. Similarly, Corollary 5.3 can be used to find asymptotic quantiles for $\rho(B)$ (and $\rho_\infty(B)$).

Additionally, explicit results relating to $\kappa_\infty(B)$ can be computed for $B \in \mathrm{B}_s(N)$ and $B \sim \mathrm{B}_s(N, \Sigma_S)$:

Proposition 5.3. (I) *Let* $B = B(\boldsymbol{\theta}) \in \mathrm{B}_s(N)$. *Then*

$$\kappa_\infty(B) = \prod_{j=1}^{n}(1 + |\sin(2\theta_j)|) \tag{5.38}$$

with $1 \leq \kappa_\infty(B) \leq N$, *where* $1 = \kappa_\infty(B)$ *iff* $\sin\theta_j = 0$ *or* $\cos\theta_j = 0$ *for all* j, $\kappa_\infty(B) = N$ *iff* $|\tan\theta_j| = 1$ *for all* j, *and strict inequalities otherwise.*

(II) *Let* $B \sim \mathrm{B}_s(N, \Sigma_S)$ *and* Y_j *iid* $\mathrm{Arcsine}(0,1)$. *Then*

$$\kappa_\infty(B) \sim \prod_{j=1}^{n}(1 + \sqrt{Y_j}) \tag{5.39}$$

with $1 \leq \kappa_\infty(B) \leq N$.

Similarly, explicit average condition numbers as well as the average product of the condition number and growth factor when using GEPP or GERP (since $\kappa_\infty \geq 1$ this product is still not integrable in the GENP case), as found in (5.24) and (5.25), can be computed:

Corollary 5.4. *Let* $B \sim \mathrm{B}_s(N, \Sigma_S)$. *Then*

$$\mathbb{E}\kappa_\infty(B) = \left(1 + \frac{2}{\pi}\right)^n = N^\gamma \tag{5.40}$$

for $\gamma = \log_2(1 + \frac{2}{\pi}) \approx 0.71071919$. *Using GEPP or GERP, then*

$$\mathbb{E}\kappa_\infty(B)\rho(B) = \left(\frac{4}{\pi}(1 + \log 2)\right)^n = N^{1+\zeta} \tag{5.41}$$

$$\mathbb{E}\kappa_\infty(B)\rho_\infty(B) = \left(2 + \frac{\log 4}{\pi}\right)^n = N^{1+\xi} \tag{5.42}$$

for $\zeta = \log_2(\frac{2}{\pi}(1 + \log 2)) \approx 0.10821126$ *and* $\xi = \log_2(1 + \frac{\log 2}{\pi}) \approx 0.28763257$.

Note since $1 \leq \rho(B) \leq \rho_\infty(B) \leq N$ for $B \in \mathrm{B}_s(N)$ by Theorem 5.3, then (5.25) provides a

Floating-point format	Precision (bit)	$4N^{3+\zeta}\epsilon$	$3N^{2+\xi}\epsilon$
half precision	11	$2^{n(3+\zeta)-9}$	$2^{n(2+\xi)+\log_2 3-11}$
single precision	24	$2^{n(3+\zeta)-22}$	$2^{n(2+\xi)+\log_2 3-24}$
double precision	53	$2^{n(3+\zeta)-51}$	$2^{n(2+\xi)+\log_2 3-53}$
quad precision	113	$2^{n(3+\zeta)-111}$	$2^{n(2+\xi)+\log_2 3-113}$

Table 5.4: Average upper bounds (from (5.25) and (5.24)) on the ℓ_∞-relative error of the computed solution using the GEPP LU factorization of $\Omega \sim \mathrm{B}_s(N, \Sigma_s)$ to solve the linear system $\Omega\mathbf{x} = \mathbf{b}$

sharper bound for the relative error in the naïve model than (5.24):

$$\frac{4N^2\epsilon\kappa_\infty(B)\rho(B)}{\gamma_{3N}\kappa_\infty(B)\rho_\infty(B)} \gtrsim \frac{4N^2\epsilon\kappa_\infty(B)\rho(B)}{3N\epsilon\kappa_\infty(B)\rho_\infty(B)} = \frac{4}{3}\frac{N\rho(B)}{\rho_\infty(B)} \geq \frac{4}{3}. \tag{5.43}$$

By computing the expected values of the right-hand bounds in (5.25) and (5.24), Table 5.4 gives a reference using different floating-point formats for the worst-case ℓ_∞-relative error in using GEPP (or GERP) to compute a solution to the linear system $\Omega\mathbf{x} = \mathbf{b}$ for $\Omega \sim \mathrm{B}_s(N, \Sigma_S)$. In particular, the last two columns of Table 5.4 give an upper bound for the ℓ_∞-relative error of the computed solution $\hat{\mathbf{x}}$.

For example, if one wanted to ensure the ℓ_∞-relative error 10-bit average accuracy in the naïve model, then this would be unattainable using half precision, while using single precision one needs $n(2+\xi) + \log_2 3 - 24 < -10$ and so $n < \frac{14-\log_2 3}{2+\xi} \approx 5.42702427$. Using double precision, one needs $n < \frac{43-\log_2 3}{2+\xi} \approx 18.10388521$ and using quad precision one needs $n < \frac{103-\log_2 3}{2+\xi} \approx 44.33187336$. In particular, if $\|\mathbf{x}\|_\infty = 1$, then each component in the computed solution $\widehat{\mathbf{x}}$ maintains an average of at least 10-bits of accuracy for an order $2^{18} = 262,144$ order butterfly matrix using double precision and an order $2^{44} \approx 1.7592 \cdot 10^{13}$ butterfly matrix using quad precision.

Although the intermediate GE matrices $B^{(k)}$ do not have convenient Kronecker product factorizations, they can be derived directly from B and L which do have these factorizations (see Lemma 5.2). In particular, the symmetry from the start partially carries through each

intermediate GE step, which results in the GEPP factorization then aligning exactly with the GERP factorization (see Proposition 5.5). However, this does not go through to the GECP factorization, as the following example illustrates:

For $B(\theta_1, \theta_2) \in \mathrm{B}_s(4)$, using GENP we have

$$B = B^{(1)} = \begin{bmatrix} \cos\theta_2 \cos\theta_1 & \cos\theta_2 \sin\theta_1 & \sin\theta_2 \cos\theta_1 & \sin\theta_2 \sin\theta_1 \\ -\cos\theta_2 \sin\theta_1 & \cos\theta_2 \cos\theta_1 & -\sin\theta_2 \sin\theta_1 & \sin\theta_2 \cos\theta_1 \\ -\sin\theta_2 \cos\theta_1 & -\sin\theta_2 \sin\theta_1 & \cos\theta_2 \cos\theta_1 & \cos\theta_2 \sin\theta_1 \\ \sin\theta_2 \sin\theta_1 & -\sin\theta_2 \cos\theta_1 & -\cos\theta_2 \sin\theta_1 & \cos\theta_2 \cos\theta_1 \end{bmatrix}$$

$$B^{(2)} = \begin{bmatrix} \cos\theta_2 \cos\theta_1 & \cos\theta_2 \sin\theta_1 & \sin\theta_2 \cos\theta_1 & \sin\theta_2 \sin\theta_1 \\ & \cos\theta_2 \sec\theta_1 & & \sin\theta_2 \sec\theta_1 \\ & & \sec\theta_2 \cos\theta_1 & \sec\theta_2 \sin\theta_1 \\ & -\sin\theta_2 \sec\theta_1 & -\sec\theta_2 \sin\theta_1 & \sec\theta_2 \sec\theta_1 (\cos^2\theta_1 - \sin^2\theta_2) \end{bmatrix}$$

$$B^{(3)} = \begin{bmatrix} \cos\theta_2 \cos\theta_1 & \cos\theta_2 \sin\theta_1 & \sin\theta_2 \cos\theta_1 & \sin\theta_2 \sin\theta_1 \\ & \cos\theta_2 \sec\theta_1 & & \sin\theta_2 \sec\theta_1 \\ & & \sec\theta_2 \cos\theta_1 & \sec\theta_2 \sin\theta_1 \\ & & -\sec\theta_2 \sin\theta_1 & \sec\theta_2 \cos\theta_1 \end{bmatrix}$$

$$U = B^{(4)} = \begin{bmatrix} \cos\theta_2 \cos\theta_1 & \cos\theta_2 \sin\theta_1 & \sin\theta_2 \cos\theta_1 & \sin\theta_2 \sin\theta_1 \\ & \cos\theta_2 \sec\theta_1 & & \sin\theta_2 \sec\theta_1 \\ & & \sec\theta_2 \cos\theta_1 & \sec\theta_2 \sin\theta_1 \\ & & & \sec\theta_2 \sec\theta_1 \end{bmatrix}$$

(using Lemma 5.2). If $|\tan\theta_1|, |\tan\theta_2| < 1$ then no pivoting would be needed using GEPP, so that the above intermediate GE steps would coincide for both the GENP and GEPP methods. Furthermore, the symmetry in the lower blocks, and in particular in the lower leading column and row, yield GERP also coincides with these other methods. However,

this does not hold for GECP. If $|\cos\theta_1| > |\cos\theta_2|$ then

$$|B_{33}^{(2)}| = |\cos\theta_1 \sec\theta_2| = |\sec\theta_1 \sec\theta_2| \cos^2\theta_1 > |\sec\theta_1 \sec\theta_2| \cos^2\theta_2 = |B_{22}^{(2)}|$$

and since $|\tan\theta_1|, |\tan\theta_2| < 1$ yields $\sin^2\theta_2 < \cos^2\theta_1$ then also

$$|B_{33}^{(2)}| = |\sec\theta_1 \sec\theta_2| \cos^2\theta_1 \geq |\sec\theta_1 \sec\theta_2|(\cos^2\theta_1 - \sin^2\theta_2) = |B_{44}^{(2)}|.$$

In this case, GECP would require an additional set of row and column pivots in the second GE step using $P_{(2\ 3)}$. Since $P = P_{(2\ 3)} \in \mathcal{P}_4$ is a perfect shuffle matrix, we see $PB(\theta_1, \theta_2)P^T = B(\theta_2, \theta_1)$, whose resulting GENP factorization is also its GECP factorization. Future work will explore the GECP factorization of butterfly matrices further.

Also, while (5.33) holds using any pivoting scheme explored here so that the max-norm of the final GE step maximizes the intermediate max-norms, empirical results suggest the max-norm increases weakly with each intermediate step in the pivoting cases. This suggests:

Conjecture 4. *Let $B \in \mathrm{B}_s(N)$. Using pivoting, then $\|B^{(k)}\|_{\max} \leq \|B^{(k+1)}\|_{\max}$ for all $1 \leq k < N$.*

This was not needed to yield our desired results, so this was not explored further in this document.

Using GENP this monotonicity does not hold. Just to quickly illustrate this point, we can return to the previous example: for $\mathrm{B}(\theta_1, \theta_2) \in \mathrm{B}_s(4)$ from before, with GENP then

$$\begin{aligned}
\|B^{(1)}\|_{\max} &= \max(|\cos\theta_1|, |\sin\theta_1|) \max(|\cos\theta_2|, |\sin\theta_2|) \\
|B^{(2)}\|_{\max} &= |\sec\theta_1 \sec\theta_2| \max(\cos^2\theta_1, |\cos\theta_1 \sin\theta_1|, \cos^2\theta_2, |\cos\theta_2 \sin\theta_2|, |\cos^2\theta_1 - \sin^2\theta_2|) \\
\|B^{(3)}\|_{\max} &= |\sec\theta_1 \sec\theta_2| \max(\cos^2\theta_1, |\cos\theta_1 \sin\theta_1|, \cos^2\theta_2, |\cos\theta_2 \sin\theta_2|)
\end{aligned}$$

$$\|B^{(4)}\|_{\max} = |\sec\theta_1 \sec\theta_2|,$$

so that

$$\|B^{(1)}\|_{\max} \leq \|B^{(3)}\|_{\max} \leq \|B^{(2)}\|_{\max} \leq \|B^{(4)}\|_{\max}.$$

If $\theta_j = \frac{\pi}{6}$ for each j, then $\|B^{(1)}\|_{\max} = \frac{3}{4}$, $\|B^{(2)}\|_{\max} = 2$, $\|B^{(3)}\|_{\max} = \sqrt{3}$ and $\|B^{(4)}\|_{\max} = 4$, so that $\|B^{(2)}\|_{\max} > \|B^{(3)}\|_{\max}$, showing monotonicity fails in the GENP case. If $|\tan\theta_j| \leq 1$ for each j, then as before these are also the intermediate GE steps using GEPP. It follows then necessarily $\cos^2\theta_1 \geq \sin^2\theta_2$, so that $|\cos^2\theta_1 - \sin^2\theta_2| = \cos^2\theta_1 - \sin^2\theta_2 \leq \cos^2\theta_1$ and hence $\|B^{(2)}\|_{\max} = \|B^{(3)}\|_{\max}$.

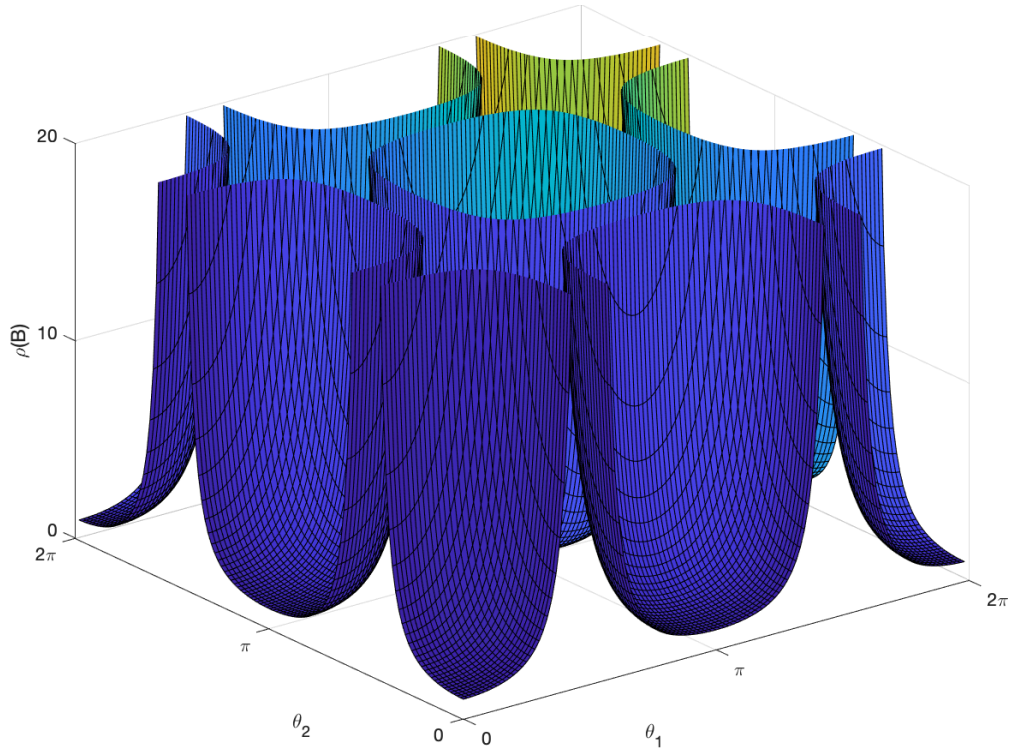


Figure 5.1: ρ using GENP on $B = B(\theta_1, \theta_2) \in \mathrm{B}_s(2)$.

Figures 5.1 and 5.2 show maps of $\rho(B)$ and $\rho_\infty(B)$ using GENP for $B = B(\theta_1, \theta_2) \in \mathrm{B}_s(2)$. Note the singularity lines for the model accounting for when $\theta_i = \pi \pm \frac{\pi}{2}$ for some i, which coincide with the case when $B_{11} = \cos\theta_1 \cos\theta_2 = 0$ and GENP factorization fails on the first step. Figures 5.3 and 5.4 show maps of $\rho(B)$ and $\rho_\infty(B)$ using GEPP or GERP for

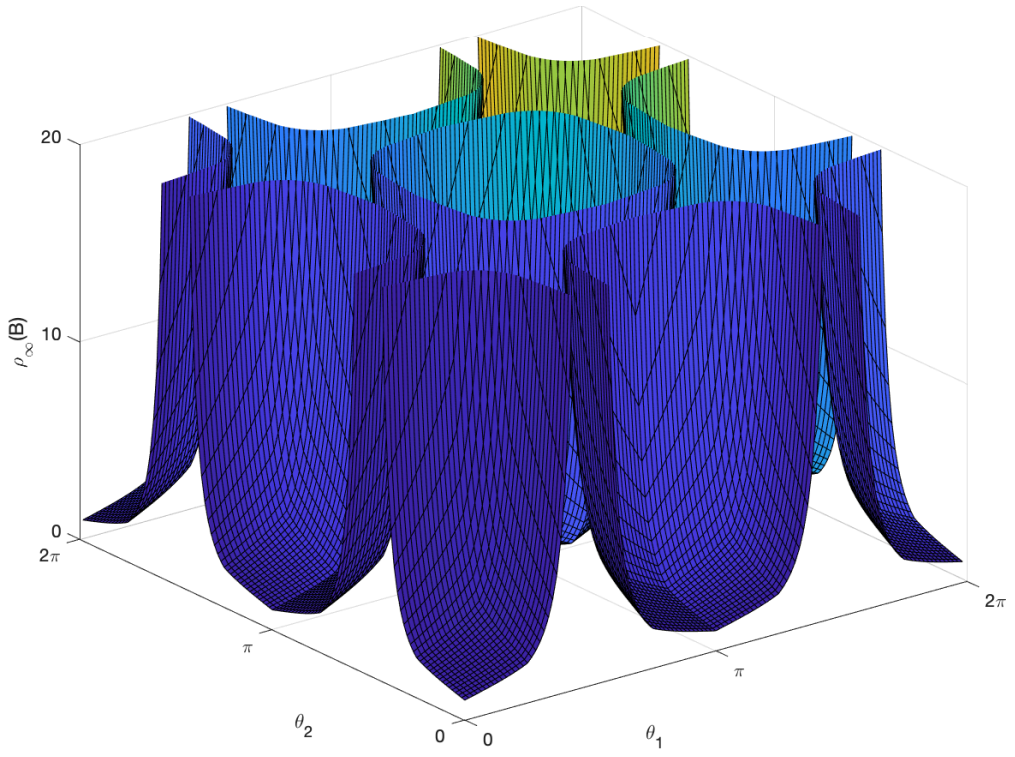


Figure 5.2: ρ_∞ using GENP on $B = B(\theta_1, \theta_2) \in \mathrm{B}_s(2)$.

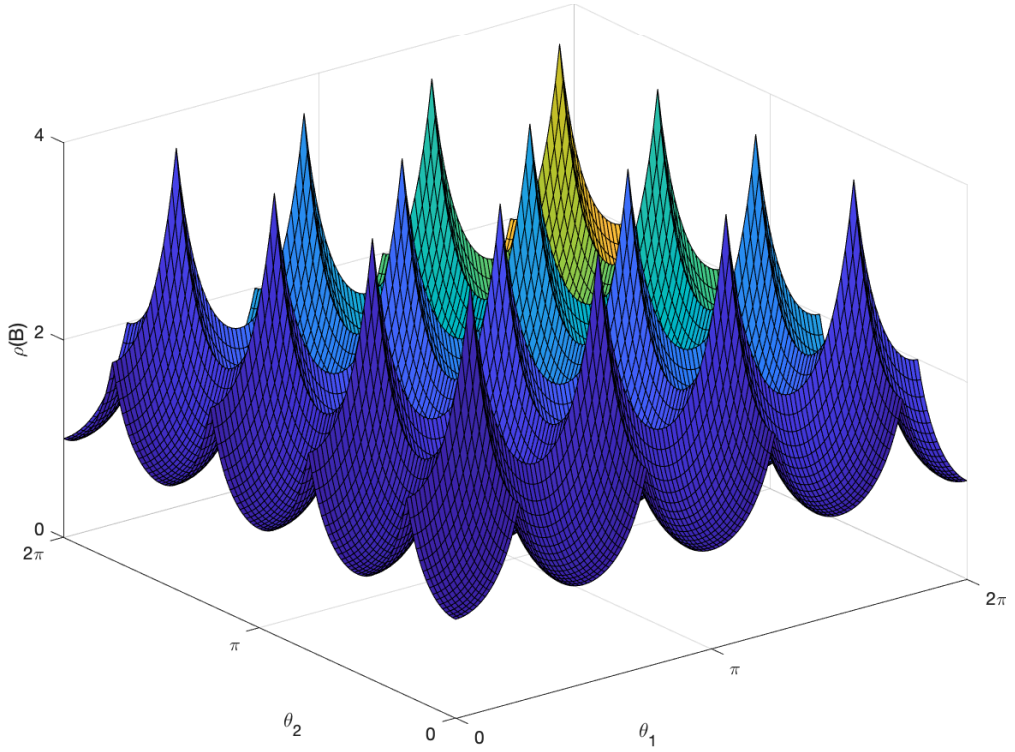


Figure 5.3: ρ using pivoting on $B = B(\theta_1, \theta_2) \in \mathrm{B}_s(2)$.

