Title

Two Geometric Results regarding Hölder-Brascamp-Lieb Inequalities, and Two Novel Algorithms for Low-Rank Approximation

Author

Rusciano, Alexander

Publication Date

2019

Peer reviewed|Thesis/dissertation

Two Geometric Results regarding Hölder-Brascamp-Lieb Inequalities, and Two Novel Algorithms for Low-Rank Approximation

by

Alexander Robert Rusciano

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Mathematics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor James Demmel, Chair
Professor Ming Gu
Assistant Professor Nikhil Srivastava
Professor Kathy Yelick

Fall 2019

Two Geometric Results regarding Hölder-Brascamp-Lieb Inequalities, and Two Novel Algorithms for Low-Rank Approximation

Abstract

Two Geometric Results regarding Hölder-Brascamp-Lieb Inequalities, and Two Novel Algorithms for Low-Rank Approximation

by

Alexander Robert Rusciano

Doctor of Philosophy in Mathematics

University of California, Berkeley

Professor James Demmel, Chair

Broadly speaking, this thesis investigates mathematical questions motivated by computer science. The involved topics include communication avoiding algorithms, classical analysis, convex geometry, and low-rank matrix approximation. In total, the thesis consists of four self-contained sections, each adapted from papers the author has been a part of.

The first two sections are both motivated by the Brascamp-Lieb inequalities, which are also often referred to as Hölder-Brascamp-Lieb inequalities. These inequalities have featured prominently in recent theoretical computer science work, due to connections to geometric complexity theory [32], harmonic analysis [10], communication-avoidance [19], and many other areas. Moreover, work generalizing the inequalities in various ways, such as to nonlinear versions, has been impactful to the study of differential equations.

Section 1 studies the application of Hölder-Brascamp-Lieb (HBL) inequalities to the design of communication optimal algorithms. In particular, it describes optimal tiling (blocking) strategies for nested loops that lack data dependencies and exhibit affine memory access patterns. The problem roughly amounts to maximizing the volume of an object provided some of its linear images have bounded volume. The methods used are algorithmic.

Another reason for the interest in these inequalities is because they are an interesting test case for non-convex optimization techniques. The optimal constant for a particular instance of the inequality is given by solving a non-convex optimization problem that is still highly structured [68, 32, 60]. Of particular relevance to this thesis is that it can be formulated as a geodesically-convex problem, considered in the context of the manifold of positive definite matrices of determinant 1 (the symmetric space SL_n/SO_n [8]). Even using the methods of Section 1, the procedure is not necessarily polynomial time, and this motivates further study

of geodesic convexity.

This lead to the work of Section 2, which discusses a notion of halfspace for Hadamard manifolds that is natural in the context of convex optimization. For this notion of halfspace, we generalize a classic result of Grünbaum, which itself is a corollary of Helly's theorem. Namely, given a probability distribution on the manifold, there is a point for which all halfspaces based at this point have at least $\frac{1}{n+1}$ of the mass, n being the dimension of the manifold. As an application, the gradient oracle complexity of geodesic convex optimization is polynomial in the parameters defining the problem. In particular it is polynomial in $log(\epsilon^{-1})$, where ϵ is the desired error. This is a step toward the open question of whether such an algorithm exists.

The remaining two sections of the paper present a different research direction, randomized numerical linear algebra. Numerical linear algebra has long been an important part of scientific computing. Due to the current trend of increasing matrix sizes and growing importance of fast, approximate solutions in industry, randomized methods are quickly increasing in popularity. Sections 3 and 4 in this thesis aim to show that randomized low-rank approximation algorithms satisfy many of the properties of classical rank-revealing factorizations.

Section 3 introduces a Generalized Randomized QR-decomposition (**RURV**) that may be applied to arbitrary products of matrices and their inverses, without needing to explicitly compute the products or inverses. This factorization is a critical part of a communication-optimal spectral divide-and-conquer algorithm for the nonsymmetric eigenvalue problem. In this paper, we establish that this randomized QR-factorization satisfies the strong rank-revealing properties. We also formally prove its stability, making it suitable in applications. Finally, we present numerical experiments which demonstrate that our theoretical bounds capture the empirical behavior of the factorization.

Section 4 concerns a Generalized LU-Factorization (**GLU**) for low-rank matrix approximation. We relate this to past approaches and extensively analyze its approximation properties. The established deterministic guarantees are combined with sketching ensembles satisfying Johnson-Lindenstrauss properties to present complete bounds. Particularly good performance is shown for the sub-sampled randomized Hadamard transform (SRHT) ensemble. Moreover, the factorization is shown to unify and generalize many past algorithms. It also helps to explain the effect of sketching on the growth factor during Gaussian Elimination.

To my family and friends,

Contents

Acknowledgments

I am indebted to my friends, who have made my time at UC Berkeley enjoyable.

I would also like to thank my adviser James Demmel, as well as the other researchers I have had the opportunity with, Grey Ballard, Ioana Dumitriu, and Laura Grigori. They are credited in the appropriate sections.

Finally, several individuals were generous with their time, discussing problems with me or reading my work. These include Alex Appleton, Richard Bamler, Andrew Hanlon, and Nikhil Srivastava.

Chapter 1

Parallelepipeds obtaining HBL upper bounds[1,2]

1.1 Introduction

Hölder-Brascamp-Lieb (HBL) Inequalities

HBL inequalities are very general, including famous inequalities such as Hölder's inequality and Young's inequality. To state the inequalities in general, fix maps $\phi_i : X \to X_i$ and measures dx, dx_i on X, X_i respectively, with $i \in \{1, \ldots, m\}$. Then for weights $\{s_i\}_{i=1}^m \geq 0$, they take the form

$$\int_{x \in X} \prod_{i=1}^{m} f_i(\phi_i(x))dx \leq C(s) \prod_{i=1}^{m} \|f_i\|_{1/s_i} \tag{1.1}$$

holding for all integrable nonnegative f_i on X_i. We have denoted by $\|\cdot\|_{1/s_i}$ the L_p quasinorm and by $C(s)$ the smallest constant for which the inequality holds for all choices of f_i. Perhaps the most studied instances are when X, X_i are vector spaces with the Lebesgue measure or abelian groups under the counting measure, and the ϕ_i are linear maps. For example, in these cases the set of all s for which $C(s) < \infty$ are known to form a convex polytope we will denote by $\mathcal{P}$. All of our considerations are restricted to these two cases, and are intimately concerned with $\mathcal{P}$. This polytope has an interesting structure, although the two cases differ slightly:

For the discrete case, [20] establishes

[1] Joint work with James Demmel

[2] Preprint [26], submitted for publication to SIAM J. Discrete Mathematics.

Theorem 1. *Given linear maps $\phi : \mathbb{Z}^d \to \mathbb{Z}^{d_i}$, a collection $s_i \geq 0$ satisfies inequality (??) if and only if they satisfy for all subgroups $H \leq \mathbb{Z}^d$,*

$$\sum_{i=1}^{m} s_i \cdot rank(\phi_i(H)) \geq rank(H)$$

Moreover, the sharpest constant is $C(s) = 1$.