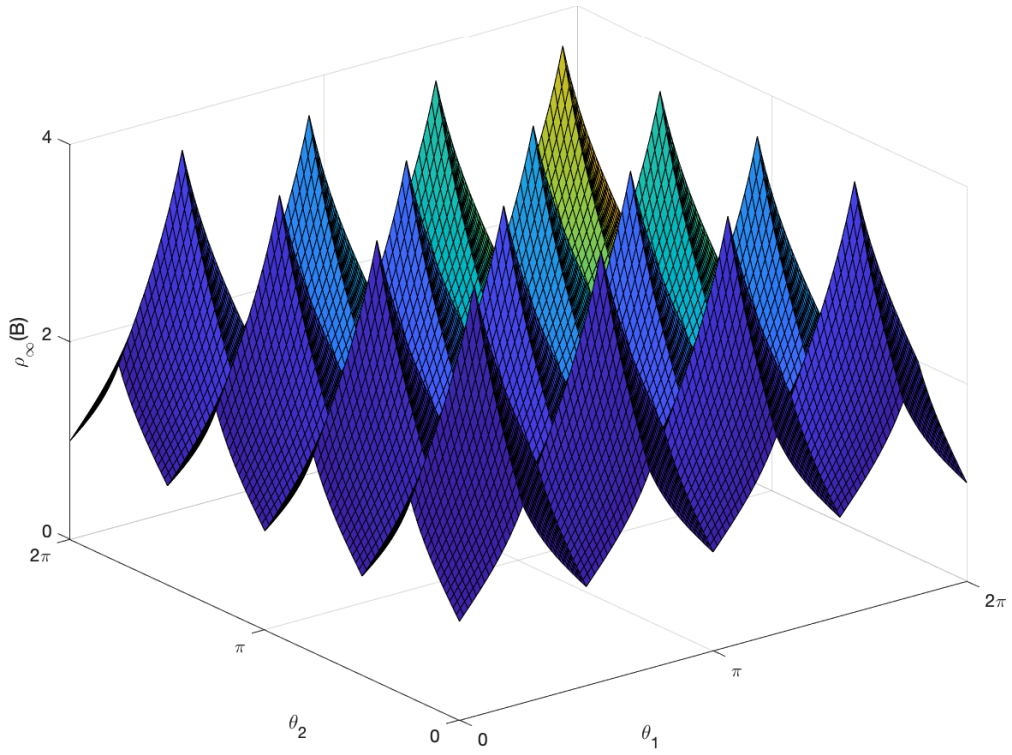


Figure 5.4: ρ_∞ using pivoting on $B = B(\theta_1, \theta_2) \in \mathrm{B}_s(2)$.

$B = B(\theta_1, \theta_2) \in \mathrm{B}_s(2)$. The relationship $1 \leq \rho(B) \leq \rho_\infty(B)$ can easily be viewed for this case, as Figure 5.1 fits inside Figure 5.2 and Figure 5.3 fits inside Figure 5.4.

From Theorem 5.3, then each of the peaks in Figures 5.3 and 5.4 occur precisely at the scaled **Hadamard matrices**, $B(\boldsymbol{\theta}) \in \mathrm{B}_s(N)$ for $\boldsymbol{\theta} \in (\frac{\pi}{4} + \frac{\pi}{2}\mathbb{Z})^n$ so that $\sqrt{N}B(\boldsymbol{\theta}) \in \{\pm 1\}^n$ with orthogonal rows and columns. As such, butterfly models can be used as a continuous approximation of Hadamard matrices to derive other desirable properties. This will be further explored in future work.

5.3.3 Worst-case model

This section will look at a specific application of butterfly matrices to decrease the growth factor in the worst-case model.

Wilkinson proved the backward stability of GEPP by showing $\rho(A) \leq 2^{N-1}$ for any non-singular $A \in \mathbb{R}^{N\times N}$ (cf. [39, 40]). In [39, pg. 202], Wilkinson further shows this bound on worst case growth factor is sharp, using the following example:

$$A_N = \mathbf{I}_N - \sum_{i>j} \mathbf{E}_{ij} + \sum_{i=1}^{N-1} \mathbf{E}_{iN}. \tag{5.44}$$

By construction, GEPP would carry out without using any pivoting, so that the LU factorizations coincide for the GENP and GEPP pivoting schemes, where

$$L = \mathbf{I}_N - \sum_{i>j} \mathbf{E}_{ij} \quad \text{and} \quad U = \mathbf{I}_N - \mathbf{E}_{NN} + \sum_{i=1}^{N} 2^{i-1}\mathbf{E}_{iN}. \tag{5.45}$$

It follows $\rho(A) = |U_{NN}| = 2^{N-1}$. For example,

$$A_4 = \begin{bmatrix} 1 & & & 1 \\ -1 & 1 & & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ -1 & -1 & 1 & \\ -1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & & & 1 \\ & 1 & & 2 \\ & & 1 & 4 \\ & & & 8 \end{bmatrix}$$

has $\rho(A_4) = 2^3 = 8$.

It happens that also $\rho_\infty(A_N) = 2^{N-1}$, although this is not the upper bound using the induced ℓ_∞-induced growth factor. Note GECP or any column pivoting scheme would result in $PAQ = LU$ for $P = \mathbf{I}$ and $Q = P_{(2\ N)(3\ N)\cdots(N-1\ N)} = P_{(2\ N\ N-1\ \cdots\ 3)}$, where $\rho(A) = 2$ (and $\rho_\infty(A) = 3$ for $N \geq 3$ with $\rho_\infty(A) = 2$ when $N = 2$). For example,

$$A_4 P_{(2\ 4\ 3)} = \begin{bmatrix} 1 & 1 & & \\ -1 & 1 & 1 & \\ -1 & 1 & -1 & 1 \\ -1 & 1 & -1 & -1 \end{bmatrix} = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ -1 & 1 & 1 & \\ -1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & & \\ & 2 & 1 & \\ & & -2 & 1 \\ & & & -2 \end{bmatrix}.$$

Moreover, a lower bound on $\kappa_2(A_N)$ can be computed. We have

$$A_N A_N^T = -\mathbf{E}_{NN} + \sum_{k=1}^{N}(k+1)\mathbf{E}_{kk} + \sum_{i>j}(j-1)(\mathbf{E}_{ij} + \mathbf{E}_{ji}). \tag{5.46}$$

Using the same example, we see

$$A_4 A_4^T = \begin{bmatrix} 2 & & & \\ & 3 & 1 & 1 \\ & 1 & 4 & 2 \\ & 1 & 2 & 4 \end{bmatrix}. \tag{5.47}$$

Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_N$ denote the singular values of A_N. Hence, σ_k^2 are the eigenvalues of $A_N A_N^T$.

We can give an explicit quadratic form of $\mathbf{x}^T A_N A_N^T \mathbf{x}$. This follows again from straightforward induction:

Lemma 5.1. *For any* $\mathbf{x} \in \mathbb{R}$,

$$\mathbf{x}^T \left(A_N A_N^T - 2\mathbf{I}_N\right) \mathbf{x} = \sum_{j=1}^{N-2} \left(\sum_{k=N-j}^{N} x_k \right)^2. \tag{5.48}$$

Proof. We will use induction on N. For $N = 2$, we have $A_2 A_2^T = 2\mathbf{I}_2$ so the result holds trivially. For $N = 3$, then

$$A_3 A_3^T - 2\mathbf{I}_2 = \begin{bmatrix} 0 & & \\ & 1 & 1 \\ & 1 & 1 \end{bmatrix} \tag{5.49}$$

so that $\mathbf{x}^T(A_3 A_3^T - 2\mathbf{I}_3)\mathbf{x} = (x_2 + x_3)^2$.

Now assume the result holds for N. Note

$$A_{N+1}A_{N+1}^T - 2\mathbf{I}_{N+1} = \begin{bmatrix} A_N A_N^T - 2\mathbf{I}_N + \mathbf{E}_{NN} & \mathbf{v} \\ \mathbf{v}^T & N-1 \end{bmatrix} \tag{5.50}$$

for

$$\mathbf{v} = \sum_{j=2}^{N}(j-1)\mathbf{e}_j. \tag{5.51}$$

For $\mathbf{x} = (\mathbf{x}', x_{N+1}) \in \mathbb{R}^{N+1}$, then

$$\begin{aligned}
&\mathbf{x}^T(A_{N+1}A_{N+1}^T - 2\mathbf{I}_{N+1})\mathbf{x} \\
&\quad = \mathbf{x}'^T(A_N A_N^T - 2\mathbf{I}_N)\mathbf{x}' + x_N^2 + 2x_{N+1}\mathbf{v}^T\mathbf{x}' + (N-1)x_{N+1}^2 \\
&\quad = \sum_{j=1}^{N-2}\left(\sum_{k=N-j}^{N} x_k\right)^2 + x_N^2 + 2x_{N+1}\sum_{j=2}^{N}(j-1)x_j + (N-1)x_{N+1}^2 \\
&\quad = \sum_{j=0}^{N-2}\left(x_{N+1}^2 + \left(\sum_{k=N-j}^{N} x_k\right)^2\right) + 2x_{N+1}\sum_{j=2}^{N}(j-1)x_j \\
&\quad = \sum_{j=0}^{N-2}\left(x_{N+1} + \sum_{k=N-j}^{N} x_k\right)^2 + 2x_{N+1}\left(\sum_{j=2}^{N}(j-1)x_j - \sum_{j=0}^{N-2}\sum_{k=N-j}^{N} x_k\right) \\
&\quad = \sum_{j=1}^{N-1}\left(\sum_{k=N+1-j}^{N+1} x_k\right)^2
\end{aligned}$$

using the inductive hypothesis for the second line along with the fact

$$\sum_{j=0}^{N-2}\sum_{k=N-j}^{N} x_k = \sum_{k=2}^{N}\sum_{j=N-k}^{N-2} x_k = \sum_{k=2}^{N}((N-2)-(N-k)+1)x_k = \sum_{j=2}^{N}(j-1)x_j.$$

□

Corollary 5.5. *For any* $\mathbf{x} \in \mathbb{R}^N$,

$$\mathbf{x}^T A_N A_N^T \mathbf{x} = \sum_{j=1}^{N-2} \left(\sum_{k=N-j}^{N} x_k \right)^2 + 2\|\mathbf{x}\|_2^2. \tag{5.52}$$

In particular, the eigenvalues of $A_N A_N^T$ are bounded below by 2: if $\mathbf{x}$ is a unit eigenvector for σ_k^2, then

$$\sigma_k^2 = \mathbf{x}^T A_N A_N^T \mathbf{x} = \sum_{j=1}^{N-2} \left(\sum_{k=N-j}^{N} x_k \right)^2 + 2\|\mathbf{x}\|_2^2 \geq 2.$$

Since $\mathbf{e}_1$ and $\mathbf{e}_{N-1} - \mathbf{e}_N$ are eigenvectors for $A_N A_N^T$ with associated eigenvalue 2, then $\sigma_N^2 = \sigma_{N-1}^2 = 2$. Since also

$$\sigma_1^2 \geq \frac{1}{N} \mathbf{1}_N^T A_N A_N^T \mathbf{1}_N = \frac{1}{N} \left(\sum_{j=1}^{N-2} (j+1)^2 + 2N \right) = \frac{N^2}{3} - \frac{N}{2} + \frac{13}{6} - \frac{1}{N},$$

it follows

$$\kappa_2(A_N) = \frac{\sigma_1}{\sigma_N} \geq \sqrt{\frac{N^2}{6} - \frac{N}{4} + \frac{13}{12} - \frac{1}{2N}} = \frac{N}{\sqrt{6}}(1 + o(1)). \tag{5.53}$$

Also using (5.52), if $\|\mathbf{x}\|_2 = 1$ is such that $\mathbf{x}^T A_N A_N^T \mathbf{x} = \sigma_1^2$ then

$$\sigma_1^2 = \mathbf{x}^T A_N A_N^T \mathbf{x} \leq \sum_{j=1}^{N-2} \left((j+1) \sum_{k=N-j}^{N} x_k^2 \right) + 2\|\mathbf{x}\|_2^2 \leq 1 + \sum_{j=1}^{N-1} j = \frac{N^2}{2} - \frac{N}{2} + 1$$

using the Cauchy-Schwarz inequality. Similarly, it follows

$$\kappa_2(A_N) \leq \sqrt{\frac{N^2}{4} - \frac{N}{4} + \frac{1}{2}} = \frac{N}{2}(1 + o(1)). \tag{5.54}$$

Figure 5.5 shows how these lower and upper bounds compares against the computed value of $\kappa_2(A_N)$.

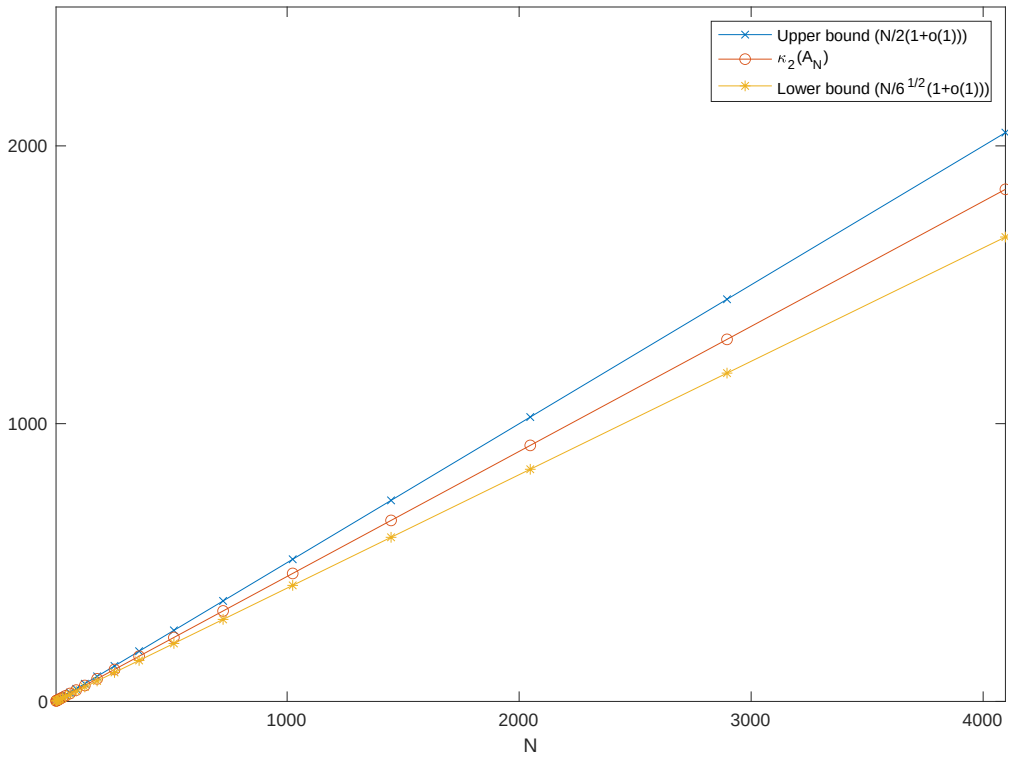


Figure 5.5: $\kappa_2(A_N)$ versus the bounds in (5.53) and (5.54)

Higham and Higham show that any square matrix $A \in \mathbb{R}^{N\times N}$ with worst case growth factor $\rho(A) = 2^{N-1}$ can be classified as follows:

Theorem 5.4 ([17]). *Let $A \in \mathbb{R}^{N\times N}$. Then $\rho(A) = 2^{N-1}$ using GEPP if and only if*

$$A = DEF \quad for \quad F = \left[\begin{array}{c|c} G & \\ \mathbf{0}^T & \theta\mathbf{v} \end{array}\right] \tag{5.55}$$

where $D = \mathrm{diag}(\pm 1)$, E is unit lower triangular with $E_{ij} = -1$ for $i > j$, G is an upper triangular nonsingular matrix of order $N-1$, $\mathbf{v} \in \mathbb{R}^{N-1}$ with $v_i = 2^{i-1}$, and $\theta = |A_{1n}| = \|A\|_{\max}$.

For example, A_N is constructed with $D = \mathbf{I}_N$, $G = \mathbf{I}_{N-1}$ and $\theta = 1$. Note $A = LU$ using GENP or GEPP holds for $L = DED$ and $U = DF$.

This model can be used to construct matrices with maximal growth factors (with respect to

GEPP) and arbitrarily large 2-condition numbers. Recently, Higham, Higham and Pranesh studied a particular model with relatively large (viz., linear) growth factors and a parameter that controls the 2-condition number [18]. This construction relies multiplying a diagonal matrix $\mathbf{I}_{N-1} \oplus \theta$ on the left and right by two independent Haar(O(N)) matrices. We will explore two separate new constructions, which can then be used for form random ensembles with maximal growth factors and arbitrarily large 2-condition numbers.

First model: $A_N(X)$

First, we can consider the case when randomness is introduced only in the θ parameter in Theorem 5.4. Note $A_N \mathbf{E}_{NN} A_N^T = \mathbf{1}_N \mathbf{1}_N^T$. Hence, taking the model $A_N(\theta)$ as in Theorem 5.4 with $D = \mathbf{I}_N, G = \mathbf{I}_{N-1}$ and $\theta \geq 1$, we have $A_N(\theta) = A_N(\mathbf{I}_{N-1} \oplus \theta)$. Since

$$(\mathbf{I}_{N-1} \oplus \theta)(\mathbf{I}_{N-1} \oplus \theta)^T = \mathbf{I}_N + (\theta^2 - 1)\mathbf{E}_{NN},$$

then

$$A_N(\theta) A_N(\theta)^T = A_N A_N^T + (\theta^2 - 1)\mathbf{1}_N \mathbf{1}_N^T. \tag{5.56}$$

Using (5.52), we have for any $\mathbf{x} \in \mathbb{R}^N$

$$\mathbf{x}^T A_N(\theta) A_N(\theta)^T \mathbf{x} = \sum_{j=1}^{N-2} \left(\sum_{k=N-j}^{N} x_k \right)^2 + 2\|\mathbf{x}\|_2^2 + (\theta^2 - 1) \left(\sum_{j=1}^{N} x_j \right)^2. \tag{5.57}$$

Now for $\sigma_k(\theta)$ the singular values of $A_N(\theta)$, we similarly have $\sigma_k(\theta)^2 \geq 2$ for all k using (5.57) along with $\theta \geq 1$. Note for $\mathbf{u} = \mathbf{e}_{N-1} - \mathbf{e}_N$ then $A_N(\theta) A_N(\theta)^T \mathbf{u} = A_N A_N^T \mathbf{u} = 2\mathbf{u}$, so that $\sigma_N(\theta)^2 = 2$. Also,

$$\sigma_1(\theta)^2 \geq \frac{1}{N} \mathbf{1}_N^T A_N(\theta) A_N(\theta)^T \mathbf{1}_N = \frac{1}{N} \mathbf{1}_N^T A_N A_N^T \mathbf{1}_N + (\theta^2 - 1) \frac{1}{N} (\mathbf{1}_N^T \mathbf{1}_N)^2$$

$$= \frac{N^2}{3} + \left(\theta^2 - \frac{3}{2}\right) N + \frac{13}{6} - \frac{1}{N}$$

Hence,

$$\kappa_2(A_N(\theta)) = \frac{\sigma_1(\theta)}{\sigma_N(\theta)} \geq \left(\frac{N}{3} + \theta^2\right)^{1/2} \sqrt{\frac{N}{2}}(1 + o(1)). \tag{5.58}$$

In particular, for $\varphi \geq 0$ and $\theta = N^\varphi$ then

$$\kappa_2(A_N(\theta)) \gtrsim \begin{cases} \frac{1}{\sqrt{6}} N & \text{if } \varphi < \frac{1}{2} \\ \sqrt{\frac{2}{3}} N & \text{if } \varphi = \frac{1}{2} \\ \frac{1}{\sqrt{2}} N^{\varphi + \frac{1}{2}} & \text{if } \varphi > \frac{1}{2}. \end{cases} \tag{5.59}$$

Now write $A_N(X) = A_N(N^{X+\frac{1}{2}})$ for X a nonnegative random variable. For example, taking $X \sim \text{Bernoulli}(p)$, then

$$\mathbb{E}\kappa_2(A_N(X)) \gtrsim \sqrt{\frac{2}{3}} N(1-p) + \frac{1}{\sqrt{2}} N^{3/2} p = \frac{p}{\sqrt{2}} N^{3/2} + O(N). \tag{5.60}$$

If $X \sim \text{Poisson}(\lambda)$, then

$$\mathbb{E}\kappa_2(A_N(X)) \gtrsim e^{-\lambda} \sqrt{\frac{2}{3}} N + e^{-\lambda} \sqrt{\frac{N}{2}} \sum_{j \geq 1} \frac{(N\lambda)^j}{j!} = e^{-\lambda} \sqrt{\frac{N}{2}} e^{N\lambda} + O(N).$$

If $X \sim \text{Uniform}(0, 1)$, then

$$\mathbb{E}\kappa_2(A_N(X)) \gtrsim \sqrt{\frac{N}{2}}(N-1) = \frac{1}{\sqrt{2}} N^{3/2} + O(\sqrt{N}). \tag{5.61}$$

If $X \sim |Z|$ for $Z \sim N(0, 1)$, then

$$\mathbb{E}\kappa_2(A_N(X)) \gtrsim \sqrt{\frac{N}{2}} \mathbb{E} e^{(\log N)|Z|} = \sqrt{\frac{N}{2}} N^{\frac{1}{2}\log N} \left(1 + \text{erf}\left(\frac{\log N}{\sqrt{2}}\right)\right) = \sqrt{2e} N^{\frac{1}{2}\log N}(1 + o(1)). \tag{5.62}$$

If $X \sim Z^2$ for $Z \sim N(0,1)$ or $X \sim \text{Cauchy}(1)$, then $\kappa_2(A_N(X))$ does not have a finite mean.