The Lebesgue case was studied in [10], and their Theorem 1.13 establishes

Theorem 2. *Given linear maps $\phi : \mathbb{R}^d \to \mathbb{R}^{d_i}$, a collection $s_i \geq 0$ satisfies inequality (??) if and only if they satisfy for all subspaces $H \leq \mathbb{R}^d$,*

$$\sum_{i=1}^{m} s_i \cdot dim(\phi_i(H)) \geq dim(H)$$

and in addition equality holds for the case $H = \mathbb{R}^d$. The sharpest constant $C(s)$ is given by the solution to the optimization problem over the set of positive definite matrices $X_i \succeq 0$

$$C(s) = \sup_{X_i \succeq 0} \left[\frac{\prod_i det(X_i)^{s_i}}{det(\sum_i \phi_i^T X_i \phi_i)} \right]^{1/2}$$

Motivation from communication-avoiding algorithms, and Notation

The goal of this subsection is to provide a brief summary of the computational model introduced in [39], to motivate the origins of the problem we study. We recommend the interested reader to find more detail and examples in [19].

For the application to communication-avoiding algorithms, the spaces X, X_i will be $\mathbb{Z}^d$, $\mathbb{Z}^{d_i}$ and the measures will be the counting measures. The proofs in the main body adopt this situation. We still note that the fundamental geometric insight of Thm ?? at the heart of the paper remains true in the case when X, X_i are real vector spaces and the measures are Lebesgue. As this could be of independent interest, Appendix ?? provides the minor modifications needed to adapt the proofs to the Lebesgue case.

To connect the HBL inequalities to communication costs of an algorithm, within some region of the space $\mathbb{Z}^d$ each point will correspond to executing one step of an algorithm. For example [39], for classical matrix multiplication $C = A * B$, $d = 3$ and the point $x = (i, j, k) \in \mathbb{Z}^3$ corresponds to the execution of

$$C(i, j) = C(i, j) + A(i, k) \cdot B(k, j)$$

To execute this, one needs to have the 3 corresponding array entries available in memory, and these are indicated by the 3 linear maps $\phi_1(x) = (i, j)$ (to identify the entry $C(i, j)$),

$\phi_2(x) = (i,k)$ and $\phi_3(x) = (k,j)$. So to perform a set of computations $S \subset \mathbb{Z}^3$, one needs the data $\phi_1(S)$, $\phi_2(S)$ and $\phi_3(S)$. To use the memory of size M as efficiently as possible, one wants to maximize the amount of work possible, the cardinality $|S|$ of S, given all the data available in memory: $|\phi_1(S)| + |\phi_2(S)| + |\phi_3(S)| \leq M$. We approximate this by maximizing $|S|$ subject to the 3 constraints $|\phi_i(S)| \leq M$. (Appendix B deals with the more precise constraint.) Theorem **??** provides this upper bound: By choosing f_i to be the indicator function of $\phi_i(S)$, we see $x \in S$ implies $\prod_{i=1}^m f_i(\phi_i(x)) = 1$, so

$$|S| \leq \int_{x \in X} \prod_{i=1}^m f_i(\phi_i(x))dx \leq \prod_{i=1}^m \|f_i\|_{1/s_i} = \prod_{i=1}^m |\phi_i(S)|^{s_i} \leq \prod_{i=1}^m M^{s_i} = M^{\vec{1}^T s} \tag{1.2}$$

As stated above, this upper bound holds for all s in a convex polytope $\mathcal{P}$, so the tighest upper bound is gotten by minimizing $\vec{1}^T s$ over all $s \in \mathcal{P}$, a linear program. As explained in the next section, we denote this minimum value of $\vec{1}^T s$ by $h_*(\vec{1})$.

References [39, 19] show how to use this upper bound on $|S|$, which happens to be $M^{h_*(\vec{1})} = M^{3/2}$ in the case of classical matrix multiplication, to get a lower bound on the amount of data that needs to be moved into and out of memory in order to execute the entire computation, which is $\Omega(n^3/M^{1/2})$ for multiplying n-by-n matrices. This is of significant practical interest because the cost (measured in time or energy) of moving data (i.e. communication) can be much larger than the cost of the arithmetic operations on data in memory.

Given this lower bound on communication costs, the practical goal is to find an algorithm that attains it. This means finding a set S that not only (approximately) attains its upper bound on $|S|$, but *tiles* all the points $x \in \mathbb{Z}^d$ that correspond to the execution of an algorithm. This means that we need to able to write $\mathbb{Z}^d$ (or appropriate subsets) as a disjoint union of shifted copies of S. In the case of classical matrix multiplication, the optimal S turns out to be a cube of side length $M^{1/2}$. Our contribution in this paper is to find such a parallelepiped S in general.

Since S depends on the memory size M, in the remainder of this paper we will use the notation $S(M)$, and derive results that hold asymptotically in M.

Main Result

Before stating the main geometric content of the paper, we generalize the context with no additional difficulty. Denote by $|\cdot|$ the Lebesgue measure or cardinality as appropriate. Suppose we require $|\phi_i(S(M))| = O(M^{\alpha_i})$ for some $\alpha_i \geq 0$, with as always the asymptotic part referring to M. Then generalizing to (coordinate-wise) $\alpha \geq 0$ in place of bound (**??**)

$$|S(M)| \leq \prod |\phi_i(S(M))|^{s_i} = O(M^{\alpha^T s}) \tag{1.3}$$

and $h_*(\alpha) := \min_{s \in \mathcal{P}} \alpha^T s$ determines the tightest upper bound.

Definition 3. *The family of sets $S(M)$, parametrized by integer M, are considered asymptotically optimal if*

$$|S(M)| = \Theta(M^{h_*(\alpha)}) \tag{1.4}$$

as well as for all i,

$$|\phi_i(S(M))| = O(M^{\alpha_i}) \tag{1.5}$$

*We say the member sets **tile** if translations of a given member injectively cover the space ($\mathbb{R}^d$ or $\mathbb{Z}^d$).*

We can now state our main contribution:

Theorem 4. *For any choice of $\alpha \geq 0$, there exists a family of asymptotically optimal shapes as defined in Definition ??. Moreover, the family can be chosen to be parallelepiped-like shapes which tile.*

The rest of this paper is organized as follows.

Section 2 introduces the primal and dual linear programs (LPs) associated with the discrete HBL inequalities, as well as the tiling shape and procedure.

Section 3 shows how to construct an optimal tiling in the general case, given an optimal solution to the HBL LP, our main result.

Section 4 draws conclusions and presents a couple possible future directions.

We include some Appendices with further details and examples:
Appendix A modifies the shape construction from the discrete to the continuous case.
Appendix B refines the definition of "optimal" tiling to reflect the bound $\sum_{i=1}^{m} |\phi_i(S)| \leq M$ rather than the approximation $|\phi_i(S)| \leq M$, $\forall i$.
Appendix C gives a numerical example of rank 1 maps in 2 dimensions.

1.2 HBL Primal and Dual

Primal

Because there are only finitely many possible values of the rank of H and its images, there exists a finite list of subgroups sufficient in Thm ??. This is why we the problem of computing $h_*(\alpha)$ can be formulated as a linear program (LP), if we can find a sufficient list of subgroups. This can be used to justify calling $\mathcal{P}$ a Polytope, and the following a linear program, even if ostensibly there are an infinite number of constraints.

Definition 5 (HBL Primal LP)**.** *The HBL primal LP is*

$$\begin{aligned} \underset{s}{minimize} \quad & \alpha^T s \\ subject\ to \quad & \sum_{i=1}^{m} s_i \cdot rank(\phi_i(H)) \geq rank(H),\ \forall H \leq \mathbb{Z}^d \\ & s \geq 0 \end{aligned}$$

Although not a focus of this work, recent progress has led to a better understanding of $\mathcal{P}$. In [20] a few things are established. For one thing, only the lattice of subgroups generated by $\ker(\phi_i)$ under sums and intersections needs to be used to generate inequality constraints in Theorem **??**. However, this lattice is often infinite in higher dimensions. Also, [20] describes a terminating algorithm which discovers all the constraints needed to formulate an equivalent LP. However, the algorithm's complexity is unknown. The results of [32] provide a number of novel insights into algorithmic computation of the Lebesgue version of $\mathcal{P}$, including a polynomial time membership and weak separation oracles. Here are a selection of special cases for which computation of $\mathcal{P}$ could be reasonably managed:

- All maps are coordinate projections [19].
- All cases when the dimension of the computation lattice is $d \leq 5$ [18].
- Each $\ker(\phi_i)$ is rank 1,2, $d-1$, or $d-2$, some mixes are allowed [64]. (Stated for Lebesgue version of $\mathcal{P}$).
- There are no more than 3 maps; then the kernel subgroup lattice is bound by 28, a classical result [23].

These cases likely cover many of the communication-avoiding application cases.

Dual

In this and the subsequent section, we show how the dual of the HBL Primal LP leads to an asymptotically optimal parallelpiped tiling of the computation lattice.

Denote the dual variable by y; its indices are in bijection with subgroups of $\mathbb{Z}^d$. We require that only finitely many of its entries are nonzero. The groups corresponding to these indices, termed the support, are grouped into the list $\mathbf{E}$ with members E_j. For linear maps L, we subsequently employ the natural shorthand

$$y^T \text{rank}(L(\mathbf{E})) := \sum_{E_j} y_{E_j} \text{rank}(L(E_j))$$

Definition 6 (Dual LP)**.** *Recall the HBL setting consists of maps ϕ_i from lattice $\mathbb{Z}^d$. A dual vector y will be considered to be indexed by all subgroups of $\mathbb{Z}^d$, but with finitely many non-zero coordinates. The non-zero coordinates are defined to be the support of y. Now define the objective value of y to be*

$$\begin{aligned} \underset{y}{\text{maximize}} \quad & val(y) := y^T rank(\mathbf{E}) \\ \text{subject to} \quad & C_i(y) := y^T rank(\phi_i(\mathbf{E})) \leq \alpha_i, \forall \phi_i \\ & y \geq 0 \end{aligned} \tag{1.6}$$

As the support list $\mathbf{E}$ changes during our algorithm, we use a few symbols in place of $\mathbf{E}$ when extra information is present. Typically we will use $\mathbf{Y}$ when the supporting subgroups are independent, and $\mathbf{U}$ when they are a flag. These definitions are covered later.

Interpretation of Dual

The dual is important because of its geometric significance. Most clearly, it is readily interpretable when the supporting subgroups $\mathbf{Y}$ of the dual vector are independent. Here independent means that $\text{rank}(\oplus_i Y_i) = \sum_i \text{rank}(Y_i)$. We need the following intuitive lemma.

Lemma 7. *Take any independent elements $e_1, \ldots, e_h$ contained in rank h subgroup $Y \subset \mathbb{Z}^d$. Define the set*

$$S := \{z \in \mathbb{Z}^d | z = \sum_i a_i e_i \text{ with } 0 \leq a_i \leq \lfloor M^k \rfloor - 1,\ a_i \in \mathbb{Z}\} \tag{1.7}$$

In this equation, k is an arbitrary fixed positive number, determining how sizes scale with M.

Then for any linear map L, $|S| = \lfloor M^k \rfloor^h$ and $|L(S)| = O(M^{kr})$ where $r := rank(L(Y))$. In applications later, L is taken to be one of the ϕ_i.

Proof. The elements in set S are $O(M^k)$ from the origin in Euclidean distance, hiding the dimensional factor d in the big O notation. By linearity, the elements of $L(S)$ are also $O(M^k)$ from the origin in $\text{im}(L)$. Therefore an r dimensional cube residing within $L(Y)$ with side lengths $O(M^k)$ can contain $L(S)$. This means that $|L(S)| = O(M^{kr})$.

Finally, from independence of the e_i, it follows that $|S| = \lfloor M^k \rfloor^h$ □

We now define the parallelepiped-like construction that will be used to create asymptotically optimal tilings. Although not the only possible way to build good tiling shapes, it is flexible and leads to the cleanest descriptions in the case of $\mathbb{Z}^d$.

Definition 8 (Product Parallelepiped)**.** *Suppose we are given a dual vector y, with support given by the list of independent subgroups $(Y_1, \ldots, Y_t)$. Form S_{Y_i} as in Eq. ??, using $k = y_{Y_i}$.*

Now define a parallelepiped shape through a Minkowski sum of sets

$$S := S_{Y_1} + \cdots + S_{Y_t} \tag{1.8}$$

The independent elements used to construct S_{Y_i} are left unspecified; the choice affects constants, but will not affect optimality in the sense of Def. **??**.

We will use the construction of Def. **??** to produce optimal tilings. Therefore we begin to relate it to the HBL problem:

Proposition 9. *Suppose we are given a dual vector y, whose non-zero values are attached to a list of independent subgroups $\mathbf{Y} = (Y_1, \ldots, Y_t)$. Form the product parallelepiped S of Def. **??**.*

Then $|S| = \Theta(M^{y^T rank(\mathbf{Y})})$. If in addition y is dual feasible, then $|\phi_i(S)| = O(M^{\alpha_i})$ holds for each ϕ_j.

Proof. By independence of the subgroups contained in $\mathbf{Y}$, it follows that $|S| = \prod_i |S_{Y_i}|$. Apply the count estimates of Lemma **??** to this:

$$|S| = \prod_i \Theta(M^{\mathrm{rank}(Y_i)\cdot y_{Y_i}}) = \Theta(M^{y^T \mathrm{rank}(\mathbf{Y})})$$

It remains to consider the images of this set under the ϕ_i in the case y is feasible. This requires a bound on $|\phi_i(S)|$. Invoking the count estimates of Lemma **??** in the second inequality below, and feasibility property (**??**) of y in the third inequality

$$|\phi_i(S)| \leq \prod_j |\phi_i(S_{Y_j})| = \prod_j O\left(M^{\mathrm{rank}(\phi_i(Y_j))y_{Y_j}}\right) = O(M^{C_i(y)}) = O(M^{\alpha_i})$$

□

It will be necessary to strengthen the bounds on the $|\phi_i(S)|$ later.

It is clear enough that parallelepipeds in $\mathbb{R}^d$ tile, but the above version is slightly non-standard because of the discreteness of the object. To make things explicit, Algorithm **??** below produces the translations needed to tile $\mathbb{Z}^d$ with set S.