Second model: $A_N(\beta)$

This model will limit randomness in how G is generated in Theorem 5.4. Take $A \sim \text{Ginibre}(N-1)$ and $\mathbf{X}$ independent of A with independent entries such that $X_k \sim \chi_k(\beta(N-k+1))$. Let $G = \text{diag}(\mathbf{X}) + \sum_{j>i} A_{ij}\mathbf{E}_{ij}$. Let

$$\theta = \left\| E \begin{bmatrix} G \\ \mathbf{0}^T \end{bmatrix} \right\|_{\max}. \tag{5.63}$$

Now write $A_N(\beta) = EF$ for $F = F(G,\theta)$ formed as in Theorem 5.4.

As mentioned in [17], implementing $A_N(\beta)$ versus $A_N(X)$ using floating-point arithmetic instead of exact arithmetic can encounter some unexpected snags in that running GEPP can inadvertently trigger pivoting. This follows since the computed product EF may no longer have leading entries of the same magnitude for each intermediate GE step. This can occur for small N even. An early introduction of pivoting on these class of matrices can have a significant reduction in the growth factor.

Example 5.6. *Using the row pivot $P_{(2\ 3)}$ on $A_5^{(2)}$ results in $\|U\|_{\max} = 4$ instead of 2^4.*

Note this potential unexpected behavior results in using GEPP with floating-point arithmetic instead of exact arithmetic with $A_N(\beta)$. This does not occur when using $A_N(X)$ since the leading untriangularized columns are not impacted by any accumulated error from earlier steps. For example, the unexpected behavior occurs as early as $N = 4$. Here is one sample

point for $A_4(1)$ that illustrates this point:

$$W = \begin{bmatrix} 2.1249 & -1.2820 & 0.9062 & 4.0689 \\ -2.1249 & 4.0689 & 0.2786 & 4.0689 \\ -2.129 & -1.5050 & -1.1367 & 4.0689 \\ -2.1249 & -1.5050 & -3.0452 & 4.0689 \end{bmatrix}. \tag{5.64}$$

Then

$$W^{(3)} = \begin{bmatrix} 2.1249 & -1.2820 & 0.9062 & 4.0689 \\ & 2.7870 & 1.1848 & 8.1379 \\ & & 0.9542 & 16.2758 \\ & & -0.9542 & 16.2758 \end{bmatrix}. \tag{5.65}$$

Using double precision in MATLAB, we have $|W_{43}| - |W_{33}| = 3.3307 \cdot 10^{-16}$, so that running LU in MATLAB results in a row pivot for the last two rows before the final step. With exact arithmetic, $W_{43} = -W_{33}$, so no final pivot is needed.

For a workaround to avoid unnecessary pivoting when running a built-in LU function, one can introduce an additional perturbation in the E factor. With an ε perturbation, we can replace $E = \mathbf{I}_N - \sum_{i>j} \mathbf{E}_{ij}$ with $\widetilde{E} = \mathbf{I}_N - (1-\varepsilon)\sum_{i>j} \mathbf{E}_{ij}$. Through experiments, this perturbation needs to grow with N, which makes this model highly unstable in the goal of preserving the maximal growth factor property when using floating-point arithmetic. For $N = 5$, $\varepsilon = 10^{-12}$ suffices, but for $N = 50$, $\varepsilon = 10^{-1}$. Again, this indicates why GEPP is often successful in practice. Even with models that would have a large growth factor, the accumulated computational errors compound to introduce a perturbation that shifts the maximal growth model to an often moderate growth factor class.

Example 5.7. *Introducing a row pivot in step 2 using*

$$A_5 = \begin{bmatrix} 1 & & & & 1 \\ -1 & 1 & & & 1 \\ -1 & -1 & 1 & & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix} \quad \textit{so that} \quad \widetilde{A}_5^{(2)} = P_{(2\ 3)} A_5^{(2)} = \begin{bmatrix} 1 & & & & 1 \\ & -1 & 1 & & 2 \\ & 1 & & & 2 \\ & -1 & -1 & 1 & 2 \\ & -1 & -1 & -1 & 2 \end{bmatrix}$$

results in the final

$$U = \begin{bmatrix} 1 & & & & 1 \\ & -1 & 1 & 2 & \\ & & -2 & 1 & \\ & & & -2 & \\ & & & & 4 \end{bmatrix}$$

such that $\|U\|_{\max} = 4$ *instead of* $2^4 = 16$.

Numerical experiments

For the worst-case model, our goal is to understand how much of a dampening impact preconditioning by butterfly matrices may have on the growth factor. We will study only *one-sided* conditioning by looking at models of the form ΩA_N for Ω a random orthogonal transformation. For comparison, we will consider $\Omega \sim \mathrm{B}_s(N, \Sigma)$, $\Omega \sim \mathrm{B}(N, \Sigma)$ for $\Sigma = \Sigma_S$ and $\Sigma = \Sigma_D$, along with $\Omega \sim \mathrm{Haar}(\mathrm{O}(N))$ and $\Omega \sim \mathrm{DCT}$, with the latter indicating a (deterministic) implementation of the Discrete Cosine Transformation (DCT) II that is then multiplied on the left by a diagonal matrix D with $\mathrm{diag}(D) \sim \mathrm{Uniform}(\{\pm 1\}^N)$. For these experiments, we will compute only $\rho_\infty(\Omega A_N)$ using GENP, GEPP and GECP. Each set of experiments consists of 10^4 trials for 2^n for $n = 2$ to 8, with the explicit results given below focused on the $n = 8$ case.

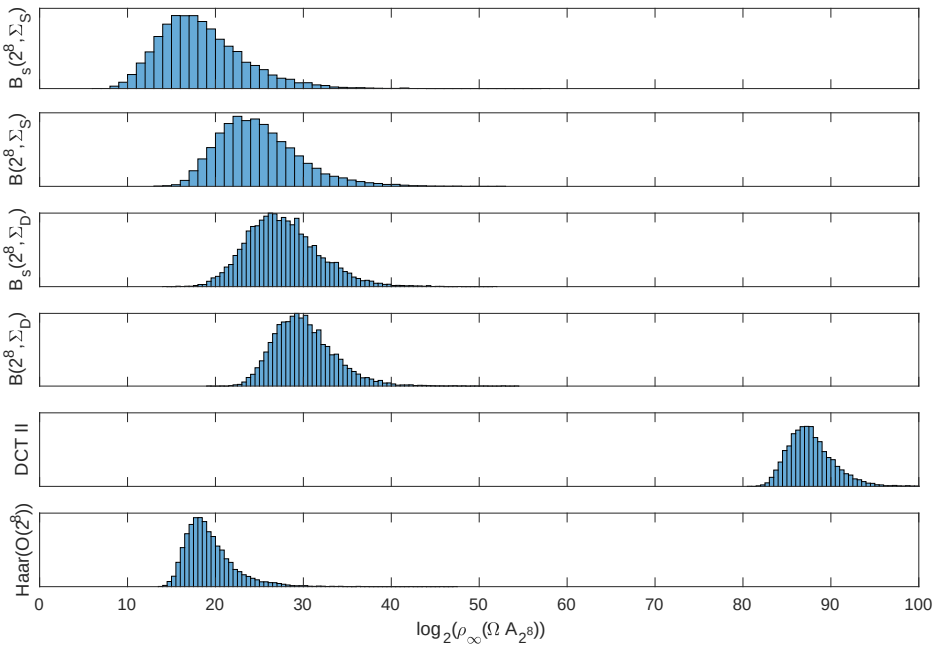


Figure 5.6: $\log_2(\rho_\infty(\Omega A_N))$ using GENP for random orthogonal Ω and $N = 2^8 = 256$, using 10^4 trials.

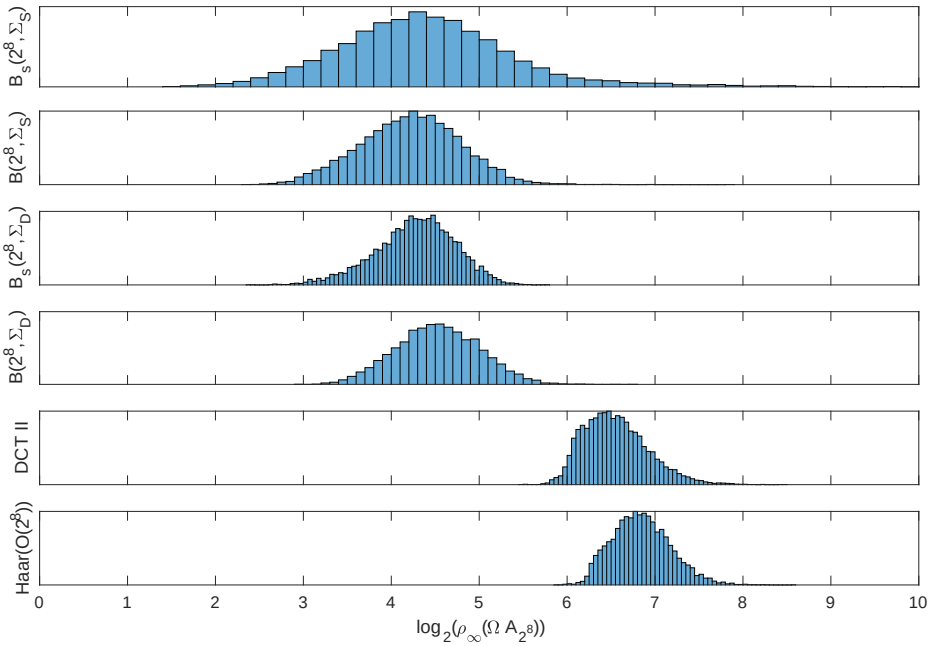


Figure 5.7: $\log_2(\rho_\infty(\Omega A_N))$ using GEPP for random orthogonal Ω and $N = 2^8 = 256$, using 10^4 trials.

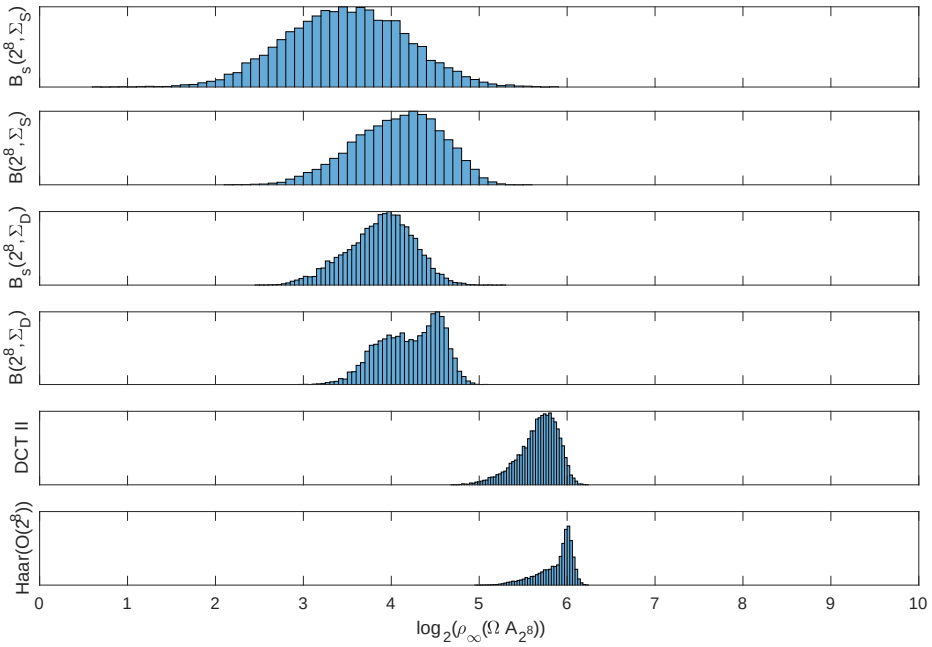


Figure 5.8: $\log_2(\rho_\infty(\Omega A_N))$ using GECP for random orthogonal Ω and $N = 2^8 = 256$, using 10^4 trials.

Recall $\rho_\infty(A_N) = 2^{N-1}$ so that $\log_2(\rho_\infty(A_N)) = N - 1$. So using Figures 5.6 and 5.7, with $N - 1 = 2^8 - 1 = 255$, each of the orthogonal models introduced have a significant dampening affect. Using GENP, the DCT had less of a dampening impact compared to the other models while using GEPP and GECP then DCT and Haar(O(N)) had similar behavior. Note Figure 5.8 is closer to the naïve model in that $\rho_\infty(A_N) = 3$ (for $N \geq 3$), so this case determines which random orthogonal transformations do the least damage. Also, it is worth noting that overall each figure provides a visual cue for the median rather than the mean of the logarithmic growth factors since we are mapping $\log_2(\rho_\infty)$.

5.3.4 Proofs of theorems

Note first using (2.12) then we can find the GENP and GEPP factorizations of simple butterfly matrices directly:

Proposition 5.4. *Let $B = B(\boldsymbol{\theta}) \in \mathrm{B}_s(N)$. If $\cos\theta_i \neq 0$ for all i, then B has GENP factorization $B = L_{\boldsymbol{\theta}} U_{\boldsymbol{\theta}}$ for*

$$L_{\boldsymbol{\theta}} = \bigotimes_{j=1}^{n} \begin{bmatrix} 1 & \\ -\tan\theta_{n-j+1} & 1 \end{bmatrix} \quad \text{and} \quad U_{\boldsymbol{\theta}} = \bigotimes_{j=1}^{n} \begin{bmatrix} \cos\theta_{n-j+1} & \sec\theta_{n-j+1} \\ & \sec\theta_{n-j+1} \end{bmatrix} \tag{5.66}$$

Moreover, for $k_n k_{n-1} \ldots k_0$ the binary representation of k where $k_j = \mathbb{1}_{[2^j,\infty)}(k \pmod{2^{j+1}})$, then

$$U_{kk} = \prod_{j=1}^{n} (\cos\theta_j)^{(-1)^{(k-1)_{j-1}}} \tag{5.67}$$

and

$$\det B_{:k,:k} = \prod_{j=1}^{k} U_{jj} = \prod_{j=1}^{n} (\cos\theta_j)^{\max(|k \pmod{2^j}|, |-k \pmod{2^j}|)}. \tag{5.68}$$

Let $\boldsymbol{\theta}'$ be such that

$$\theta_i' = \begin{cases} \theta_i & \text{if } |\tan\theta_i| \leq 1 \\ \frac{\pi}{2} - \theta_i & \text{if } |\tan\theta_i| > 1. \end{cases} \tag{5.69}$$

The GEPP factorization of B is $PB = LU$ where

$$P = P_{\boldsymbol{\theta}} = \bigotimes_{j=1}^{n} P_{(1\ 2)}^{\mathbb{1}_{(1,\infty)}(|\tan\theta_{n-j+1}|)} \quad \text{and} \quad D_{\boldsymbol{\theta}} = \bigotimes_{j=1}^{n} \left((-1)^{\mathbb{1}_{(1,\infty)}(|\tan\theta_{n-j+1}|)} \oplus 1\right) \tag{5.70}$$

and $L = L_{\boldsymbol{\theta}'}$, $U = U_{\boldsymbol{\theta}'} D_{\boldsymbol{\theta}}$. Moreover, for $B(\boldsymbol{\theta}') \in \mathrm{B}_s(N)$ then $(PB)^{(k)} = B(\boldsymbol{\theta}')^{(k)} D$ for all k.

Proof. First consider the GENP case. Note if $\cos\theta \neq 0$, then $B(\theta) \in \mathrm{SO}(2)$ has an LU

factorization with

$$B(\theta) = \begin{bmatrix} 1 & \\ -\tan\theta & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ & \sec\theta \end{bmatrix} = L_\theta U_\theta. \tag{5.71}$$

The result then follows by Lemma 1.7 and (2.12). (5.67) follows directly form (5.66). Next, write

$$L = \begin{bmatrix} L_{11} & \mathbf{0} \\ L_{21} & L_{22} \end{bmatrix}$$

for $L_{11} \in \mathbb{R}^{k\times k}$ and $\det L_{11} = 1$, so that

$$\begin{aligned} \det B_{:k,:k} &= \det\left(\begin{bmatrix} \mathbf{I}_k & \mathbf{0} \end{bmatrix} LU \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0} \end{bmatrix} \right) = \det\left(L_{11} \begin{bmatrix} \mathbf{I}_k & \mathbf{0} \end{bmatrix} U \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0} \end{bmatrix} \right) \\ &= \det U_{:k,:k} = \prod_{j=1}^{k} U_{jj}. \end{aligned}$$

The last equality in (5.68) follows directly from (5.67).

For the GEPP case: Let $PB = LU$ be the GEPP factorization of B. Note first using GEPP for $B(\theta) \in \mathrm{SO}(2)$, then a row pivot is needed only if $|\tan\theta| > 1$. The format for P then follows immediately from Lemma 1.7. Let $e_j = \mathbb{1}_{[|\tan\theta_j|>1]}$. Note

$$P_{(1\ 2)} = \begin{bmatrix} & 1 \\ 1 & \end{bmatrix} = \begin{bmatrix} & 1 \\ -1 & \end{bmatrix} \begin{bmatrix} -1 & \\ & 1 \end{bmatrix} = B\left(\frac{\pi}{2}\right)(-1 \oplus 1).$$

so that

$$P_{(1\ 2)}^{e_j} = B\left(\frac{\pi}{2} e_j\right)((-1)^{e_j} \oplus 1). \tag{5.72}$$

Note also

$$(-1\oplus 1)B(\theta)(-1\oplus 1) = \begin{bmatrix} -1 & \\ & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} -1 & \\ & 1 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = B(-\theta). \tag{5.73}$$

Using the mixed-product property (1.44) and (2.12), we have $PB = B'D$ for $D = D_{\boldsymbol{\theta}}$ diagonal with diagonal entries in $\{\pm 1\}$ and $B' = B(\boldsymbol{\theta}') \in \mathrm{B}_s(N)$ with $\boldsymbol{\theta}'$ such that

$$\theta'_j = \frac{\pi}{2}e_j + (-1)^{e_j}\theta_j = \begin{cases} \theta_j & \text{if } |\tan\theta_j| \leq 1 \\ \frac{\pi}{2} - \theta_j & \text{if } |\tan\theta_j| > 1. \end{cases}$$

If $B' = L'U'$ is the GENP factorization of B' then $B'D = L'(U'D)$ is the GENP factorization of $B'D$. It follows

$$(PB)^{(k)} = (B'D)^{(k)} = (L')^{(k)}B'D = B'^{(k)}D. \tag{5.74}$$

The final factorization follows from the GENP case applied to B'. □

Having this explicit LU factorization of $B \in \mathrm{B}_s(N)$ allows us to also construct each of the intermediate matrices $B^{(k)}$ as well.

Lemma 5.2. *Suppose $A \in \mathbb{R}^{N/2\times N/2}$ has an LU factorization using GENP. Let*

$$B = \begin{bmatrix} \cos\theta A & \sin\theta A \\ -\sin\theta A & \cos\theta A \end{bmatrix} = B(\theta) \otimes A$$

for $\cos\theta \neq 0$. *Then* B *has an LU factorization using GENP. Moreover, if* $k \leq N/2$, *then*

$$B^{(k)} = \begin{bmatrix} \cos\theta A^{(k)} & \sin\theta A^{(k)} \\ -\sin\theta \begin{bmatrix} \mathbf{0} & \\ & \mathbf{I}_{N/2-k+1} \end{bmatrix} A^{(k)} & \sec\theta \left(A - \sin^2\theta \begin{bmatrix} \mathbf{0} & \\ & \mathbf{I}_{N/2-k+1} \end{bmatrix} A^{(k)} \right) \end{bmatrix}. \quad (5.75)$$

If $k = N/2 + j$ *for* $j \geq 1$, *then*

$$B^{(k)} = \begin{bmatrix} \cos\theta A^{(N/2)} & \sin\theta A^{(N/2)} \\ & \sec\theta A^{(j)} \end{bmatrix}. \quad (5.76)$$

Proof. Let $A = L'U'$ be the GENP factorization of A. Note first

$$B(\theta) = \begin{bmatrix} 1 & \\ -\tan\theta & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ & \sec\theta \end{bmatrix} = L_\theta U_\theta. \quad (5.77)$$

Then B has an LU factorization using GENP, and let $B = LU$ where

$$L = L_\theta \otimes L' = \begin{bmatrix} L' & \\ -\tan\theta L' & L' \end{bmatrix} \quad \text{and} \quad U = U_\theta \otimes U' = \begin{bmatrix} \cos\theta U' & \sin\theta U' \\ & \sec\theta U' \end{bmatrix} \quad (5.78)$$

by Lemma 1.7. Recall

$$B^{(k)} = L_k B^{(k-1)} = L_k L_{k-1} \cdots L_1 B =: L^{(k)} B, \quad (5.79)$$

where $L^{(1)} := \mathbf{I}$, $L^{(N)} = L^{-1}$, and for $1 \leq k < N$

$$L_k = \mathbf{I} - \sum_{i \geq k} L_{i,k-1} \mathbf{E}_{i,k-1}.$$

It follows

$$
\begin{aligned}
L^{(k)^{-1}} &= L_1^{-1}\cdots L_k^{-1} = (\mathbf{I}+\sum_{i>1} L_{i1}\mathbf{E}_{i1})\cdots(\mathbf{I}+\sum_{i>k-1} L_{i,k-1}\mathbf{E}_{i,k-1}) \\
&= \mathbf{I}+\sum_{i>j,k>j} L_{ij}\mathbf{E}_{ij} \\
&= \left[\begin{array}{c|c} L_{:,:k-1} & \begin{array}{c}\mathbf{0}\\ \hline \mathbf{I}_{N-k+1}\end{array} \end{array}\right]. \qquad (5.80)
\end{aligned}
$$

If $k \le N/2$, then by (5.78) we have

$$
\begin{aligned}
L^{(k)^{-1}} &= \left[\begin{array}{c|c|c} L'_{:,:k-1} & \begin{array}{c}\mathbf{0}\\ \hline \mathbf{I}_{N/2-k+1}\end{array} & \\ \hline -\tan\theta L'_{:,:k-1} & \mathbf{0} & \mathbf{I}_{N/2} \end{array}\right] \\
&= \left[\begin{array}{c|c} L'^{(k)^{-1}} & \\ \hline -\tan\theta L'^{(k)^{-1}} \begin{bmatrix}\mathbf{I}_{k-1} & \\ & \mathbf{0}\end{bmatrix} & \mathbf{I}_{N/2} \end{array}\right] \\
&= \begin{bmatrix} L'^{(k)^{-1}} & \\ & L'^{(k)^{-1}} \end{bmatrix} \begin{bmatrix} & \mathbf{I}_{N/2} \\ -\tan\theta \begin{bmatrix}\mathbf{I}_{k-1} & \\ & \mathbf{0}\end{bmatrix} & \mathbf{I}_{N/2} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{N/2} & \\ & L'^{(k)} \end{bmatrix},
\end{aligned}
$$

and so

$$
L^{(k)} = \begin{bmatrix} \mathbf{I}_{N/2} & \\ & L'^{(k)^{-1}} \end{bmatrix} \begin{bmatrix} & \mathbf{I}_{N/2} \\ \tan\theta \begin{bmatrix}\mathbf{I}_{k-1} & \\ & \mathbf{0}\end{bmatrix} & \mathbf{I}_{N/2} \end{bmatrix} \begin{bmatrix} L'^{(k)} & \\ & L'^{(k)} \end{bmatrix}
$$

$$= \begin{bmatrix} & L'^{(k)} & \\ \tan\theta L'^{(k)-1} \begin{bmatrix} \mathbf{I}_{k-1} & \\ & \mathbf{0} \end{bmatrix} L'^{(k)} & & \mathbf{I}_{N/2} \end{bmatrix}.$$

It follows

$$\begin{aligned}
B^{(k)} = L^{(k)}B &= \begin{bmatrix} \mathbf{I} & \\ & L'^{(k)-1} \end{bmatrix} \begin{bmatrix} & \mathbf{I} & \\ \tan\theta \begin{bmatrix} \mathbf{I}_{k-1} & \\ & \mathbf{0} \end{bmatrix} & & \mathbf{I} \end{bmatrix} \begin{bmatrix} L'^{(k)} & \\ & L'^{(k)} \end{bmatrix} B \\
&= \begin{bmatrix} \mathbf{I} & \\ & L'^{(k)-1} \end{bmatrix} \begin{bmatrix} & \mathbf{I} & \\ \tan\theta \begin{bmatrix} \mathbf{I}_{k-1} & \\ & \mathbf{0} \end{bmatrix} & & \mathbf{I} \end{bmatrix} \begin{bmatrix} \cos\theta A^{(k)} & \sin\theta A^{(k)} \\ -\sin\theta A^{(k)} & \cos\theta A^{(k)} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{I} & \\ & L'^{(k)-1} \end{bmatrix} \begin{bmatrix} \cos\theta A^{(k)} & \sin\theta A^{(k)} \\ \sin\theta \left(\begin{bmatrix} \mathbf{I}_{k-1} & \\ & \mathbf{0} \end{bmatrix} - \mathbf{I} \right) A^{(k)} & \sec\theta \left(\sin^2\theta \begin{bmatrix} \mathbf{I}_{k-1} & \\ & \mathbf{0} \end{bmatrix} + \cos^2\theta \mathbf{I} \right) A^{(k)} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{I} & \\ & L'^{(k)-1} \end{bmatrix} \begin{bmatrix} \cos\theta A^{(k)} & \sin\theta A^{(k)} \\ -\sin\theta \begin{bmatrix} \mathbf{0} & \\ & \mathbf{I}_{N/2-k+1} \end{bmatrix} A^{(k)} & \sec\theta \begin{bmatrix} \mathbf{I}_{k-1} & \\ & \cos^2\theta \mathbf{I}_{N/2-k+1} \end{bmatrix} A^{(k)} \end{bmatrix} \\
&= \begin{bmatrix} \cos\theta A^{(k)} & \sin\theta A^{(k)} \\ -\sin\theta L'^{(k)-1} \begin{bmatrix} \mathbf{0} & \\ & \mathbf{I}_{N/2-k+1} \end{bmatrix} A^{(k)} & \sec\theta L'^{(k)-1} \begin{bmatrix} \mathbf{I}_{k-1} & \\ & \cos^2\theta \mathbf{I}_{N/2-k+1} \end{bmatrix} A^{(k)} \end{bmatrix}.
\end{aligned} \tag{5.81}$$

We can write

$$L'^{(k)} = \begin{bmatrix} L_1 & \\ L_2 & \mathbf{I}_{N/2-k+1} \end{bmatrix} = \begin{bmatrix} L_1 & \\ & \mathbf{I}_{N/2-k+1} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{k-1} & \\ L_2 & \mathbf{I}_{N/2-k+1} \end{bmatrix}$$

and so

$$L'^{(k)-1} = \begin{bmatrix} L_1^{-1} & \\ -L_2 L_1^{-1} & \mathbf{I}_{N/2-k+1} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{k-1} & \\ -L_2 & \mathbf{I}_{N/2-k+1} \end{bmatrix} \begin{bmatrix} L_1^{-1} & \\ & \mathbf{I}_{N/2-k+1} \end{bmatrix}.$$

First, we see

$$L'^{(k)-1} \begin{bmatrix} \mathbf{0} & \\ & \mathbf{I}_{N/2-k+1} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \\ & \mathbf{I}_{N/2-k+1} \end{bmatrix}. \tag{5.82}$$

Next, note

$$A^{(k)} = L'^{(k)} A = \begin{bmatrix} L_1 & \\ L_2 & \mathbf{I} \end{bmatrix} \begin{bmatrix} A_{:k-1,:k-1} & A_{:k-1,k:} \\ A_{k:,:k-1} & A_{k:,k:} \end{bmatrix} = \begin{bmatrix} L_1 A_{:k-1,:k-1} & L_2 A_{:k-1,k:} \\ & L_2 A_{:k-1,k:} + A_{k:,k:} \end{bmatrix},$$

where we further note

$$A^{(k)}_{k:,:k-1} = L_2 A_{:k-1,:k-1} + A_{k:,:k-1} = \mathbf{0}$$

since $A^{(k)}$ has zeros below the first $k-1$ diagonals. It follows

$$\begin{aligned} L'^{(k)-1} \begin{bmatrix} \mathbf{I} & \\ & \cos^2\theta \mathbf{I}_{N/2-k+1} \end{bmatrix} A^{(k)} &= \begin{bmatrix} \mathbf{I} & \\ -L_2 & \mathbf{I} \end{bmatrix} \begin{bmatrix} L_1^{-1} & \\ & \mathbf{I} \end{bmatrix} \begin{bmatrix} L_1 A_{:k-1,:k-1} & L_2 A_{:k-1,k:} \\ & \cos^2\theta A^{(k)}_{k:,k:} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I} & \\ -L_2 & \mathbf{I} \end{bmatrix} \begin{bmatrix} A_{:k-1,:k-1} & A_{:k-1,k:} \\ & \cos^2\theta A^{(k)}_{k:,k:} \end{bmatrix} \\ &= \begin{bmatrix} A_{:k-1,:k-1} & A_{:k-1,k:} \\ -L_2 A_{:k-1,:k-1} & -L_2 A_{:k-1,k:} + \cos^2\theta A^{(k)}_{k:,k:} \end{bmatrix} \\ &= \begin{bmatrix} A_{:k-1,:k-1} & A_{:k-1,k:} \\ A_{k:,:k-1} & A_{k:,k:} - \sin^2\theta A^{(k)}_{k:,k:} \end{bmatrix} \end{aligned}$$

$$= A - \sin^2\theta \begin{bmatrix} \mathbf{0} & \\ & \mathbf{I}_{N/2-k+1} \end{bmatrix} A^{(k)}, \tag{5.83}$$

using also $-L_2 A_{:k-1,k:} = A_{k:,k:} - A^{(k)}_{k:,k:}$. Combining (5.81), (5.82) and (5.83) then yields (5.75).

If $k = N/2 + j$ for $j \geq 1$, then again using (5.78) we have

$$\begin{aligned} {L^{(k)}}^{-1} &= \left[\begin{array}{c|c|c} L' & & \\ \hline -\tan\theta L' & L'_{:,1:j-1} & \begin{array}{c} \mathbf{0} \\ \hline \mathbf{I}_{N/2-j+1} \end{array} \end{array}\right] \\ &= \begin{bmatrix} L' & \\ -\tan\theta L' & {L'^{(j)}}^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I} & \\ -\tan\theta\mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} L' & \\ & {L'^{(j)}}^{-1} \end{bmatrix} \end{aligned}$$

so that

$$L^{(k)} = \begin{bmatrix} {L'}^{-1} & \\ & L'^{(j)} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \\ \tan\theta\mathbf{I} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} {L'}^{-1} & \\ \tan\theta L'^{(j)} & L'^{(j)} \end{bmatrix}$$

and hence

$$B^{(k)} = L^{(k)} B = \begin{bmatrix} {L'}^{-1} & \\ & L'^{(j)} \end{bmatrix} \begin{bmatrix} \cos\theta A & \sin\theta A \\ & \sec\theta A \end{bmatrix} = \begin{bmatrix} \cos\theta U' & \sin\theta U' \\ & \sec\theta A^{(j)} \end{bmatrix}. \tag{5.84}$$

(5.76) follows then by noting $U' = A^{(N/2)}$. □

Recall $|A|$ denotes the matrix with $|A|_{ij} = |A_{ij}|$. Note $|A|^T = |A^T|$.

Proposition 5.5. *Let $B \in \mathrm{B}_s(N)$ with $PB = LU$ the factorization of B using GENP (with $P = \mathbf{I}$) or GEPP. Let $\eta, \varepsilon \in \mathbb{R}$ such that $|\eta|, |\varepsilon|, |\eta - \varepsilon| \leq 1$. Then*

$$|\eta PB - \varepsilon(PB)^{(k)}|_{k:,k:} = \begin{bmatrix} \mathbf{0} & \mathbf{I}_{N-k+1} \end{bmatrix} \left| \eta PB - \varepsilon(PB)^{(k)} \right| \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{N-k+1} \end{bmatrix} \tag{5.85}$$

is symmetric. In particular, if $PB = LU$ is the GEPP factorization of B, then this is also the GERP factorization of B.

Note this depends on GERP prioritizing column pivot scans over row scans.

Proof. We can reduce to the case $P = \mathbf{I}$: there exist diagonal D with diagonal entries in $\{\pm 1\}$ and $B' \in \mathrm{B}_s(N)$ such that $PB = B'D$ and $(PB)^{(k)} = B'^{(k)}D$ by Proposition 5.4, while then

$$|\eta PB - \varepsilon(PB)^{(k)}| = |(\eta B' - \varepsilon B'^{(k)})D| = |\eta B' - \varepsilon B'^{(k)}|,$$

so the result for general P follows directly from the result for $P = \mathbf{I}$.

Now note how the GEPP and GERP factorizations necessarily align: Since $B = LU$ is the GEPP factorization, then $|B_{kk}^{(k)}| \geq |B_{ik}^{(k)}|$ for all $i \geq k$. By (5.85) with $\eta = 0$ and $\varepsilon = -1$, we have $|B^{(k)}|_{k:,k:}$ is symmetric so that also $|B_{kk}^{(k)}| \geq |B_{ik}^{(k)}| = |B_{ki}^{(k)}|$ for all $i \geq k$. It follows then GERP would not yield any column swaps so that the GEPP and GERP factorizations would align.

To prove (5.85), we will use induction on n. Note $B = B(\theta)$ satisfies

$$|B| = \begin{bmatrix} |\cos\theta| & |\sin\theta| \\ |\sin\theta| & |\cos\theta| \end{bmatrix} = |B(-\theta)| = |B|^T. \tag{5.86}$$

so that for $B = B^{(1)}$ we always have

$$|B| = \left|\bigotimes_{j=1}^{n} B(\theta_{n-j+1})\right| = \bigotimes_{j=1}^{n} |B(\theta_{n-j+1})| = \bigotimes_{j=1}^{n} |B(\theta_{n-j+1})|^T = |B|^T. \tag{5.87}$$

It suffices to consider only $k \geq 2$. If $n = 1$, for $k = 2$ then $|\eta B - \varepsilon B^{(2)}|_{2:,2:} = |\eta \cos\theta - \varepsilon \sec\theta|$ is an order 1 matrix and so is trivially symmetric.

Now assume the result holds for $\mathrm{B}_s(N/2)$ and $n \geq 2$, and let $B = B(\theta, A) \in \mathrm{B}_s(N)$ with $A \in \mathrm{B}_s(N/2)$, which also necessarily has an LU factorization using GENP. For $k \leq N/2$, then for $\mathbf{I} = \mathbf{I}_{N/2-k+1}$ when not indicated otherwise, we have

$$\begin{aligned}
&\begin{bmatrix} \mathbf{0} & \mathbf{I}_{N-k+1} \end{bmatrix} \left|\eta B - \varepsilon B^{(k)}\right| \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{N-k+1} \end{bmatrix} \\
&= \begin{bmatrix} |\cos\theta| \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} \left|\eta A - \varepsilon A^{(k)}\right| \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} & |\sin\theta| \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} \left|\eta A - \varepsilon A^{(k)}\right| \\ |\sin\theta| \left|\eta A - \varepsilon \begin{bmatrix} \mathbf{0} & \\ & \mathbf{I} \end{bmatrix} A^{(k)}\right| \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} & |\sec\theta| \left|(\varepsilon - \eta\cos^2\theta)A - \varepsilon\sin^2\theta \begin{bmatrix} \mathbf{0} & \\ & \mathbf{I} \end{bmatrix} A^{(k)}\right| \end{bmatrix}
\end{aligned}$$

using Lemma 5.2. For $\eta' = \varepsilon - \eta\cos^2\theta$ and $\varepsilon' = \sin^2\theta\varepsilon$, then $|\varepsilon'| = \sin^2\theta|\varepsilon| \leq 1$, $|\eta'| = |\eta\cos^2\theta - \varepsilon| \leq 1$ (since if $\eta > 0$ then $-1 \leq -\varepsilon \leq \eta\cos^2 -\varepsilon \leq \eta - \varepsilon \leq 1$ and similarly if $\eta \leq 0$), while also $|\eta' - \varepsilon'| = \cos^2\theta|\eta - \varepsilon| \leq 1$, then

$$\begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} \left|\eta A - \varepsilon A^{(k)}\right| \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} \left|(\varepsilon - \eta\cos^2\theta)A - \varepsilon\sin^2\theta A^{(k)}\right| \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix}$$

are symmetric by the inductive hypothesis. Hence,

$$\left|(\varepsilon - \eta\cos^2\theta)A - \varepsilon\sin^2\theta \begin{bmatrix} \mathbf{0} & \\ & \mathbf{I} \end{bmatrix} A^{(k)}\right|$$

$$= \begin{bmatrix} |\varepsilon - \eta\cos^2\theta||A_{:k-1,:k-1}| & |\varepsilon - \eta\cos^2\theta||A_{:k-1,k:}| \\ |\varepsilon - \eta\cos^2\theta||A_{k:,:k-1}| & \begin{bmatrix}\mathbf{0} & \mathbf{I}\end{bmatrix} \left|(\varepsilon - \eta\cos^2\theta)A - \varepsilon\sin^2\theta A^{(k)}\right| \begin{bmatrix}\mathbf{0}\\ \mathbf{I}\end{bmatrix} \end{bmatrix}$$

is symmetric, using also the fact $|A|^T = |A|$ by (5.87). Since also

$$\left|\eta A - \varepsilon \begin{bmatrix}\mathbf{0} & \\ & \mathbf{I}\end{bmatrix} A^{(k)}\right| \begin{bmatrix}\mathbf{0}\\ \mathbf{I}\end{bmatrix} = \left| \begin{bmatrix} |\eta||A_{:k-1,:k-1}| \\ \begin{bmatrix}\mathbf{0} & \mathbf{I}\end{bmatrix} \left|\eta A - \varepsilon A^{(k)}\right| \begin{bmatrix}\mathbf{0}\\ \mathbf{I}\end{bmatrix} \end{bmatrix} \right| = \left(\begin{bmatrix}\mathbf{0} & \mathbf{I}\end{bmatrix} |\eta A - \varepsilon A^{(k)}| \right)^T$$

then the result follows.

For $k = N/2+j$ and $j \geq 1$ so that $N-k+1 = N/2-j+1$, then writing $\mathbf{I} = \mathbf{I}_{N-k+1} = \mathbf{I}_{N/2-j+1}$, we have

$$\begin{bmatrix}\mathbf{0} & \mathbf{I}\end{bmatrix} \left|\eta B - \varepsilon B^{(k)}\right| \begin{bmatrix}\mathbf{0}\\ \mathbf{I}\end{bmatrix} = |\sec\theta| \begin{bmatrix}\mathbf{0} & \mathbf{I}\end{bmatrix} \left|\eta\cos^2\theta A - \varepsilon A^{(j)}\right| \begin{bmatrix}\mathbf{0}\\ \mathbf{I}\end{bmatrix}$$

is symmetric by the inductive hypothesis since $\eta' = \eta\cos^2\theta$ and $\varepsilon' = \varepsilon$ satisfy the hypotheses.

□

Now we will establish main tool to show the max norm of the intermediate GE steps is maximized by the final U factor.

Proposition 5.6. *Let $B \in \mathrm{B}_s(N)$, $\eta, \varepsilon \in \mathbb{R}$ such that $|\eta|, |\varepsilon|, |\eta - \varepsilon| \leq 1$. Let $PBQ = LU$ the LU factorization of B using GENP (with $P = Q = \mathbf{I}$), GEPP (with $Q = \mathbf{I}$) or GERP. Then for all k,*

$$\left\| \begin{bmatrix}\mathbf{0} & \mathbf{I}_{N-k+1}\end{bmatrix} (\eta PBQ - \varepsilon(PBQ)^{(k)}) \begin{bmatrix}\mathbf{0}\\ \mathbf{I}_{N-k+1}\end{bmatrix} \right\|_{\max} \leq \|U\|_{\max}. \tag{5.88}$$

In particular,

$$\max_k \|(PBQ)^{(k)}\|_{\max} = \|U\|_{\max}. \tag{5.89}$$

Proof. First, note it suffices to show this result for $P = Q = \mathbf{I}$: We have $PBQ = B'D$ for diagonal D with diagonal entries in $\{\pm 1\}$ and $B' \in \mathrm{B}_s(N)$ with $(PBQ)^{(k)} = B'^{(k)}D$ using Proposition 5.4 if using GEPP and Proposition 5.5 if using GERP. It follows then

$$\begin{aligned}
&\left\| \begin{bmatrix} \mathbf{0} & \mathbf{I}_{N-k+1} \end{bmatrix} (\eta PBQ - \varepsilon (PBQ)^{(k)}) \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{N-k+1} \end{bmatrix} \right\|_{\max} \\
&\quad = \left\| \begin{bmatrix} \mathbf{0} & \mathbf{I}_{N-k+1} \end{bmatrix} (\eta B' - \varepsilon B'^{(k)}) D \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{N-k+1} \end{bmatrix} \right\|_{\max} \\
&\quad = \left\| \begin{bmatrix} \mathbf{0} & \mathbf{I}_{N-k+1} \end{bmatrix} (\eta B' - \varepsilon B'^{(k)}) \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{N-k+1} \end{bmatrix} \right\|_{\max}
\end{aligned}$$

using the fact $\|\cdot\|_{\max}$ is invariant under unit multiples of rows or columns. Hence, it suffices to consider only the case $P = Q = \mathbf{I}$, so assume this holds for the remainder of this proof.

First note how (5.89) follows: since

$$B^{(k)} = \begin{bmatrix} U_{:k-1,:k-1} & U_{:k-1,k:} \\ \mathbf{0} & \begin{bmatrix} \mathbf{0} & \mathbf{I}_{N-k+1} \end{bmatrix} B^{(k)} \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{N-k+1} \end{bmatrix} \end{bmatrix}$$

then $\|B^{(k)}\|_{\max} \leq \|U\|_{\max}$ using (5.88) with $\eta = 0$ and $\varepsilon = -1$.