In the algorithm, we employ a matrix factorization for linear maps between abelian groups (or more generally between modules over a principle ideal domain) known as the Smith Normal Form. The Smith Normal Form of a matrix A with integer entries is of the form $A = UDV^{-1}$ where U, V are unimodular and D is diagonal with non-negative integer entries. Its diagonal entries $d_i := D_{ii}$ are uniquely defined by requiring $d_i | d_{i+1}$.

Algorithm 1 Construct a tile S and its translations T that tile $\mathbb{Z}^d$

Input: Memory size parameter M
Input: For each $i = 1, \ldots, t$: subgroup Y_i represented by a matrix with independent columns
Input: Values y_{Y_i} associated to each Y_i
$Y \leftarrow (Y_1, \ldots, Y_t)$
$S \leftarrow \{Y \cdot (a_{11}, a_{12}, \ldots, a_{th_t})^T \mid a_{ij} \in \{0, \ldots, \lfloor M^{y_{Y_i}} \rfloor - 1\}$
$m \leftarrow \sum h_i$
$(U, D, V) \leftarrow$ Smith Normal Form(Y)
$U' \leftarrow$ last $d - m$ columns of U
$U'' \leftarrow$ first m columns of U
$T_1 \leftarrow \{Y \cdot (a_{11}, a_{12}, \ldots, a_{th_t})^T \mid a_{ij} \in \lfloor M^{y_{Y_i}} \rfloor \cdot \mathbb{Z}\}$
$T_2 \leftarrow \{U' \cdot (a_1, \ldots, a_{d-m})^T \mid a_i \in \mathbb{Z}\}$
$T_3 \leftarrow \{U'' \cdot (b_1, \ldots, b_m)^T \mid b_i \in \{0, \ldots, d_i - 1\}\}$
$T \leftarrow$ Minkowski sum $T_1 + T_2 + T_3$
Return S, T

The set S returned by the algorithm exactly follows Def. **??**. The translations T come from two sources: T_1 accounts for the finite size of M while tiling the subgroup generated by the columns of Y under integer linear combinations. In the remainder of this section, we will write this subgroup as $\langle Y \rangle$, to differentiate between the subgroup and the matrix. The sets T_2, T_3 account for the need to tile each coset in $\mathbb{Z}^d / \langle Y \rangle$.

Proposition 10. *Algorithm* **??** *correctly outputs a parallelepiped set S which under translation by T tiles $\mathbb{Z}^d$. This holds for any input: that is, for any selection of independent subgroups Y_i, choice of independent elements within these subgroups, associated values y_{Y_i}, and memory size M.*

Proof. Let us begin with the image of Y, by considering $x = Y \cdot a$. It will further be convenient to let e_{ij} enumerate the columns of block Y_i of Y, so that a is indexed by a_{ij}. We now observe that there is exactly one member $t_1 \in T_1$ that yields $x - t_1 \in S$. Indeed, the set S only uses scalings of 0 to $\lfloor M^{y_{Y_i}} \rfloor - 1$ for each e_{ij}. As Y is injective, for $x - t_1 \in S$ to be true, t_1 must be produced by scaling e_{ij} by an amount in the range

$$[a_{ij} - \lfloor M^{y_{Y_i}} \rfloor + 1, a_{ij}]$$

As elements of T_1 are by construction required to scale e_{ij} by multiples of $\lfloor M^{y_{Y_i}} \rfloor$, the only such member of T_1 uses a scaling of $\lfloor a_{ij} / \lfloor M^{y_{Y_i}} \rfloor \rfloor \cdot \lfloor M^{y_{Y_i}} \rfloor$ for e_{ij}.

This shows that each $T_1 + S$ is exactly the image of Y. We now complete the proof by showing that $T_2 + T_3$ contains exactly one element from each coset of

$$\mathbb{Z}^d/\langle Y\rangle \simeq \mathbb{Z}^{d-m} \oplus (\bigoplus_{i=d-m}^{d} \mathbb{Z}/d_i\mathbb{Z})$$

To be specific, T_2 accounts for the free component, and T_3 for the torsion component.

Let $U'a + U''b$ and $U'a' + U''b'$ be two distinct elements of $T_2 + T_3$. Saying they are in the same coset is exactly saying their difference lies in $\mathrm{im}(Y)$. Writing $Y = UDV^{-1}$ as in Algorithm **??** and noting V is unimodular, it is clear that $\mathrm{im}(Y) = \mathrm{im}(UD)$. Conclude that lying in the same coset is equivalent to

$$U'a + U''b - U'a' - U''b' \in \mathrm{im}(UD)$$

As U is unimodular, This means for some $c \in \mathbb{Z}^d$

$$(b; a)^T - (b'; a')^T = Dc$$

Because D is d-by-m, the last $d - m$ coordinates of Dc are 0. This means $a = a'$. Also for b, b' to be used in T_3, they must satisfy $0 \leq b_i, b_i' < d_i$. But then

$$-d_i < b_i - b_i' = d_i c_i < d_i$$

which is only possible if $b_i = b_i'$.

To conclude that all cosets are represented, we need to show that for any $x \in \mathbb{Z}^d$, there are $U'a, U''b$ such that $x - U(b; a)^T \in \mathrm{im}(Y)$. Simply take $b_i = (U^{-1}x)_i \bmod d_i$ and $a_i = (U^{-1}x)_{m+i}$.

□

We include an example in Appendix **??** to illustrate this approach.

1.3 Construction of Optimal Shape

In this section, we describe an algorithm for producing an asymptotically optimal tiling. The recipe is to formulate the primal, solve the corresponding dual, and iteratively modify the solution of the dual to something geometrically interpretable. Consequently, at least asymptotically, the HBL upper bounds are attainable by a parallelepiped, and to do so is essentially no harder than describing the Brascamp-Lieb polytope $\mathcal{P}$.

Solutions Supported on Flags

It might not be possible to find a dual vector supported on independent subgroups that obtains the optimal value. However, it turns out that it is possible to find one supported on what we here define to be a flag.

Definition 11. *A flag of the lattice $\mathbb{Z}^d$ is a sequence* $\mathbf{U}$ *of strictly nested subgroups*

$$\{0\} < U_1 < \cdots < U_t = \mathbb{Z}^d$$

Our approach is to establish that an arbitrary optimal solution to the dual can be transformed to an optimal one supported on a flag. The following is a simple but important property in accomplishing this goal. It was also helpful in [20] and [64], the latter of which found flags useful in studying the vertices of the Brascamp-Lieb polyhedron. Indeed, one could view Thm. **??** as an algorithmic, dual version of [64]'s insight that vertices of $\mathcal{P}$ correspond to certain flags.