To prove (5.88), we will once again use induction on n. Note first the result always holds for

$k = 1$ since then

$$\|\mathbf{I}_N(\eta B - \varepsilon B^{(1)})\mathbf{I}_N\|_{\max} = |\eta - \varepsilon|\|B\|_{\max} \le \|U\|_{\max} \tag{5.90}$$

since $\|B\|_{\max} \le 1 \le \|U\|_{\max}$. So we can consider only $k \ge 2$. For $n = 1$ for $k = 2$,

$$\left\| \begin{bmatrix} 0 & 1 \end{bmatrix} (\eta B - \varepsilon B^{(2)}) \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\|_{\max} = |\eta \cos^2\theta - \varepsilon||\sec\theta| \le |\sec\theta| = \|U\|_{\max},$$

where we note $|\eta\cos^2 - \varepsilon| \le 1$ as before.

Now assume the result holds for $\mathrm{B}_s(N/2)$ and $n \ge 2$, and let $B = B(\theta, A) \in \mathrm{B}_s(N)$ with $A = L'U' \in \mathrm{B}_s(N/2)$. Note

$$\|A\|_{\max} = \|A\mathbf{e}_1\|_\infty \le 1 \le \|U'\|_{\max} \le |\sec\theta|\|U'\|_{\max} = \|U\|_{\max}, \tag{5.91}$$

using Proposition 5.4 for the last equality.

For $k \le N/2$, then for $\mathbf{I} = \mathbf{I}_{N/2-k+1}$ when not indicated otherwise, we have

$$\begin{aligned}
&\begin{bmatrix} \mathbf{0} & \mathbf{I}_{N-k+1} \end{bmatrix} (\eta B - \varepsilon B^{(k)}) \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{N-k+1} \end{bmatrix} \\
&= \begin{bmatrix} \cos\theta \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} (\eta A - \varepsilon A^{(k)}) \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} & \sin\theta \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} (\eta A - \varepsilon A^{(k)}) \\ -\sin\theta \left(\eta A - \varepsilon \begin{bmatrix} \mathbf{0} & \\ & \mathbf{I} \end{bmatrix} A^{(k)} \right) \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} & -\sec\theta \left((\varepsilon - \eta\cos^2\theta)A - \varepsilon\sin^2\theta \begin{bmatrix} \mathbf{0} & \\ & \mathbf{I} \end{bmatrix} A^{(k)} \right) \end{bmatrix}
\end{aligned}$$

using Lemma 5.2. Let $\eta' = \varepsilon - \eta\cos^2\theta$ and $\varepsilon' = \varepsilon\sin^2\theta$. As above, we have $|\eta'| \le 1$ while also $|\varepsilon'| = \sin^2\theta|\varepsilon| \le 1$ and $|\eta' - \varepsilon'| = \cos^2\theta|\eta - \varepsilon| \le 1$. By the inductive hypothesis, we

have

$$\left\| \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} (\eta A - \varepsilon A^{(k)}) \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \right\|_{\max} \leq \|U'\|_{\max} \leq \|U\|_{\max} \quad \text{and}$$

$$|\sec\theta| \left\| \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} \left(\eta' A - \varepsilon' A^{(k)}\right) \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \right\|_{\max} \leq |\sec\theta| \|U'\|_{\max} = \|U\|_{\max}.$$

Moreover, $|\eta \sin\theta| \|A\|_{\max} \leq \|U'\|_{\max} \leq \|U\|_{\max}$ and

$$|\sec\theta||\varepsilon - \eta\cos^2\theta| \|A\|_{\max} \leq |\sec\theta| \|U'\|_{\max} = \|U\|_{\max} \tag{5.92}$$

by (5.91). It follows

$$\left\| \begin{bmatrix} \mathbf{0} & \mathbf{I}_{N-k+1} \end{bmatrix} (\eta B - \varepsilon B^{(k)}) \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{N-k+1} \end{bmatrix} \right\|_{\max} \tag{5.93}$$

$$= \max \left(\begin{array}{c} |\eta \sin\theta| \|A_{N/2-k+1:,:k-1}\|_{\max}, \\ |\eta \sin\theta| \|A_{:k-1,N/2-k+1:}\|_{\max}, \\ \max(|\cos\theta|, |\sin\theta|) \left\| \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} (\eta A - \varepsilon A^{(k)}) \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \right\|_{\max}, \\ |\sec\theta||\varepsilon - \eta\cos^2\theta| \|A\|_{\max}, \\ |\sec\theta| \left\| \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} \left((\varepsilon - \cos^2\theta\eta)A - \varepsilon\sin^2\theta A^{(k)}\right) \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \right\|_{\max} \end{array} \right) \tag{5.94}$$

$$\leq \|U\|_{\max}. \tag{5.95}$$

For $k = N/2 + j$ and $j \geq 1$ so that $N - k + 1 = N/2 - j + 1$, let $\eta' = \eta\cos^2\theta$ and $\varepsilon' = \varepsilon$

then $|\eta'|, |\varepsilon'|, |\eta' - \varepsilon'| \leq 1$. Hence,

$$\begin{aligned}
&\left\| \begin{bmatrix} \mathbf{0} & \mathbf{I}_{N-k+1} \end{bmatrix} (\eta B - \varepsilon B^{(k)}) \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{N-k+1} \end{bmatrix} \right\|_{\max} \\
&\quad = |\sec\theta| \left\| \begin{bmatrix} \mathbf{0} & \mathbf{I}_{N/2-j+1} \end{bmatrix} (\eta \cos^2\theta A - \varepsilon A^{(j)}) \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{N/2-j+1} \end{bmatrix} \right\|_{\max} \\
&\quad \leq |\sec\theta| \|U'\|_{\max} = \|U\|_{\max}
\end{aligned}$$

by the inductive hypothesis. □

We apply this result to show the growth factors of Haar-butterfly matrices are multiplicative with respect to the Kronecker product factors.

Lemma 5.3. *If $B = B(\boldsymbol{\theta}) \in \mathrm{B}_s(N)$, then*

$$\kappa_\infty(B) = \prod_{j=1}^{n} \kappa_\infty(B(\theta_j)). \tag{5.96}$$

Using GENP, GEPP, or GERP, then

$$\rho(B) = \prod_{j=1}^{n} \rho(B(\theta_j)) \tag{5.97}$$

$$\rho_\infty(B) = \prod_{j=1}^{n} \rho_\infty(B(\theta_j)). \tag{5.98}$$

Proof. Let $PBQ = LU$ be the LU factorization of B using GENP, GEPP, or GERP. By Proposition 5.6, then

$$\rho(B) = \frac{\|L\|_{\max} \cdot \max_k \|B^{(k)}\|_{\max}}{\|B\|_{\max}} = \frac{\|L\|_{\max}\|U\|_{\max}}{\|B\|_{\max}}. \tag{5.99}$$

Since $B^{-1} = \bigotimes_{j=1}^{n} B(\theta_{n-j+1})^{-1}$, then the result follows from Lemmas 1.6 and 1.7. □

It thus remains only to establish the case for $n = 1$. Let

$$B = B(\theta) = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \in \mathrm{B}_s(2) \tag{5.100}$$

then

$$B = \begin{bmatrix} 1 & \\ -\tan\theta & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ & \sec\theta \end{bmatrix} = L_1 U_1 \quad \text{and} \tag{5.101}$$

$$PB = \begin{bmatrix} 1 & \\ -\cot\theta & 1 \end{bmatrix} \begin{bmatrix} -\sin\theta & \cos\theta \\ & \csc\theta \end{bmatrix} = L_2 U_2 \tag{5.102}$$

for $P = P_{(1\ 2)}$. It follows

$$\|B\|_{\max} = \max(|\cos\theta|, |\sin\theta|) = |\cos\theta| \max(1, |\tan\theta|)$$
$$\|L_1\|_{\max} = \max(1, |\tan\theta|)$$
$$\|U_1\|_{\max} = |\sec\theta|$$

and

$$\|B\|_\infty = \|B^{-1}\|_\infty = |\cos\theta| + |\sin\theta| = |\cos\theta|(1 + |\tan\theta|)$$
$$\|L_1\|_\infty = 1 + |\tan\theta|$$
$$\|U_1\|_\infty = \max(|\cos\theta| + |\sin\theta|, |\sec\theta|) = |\cos\theta|(1 + \max(|\tan\theta|, \tan^2\theta)).$$

and similarly using L_2, U_2. It follows directly

Lemma 5.4. *Let $B = B(\theta) \in \mathrm{B}_s(2)$. Using GENP, then*

$$\rho(B) = 1 + \tan^2\theta \tag{5.103}$$

$$\rho_\infty(B) = 1 + \max(|\tan\theta|, \tan^2\theta) \tag{5.104}$$

for all θ*. Moreover,* $1 \leq \rho(B) \leq \rho_\infty(B)$ *with* $1 = \rho(B) = \rho_\infty(B)$ *for* $|\cos\theta| = 1$*,* $\rho(B) = \rho_\infty(B)$ *if* $|\tan\theta| \geq 1$*, and strict inequalities otherwise.*

Using GEPP, GERP or GECP, then

$$\rho(B) = 1 + \min(\tan^2\theta, \cot^2\theta) \tag{5.105}$$

$$\rho_\infty(B) = 1 + \min(|\tan\theta|, |\cot\theta|) \tag{5.106}$$

for all θ*. Moreover,* $1 \leq \rho(B) \leq \rho_\infty(B) \leq 2$ *with* $1 = \rho(B) = \rho_\infty(B)$ *for* $\cos\theta = 0$ *or* $\sin\theta = 0$*,* $\rho(B) = \rho_\infty(B) = 2$ *for* $|\tan\theta| = 1$*, and strict inequalities otherwise.*

Proof. The GENP case is obvious using the Pythagorean identity $\sec^2\theta = 1 + \tan^2\theta$, where we note $\|B\|_{\max} \leq 1 \leq \|U\|_{\max}$. For GE with pivoting schemes, pivoting will occur only if $|\sin\theta| > |\cos\theta|$, which will result in using the factors L_2, U_2 instead, where we note then $\rho^{\text{GEPP}}(B(\theta)) = \rho^{\text{GENP}}(B(\theta'))$ and similarly for ρ_∞ where $\theta' = \theta$ if $|\tan\theta| \leq 1$ and $\theta' = \frac{\pi}{2} - \theta$ if $|\tan\theta| > 1$. □

Lemma 5.5. *Let* $B = B(\theta) \in \mathrm{B}_s(2)$*. Then*

$$\kappa_\infty(B) = 1 + |\sin(2\theta)|. \tag{5.107}$$

Proof. Since $\|B(\theta)^{-1}\|_\infty = \|B(-\theta)\|_\infty = \|B(\theta)\|_\infty$, we can compute directly

$$\kappa_\infty(B) = \|B\|_\infty\|B^{-1}\|_\infty = \|B\|_\infty^2 = (|\cos\theta| + |\sin\theta|)^2 = 1 + |\sin(2\theta)|.$$

□

It follows:

Lemma 5.6. *Let $B \sim \mathrm{B}_s(2, \Sigma_S)$ and $X \sim \mathrm{Cauchy}(1)$. Using GENP, then*

$$\rho(B) \sim 1 + |X|^2 \tag{5.108}$$

$$\rho_\infty(B) \sim 1 + \max(|X|, |X|^2). \tag{5.109}$$

Using GEPP, GERP or GECP,

$$\rho(B) \sim 1 + |X|^2 \mid |X| \leq 1 \tag{5.110}$$

$$\rho_\infty(B) \sim 1 + |X| \mid |X| \leq 1. \tag{5.111}$$

Proof. This follows directly from Lemmas 1.10, 1.11 and 5.4 □

Corollary 5.6. *Let $B \sim \mathrm{B}_s(2, \Sigma_S)$. Using GEPP, GERP or GECP, then*

$$\mathbb{E}\rho(B) = \frac{4}{\pi} \tag{5.112}$$

$$\mathbb{E}\rho_\infty(B) = 1 + \frac{\log 4}{\pi}. \tag{5.113}$$

Proof. We compute

$$\mathbb{E}\rho(B) = \frac{1}{2\pi} \int_0^{2\pi} (1 + \min(\tan^2\theta, \cot^2\theta))\ d\theta = \frac{4}{\pi} \int_0^{\pi/4} \sec^2\theta\ d\theta = \frac{4}{\pi}$$

and

$$\begin{aligned}\mathbb{E}\rho_\infty(B) &= 1 + \mathbb{E}\min(|\tan\theta|, |\cot\theta|) = 1 + \frac{1}{2\pi} \int_0^{2\pi} \min(|\tan\theta|, |\cot\theta|)\ d\theta \\ &= 1 + \frac{4}{\pi} \int_0^{\pi/4} \tan\theta\ d\theta = 1 + \frac{\log 4}{\pi}.\end{aligned}$$

□

Additionally, we can say more directly about the ∞-condition numbers of random butterfly

matrices.

Lemma 5.7. *Let* $B \sim \mathrm{B}_s(2, \Sigma_S)$ *and* $Y \sim \mathrm{Arcsine}(0, 1)$*, then*

$$\kappa_\infty(B) \sim 1 + \sqrt{Y}.$$

Proof. Use Lemmas 1.9 and 5.5. □

In particular, we can explicitly compute averages for the condition numbers as well as, in light of (5.24) and (5.25), $\kappa_\infty(B)\rho(B)$ and $\kappa_\infty(B)\rho_\infty(B)$ in the GE with pivoting scheme cases. (The GENP cases have no moments of any order $k \geq 1$ again since $\kappa_\infty(B) \geq 1$.)

Corollary 5.7. *Let* $B \sim \mathrm{B}_s(2, \Sigma_S)$*. Then*

$$\mathbb{E}\kappa_\infty(B) = 1 + \frac{2}{\pi}. \tag{5.114}$$

Using GEPP, GERP or GECP, we have

$$\mathbb{E}\kappa_\infty(B)\rho(B) = \frac{4}{\pi}(1 + \log 2) \tag{5.115}$$

$$\mathbb{E}\kappa_\infty(B)\rho_\infty(B) = 2 + \frac{\log 4}{\pi} \tag{5.116}$$

Proof. Using Lemma 5.5, we can compute

$$\mathbb{E}\kappa_\infty(B) = 1 + \frac{1}{2\pi}\int_0^{2\pi} |\sin(2\theta)|\, d\theta = 1 + \frac{2}{\pi}\int_0^{\pi/2} \sin(2\theta)\, d\theta = 1 + \frac{2}{\pi}. \tag{5.117}$$

Using Lemmas 5.4 and 5.5, then

$$\begin{aligned}\kappa_\infty(B)\rho(B) &= 1 + |\sin(2\theta)| + \min(\tan^2\theta, \cot^2\theta) + |\sin 2\theta| \min(\tan^2\theta, \cot^2\theta) \\ &= \kappa_\infty(B) + \rho(B) - 1 + 2\min(|\sin^3\theta \sec\theta|, |\cos^3\theta \csc\theta|)\end{aligned}$$

$$\begin{aligned}\kappa_\infty(B)\rho_\infty(B) &= 1 + |\sin(2\theta)| + \min(|\tan\theta|, |\cot\theta|) + |\sin 2\theta| \min(|\tan\theta|, |\cot\theta|) \\ &= \kappa_\infty(B) + \rho_\infty(B) - 1 + \min(|2\sin^2\theta|, |2\cos^2\theta|) \\ &= \kappa_\infty(B) + \rho_\infty(B) - 1 + \min(1-\cos\theta, 1+\cos\theta) \\ &= \kappa_\infty(B) + \rho_\infty(B) - |\cos\theta|\end{aligned}$$

Note

$$\begin{aligned}&\mathbb{E}\min(|\sin^3\theta\sec\theta|, |\cos^3\theta\csc\theta|) \\ &= \frac{1}{2\pi}\int_0^{2\pi} \min(|\sin^3\theta\sec\theta|, |\cos^3\theta\csc\theta|)\ d\theta \\ &= \frac{4}{\pi}\int_0^{\pi/4} \sin^3\theta\sec\theta\ d\theta = \frac{4}{\pi}\int_0^{\pi/4} (\tan\theta - \frac{1}{2}\sin(2\theta))\ d\theta \\ &= \frac{\log 4 - 1}{\pi}\end{aligned}$$

and

$$\mathbb{E}|\cos\theta| = \mathbb{E}|\sin\theta| = \frac{1}{2\pi}\int_0^{2\pi} |\sin\theta|\ d\theta = \frac{1}{\pi}\int_0^{\pi} \sin\theta\ d\theta = \frac{2}{\pi}.$$

The results for $\mathbb{E}\kappa_\infty(B)\rho(B)$ and $\mathbb{E}\kappa_\infty(B)\rho_\infty(B)$ follow then by combining these with Corollaries 5.6 and 5.7. □

Now we can sum up these results to establish the main statements from Section 5.3.2.

Proof of Theorem 5.3. Use Lemmas 5.3, 5.4 and 5.6 with the uniqueness results in the random models following from Theorems 1.7 and 1.10. □

Proof of Corollary 5.2. Use Theorem 5.3 and corollary 5.6. □

Proof of Proposition 5.3. Use Lemmas 5.3 and 5.7 □

Proof of Corollary 5.4. Use Lemma 5.3 and corollary 5.7. □

5.3.5 GECP Growth Factor

Some additional results relating to the GECP growth factor can be established for butterfly matrices. These results are explicit weaker than for the previously visited pivoting schemes. Future work can revisit methods to strengthen these results.

First, we will establish this straightforward building block.

Lemma 5.8. *Let $\alpha, \beta, \theta, \eta, \varepsilon \in \mathbb{R}$ such that $|\varepsilon| \leq |\eta - \varepsilon|$. Then*

$$|\eta\alpha \cos^2\theta - \varepsilon\beta| \leq \cos^2\theta|\eta\alpha - \varepsilon\beta| + \sin^2\theta|\eta - \varepsilon||\beta| \tag{5.118}$$

Proof. Since $\eta\alpha\cos^2\theta - \varepsilon\beta = \cos^2\theta(\eta\alpha - \varepsilon\beta) - \sin^2\theta\varepsilon\beta$, this follows directly by the triangle inequality and $|\varepsilon| \leq |\eta - \varepsilon|$. □

Now we can introduce a weaker result using GECP.

Proposition 5.7. *Let $B = B(\boldsymbol{\theta}) \in \mathrm{B}_s(N)$ such that $|\cos\theta_{i+1}| \geq |\cos\theta_i| > |\sin\theta_1|$ for all i, and let $B = LU$ be the LU factorization of B using GENP. Let $\eta, \varepsilon \in \mathbb{R}$ such that $|\varepsilon| \leq |\eta - \varepsilon|$. Then for all k*

$$\left\| \begin{bmatrix} \mathbf{0} & \mathbf{I}_{N-k+1} \end{bmatrix} (\eta B - \varepsilon B^{(k)}) \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{N-k+1} \end{bmatrix} \right\|_{\max} \leq |\eta - \varepsilon||U_{kk}|. \tag{5.119}$$

In particular, then $B = LU$ is also the LU factorization of B using GECP.

Proof. First, note the last implication follows since then

$$|B_{kk}^{(k)}| \leq \left\| \begin{bmatrix} \mathbf{0} & \mathbf{I}_{N-k+1} \end{bmatrix} B^{(k)} \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{N-k+1} \end{bmatrix} \right\|_{\max} \leq |U_{kk}| = |B_{kk}^{(k)}|$$

for all k using (5.119) with $\eta = 0$ and $\varepsilon = -1$, so that no row or column swaps would be needed at any intermediate step using GECP.

To prove (5.119), we will again use induction on n. Note first the result is immediate for $k = 1$ since then

$$\|\mathbf{I}_N(\eta B - \varepsilon B^{(1)})\mathbf{I}_N\|_{\max} = |\eta - \varepsilon|\|B\|_{\max} = |\eta - \varepsilon||U_{11}|.$$

So it suffices to assume $k \geq 2$. If $n = 1$, then using (5.77) with $B^{(2)} = U = U_\theta$, we see for $k = 2$ then

$$\left\| \begin{bmatrix} 0 & 1 \end{bmatrix} (\eta B - \varepsilon B^{(2)}) \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\|_{\max} = |\eta \cos\theta - \varepsilon \sec\theta| = |\eta \cos^2\theta - \varepsilon||U_{22}| \leq |\eta - \varepsilon||U_{22}|$$

using Lemma 5.8 with $\alpha = \beta = 1$.