Lemma 12. *For any linear map L on $\mathbb{Z}^d$ and subgroups V, W,*

$$\mathit{rank}(L(V)) \geq \mathit{rank}(L(V \cap W)) + \mathit{rank}(L(V + W)) - \mathit{rank}(L(W))$$

On the other hand,

$$\mathit{rank}(V) = \mathit{rank}(V \cap W) + \mathit{rank}(V + W) - \mathit{rank}(W)$$

Proof. The claimed equality in the lemma follows by writing a basis for $V \cap W$ and completing it to a basis for W with a second set of independent basis elements. Call the subgroup spanned by the second set P. Observe P has trivial intersection with V, and the rank of P is $\text{rank}(W) - \text{rank}(V \cap W)$. Applying these observations to $W + V = P + V$,

$$\text{rank}(W + V) = \text{rank}(P + V) = \text{rank}(P) + \text{rank}(V) = \text{rank}(W) - \text{rank}(W \cap V) + \text{rank}(V)$$

establishing the result. To prove the inequality, apply the equality to subspaces $L(V)$, $L(W)$, and then observe

$$L(V \cap W) \subseteq L(V) \cap L(W), \text{ while } L(V + W) = L(V) + L(W)$$

The reason for the possible inequality is that maybe there are different elements $v \in V$ and $w \in W$, but $L(v) = L(w)$. □

We employ this observation repeatedly to shift the support of a dual vector onto a flag, through the following procedure. It takes as input a feasible y supported on an arbitrary list $\mathbf{E}$ and outputs a feasible y' supported on a flag $\mathbf{U}$ with the same objective value. Recall by feasible we mean $y \geq 0$ and Eq. **??** are satisfied.

Algorithm 2 Find feasible y' supported on flag $U_1, \dots U_t$ with same objective value

Input: feasible vector y supported on $E_1, \dots, E_m$
Initialize y' as y
while y' is not supported on a flag **do**
  $V, W \leftarrow$ any pair in the support of y' NOT satisfying $V \subset W$ or $W \subset V$
  Let V be the member of the pair with $y'_V \leq y'_W$
  $y'_W \leftarrow y'_W - y'_V$
  $y'_{V+W} \leftarrow y'_{V+W} + y'_V$
  $y'_{V \cap W} \leftarrow y'_{V \cap W} + y'_V$ (if $V \cap W \neq \{0\}$)
  $y'_V \leftarrow 0$
end while
Return y' and its support, denoted $U_1, \dots, U_t$

Theorem 13. *Algorithm* **??** *is correct: given input a dual feasible vector y supported on $E_1, \dots, E_m$, it outputs a dual feasible y' supported on a flag $\mathbf{U} = (U_1, \dots U_t)$ with the same objective value as y.*

Proof. The existence of the pair V, W is equivalent to the support of y' not being totally ordered, which is equivalent to the support of y' not being a flag. So if the algorithm does terminate, the support will be a flag. We must show the algorithm terminates, and that y' maintains the objective value and feasibility.

Induction establishes that y' is always non-negative. Indeed, inside the while loop, the only danger is $y'_W - y'_V$. But y'_V is the smaller of the two by construction. So $y' \geq 0$ is maintained.

Let y'' denote the value of y' after another pass through the while loop. We examine the effect of the iteration on Eq. **??**. In the case $V \cap W \neq \{0\}$, the new value $C_i(y'')$ is

$$C_i(y') - y'_V \left[\text{rank}(\phi_i(W)) - \text{rank}(\phi_i(V \cap W)) - \text{rank}(\phi_i(V + W)) + \text{rank}(\phi_i(V))\right]$$

The bracketed quantity is non-negative by Lemma **??**, meaning Eq. **??** still holds. If $V \cap W = \{0\}$, then the new value $C_i(y'')$ is

$$C_i(y') - y'_V \left[\text{rank}(\phi_i(W)) - \text{rank}(\phi_i(V + W)) + \text{rank}(\phi_i(V))\right]$$

but using $\text{rank}(\phi_i(V \cap W)) = 0$ this can be written again as

$$C_i(y') - y'_V \left[\text{rank}(\phi_i(W)) - \text{rank}(\phi_i(V \cap W)) - \text{rank}(\phi_i(V + W)) + \text{rank}(\phi_i(V))\right]$$

Consequently Lemma **??** applies again. Similarly, the objective value is preserved: in the case of $W \cap V \neq \{0\}$, the new $\text{val}(y'')$ is

$$\text{val}(y') - y'_V \left[\text{rank}(W) - \text{rank}(V \cap W) - \text{rank}(V + W) + \text{rank}(V)\right]$$

with the bracketed quantity being 0 by Lemma **??**. As before, the same follows in the case $V \cap W = \{0\}$ by noting $\text{rank}(V \cap W) = 0$.

It remains to establish that the algorithm will terminate. At first glance, it appears that the y' might cycle in the algorithm. However, each iteration is increasing the dual variables on $V + W$ and $V \cap W$, so the dual vector seems to be shifting towards the high and low rank subgroups.

To capture this intuition, we define a simple measure of extremeness on dual vectors. Recall all groups reside in $\mathbb{Z}^d$. To a dual vector y we assign a list $w(y)$ of length d. To do this, set

$$w(y)_i = \sum_{U \in \text{ support}(y),\ \text{rank}(U)=i} y_U$$

For example, if y is supported on $\langle e_1, e_2\rangle, \langle e_1\rangle, \langle e_2\rangle$ with values $1, .5, 2$, and $d = 3$, then $w(y) = (2.5, 1, 0)$. We say y' is *more extreme* than y'' if $w(y')$ is reverse lexicographically more than $w(y'')$. Every iteration of the while loop makes y' more extreme; indeed, the value y_{V+W} increases and $V + W$ is of strictly larger rank than V or W.

Now we show that $w(y')$ can take on only finitely many values, completing the proof. Observe that $1^T y'$ stays the same or decreases each iteration, so coordinates of $w(y)$ are bound by $\vec{1}^T y'$. Also, the values produced by the algorithm come from performing only addition and subtraction operations on the the coordinates of y, which are rational. Consequently coordinates of $w(y')$ lie in the finite set

$$\text{span}_{\mathbb{Z}}(y_{E_1}, \ldots, y_{E_m}) \cap [0, 1^T]$$

whose size may be conservatively bounded by the least common denominator of all the y_{E_i}. □

Parallelepiped Tilings from Flags

The main theorem of the previous section allows us to transform an optimal dual vector into another optimal dual vector supported on a flag. Now we convert the flag subgroups into independent subgroups in the natural manner in order to produce a tiling shape.

Definition 14 (Flag Parallelepiped)**.** *Suppose y is supported on flag $\mathbf{U}$. Let $\mathbf{Y}$ be a sequence of independent subgroups such that $Y_1 + \cdots + Y_i = U_i$. Define the dual vector y' supported on $\mathbf{Y}$ by*

$$y'_{Y_i} = y_{U_i} + \cdots + y_{U_t}$$

*Form a product parallelepiped S of Def. **??** from y'. We will call S the flag parallelepiped of y, and y' its associated dual vector.*

Here let's briefly summarize the progress so far, and what we still need to accomplish. Provided we formulated the HBL Primal LP and solved its dual, we found a feasible y with objective value $h_*(\alpha)$. From Thm **??**, we can assume y is supported on some flag. Next apply the flag parallelepiped construction of Def. **??** to y' to create a tile S and its associated y'. Proposition **??** implies that S includes $\Theta(M^{h_*(\alpha)})$ lattice points. However, y' might no longer satisfy Eq. **??**, so Proposition **??** does not show that $|\phi_i(S)| = O(M^{\alpha_i})$. We need to expand the analysis of Lemma **??** to the case of parallelepipeds instead of cubes.

Lemma 15. *Consider independent subgroups $Y_1, \ldots, Y_t$ with corresponding dual values y_{Y_i}. Construct the product parallelepiped as in Def.* **??** *from these independent spaces and dual values. Assume the subgroups are ordered so that y_{Y_i} monotonically decreases with i. In keeping with Def.* **??** *have $U_i := Y_1 + \cdots + Y_i$ for $i = 1, \ldots, t$, and for convenience $U_0 := \{0\}$. For any linear map L, set*

$$d_i := \mathit{rank}(L(U_i)) - \mathit{rank}(L(U_{i-1}))$$

Then we have the bound

$$|L(S)| = O\left(\prod_{i=1}^{t} M^{y_{Y_i} \cdot d_i}\right)$$

In particular, this holds for L chosen to be any of the ϕ_j.

Before beginning the proof, we remark on the significance. The weaker bound used in Proposition **??** was

$$|L(S)| \leq \prod_{i=1}^{t} |L(S_{Y_i})| = O\left(\prod_{i=1}^{t} M^{y_{Y_i} \cdot a_i}\right)$$

with $a_i := \mathrm{rank}(L(Y_i))$. From independence of the subgroups Y_j, it is immediate that $d_i \leq a_i$. For example, when L is the identity, $d_i = a_i$. However, when $L(Y_i)$ is not independent of $L(U_{i-1})$, it is always the case that $d_i < a_i$.

Proof. The goal is to propose a rectangular prism T containing $L(S)$. Of the defining edges, d_i of them will be length $O(M^{y_{Y_i}})$. This would prove the needed bound.

Intuitively, we just need to make the d_1 dimensions coming from $L(Y_1)$ have the largest size $O(M^{y_{Y_1}})$, and the next d_2 dimensions coming from Y_2 will need to have length $O(M^{y_{Y_2}})$ and so forth. To formally show this by constructing T, it is convenient to interpret all subgroups instead as subspaces of $\mathbb{Q}^d$ with the standard Euclidean inner product and its induced norm. Now apply a Gram-Schmidt orthogonalization procedure to the sequence $L(Y_1), L(Y_2), \ldots, L(Y_t)$. This yields subspaces $E_1, \ldots E_t$ satisfying

$$E_1 = L(Y_1),\ E_1 + \cdots + E_i = L(Y_1) + \cdots + L(Y_i),\ E_i \perp E_j \text{ for } i \neq j$$

Take T to be the Minkowski sum formed by cubes T_i of side length $O(M^{y_{Y_i}})$ growing in the spaces E_i. It is readily observed that $|T| = O\left(\prod_{i=1}^{t} M^{y_{Y_i} \cdot d_i}\right)$. Denote by P_{E_i} the orthogonal projection onto E_i. If we can show $P_{E_i}(L(S)) \subset T_i$ for each i, then $L(S) \subset T$. The proof would then be complete.

Select an arbitrary $x \in S$. That is,

$$L(x) = L(x_{Y_1}) + \cdots + L(x_{Y_t})$$

where $x_{Y_j} \in S_{Y_j}$. Observe that $L(x_{Y_i}) \in \ker(P_{E_j})$ for $i < j$. Also $\|L(x_{Y_j})\|_2 = O(M^{y_{Y_j}})$. This implies

$$P_{E_i}(L(x)) = P_{E_i}(L(x_{Y_i})) + \cdots + P_{E_i}(L(x_{Y_t}))$$

and therefore

$$\|P_{E_i}(L(x))\|_2 = O(M^{y_{Y_i}}) + \cdots + O(M^{y_{Y_t}}) = O(M^{y_{Y_i}})$$

As T is permitted to be $O(M^{y_{Y_i}})$ in $E_i = \mathrm{im}(P_i)$, we conclude that $P_i(S) \subset T$ if the hidden constant for T large enough. □

This readily applies to the construction of Def. **??**:

Theorem 16. *From an optimal dual feasible vector y supported on a flag $\mathbf{U}$, form a flag parallelepiped S. Then $|\phi_j(S)| = O(M^{\alpha_j})$ for each ϕ_j in the HBL problem, and $|S| = \Theta(M^{h_*(\alpha)})$.*

Proof. Let y' be the associated dual vector of S with independent subgroups $Y_1, \ldots, Y_t$ as described in Def. **??**. As y has positive entries, y'_{Y_i} are monotonically decreasing. Consequently, Lemma **??** applies. It implies that

$$\begin{aligned} |\phi_j(S)| = O(\prod_{i=1}^{t} M^{y'_{Y_i} \cdot d_i}) &= O(\prod_{i=1}^{t} M^{(y_{Y_i} + \cdots + y_{Y_t}) \cdot d_i}) \\ &= O(M^{\sum_i y_{Y_i} \cdot (d_1 + \cdots + d_i)}) = O(M^{y^T \operatorname{rank}(\phi_j(\mathbf{U}))}) = O(M^{\alpha_j}) \end{aligned}$$

That $|S| = M^{\Theta(h_*(\alpha))}$ follows from Proposition **??**. □

This is the major theoretical result. Combined with earlier results, it notably establishes Thm **??**.

1.4 Conclusion

In this paper, we showed how to a maximize the volume of a shape, while requiring several of its linear images to satisfy a volume bound. To solve this problem, we extracted important geometric information from the HBL inequalities. This construction was inspired by the blocking strategies used inside of the nested for-loops of matrix and tensor computation algorithms. We believe our result places those strategies into a more general context, and would be of interest to future algorithm designers. Besides this, the construction could be of mathematical use, as already found in [53].

There are several questions of a theoretical nature that could be interesting to explore. When we construct tilings with size $|S| = O(M^{h_*(\alpha)})$, the hidden constant is complicated and possibly large. It is unclear what the size of this hidden constant is, or how close it is to optimal. We provide an exact formulation in Appendix **??**, and wonder if or when the sharpened bound could likewise be attained. Also, the complexity of solving for h_* and for the volume-optimal shape is unknown and connected to other computational complexity questions through its connection to the Brascamp-Liebb inequalities.

Chapter 2

A Riemannian Corollary of Helly's Theorem [1]

2.1 Overview

Introduction

The extrema of functions are of fundamental importance in mathematics and its applications. Much of numerical optimization studies this topic. Most of the theory focuses on convex functions, as it has proven hard to find other classes that are both useful and tractable. The motivation for this paper comes from the desire to expand the boundaries of this class of tractable functions.

Rigorous study of convergence rates was initiated in [67] for first order methods for convex functions on Hadamard manifolds. That is, they studied gradient descent methods for simply connected manifolds of non-positive sectional curvature. Such manifolds are diffeomorphic to $\mathbb{R}^n$ and exhibit natural convex functions. In a sense, they give new classes of functions for which optimization is tractable.

Still, as far as the author is aware, all known algorithms for general convex optimization on Riemannian manifolds have iteration complexity depending polynomially on ϵ^{-1}, where ϵ is the desired accuracy. To achieve better convergence rates, further conditions are added such as strong convexity, dominated gradients, or recently robust second-order [67], [66], [68]. One major unresolved question for Hadamard manifolds like SL_n/SO_n is, does convexity enable algorithms whose time complexity depends polynomially on $\log(\epsilon^{-1})$? The answer to this is still open, and we make progress by establishing that only a polynomial number of gradient oracle accesses are required.