Now assume the result holds for $\mathrm{B}_s(N/2)$ and $n \geq 2$, and let $B = B(\theta, A) \in \mathrm{B}_s(N)$ for $A = B(\boldsymbol{\theta}') \in \mathrm{B}_s(N/2)$ such that $\boldsymbol{\theta} = (\boldsymbol{\theta}', \theta)$ with $\theta = \theta_n$, where we note also $\boldsymbol{\theta}'$ still satisfies the condition $|\cos\theta'_{i+1}| \geq |\cos\theta'_i| > |\sin\theta'_1|$ for all i. Since B has an LU factorization using GENP, then necessarily A does also. Let $A = L'U'$ be this factorization, where we further note by Proposition 5.4 then

$$U_{kk} = \begin{cases} \cos\theta U'_{kk} & \text{if } k \leq N/2 \\ \sec\theta U'_{jj} & \text{if } k = N/2 + j \text{ for } j \geq 1. \end{cases} \tag{5.120}$$

For $1 < k \leq N/2$, then for $\mathbf{I} = \mathbf{I}_{N/2-k+1}$ when not indicated otherwise, we have

$$\begin{aligned}
&\begin{bmatrix} \mathbf{0} & \mathbf{I}_{N-k+1} \end{bmatrix} (\eta B - \varepsilon B^{(k)}) \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{N-k+1} \end{bmatrix} \\
&= \begin{bmatrix} \cos\theta \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} (\eta A - \varepsilon A^{(k)}) \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} & \sin\theta \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} (\eta A - \varepsilon A^{(k)}) \\ -\sin\theta \left(\eta A - \varepsilon \begin{bmatrix} \mathbf{0} & \\ & \mathbf{I} \end{bmatrix} A^{(k)} \right) \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} & -\sec\theta \left((\varepsilon - \eta\cos^2\theta) A - \varepsilon \sin^2\theta \begin{bmatrix} \mathbf{0} & \\ & \mathbf{I} \end{bmatrix} A^{(k)} \right) \end{bmatrix}
\end{aligned} \tag{5.121}$$

using Lemma 5.2. Let $\eta' = \varepsilon - \eta\cos^2\theta$ and $\varepsilon' = \varepsilon\sin^2\theta$ where we note

$$|\varepsilon'| = \sin^2\theta|\varepsilon| \leq \cos^2\theta|\eta - \varepsilon| = |\eta' - \varepsilon'|$$

since $\sin^2\theta \leq \cos^2\theta$ and $|\varepsilon| \leq |\eta - \varepsilon|$. By the inductive hypothesis, we have

$$\begin{aligned}
\left\| \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} (\eta' A - \varepsilon' A^{(k)}) \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \right\|_{\max} &\leq |\eta' - \varepsilon'||U'_{kk}| = |\cos\theta||\eta - \varepsilon||U_{kk}| \quad \text{and} \\
\left\| \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} (\eta A - \varepsilon A^{(k)}) \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \right\|_{\max} &\leq |\eta - \varepsilon||U'_{kk}| = |\sec\theta||\eta - \varepsilon||U_{kk}|
\end{aligned}$$

using (5.120). Also, since $|\cos\theta_{i+1}| \geq |\cos\theta_i|$ for all i, then

$$|\sec\theta|\|A\|_{\max} = |\sec\theta_n||U'_{11}| = \frac{\prod_{j=1}^{n-1}|\cos\theta_j|}{|\cos\theta_n|} \leq \frac{\prod_{j\in\mathcal{J}_k}|\cos\theta_j|}{\prod_{j\in[n]\setminus\mathcal{J}_k}|\cos\theta_j|} = |U_{kk}|$$

for some $\varnothing \neq \mathcal{J}_k \subset [n]$ when $k > 1$ by Proposition 5.4. Next, we see

$$|\eta\sin\theta|\|A\|_{\max} \leq |\eta - \varepsilon||2\sin\theta|\|A\|_{\max} \leq |\eta - \varepsilon||\sec\theta|\|A\|_{\max} \leq |\eta - \varepsilon||U_{kk}|$$

using $|\eta| \leq |\eta-\varepsilon|+|\varepsilon| \leq 2|\eta-\varepsilon|$ and $|2\sin\theta| \leq |\sec\theta|$ (since $|2\sin\theta\cos\theta| = |\sin(2\theta)| \leq 1$). It follows

$$\left\| \begin{bmatrix} \mathbf{0} & \mathbf{I}_{N-k+1} \end{bmatrix} (\eta B - \varepsilon B^{(k)}) \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{N-k+1} \end{bmatrix} \right\|_{\max} \tag{5.122}$$

$$= \max \left(\begin{array}{c} |\eta \sin\theta| \| A_{N/2-k+1:,:k-1} \|_{\max}, \\ |\eta \sin\theta| \| A_{:k-1,N/2-k+1:} \|_{\max}, \\ |\cos\theta| \left\| \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} (\eta A - \varepsilon A^{(k)}) \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \right\|_{\max}, \\ |\sec\theta| |\varepsilon - \eta\cos^2\theta| \|A\|_{\max}, \\ |\sec\theta| \left\| \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} \left((\varepsilon - \eta\cos^2\theta) A - \varepsilon \sin^2\theta A^{(k)} \right) \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \right\|_{\max} \end{array} \right) \tag{5.123}$$

$$\leq |\eta - \varepsilon| |U_{kk}| \tag{5.124}$$

using again Lemma 5.8 so that $|\varepsilon - \eta\cos^2\theta| \leq |\eta - \varepsilon|$.

For $k = N/2+j$ and $j \geq 1$ so that $N-k+1 = N/2-j+1$, writing now $\mathbf{I} = \mathbf{I}_{N-k+1} = \mathbf{I}_{N/2-j+1}$, we have

$$\left\| \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} (\eta B - \varepsilon B^{(k)}) \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \right\|_{\max}$$

$$= |\sec\theta| \left\| \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} (\eta\cos^2\theta A - \varepsilon A^{(j)}) \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \right\|_{\max}$$

$$= |\sec\theta| \left\| \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} \left| \eta\cos^2\theta A - \varepsilon A^{(j)} \right| \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \right\|_{\max}$$

$$\leq |\sec\theta| \left(\cos^2\theta \left\| \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} \left| \eta A - \varepsilon A^{(j)} \right| \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \right\|_{\max} + \sin^2\theta |\eta - \varepsilon| \left\| \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} |A^{(j)}| \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \right\|_{\max} \right)$$

$$
\begin{aligned}
&= |\sec\theta| \left(\cos^2\theta \left\| \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} (\eta A - \varepsilon A^{(j)}) \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \right\|_{\max} + \sin^2\theta |\eta - \varepsilon| \left\| \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} A^{(j)} \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \right\|_{\max} \right) \\
&\leq |\sec\theta| |\eta - \varepsilon| |U'_{jj}| = |\eta - \varepsilon| |U_{kk}|
\end{aligned}
$$

using Lemma 5.2 for the first equality, Lemma 5.8 for the first inequality (applied componentwise with $\alpha = A_{i'j'}$ and $\beta = A^{(j)}_{i'j'}$), the inductive hypothesis for the last inequality (with $\eta = 0$ and $\varepsilon = 1$ for the second term), (5.120) for the last equality, and the fact $\|A\|_{\max} = \||A|\|_{\max}$ for the remaining steps. □

This does not give the full picture in terms of the GECP growth factor of Haar-butterfly matrices. Empirical results indicate the distribution of the GECP max-norm growth factor ρ matches that on the GENP, GEPP and GERP max-norm growth factors. However, ρ_∞ does differ: experiments indicate $\rho_\infty^{\text{GECP}}(B) < \rho_\infty^{\text{GEPP}}(B)$ for $B \sim \text{B}_s(N, \Sigma_S)$. The LU factorization of Haar-butterfly matrices can prove to be sensitive to multiplication on the right by a permutation matrix, which is true starting for $N \geq 8$, as seen in the following example.

Example 5.8. *Let*

$$
B = \begin{bmatrix}
-0.02 & 0.03 & -0.41 & 0.64 & 0.02 & -0.03 & 0.35 & -0.55 \\
-0.03 & -0.02 & -0.64 & -0.41 & 0.03 & 0.02 & 0.55 & 0.35 \\
0.41 & -0.64 & -0.02 & 0.03 & -0.35 & 0.55 & 0.02 & -0.03 \\
0.64 & 0.41 & -0.03 & -0.02 & -0.55 & -0.35 & 0.03 & 0.02 \\
-0.02 & 0.03 & -0.35 & 0.55 & -0.02 & 0.03 & -0.41 & 0.64 \\
-0.03 & -0.02 & -0.55 & -0.35 & -0.03 & -0.02 & -0.64 & -0.41 \\
0.35 & -0.55 & -0.02 & 0.03 & 0.41 & -0.64 & -0.02 & 0.03 \\
0.55 & 0.35 & -0.03 & -0.02 & 0.64 & 0.41 & -0.03 & -0.02
\end{bmatrix} \tag{5.125}
$$

with $B \in \mathrm{B}_s(8)$. *Then using GECP, we have* $PBQ^T = LU$ *where, in particular,*

$$U = \begin{bmatrix} 0.64 & -0.55 & -0.35 & 0.41 & -0.03 & 0.03 & -0.02 & 0.02 \\ & 1.12 & 0.71 & & & -0.05 & & -0.04 \\ & & -0.89 & -0.78 & & & 0.04 & 0.04 \\ & & & -1.57 & & & 0.08 & \\ & & & & -0.64 & 0.55 & -0.41 & 0.35 \\ & & & & & -1.12 & & -0.71 \\ & & & & & & 0.90 & -0.78 \\ & & & & & & & 1.57 \end{bmatrix}. \tag{5.126}$$

U does not have a Kronecker product form any more as a result of introducing column pivots. For instance,

$$U_{1:2,3:4} = \begin{bmatrix} -0.35 & 0.41 \\ 0.71 & \end{bmatrix} \tag{5.127}$$

is no longer upper triangular.

Future work can work on determining whether a full distribution on the max-norm growth factor for Haar-butterfly matrices using GECP can be attained.

Growth factors of Hadamard matrices

As evidenced in Section 5.3.2, the butterfly Hadamard matrices maximized ρ and ρ_∞ when using GENP, GEPP, or GERP. Empirically, this holds even when using GECP. The question about the growth factors of Hadamard matrices when using GECP remains an open question. All evidence continues to point to:

Conjecture 5. *Let H be an order N Hadamard matrix. Using GE with any row or column*

pivoting scheme, then $\rho(H) = N$.

Cryer made this conjecture in 1968 when reviewing Wilkinson's work on the GEPP max-norm growth factor [5]. Only information on particular equivalence classes of Hadamard matrices have been attained as of now. In particular, this has been proven already for Sylvester Hadamard matrices, which includes the butterfly Hadamard matrices by Proposition 2.14. In general, the growing complexity on total number of Hadamard matrices for a given order prevents direct approaches from being computationally feasible. As of now, the highest order where it has been proven that the GECP max-norm growth factors of Hadamard matrices matches its order is 16 [23]. The most recent progress focused on studying specifically the possible patterns in pivots one can encounter.

Butterfly matrices can potentially be used to give an alternative proof of the GECP growth factors of Sylvester Hadamard matrices.

Bibliography

[1] K. Alizadeh-Vahid, A. Farhadi, and M. Rastegari. Butterfly transform: An efficient fft based neural architecture design. *2020 IEEE/CVF Conf. on CVPR*, pages 12021–12030, 2020.

[2] M. Baboulin, X. Li, and F. Rouet. Using random butterfly transformations to avoid pivoting in sparse direct methods. *In: Proc. of Int. Con. on Vector and Par. Proc.*, 2014.

[3] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19:297–301, 1965.

[4] V. Cortés and J. Peña. Growth factor and expected growth factor of some pivoting strategies. *Journal of Comp. and App. Math.*, 202:292–303, may 2007.

[5] C. W. Cryer. Pivot size in Gaussian elimination. *Numer. Math.*, 12:335–345, 1968.

[6] T. Dao, A. Gu, M. Eichhorn, A. Rudra, and C. Re. Learning fast algorithms for linear transforms using butterfly factorizations. *Proc. Mach. Learn. Res.*, pages 1517–1527, June 2019.

[7] P. Diaconis and S. N. Evans. Linear functionals of eigenvalues of random matrices. *Trans. Amer. Math. Soc.*, 353(7):2615–2633, 2001.

[8] P. Diaconis, R. L. Graham, and W. M. Kantor. The mathematics of perfect shuffles. *Adv. in App. Math.*, 4(2):175–196, 1983.

[9] P. Diaconis and M. Shahshahani. The subgroup algorithm for generating uniform random variables. *Prob. in the Eng. and Info. Sci.*, 1:15–32, 1987.

[10] P. Diaconis and M. Shahshahani. On the eigenvalues of random matrices. *J. Appl. Probab.*, 31A:49–62, 1994.

[11] D. Dokovic. Hadamard matrices of order 764 exist. *Combinatorica*, 28:487–489, 2008.

[12] J. Dongarra and F. Sullivan. Guest editors introduction to the top 10 algorithms. *Computing in Science Engineering*, 2(1):22–23, 2000.

[13] A. Edelman. The complete pivoting conjecture for gaussian elimination is false. *The Mathematica Journal*, 2:58–61, 1992.

[14] A. Ferber, V. Jain, and Y. Zhao. On the number of hadamard matrices via anti-concentration, 2018.

[15] W. Givens. Computation of plane unitary rotations transforming a general matrix to triangular form. *J. Soc. Indust. Appl. Math.*, 6:26–50, 1958.

[16] N. J. Higham. *Accuracy and Stability of Numerical Algorithms, Second Edition*. SIAM, Philadelphia, PA, 2002.

[17] N. J. Higham and D. Higham. Large growth factors in gaussian elimination with pivoting. *SIAM Journal on Matrix Anal. and App.*, 10:155–164, Apr. 1989.

[18] N. J. Higham, D. Higham, and S. Pranesh. Random matrices generating large growth in lu factorization with pivoting. *SIAM J. Matrix Anal. Appl.*, 42:185–201, Oct. 2020.

[19] C. P. Hughes and Z. Rudnick. Mock-gaussian behaviour for linear statistics of classical compact groups. *Journal of Physics A: Mathematical and General*, 36(12):2919–2932, mar 2003.

[20] A. Jagganath and T. Trogdon. Random matrices and the new york city subway system. *Phys. Rev. E*, 96, 2017.

[21] Y. Jiao, T. Lau, H. Hatzikirou, M. Meyer-Hermann, J. C. Corbo, and S. Torquato. Avian photoreceptor patterns represent a disordered hyperuniform solution to a multiscale packing problem. *Phys. Rev. E*, 89:022721, Feb 2014.

[22] I. Koren. *Computer Arithmetic Algorithms, Second Edition*. A K Peters, Natick, MA, 2002.

[23] C. Kravvaritis and M. Mitrouli. The growth factor of a Hadamard matrix of order 16 is 16. *Numer. Linear Algebra Appl.*, 16(9):715–743, 2009.

[24] M. Krbálek. Inter-vehicle gap statistics on signal-controlled crossroads. *J. Phys. A Math. Theor.*, 41, 2008.

[25] M. Krbálek and P. Seba. The statistical properties of the city transport in cuernavaca (mexico) and random matrix ensembles. *J. Phys. A: Math. Gen.*, 33(26), 2000.

[26] P. Lancaster and H. K. Farahat. Norms on direct sums and tensor products. *Math. of Comp.*, 26(118):401–414, Apr. 1972.

[27] Y. Li, X. Cheng, and J. Lu. Butterfly-net: Optimal function representation based on convolutional neural networks, 2018.

[28] M. Mehta. *Random Matrices*. Academic Press, 3rd edition, 2004.

[29] A. Oppenheim and R. Schafer. *Digital Signal Processing*. Prentice Hall international editions. Prentice-Hall, 1975.

[30] R. Paley. On orthogonal matrices. *J. of Math. and Phys.*, 12:311–320, 1933.

[31] D. Parker. Random butterfly transformations with applications in computational linear algebra. *Tech. rep., UCLA*, 1995.

[32] G. Poole and L. Neal. The rook's pivoting strategy. *J. of Comp. and App. Math.*, 123:353–369, 2000.

[33] G. Stewart. The efficient generation of random orthogonal matrices with an application to condition estimators. *SIAM J. Numer. Anal.*, 17:403–409, 1980.

[34] J. Thomson, A. Brown, T. Ozkan-Holler, A. Ellenson, and M. Haller. Extreme conditions at wave energy sites. *Tech. report*, 2016.

[35] L. Trefethen and D. Bau. *Numerical Linear Algebra*. SIAM, 1997.

[36] L. Trefethen and R. Schreiber. Average case stability of gaussian elimination. *SIAM J. Matrix Anal. Appl.*, 11:335–360, 1990.

[37] T. Trogdon. On spectral and numerical properties of random butterfly matrices. *Applied Math. Letters*, 95(4):48–58, Sept. 2019.

[38] A. Weil. *L'intégration dans les groupes topologiques et ses applications*, volume 4. l'Institut de mathématiques de Clermont-Ferrand, Paris, 1940.

[39] J. Wilkinson. Error analysis of direct methods of matrix inversion. *J. Assoc. Comput. Mach.*, 8:281–330, jul 1961.

[40] J. Wilkinson. *The Algebraic Eigenvalue Problem*. Oxford University Press, London, UK, 1965.

[41] J. Williamson. Hadamard's determinant theorem and the sum of four squares. *Duke Math. J.*, 11:65–81, 1944.

Appendix A

Divisibility of random integers

This is a straightforward result using basic results in probability, which can be used to produce previous results regarding divisibility of integer valued random variables with a different approach. This has applications to questions at the intersection of probability and number theory.

Proposition A.1. *For X a discrete random variable with support on $\mathbb{Z}$, then*

$$\mathbb{P}(d \mid X) = \frac{1}{d} \sum_{k=1}^{d} \varphi_X \left(\frac{2\pi k}{d} \right). \tag{A.1}$$

Proof. Since $\operatorname{supp}(X) \subset \mathbb{Z}$ then

$$\varphi_X(t) = \mathbb{E}e^{itX} = \sum_{j \in \mathbb{Z}} e^{itj} \mathbb{P}(X = j).$$

A straightforward computation shows

$$\frac{1}{d} \sum_{k=1}^{d} \varphi_X \left(\frac{2\pi k}{d} \right) = \frac{1}{d} \sum_{k=1}^{d} \sum_{j \in \mathbb{Z}} e^{i2\pi kj/d} \mathbb{P}(X = j)$$

$$= \sum_{j\in\mathbb{Z}} \mathbb{P}(X=j)\left(\frac{1}{d}\sum_{k=1}^{d} e^{i2\pi kj/d}\right)$$
$$= \sum_{d|j} \mathbb{P}(X=j)$$
$$= \mathbb{P}(d \mid X)$$

using (1.9) for the penultimate equality. □

Corollary A.1. *For X a discrete random variable with support on $\mathbb{Z}$, then for $a \in [d]$,*

$$\mathbb{P}(X \equiv a \pmod d) = \frac{1}{d}\sum_{k=1}^{d} \varphi_X\left(\frac{2\pi k}{d}\right) e^{-i2\pi ka/d}. \tag{A.2}$$

Proof. Since $\mathbb{P}(X \equiv a \pmod d) = \mathbb{P}(d \mid X-a)$ and $\varphi_{X-a}(t) = e^{-ita}\varphi_X(t)$, this follows immediately from Proposition A.1. □

This can be used to give a different proof of some known results:

Proposition A.2. *For $X \sim \text{Binomial}(n,q)$ and $q \in (0,1)$, then $\mathbb{P}(X \equiv a \pmod d) = \frac{1}{d} + O(e^{-cn})$ for some $c > 0$ depending only on q, d.*

Proof. Recall

$$\varphi_X(t) = (e^{it}q + 1 - q)^n.$$

If $d = 1$, then $\mathbb{P}(1 \mid X) = 1$, so we can take $c = 1$. For $d \geq 2$, using Proposition A.1, we have

$$\mathbb{P}(X \equiv a \pmod d) = \frac{1}{d}\sum_{k=1}^{d} (e^{i2\pi k/d}q + 1 - q)^n e^{-2\pi ka/d}$$
$$= \frac{1}{d} + \frac{1}{d}\sum_{k=1}^{d-1} (e^{i2\pi k/d}q + 1 - q)^n e^{-2\pi ka/d}.$$

Since $z_k := e^{i2\pi k/d}q + (1-q)$ is a convex combination of $e^{i2\pi k/d}$ and 1, which are both on the boundary of the unit circle, then z_k is strictly within the unit circle if $k = 1, 2, \ldots, d-1$,

and closest to the boundary when $k = 1$, i.e.,

$$|z_k|^2 \le |z_1|^2 = q^2 + 2q(1-q)\cos\left(\frac{2\pi}{d}\right) + (1-q)^2 = 1 - 2q(1-q)\left(1 - \cos\left(\frac{2\pi}{d}\right)\right).$$

If $d = 2$, then $|z_1|^2 = (1-2q)^2 < 1$. We see $z_1 = 0$ if and only if $q = \frac{1}{2}$, which then shows $\mathbb{P}(2 \mid X) = \mathbb{P}(2 \nmid X) = \frac{1}{2}$. Hence, we can take $c = 1$ if $d = 2$ and $q = \frac{1}{2}$. For $d > 2$ or $q \neq \frac{1}{2}$, then $|z_1|^2 < 1$. Since

$$0 \le \frac{1}{d}\left|\sum_{k=1}^{d-1} z_k^n e^{-2\pi k a/d}\right| < |z_1|^n,$$

the result follows for

$$c = -\frac{1}{2}\ln\left(1 - 2q(1-q)\left(1 - \cos\left(\frac{2\pi}{d}\right)\right)\right) > 0.$$

□

Lemma A.1. *If $X \sim \text{Uniform}(\mathbb{Z}/n\mathbb{Z})$ and Y is independent of X with support in $\mathbb{Z}$, then $X + Y \pmod n \sim \text{Uniform}(\mathbb{Z}/n\mathbb{Z})$. For $d \le n$, then for $a = 0, \ldots, d-1$,*

$$\mathbb{P}(X + Y \equiv a \pmod d) = \frac{1}{d} + \frac{\mathbb{1}_{a<r} - r/d}{n} \tag{A.3}$$

for $0 \le r < d$ such that $n \equiv r \pmod d$. In particular, if $d \mid n$ then $X + Y \pmod d \sim \text{Uniform}(\mathbb{Z}/d\mathbb{Z})$

Proof. Note

$$\varphi_X(t) = \frac{1}{n}\sum_{j=]}^{n-1} e^{itj} \quad \text{while} \quad \varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t).$$

By (1.9), we have $\varphi_X(\frac{2\pi k}{n}) = \delta_{kn}$. It follows

$$\mathbb{P}(X + Y \equiv a \pmod n) = \frac{1}{n}\sum_{k=1}^{n} \varphi_{X+Y}\left(\frac{2\pi k}{n}\right) e^{-i2\pi k a/n} = \frac{1}{n}.$$

Hence, $X + Y \pmod n \sim \text{Uniform}(\mathbb{Z}/n\mathbb{Z})$. (Another quick argument can follow by conditioning on Y.) If $d < n$, then the residue class of $\{0, 1, \ldots, n\}$ modulo d are of size $\lfloor n/d \rfloor + 1$ the smallest element in the class is smaller than r and of size $\lfloor n/d \rfloor$ if the smallest element is at least r. □

Proposition A.3. *Let W_j be iid* Uniform$(\mathbb{F}_q)$ *for $q = p^k$ for some prime p and $W = \sum_{j=1}^n W_j$ and suppose $d > q$. Then for some $c > 0$ depending only on q, d, we have*

$$\mathbb{P}(W \equiv a \pmod d) = \frac{1}{d} + O(e^{-cn}).$$

Proof. Note for each j we have

$$\varphi_{W_j}(t) = \varphi_{W_1}(t) = \frac{1}{p}\sum_{\ell=0}^{q-1} e^{it\ell} \quad \text{while} \quad \varphi_W(t) = \prod_{j=1}^n \varphi_{W_j}(t) = \varphi_{W_1}(t)^n.$$

It follows $\varphi_{W_1}(\frac{2\pi k}{d})$ is a convex combination of p points $e^{i2\pi k\ell/d}$ on the boundary of the unit circle, so $|\varphi_{W_1}(\frac{2\pi k}{d})| \le 1$. Moreover, since 1 is included among these points (for $\ell = 0$), then we have $|\varphi_{W_j}(t)| = 1$ if and only if $\varphi_{W_j}(t) = 1$ if and only if $e^{it} = 1$ if and only if $\frac{t}{2\pi} \in \mathbb{Z}$. Hence, we have $|\varphi_{W_1}(\frac{2\pi k}{d})| < 1$ for each $k = 1, \ldots, d-1$. Moreover, we have $|\varphi_{W_1}(\frac{2\pi k}{d})| > 0$ since the uniform average is the center of the convex hull containing the points, which is never 0 since $q < d$. (For instance, the uniform average of the imaginary parts of these points is strictly positive since $q < d$.) Hence

$$0 < \max\left\{\left|\varphi_{W_1}\left(\frac{2\pi k}{d}\right)\right| : k = 1, \ldots, d-1\right\} < 1.$$

It follows

$$\mathbb{P}(W \equiv a \pmod d) = \frac{1}{d} + \frac{1}{d}\sum_{k=1}^{d-1} \varphi_{W_1}\left(\frac{2\pi k}{d}\right)^n e^{-i2\pi ka/d} = \frac{1}{d} + O(e^{-cn})$$

where

$$c = -\ln\left(\max\left\{\left|\varphi_{W_1}\left(\frac{2\pi k}{d}\right)\right| : k = 1, \ldots, d-1\right\}\right) > 0.$$

□

Appendix B

RMT statistics in ocean wave spacings

RMT statistics have recently been highlighted in transportation systems. Of particular interest is the bus system in Cuernavaca, Mexico [25]. Since the bus system is privatized, drivers implemented a notification system to prevent small bus arrival gaps. This resulted in a natural repulsion between arrival times that has been shown to be sufficiently modeled by the ($\beta = 2$) Wigner surmise. Jagganath and Trogdon showed arrival times at certain stops in the NYC subway system also follow the Wigner surmise [20]. Building off of these results, we were interested in exploring other physical systems that exhibit RMT statistics.

The most recent focus has been on the spacings between normalized ocean wave peaks. An early ad hoc study I carried out, that involved me sitting at Little Corona del Mar beach 15 minutes away from campus for a few hours recording wave arrival times on my phone, indicated wave spacings were a good candidate to focus on for RMT modeling.

Question 1. *Do normalized spacings between ocean waves satisfy the Wigner surmise?*

In early 2020, we identified a potential dataset from SWIFT buoy data from a project on extreme wave statistics in deep sea waters [34]. Multiple buoys were dropped off the Oregon coast from a helicopter and then were allowed to drift naturally with the current. Buoy

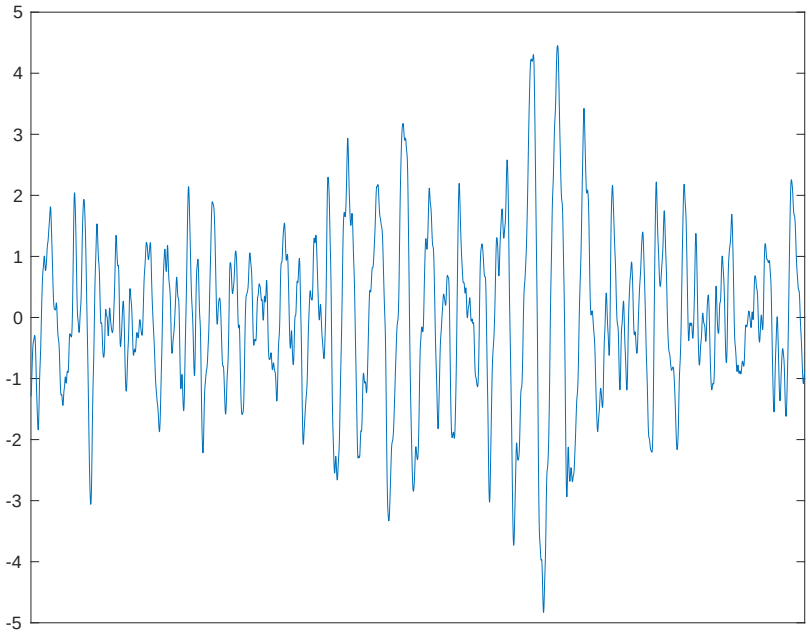


Figure B.1: Sample wave heights in meters during a 10 minute burst at 25 Hz

motion data was recorded in 10 minute bursts at 25 Hz during December 2015, and would repeat until the buoy was recovered on the coast line. With 516 total bursts of wave data, with each including over 10,000 wave height recordings, this is a rich dataset for our analysis. Figure B.1 shows a sample burst output of wave heights during a 10 minute interval.

Our analysis focused on recorded height data statistics to test the Wigner surmise against. We first considered the raw *zero-crossing* wave period data, which recorded the distance between two successive upcrossings across the zero plane. The second focus, with which we are able to give a positive answer to Question 1, used the distance between two successive peaks of zero-crossing waves.

Let the *Kolmogorov-Smirnov (*KS*) distance* between two measures be the maximal distance between the associated cumulative distribution functions. For our analysis, we compared μ_{WS} from (1.75) against the normalized empirical measure $\tau = T/\langle T \rangle$ with T the set of wave spacings. We then used this to compute the *scaled KS distance α-test statistic*, which also

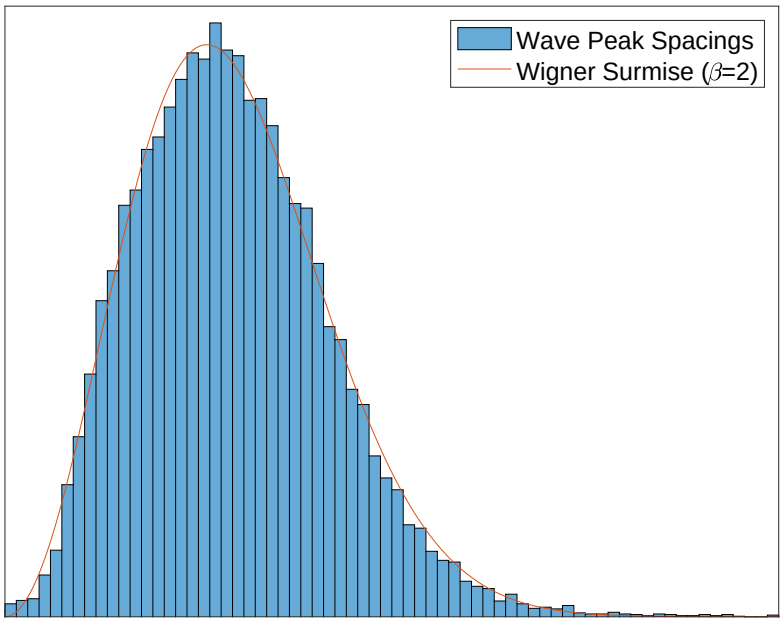


Figure B.2: Normalized Peak Wave Spacings compared to the $\beta = 2$ Wigner surmise

accounts for the test sample size, given by

$$\sqrt{\#}\,\mathrm{KS}(\mu_{\mathrm{WS}}, \tau)$$

Standard cut-offs for statistical significance tests are at scaled KS values less than 1.63, 1.36, and 1.225, respectively, which model low, medium and high significance.

In following with the work on transportation statistics, we find RMT statistics if we, after thresholding appropriately, look at the spacings between wave peaks. Figure B.2 shows the best fit model from time domain filtering, with a scaled KS statistic of 1.1397 using waves with a minimal peak to trough height of 1 meter with a 1.2 second time shift to account for minimal time to move a large body of water a fixed height. To connect this with other filters in the literature, we can roughly construct the same (in a statistical sense) histogram by using filters on the frequency side. When small waves are removed by whichever means, a universal histogram emerges. It just turns out that filtering in the time domain produces

a remarkable fit.

An interesting inverse question can be posed:

Question 2. *Can random matrices be used to simulate ocean wave data?*

Using our results, one can potentially sample random wave data by sampling the eigenvalues for a GUE(n) matrix to generate wave spacings, and sample the components of an eigenvector to generate the associated amplitudes.

Appendix C

Other Group properties of butterfly matrices

Consider the group generated by $\mathrm{B}(N)$:

Definition C.1. *Let* $\mathrm{B}_G(N)$ *denote the subgroup* $\langle \mathrm{B}(N) \rangle = \langle \mathrm{B}^R(N) \rangle$ *of* $\mathrm{SO}(N)$. *Equivalently,*

$$\mathrm{B}_G(N) = \langle \mathcal{D}^{2^j}(N) : j = \mathbf{0}, \ldots, n-1 \rangle, \tag{C.1}$$

or, alternatively, $B_G(N)$ *is generated by the scalar rotation matrices and block diagonal matrices with* $N/2$ *order diagonal entries belonging to* $\mathrm{B}_G(N/2)$, *with* $\mathrm{B}_G(1) = \{1\}$.

C.1 Orbit stabilizer

Now recall the following standard definitions from group theory:

Definition C.2. *Let* G *be a group with identity* e *and* X *a set. We say* X *is a* ***left*** G***-set***

if G acts on X, i.e., we have a map $G \cdot X \to X$, such that for all $g, h \in G$ and $x \in X$, we have $g \cdot x \in X$ and

(i) $e \cdot x = x$ *and*

(ii) $g \cdot (h \cdot x) = (gh) \cdot x$.

An analogous definition works for a ***right*** G***-set.***

For a left G-set X and $x \in X$, the ***stabilizer*** *of x is the subset of G that fixes x under the G-action, that is,*

$$G_x = \mathrm{Stab}_G(x) = \{g \in G : g \cdot x = x\},$$

and the ***orbit*** *of x is the subset of X that G sends x to under its action, that is,*

$$G \cdot x = \mathrm{Orb}_G(x) = \{g \cdot x \in X : g \in G\}.$$

We say the group action of G on X is ***transitive*** *if there is exactly one orbit, that is, $G \cdot x = X$ for all $x \in X$, or equivalently, for any $x, y \in X$ there exists a $g \in G$ such that $g \cdot x = y$.*

If G is both a left and right G-set, then we can define the ***centralizer*** *of X to be the subgroup of G that commutes with every element of X, that is,*

$$C_G(X) = \{g \in G : g \cdot x = x \cdot g \text{ for all } x \in X\}.$$

For G a group, the ***center*** *of the set G, which is itself naturally a G-set, is $Z(G) = C_G(G)$.*

Note a group G always acts transitively on itself (for any $g, h \in G$, then $gh^{-1} \cdot h = gh^{-1}h = g$). Also, any subgroup of G acts on G. A simple exercise follows:

Lemma C.1. *If X is a G-set, then the power set of X, 2^X, is also a G-set.*

Proof. For $Y \subset X$, define $g \cdot Y = \{g \cdot y : y \in Y\} \subset X$. Clearly, this now defines a map $G \times 2^X \to 2^X$, and it follows directly $e \cdot Y = Y$ and $g \cdot (h \cdot Y) = gh \cdot Y$. □

Also, recall the following result with accompanying standard proof (for completeness sake), also from any standard introduction to group theory:

Theorem C.1 (Orbit Stabilizer)**.** *If X is a G-set, then for any $x \in X$ we have G_x is a subgroup of G and $[G : G_x] = |G|/|G_x| = |G \cdot x|$.*

Proof. Note $e \in G_x$. For any $g, h \in G_x$, we have $h^{-1} \cdot x = h^{-1} \cdot (h \cdot x) = h^{-1}h \cdot x = e \cdot x = x$, and so $gh^{-1} \cdot x = g \cdot (h^{-1} \cdot x) = g \cdot x = x$, showing $gh^{-1} \in G_x$. It follows G_x is a subgroup of G, so that $g \cdot x \mapsto gG_x$ is a well-defined one-to-one correspondence between $G \cdot x$ and the set of left cosets of G_x in G. □

C.2 Stabilizer and centralizer in $\mathrm{B}_G(N)$

Since a group acts on itself, then $\mathrm{B}_G(N)$ acts naturally on $\mathrm{B}_G(N)$, while Lemma C.1 then shows $\mathrm{B}_G(N)$ acts naturally on the subsets of $\mathrm{B}_G(N)$, and hence so does $\mathrm{B}_s(N)$. In particular, these act on $\mathrm{B}(N)$ and $\mathrm{B}_s(N)$. So now one might ask what else we can say about these objects.

Proposition C.1. *Let $\mathrm{Stab}_n^L := \mathrm{Stab}_{\mathrm{B}_{G(N)}}(\mathrm{B}(N)) \cap \mathrm{B}(N)$ when viewing the power set of $\mathrm{B}_G(N)$ as a left $\mathrm{B}_G(N)$-set, and $\mathrm{Stab}_n^R := \mathrm{Stab}_{\mathrm{B}_{G(N)}}(\mathrm{B}(N)) \cap \mathrm{B}(N)$ when viewing the power set as a right $\mathrm{B}_G(N)$-set. Let $C_n := C_{\mathrm{B}_{G(N)}}(\mathrm{B}_s(N))$. For $n \leq 1$, we have $\mathrm{B}_s(N) = \mathrm{B}(N) = \mathrm{B}_G(N) = \mathrm{Stab}_n^L = \mathrm{Stab}_n^R$, while for $n \geq 2$, we have*

$$\mathrm{B}_s(N) \subset \mathrm{Stab}_n^L \cap \mathrm{B}(N) = \{R(A \oplus (\pm A)) : R \in \mathcal{R}_N, A \in \mathrm{Stab}_{n-1}^L\} \tag{C.2}$$

and

$$\mathrm{B}_s(N) \not\subset \mathrm{Stab}_n^R \cap \mathrm{B}(N) = \left\{ R(A_1 \oplus A_2) : \pm R \in \left\{ \mathbf{I}, \begin{bmatrix} & \mathbf{I} \\ -\mathbf{I} & \end{bmatrix} \right\}, A_1, A_2 \in \mathrm{Stab}_{n-1}^R \right\}. \tag{C.3}$$

In particular, the last form allows butterfly matrices formed starting with $SO(2)$ blocks followed exclusively with rotations of the form

$$\pm\mathbf{I}, \pm \begin{bmatrix} & \mathbf{I} \\ -\mathbf{I} & \end{bmatrix}.$$

Also,

$$\mathrm{B}_s(N) = C_n \cap \mathrm{B}(N) \tag{C.4}$$

for all n, but $\mathrm{B}_s(N) \neq C_n$ for $n \geq 2$.

Note first an immediate consequence of this proposition, and in particular the result $\mathrm{B}(N)^2 \not\subset \mathrm{B}(N)$, which shows $\mathrm{B}(N)$ is not multiplicatively closed. So this gives an alternative proof that $\mathrm{B}(N)$ is not a group.

It also follows $\mathrm{B}_s(N)\,\mathrm{B}(N) = \mathrm{B}(N)$ and $\mathrm{B}(N)\,\mathrm{B}_s(N) \not\subset \mathrm{B}(N)$, while the only butterfly matrices that commute with simple butterfly matrices are themselves simple butterfly matrices.

Proof of Proposition C.1. Note the statement for $n \leq 1$ is trivial since $\mathrm{B}_s(N) = \mathrm{B}(N) = \mathrm{SO}(N)$.

First, we prove (C.2). Write

$$\mathcal{S}_L := \{R(A \oplus (\pm A)) : R \in \mathcal{R}_N, A \in \mathrm{Stab}^L_{n-1}\}.$$

Note the first inclusion follows from the second equality in (C.2), using induction on n, since simple scalar butterfly matrices satisfy this form (when $A \in \mathrm{B}_s(N/2) \subset \mathrm{Stab}^L_{n-1}$). Also, the inclusion $\mathcal{S}_L \subset \mathrm{Stab}^L_n$ is immediate, using the relation

$$\begin{bmatrix} \mathbf{I} & \\ & -\mathbf{I} \end{bmatrix} \begin{bmatrix} C & S \\ -S & C \end{bmatrix} = \begin{bmatrix} C & -S \\ S & C \end{bmatrix} \begin{bmatrix} \mathbf{I} & \\ & -\mathbf{I} \end{bmatrix} = \begin{bmatrix} C & S \\ -S & C \end{bmatrix}^T \begin{bmatrix} \mathbf{I} & \\ & -\mathbf{I} \end{bmatrix} \tag{C.5}$$

along with Lemma 2.1. (Note (C.5) can be used to show $\mathcal{S}_L$ is a group.) It now remains to show $\mathrm{Stab}^L_n \subset \mathcal{S}_L$.

Fix $B_1 \in \mathrm{B}(N)$ and $B_2 \in \mathrm{Stab}^L_n$. Suppose R_i is a scalar rotation matrix and D_i a block diagonal scalar butterfly matrix, such that $B_i = R_i D_i$ is a scalar butterfly matrix. We see $B_2 B_1 \in \mathrm{B}(N)$ only when

$$B_2 B_1 = (R_2 D_2)(R_1 D_1) = R_2(D_2 R_1 D_1) = R_2(R_3^T D_3) = (R_2 R_3^T) D_3, \tag{C.6}$$

for some scalar rotation matrix R_3 and some block diagonal scalar butterfly matrix D_3, depending on R_1, D_1, D_2. In particular, it follows $D_2 R_1 D_1 = R_3^T D_3$ and hence $R_3 D_2 R_1 D_1 = D_3$ is a block diagonal scalar butterfly matrix. Using now

$$(C_i, S_i) = (\cos\theta_i, \sin\theta_i) \in S^1 \tag{C.7}$$

for the corresponding generators of the scalar rotation matrix R_i, and writing $D_1 = A_1 \oplus A_2$,

$D_2 = A_3 \oplus A_4$ for $A_i \in \mathrm{B}(N/2)$, we see

$$D_3 = R_3 D_2 R_1 D_1 = \begin{bmatrix} C_3 & S_3 \\ -S_3 & C_3 \end{bmatrix} \begin{bmatrix} A_3 & \\ & A_4 \end{bmatrix} \begin{bmatrix} C_1 & S_1 \\ -S_1 & C_1 \end{bmatrix} \begin{bmatrix} A_1 & \\ & A_2 \end{bmatrix}$$
$$= \begin{bmatrix} (C_1C_3A_3 - S_1S_3A_4)A_1 & (S_1C_3A_3 + C_1S_3A_4)A_2 \\ -(C_1S_3A_3 + S_1C_3A_4)A_1 & (-S_1S_3A_3 + C_1C_3A_4)A_2 \end{bmatrix} \tag{C.8}$$

is a block diagonal matrix with diagonal entries in $\mathrm{B}(N/2)$. It follows

$$C_1C_3A_3 - S_1S_3A_4,\ C_1C_3A_3 - S_1S_3A_4 \in \mathrm{Stab}^L_{\mathrm{B}_G(N/2)}(\mathrm{B}(N/2)) \tag{C.9}$$

since $A_1, A_2 \in \mathrm{B}(N/2)$ are arbitrary, and

$$C_1S_3A_4 = -S_1C_3A_3, \quad C_1S_3A_3 = -S_1C_3A_4. \tag{C.10}$$

Since (C_1, S_1) is arbitrary, then we can choose $C_1S_1 \neq 0$, with (C.10) then yielding necessarily $C_3S_3 \neq 0$ since $A_i \in SO(N/2)$, and hence also

$$-\frac{S_1C_3}{C_1S_3} = \pm 1$$

so that $A_3 = \pm A_4$. Next, we can choose $C_1S_1 = 0$, so that (C.10) yields

$$(C_1C_3, S_1S_2) \in \{(\pm 1, 0), (0, \pm 1)\},$$

with then (C.9) yielding $A_3 = \pm A_4 \in \mathrm{Stab}^L_{n-1}$. It follows then $B_2 \in \mathcal{S}_L$ and hence $\mathrm{Stab}^L_n = \mathcal{S}_L$.