For Euclidean optimization, cutting plane methods are the standard, general approach to

[1] Preprint [57],submitted for publication to Journal of Convex Analysis.

get $\log(\epsilon^{-1})$ complexity. It is well known that the minimum of a convex function lies in the halfspace opposite the subgradient direction. Cutting plane methods use this fact to reduce the feasible set. One feature of $\mathbb{R}^n$ that enables this approach to succeed is the existence of what are commonly termed centerpoints. A precise definition of centerpoint is given in Definition **??**. Roughly speaking, if c is to be a centerpoint for a set S, then no halfspace based at c should contain too large (or small) a fraction of the volume of S. Ellipsoid methods explicitly maintain a radially symmetric set, so the center of the current ellipse provides a perfect centerpoint. Thus a subgradient at the center of an ellipse allows one to eliminate half of the ellipse. For more general subsets $S \subset \mathbb{R}^n$ than ellipses, Grünbaum's result in [35] shows the existence of a centerpoint c for which any halfspace based at c contains at most a $\frac{n}{n+1}$-fraction of the mass of S.

As generalizing this result of Grünbaum is the main goal of this work and we fundamentally build from it in the proof, we summarize his result as Theorem **??**. Note the statement is for a general probability measure, a fact we will make use of by applying it to the Riemannian volume measure.

Theorem 17. *A $\frac{1}{n+1}$-centerpoint c exists for any probability measure on $\mathbb{R}^n$, endowed with the usual Borel σ-algebra. Here c being a $\frac{1}{n+1}$-centerpoint means any halfspace based at c contains at least a $\frac{1}{n+1}$-fraction of the mass of the probability distribution.*

We replicate this result in the more general setting of Hadamard manifolds in the hope that others find the result encouraging, useful, or intrinsically interesting. Examples illustrating the applicability of our results to interesting problems can be found below Definition **??**. Though we present the results here to provide motivation, the definitions in the Section **??** make the statements precise. The main result is from Section **??**,

Theorem 18. *Suppose μ is a probability distribution on a Hadamard manifold M of dimension n, and μ is absolutely continuous with respect to the Riemannian volume measure vol_g. Then there exists a $\frac{1}{n+1}$-centerpoint c for the measure μ. If we assume the support of μ is contained in a (geodesic) convex set S, then $c \in S$. Moreover, even for the uniform measure on convex and compact S, this value $\frac{1}{n+1}$ cannot in general be improved.*

We noted the sharpness of $\frac{1}{n+1}$ in the above theorem because this contrasts with the guaranteed existence of $\frac{1}{e}$-centerpoints for the uniform measure on convex subsets of Euclidean space [35].

The theorem leads to a bound on the number of subgradient oracle calls needed to optimize a function.

Theorem 19. *Suppose a subset S of Hadamard manifold M of dimension n is (geodesic) convex, and that $f : S \to \mathbb{R}$ is a (geodesic) convex L-Lipschitz function. Additionally assume the minimum of f, denote by x_*, is in the ϵ-interior of S, meaning that the open ball centered*

at x_* of Riemannian radius ϵ is contained within S. Then it is possible to find a point $x \in S$ such that $f(x) - f(x_*) \leq \epsilon$ using $O(n^2 \log(n\, L\, vol_g(S)\epsilon^{-1}))$ subgradient oracle calls, where $vol_g(S)$ denotes the Riemannian volume of S.

Definitions and Notation

In this section we give the definitions needed to frame the problem and results. Only basic notions of Riemannian geometry are needed in this paper; these are surveyed in Appendix **??**, with the present section mainly providing non-standard or less common definitions.

For the remainder of this paper we study triples (M, g, μ), where M is an n-dimensional, simply-connected manifold equipped with a complete Riemannian metric g of non-positive sectional curvature. Such Riemannian manfolds (M, g) are called Hadamard manifolds. For each point $x \in M$ we denote by $\langle \cdot, \cdot \rangle_x$ the inner product, defined by g, on the tangent space T_xM. The metric g also induces the Riemannian volume measure, denoted by vol_g. In addition to this, we consider a probability measure μ, which we assume to be absolutely continuous with respect to vol_g. For the motivating application μ is taken to be the Riemannian volume measure restricted to a subset $S \subset M$, i.e. $\mu = \frac{\mathbb{1}_S}{\mathrm{vol}_g(S)} \cdot \mathrm{vol}_g$. The metric g further induces a metric between points on the manifold, allowing us to define the notion of geodesics, which are locally length minimizing paths. For $x \in M$ we denote by $\exp_x : T_xM \to M$ the exponential map based at x, which maps tangent vectors to geodesics passing through x. As explained in the Riemannian geometry overview in Appendix **??**, the exponential map is a diffeomorphism when (M, g) is a Hadamard manifold.

We proceed with definitions related to convexity,

Definition 20. *We say that $S \subset M$ is **convex** if points $x, y \in S$ are joined by a unique length-minimizing geodesic contained in S.*

Definition 21. *A function $f : M \to \mathbb{R}$ is **convex** on its domain if its restrictions to geodesics are convex in t. That is, $f(\exp_x(tv)) : \mathbb{R} \to \mathbb{R}$ is a convex function in t.*

*For such a convex function f, a tangent vector $w \in T_xM$ is said to be a **subgradient** at x if for any $v \in T_xM$,*

$$f(\exp_x(tv)) \geq f(x) + t\langle w, v \rangle_x \,.$$

*The set of subgradients at x is known as the **subdifferential** at x, and is denoted by ∂f_x.*

We note Theorem 4.5 in [63] proves that convex functions have a non-empty subdifferential at all points. Accordingly, when the convex function f is discussed, we will assume a subgradient oracle that for any x outputs some $w \in \partial f_x$. As explained in Appendix **??**, the gradient of a differentiable convex function is a subgradient. Therefore the subradient oracle for such a function can be explicit.

Let us present two examples of convex functions for motivation. The first is general, the second specific.

- The distance to a convex subset (Definition **??**) of a Hadamard manifold is convex [8]. Thus finding the point minimizing the mean distance or mean squared-distance to a set of points is a convex optimization problem.

- Identify SL_n/SO_n with the set of positive-definite matrices of determinant 1 [8]. Then for arbitrary $B_i \in GL_n$, the function

$$\log\det\left(\sum_{i=1}^{m} B_i^T X B_i\right)$$

  defined on SL_n/SO_n is convex [60]. Minimizing such a function can be used to find the optimal Brascamp-Lieb constant [10]. Up to scaling, all symmetric spaces of non-compact type (examples of Hadamard manifolds) embed as totally geodesic submanifolds of these spaces. Therefore restrictions of this function to such submanifolds give many more examples. Minimizing this function has figured prominently in recent theoretical computer science research, notably in [68]. Their work succeeds in developing an optimizaton procedure depending polynomially on $\log(\epsilon^{-1})$ for such functions, but the approach relies on special properties of this family of functions.

With these examples in mind, let us return for two more important definitions,

Definition 22. *An open* ***halfspace*** *based at $x \in M$ is formed by applying $\exp_x(\cdot)$ to a halfspace of T_xM. We denote halfspaces by*

$$H_x(v) := \{\exp_x(w) \,|\, w \in T_xM,\, \langle w, v\rangle_x < 0\}\,,$$

for a given $v \in T_xM$.