Next, we will prove (C.3). Write

$$\mathcal{S}_R = \left\{ R(A_1 \oplus A_2) : \pm R \in \left\{ \mathbf{I}, \begin{bmatrix} & \mathbf{I} \\ -\mathbf{I} & \end{bmatrix} \right\} ; A_1, A_2 \in \mathrm{Stab}_{n-1}^R \, . \right\}$$

As noted in the statement, these consists of butterfly matrices formed by SO(2) blocks followed exclusively by rotations using angles $\theta_j \equiv 0 \pmod{\frac{\pi}{2}}$.

Note the statement $\mathrm{B}_s(N) \not\subset \mathcal{S}_R$ is immediate since not every simple scalar butterfly matrix is of this form. Also, the inclusion $\mathcal{S}_R \subset \mathrm{Stab}_n^R$ is immediate after noting the relation

$$\begin{bmatrix} & \mathbf{I} \\ -\mathbf{I} & \end{bmatrix} \begin{bmatrix} A_1 & \\ & A_2 \end{bmatrix} = \begin{bmatrix} A_2 & \\ & A_1 \end{bmatrix} \begin{bmatrix} & \mathbf{I} \\ -\mathbf{I} & \end{bmatrix} .$$

So now it remains to show $\mathrm{Stab}_n^R \subset \mathcal{S}_R$.

Now let us start from (C.6), where we are now using $B_1 \in \mathrm{Stab}_n^R \cap \mathrm{B}(N)$ and $B_2 \in \mathrm{B}(N)$ is arbitrary — in particular, $D_2 = A_3 \oplus A_4$ is arbitrary. Our goal is now to show $B_1 \in \mathcal{S}_R$.

From (C.8), we have

$$(C_1C_3A_3 - S_1S_3A_4)A_1, (C_1C_3A_4 - S_1S_3A_3)A_2 \in \mathrm{B}(N/2). \tag{C.11}$$

Since $A_3, A_4 \in \mathrm{B}(N/2)$ are arbitrary, we can choose $A_3 = A_4$ so that (C.11) yields

$$A_1, A_2 \in \mathrm{Stab}_{n-1}^R \, . \tag{C.12}$$

It follows then from (C.6) that now it suffices to find R_4, D_4 (defined as before) depending only on D_2, R_1 such that

$$D_2R_1 = R_4D_4.$$

This is possible since (C.12) yields then D_4D_1 is still a block diagonal butterfly matrix, with then

$$B_2B_1 = R_2(D_2R_1)D_1 = R_2(R_4D_4)D_1 = (R_2R_4)(D_4D_1) \in \mathrm{B}(N).$$

Now writing $D_4 = A_5 \oplus A_6$ for $A_i \in \mathrm{B}(N/2)$, we compute

$$\begin{bmatrix} C_1A_3 & S_1A_3 \\ -S_1A_4 & C_1A_4 \end{bmatrix} = D_2R_1 = R_4D_4 = \begin{bmatrix} C_4A_5 & S_4A_6 \\ -S_4A_5 & C_4A_6 \end{bmatrix}$$

so that

$$C_1A_3 = C_4A_5, \quad S_1A_3 = S_4A_6, \quad S_1A_4 = S_4A_5, \quad \text{and} \quad C_1A_4 = C_4A_6. \tag{C.13}$$

Since $A_i \in \mathrm{SO}(N/2)$, then (C.13) implies $(C_1, S_1) = (\pm C_4, \pm S_4)$. Also, from (C.13) we have if $C_1 \neq 0$, then $A_3 = \pm A_5, A_4 = \pm A_6$, while if $S_1 \neq 0$, then $A_3 = \pm A_6, A_4 = \pm A_5$. Hence, if both $C_1 \neq 0$ and $S_1 \neq 0$, then we must have $A_3 = \pm A_4$, but we are free to choose A_3, A_4 arbitrarily in $\mathrm{B}(N/2)$. It follows then necessarily $(C_1, S_1) \in \{(\pm 1, 0), (0, \pm 1)\}$ so that

$$\pm R_1 \in \left\{ \mathbf{I}, \begin{bmatrix} & \mathbf{I} \\ -\mathbf{I} & \end{bmatrix} \right\}, \tag{C.14}$$

and hence $B_1 = R_1D_1 \in \mathcal{S}_R$ by (C.12), and (C.14), showing now $\mathrm{Stab}_n^R \cap \mathrm{B}(N) = \mathcal{S}_R$.

And last, we will prove (C.4). Using the fact $\mathrm{B}_G(N)$ inherits a ring structure, we can write

$$C_n = \{M \in \mathrm{B}_G(N) : [B, M] = \mathbf{0} \text{ for all } B \in \mathrm{B}_s(N)\}.$$

The first inclusion $\mathrm{B}_s(N) \subset C_n$ is trivial since $\mathrm{B}_s(N)$ is abelian. So it remains to establish the reverse inclusion, $C_n \cap \mathrm{B}(N) \subset \mathrm{B}_s(N)$. We will again use induction on n. The result is trivial for $n \leq 1$, since $\mathrm{B}_G(2) = \mathrm{B}(2) = \mathrm{B}_s(2) = \mathrm{SO}(2)$. Assume $\mathrm{B}_s(N/2) = C_{n-1} \cap \mathrm{B}(N/2)$.

Suppose first $M \in C_n \cap \mathrm{B}(N)$, and fix $B = B(\theta) \otimes A \in \mathrm{B}_s(N)$ for $A \in \mathrm{B}_s(N/2)$. First, let square P, Q, U, V be such that

$$M = \begin{bmatrix} P & Q \\ U & V \end{bmatrix}.$$

Now it follows

$$\mathbf{0} = [B, M] = \begin{bmatrix} \cos\theta[A, P] + \sin\theta(AU + QA) & \cos\theta[A, Q] + \sin\theta(AV - PA) \\ \cos\theta[A, U] + \sin\theta(VA - AP) & \cos\theta[A, V] - \sin\theta(AQ + UA) \end{bmatrix}.$$

Since θ is arbitrary, choosing $\theta = 0$ yields $[A, P] = [A, Q] = [A, U] = [A, V] = \mathbf{0}$, and then choosing $\theta = \frac{\pi}{2}$ yields also $AU = -QA = -AQ$ and $AV = PA = AP$ (using $[A, P] = [A, Q] = \mathbf{0}$), and hence $U = -Q$ and $V = P$, so that now

$$M = \begin{bmatrix} P & Q \\ -Q & P \end{bmatrix} = \begin{bmatrix} \cos\varphi A_1 & \sin\varphi A_2 \\ -\sin\varphi A_1 & \cos\varphi A_2 \end{bmatrix},$$

using now also the fact $M \in \mathrm{B}(N)$. It follows

$$\cos\varphi(A_1 - A_2) = \mathbf{0} = \sin\varphi(A_1 - A_2),$$

and hence $A_1 = A_2$. Also, since then

$$[A, P] = \cos\varphi[A, A_1] = \mathbf{0} = \sin\varphi[A, A_1] = [A, Q]$$

so that $[A, A_1] = \mathbf{0}$ while A is arbitrary, we have also $A_1 \in C_{n-1} \cap \mathrm{B}(N/2) = \mathrm{B}_s(N/2)$, using the inductive hypothesis, establishing $M \in \mathrm{B}_s(N)$ and hence $C_n \cap \mathrm{B}(N) = \mathrm{B}_s(N)$, showing (C.4) holds.

□

Note the previous proposition is useful in establishing that $\mathrm{B}(N)$ is not a subgroup of $\mathrm{SO}(N)$. The next question is whether $\mathrm{B}_G(N) = \mathrm{SO}(N)$. This is trivially true for $n \leq 1$ since then $\mathrm{SO}(N) = \mathrm{B}(N) \subset \mathrm{B}_G(N) \subset \mathrm{SO}(N)$. I will show that this is false for $n \geq 2$, but we get equality if we add conjugation by permutation matrices.

Also, note the Orbit Stabilizer Theorem (C.1) doesn't say anything too substantial here. Since $\mathrm{B}_G(N) \cdot \mathrm{B}(N) = 2^{\mathrm{B}_G(N)} = \mathrm{B}(N) \cdot \mathrm{B}_G(N)$ isn't finite (in fact, it has cardinality strictly greater than the continuum — also, recall we are viewing these group actions on the powerset of $\mathrm{B}_G(N)$), then Stab_n^L and Stab_n^R have infinite index in $\mathrm{B}_G(N)$. However, since $\mathrm{B}_s(N) \cdot \mathrm{B}(N) = \{\mathrm{B}(N)\}$ (the singleton consisting only of the element $\mathrm{B}(N)$), then $\mathrm{Stab}^L_{\mathrm{B}_s(N)}(\mathrm{B}(N)) = \mathrm{B}_s(N)$ since this has index 1 in $\mathrm{B}_s(N)$. Note though

$$H := \mathrm{Stab}^R_{\mathrm{B}_s(N)}(\mathrm{B}(N)) = \mathrm{Stab}_n^R \cap \mathrm{B}_s(N)$$

has infinite index in $\mathrm{B}_s(N)$: we see $B(\boldsymbol{\theta})H = B(\boldsymbol{\varphi})H$ if and only if $B(\boldsymbol{\varphi})^T B(\boldsymbol{\theta})H = B(\boldsymbol{\theta} - \boldsymbol{\varphi})H = H$ if and only if $B(\boldsymbol{\theta} - \boldsymbol{\varphi}) \in H$ if and only if $\theta_j \equiv \varphi_j \pmod{\frac{\pi}{2}}$ for all j, so that the cosets of H are in a one-to-one correspondence with $[0, \frac{\pi}{2}]^N$.

Appendix D

Butterfly sectors

This will outline some specific properties that distinguish butterfly matrices from Haar orthogonal matrices.

Definition D.1. *Let* $E_1^N = \{-1,1\}^N$ *and for* $2 \leq j \leq n$,

$$E_j^N = E_{j-1}^N \cap \{(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}) \in \{-1,1\}^N : \mathbf{x}_1, \mathbf{x}_2 \in \{-1,1\}^{N2^{1-j}}, \mathbf{x}_1 = \pm \mathbf{x}_2\}.$$

Note the construction of E_n^N is such that the signs of the first $2m$ components are determined by the signs of the first m components, and so on for $m = 1, \ldots, N/2$. I will refer to such $\mathbf{x} \in \{-1,1\}^N$ as a **sector** of S^{N-1} by implicitly referring to the associated open region in S^{N-1} with that particular sign combination of its components, that is,

$$\operatorname{sgn}^{-1}(\mathbf{x}) \cap S^{N-1} = \{\mathbf{y} \in S^{N-1} : \operatorname{sgn}(\mathbf{y}) = \mathbf{x}\}$$

is the sector associated with $\mathbf{x} \in \{-1,1\}^N$.

Proposition D.1. *For any* $\mathbf{x} \in \{-1,1\}^N$, $B \sim \mathrm{B}(N, \Sigma_S)$, *and* $j \in [N]$, *we have*

$$\mathbb{P}(\mathrm{sgn}(B\mathbf{e}_j) = \mathbf{x}) = \frac{1}{2N} \mathbb{1}_{E_n^N}(\mathbf{x}). \tag{D.1}$$

In particular, the columns of $B \sim \mathrm{B}(N, \Sigma_S)$ *have signs of its components distributed uniformly on* E_n^N.

Since Haar orthogonal or Haar simple orthogonal matrices have columns uniformly distributed on S^{N-1}, then this result shows a clear distinction between random butterfly matrices and Haar orthogonal matrices for $n \geq 2$, since only $2N = 2^{n+1}$ of the 2^N sectors are covered by the columns of $\mathrm{B}(N)$.

Proof. I first claim the result is independent of j. This follows from noting the columns of $B \sim \mathrm{B}(N, \Sigma_S)$ are equal in distribution. For example, to see the first two columns are equal in distribution, we first note $B(\theta_1, \varphi)\mathbf{e}_1 = B(\theta_1 + \frac{\pi}{2} \pmod{2\pi}, \varphi)\mathbf{e}_2$, since the transformation that sends $\begin{bmatrix} \cos\theta_1 \\ -\sin\theta_1 \end{bmatrix}$ to $\begin{bmatrix} \sin\theta_1 \\ \cos\theta_1 \end{bmatrix}$ involves first the maps $\theta_1 \mapsto -\theta_1$ followed by a reflection about $\theta_1 = \frac{\pi}{4}$, which is achieved by $\theta_1 \mapsto \theta_1 - 2(\theta_1 - \frac{\pi}{4}) = \frac{\pi}{2} - \theta_1$. The result then follows since $B(\theta_1, \varphi) \sim B(\theta_1 + \frac{\pi}{2} \pmod{2\pi}, \varphi)$, using also Lemma 1.8.

For general n and j, first note $B \sim \mathrm{Haar}(\mathcal{D}^j(N))$ are invariant under permutation of the underlying block diagonal rotation matrices. This follows since if $\sigma \in S_N$ is a permutation matrix that corresponds to a permutation of two block diagonal entries, then $\sigma \in A_N$ since there need be $N/2^j$ transpositions in σ. Hence, for

$$B = D_1 D_2 \cdots D_n \in \mathrm{B}(n),$$

then there exist σ_i such that

$$B' = D_1' D_2' \cdots D_n' \in \mathrm{B}(n)$$

with $D_i' = P_{\sigma_i} f_i(D_i) P_{\sigma_i}^T$ for $f_i \in \mathrm{Aut}(\mathcal{D}^i(N))$ that yields the map on each block of $\theta \mapsto \theta + \frac{\pi}{2} \pmod{2\pi}$ depending on the parity of j. This combines to yield $B\mathbf{e}_j = B'\mathbf{e}_1$.

A similar argument, making use of (2.10), involves another transformation to construct another $B' \sim \mathrm{B}(N, \Sigma_S)$ such that $B\mathbf{e}_j = B'\mathbf{e}_1$, along with the result $B(\theta_1, \ldots, \theta_{\alpha_n}) \sim B(\theta_{\pi(1)}, \ldots, \theta_{\pi(\alpha_n)})$ for any $\pi \in S_{\alpha_n}$.

Now it suffices to show the result for $j = 1$, and we can assume almost surely $B\mathbf{e}_1$ has nonzero components. I will use induction on $n = \log_2 N$. Let $n = 1$ and fix $\mathbf{x} \in \{-1, 1\}^2$. Then $E_1^2 = \{-1, 1\}^2$ has $|E_1^2| = 4$, while for

$$B = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \sim \mathrm{B}(2, \Sigma_S),$$

we have $B\mathbf{e}_1$ is uniformly distributed on S^1, and hence

$$\mathbb{P}(\mathrm{sgn}(B\mathbf{e}_1) = \mathbf{x}) = \frac{1}{4} = \frac{1}{4}\mathbb{1}_{E_1}(\mathbf{x}).$$

Now assume the result holds for $k \leq n$. Now for $B \sim \mathrm{B}(2N, \Sigma_S)$, let independent $A_1, A_2 \sim \mathrm{B}(N, \Sigma_S)$ and $\theta \sim \mathrm{Uniform}[0, 2\pi)$ be such that

$$B = \begin{bmatrix} \cos\theta A_1 & \sin\theta A_2 \\ -\sin\theta A_1 & \cos\theta A_2 \end{bmatrix}.$$

Let $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ for $\mathbf{x}_i \in \{-1, 1\}^N$. In particular, the inductive hypothesis yields

$$\mathbb{P}(\operatorname{sgn}(A_1\mathbf{e}_1) = \mathbf{x}_1) = \mathbb{P}(\operatorname{sgn}(A_1\mathbf{e}_1) = -\mathbf{x}_1) = \frac{1}{2^{n+1}}\mathbb{1}_{E_n^N}(\mathbf{x}_1). \tag{D.2}$$

Also, for reference, note by definition

$$\mathbb{1}_{E_{n+1}^{2N}}(\mathbf{x}) = \mathbb{1}_{E_n^{2N}}(\mathbf{x})\mathbb{1}_{[\mathbf{x}_1=\pm\mathbf{x}_2]}(\mathbf{x}) = \mathbb{1}_{E_n^N}(\mathbf{x}_1)\mathbb{1}_{[\mathbf{x}_1=\pm\mathbf{x}_2]}(\mathbf{x}). \tag{D.3}$$

We now calculate

$$\begin{aligned}
&\mathbb{P}(\operatorname{sgn}(B\mathbf{e}_1 = \mathbf{x}) = \mathbb{P}(\operatorname{sgn}(\cos\theta A_1\mathbf{e}_1) = \mathbf{x}_1, \operatorname{sgn}(\sin\theta A_2\mathbf{e}_1) = \mathbf{x}_2)) \\
&\quad = \mathbb{P}(\operatorname{sgn}(\cos\theta A_1\mathbf{e}_1) = \mathbf{x}_1, \operatorname{sgn}(\sin\theta A_1\mathbf{e}_1) = \mathbf{x}_2 \mid \theta \in [0, \pi/2))\mathbb{P}(\theta \in [0, \pi/2)) \\
&\qquad + \mathbb{P}(\operatorname{sgn}(\cos\theta A_1\mathbf{e}_1) = \mathbf{x}_1, \operatorname{sgn}(\sin\theta A_1\mathbf{e}_1) = \mathbf{x}_2 \mid \theta \in [\pi/2, \pi))\mathbb{P}(\theta \in [\pi/2, \pi)) \\
&\qquad + \mathbb{P}(\operatorname{sgn}(\cos\theta A_1\mathbf{e}_1) = \mathbf{x}_1, \operatorname{sgn}(\sin\theta A_1\mathbf{e}_1) = \mathbf{x}_2 \mid \theta \in [\pi, 3\pi/2))\mathbb{P}(\theta \in [\pi, 3\pi/2)) \\
&\qquad + \mathbb{P}(\operatorname{sgn}(\cos\theta A_1\mathbf{e}_1) = \mathbf{x}_1, \operatorname{sgn}(\sin\theta A_1\mathbf{e}_1) = \mathbf{x}_2 \mid \theta \in [3\pi/2, 2\pi))\mathbb{P}(\theta \in [3\pi/2, 2\pi)) \\
&\quad = \frac{1}{4}\mathbb{P}(\operatorname{sgn}(A_1\mathbf{e}_1) = \mathbf{x}_1)\mathbb{1}_{[\mathbf{x}_1=\mathbf{x}_2]}(\mathbf{x}) + \frac{1}{4}\mathbb{P}(\operatorname{sgn}(A_1\mathbf{e}_1) = -\mathbf{x}_1)\mathbb{1}_{[\mathbf{x}_1=-\mathbf{x}_2]}(\mathbf{x}) \\
&\qquad + \frac{1}{4}\mathbb{P}(\operatorname{sgn}(A_1\mathbf{e}_1) = -\mathbf{x}_1)\mathbb{1}_{[\mathbf{x}_1=\mathbf{x}_2]}(\mathbf{x}) + \frac{1}{4}\mathbb{P}(\operatorname{sgn}(A_1\mathbf{e}_1) = \mathbf{x}_1)\mathbb{1}_{[\mathbf{x}_1=-\mathbf{x}_2]}(\mathbf{x}) \\
&\quad = \frac{1}{4}\mathbb{P}(\operatorname{sgn}(A_1\mathbf{e}_1) = \mathbf{x}_1)\mathbb{1}_{[\mathbf{x}_1=\pm\mathbf{x}_2]}(\mathbf{x}) + \frac{1}{4}\mathbb{P}(\operatorname{sgn}(A_1\mathbf{e}_1) = -\mathbf{x}_1)\mathbb{1}_{[\mathbf{x}_1=\pm\mathbf{x}_2]}(\mathbf{x}) \\
&\quad = \frac{1}{2}\frac{1}{2^{n+1}}\mathbb{1}_{E_n^N}(\mathbf{x}_1)\mathbb{1}_{[\mathbf{x}_1=\pm\mathbf{x}_2]}(\mathbf{x})) && \text{(by (D.2))} \\
&\quad = \frac{1}{2^{n+2}}\mathbb{1}_{E_{n+1}^{2N}}(\mathbf{x}), && \text{(by (D.3))}
\end{aligned}$$

so the result follows.

To see the columns of $B \sim \mathrm{B}(N, \Sigma_S)$ have signs distributed uniformly on E_n^N, it is enough to note $|E_n^N| = 2^{n+1}$. This follows from induction and (D.3), which gives $|E_n^N| = 2|E_{n-1}^{N/2}|$, while $|E_1^2| = 4$. (Note (D.1) only yields $|E_n^N| \geq 2^{n+1}$ since summing this probability over all sign combinations must add up to 1.) $\square$