Although such halfspaces are not convex sets in the general setting of Hadamard manifolds, they are naturally produced by cutting planes for convex functions. This notion of cutting plane is justified by the following lemma,

Lemma 23. *Consider a convex function $f : S \to \mathbb{R}$, where S is a convex subset of Hadamard manifold M. Then for any $x \in S$ and any subgradient $v \in \partial f_x$, the minimum of f within S is either attained at x or lies within $H_x(v) \cap S$. Moreover, if $y \in S \backslash H_x(v)$, then $f(y) \geq f(x)$.*

Proof. If $y \notin H_x(v)$, the corresponding $v' = \exp_x^{-1}(y)$ satisfies

$$\langle v', v\rangle_x \geq 0\,,$$

and we have

$$f(y) \geq f(c) + \langle v, v'\rangle_x \geq f(x).$$

□

Cutting plane methods need to find a point for which no halfspace based at that point has too much of the feasible set's volume. This can be captured through the notion of a centerpoint. Our definition of centerpoint technically could be applied to any probability measure on any space with a notion of halfspace. However, in proving Theorem **??**, we further restrict to the halfspaces we defined for Hadamard manifolds, and require the probability measure to be absolutely continuous with respect to the Riemannian volume measure vol_g.

Definition 24. *A β-**centerpoint** of the probability measure μ on a Hadamard manifold M is a point c such that*

$$\mu(H_c(v)) \leq 1 - \beta\,,$$

for all $v \in T_cM$.

Theorem **??** claimed that even for the uniform measure on convex subsets of M, a $\frac{1}{n+1}$-centerpoint is the best we can guarantee. We included this comment both to contrast with $\mathbb{R}^n$, as well as to include a concrete illustration. It is not difficult to present such an example through studying $\mathbb{H}^n$, the model space of constant -1 sectional curvature.

Towards this end, let us briefly state the important features of the Klein model of $\mathbb{H}^n$ that we require. We identify $\mathbb{H}^n$ with the open Euclidean unit ball $B(1) \subset \mathbb{R}^n$, which we take to be centered at the origin. This set is equipped with the metric $g = \frac{dx_1^2+\cdots+dx_n^2}{1-x_1^2-\cdots-x_n^2}$. This leads to a volume form of

$$\text{vol}_g = \frac{1}{(1-x_1^2-\cdots-x_n^2)^{\frac{n+1}{2}}} dx^1 \wedge \cdots \wedge dx^n\,.$$

Critically for our exposition, this model of hyperbolic space has its geodesics appear as Euclidean lines; thus Riemannian halfspaces appear as halfspaces intersected with $B(1)$. Other work such as [9] has found this useful in studying convex objects in $\mathbb{H}^n$.

For the construction demonstrating that $\frac{1}{n+1}$ cannot in general be improved in Theorem **??**, the idea is simply that convex polyhedra in $\mathbb{H}^n$ have their volume concentrated towards the vertices. Let $T(1)$ be the closed, regular n-simplex inscribed in $B(1)$. It can be checked that $T(1)$ has finite volume; such objects are called ideal polyhedra and have been studied extensively. By symmetry, we will see that the origin $\vec{0}$ is the optimal centerpoint for $T(1)$. However, note that a hyperplane through $\vec{0}$ parallel to any of the faces will contain exactly 1 of the $n+1$ vertices. An application of the Gauss-Bonnet theorem can be used to check that in the case of $\mathbb{H}^2$, the area of the halfspace containing a single vertex is $\frac{\pi}{3}$. The halfspace containing the other two vertices is of area $\frac{2\pi}{3}$. The author found this calculation to be easier to carry out in a conformal model such as the upper half-plane model, and conjectures the result to hold as well for higher dimensions.

For our purpose, it is more direct to modify the example slightly,

Proposition 25. *Let $T(1+\delta)$ be a closed, regular n-simplex inscribed in $B(1+\delta)$. We take $\delta > 0$ small enough so that $B(1)$ is not inscribed in $T(1+\delta)$. Also define $S_\epsilon = T(1+\delta) \cap B(1-\epsilon)$, which we observe to be convex and compact. Then the optimal centerpoint of S_ϵ approaches being a $\frac{1}{n+1}$-centerpoint as $\epsilon \to 0$.*

Proof. It is clear that $S = T(1+\delta) \cap B(1)$ has unbound Riemannian volume, because each of the $n+1$ vertices of $T(1+\delta)$ lies outside of $B(1)$. However, $S_\epsilon = T(1+\delta) \cap B(1-\epsilon)$ has finite volume for any $\epsilon > 0$, and is convex and compact. Define the set of probability measures consisting of the uniform measure restricted to S_ϵ, i.e. $\mu_\epsilon = \frac{\mathbb{1}_{S_\epsilon}}{\mathrm{vol}_g(S_\epsilon)} \cdot \mathrm{vol}_g$.

By symmetry, the origin $\vec{0}$ is the optimal centerpoint for each μ_ϵ. In more detail, the optimal centerpoint for S_ϵ is the solution to minimizing $G(y) := \sup_{\hat{v} \in S^{n-1}} \mu_\epsilon(H_y(\hat{v})$. Viewing $G(y)$ as a function on $S_\epsilon \subset \mathbb{R}^n$, Lemma **??** establishes that it is quasi-convex. Again, this is possible because in this model of $\mathbb{H}^n$, geodesics appear as Euclidean straight lines. One characterization of quasi-convex is $f(tx + (1-t)y) \leq \max(f(x), f(y))$ for all $0 \leq t \leq 1$. Suppose $x_* \neq \vec{0}$ is the optimal centerpoint. Because S_ϵ exhibits tetrahedral symmetry, we see there are $n+1$ optimal centerpoints, and their convex hull includes $\vec{0}$. Using quasi-convexity, any points in the convex hull of these $n+1$ optimal centerpoints must also be optimal. This proves $\vec{0}$ is the optimal centerpoint, using the symmetry of the set S_ϵ and quasi-convexity of the centerpoint function $G(y)$.

Now consider a hyperplane through $\vec{0}$ parallel to one of the faces of S_ϵ, and denote by H^+ the resulting halfspace containing only one of the vertices of $T(1+\delta)$. Because the volumes close to the vertices of $T(1+\delta)$ diverge at equal rates, it follows that as $\epsilon \to 0$,

$$\mu_\epsilon(H^+) = \mathrm{vol}_g(H^+ \cap S_\epsilon)/\mathrm{vol}_g(S_\epsilon) \to \frac{1}{n+1}.$$

□

This concludes our proof of the sharpness of Theorem **??**.

Overview and Conclusion

The remainder of this paper is organized as follows:

- Section **??** analyzes the existence of centerpoints on Hadamard manifolds.
- Section **??** presents the brief application of the above to upper bound subgradient oracle complexity.
- Appendix **??** recalls the relevant notions of Riemannian geometry and provides references.

To be clear, the problem of developing an efficient optimization procedure is far from resolved. However, our results show that there is not an information theoretic obstacle to developing cutting plane methods for Hadamard manifolds.

We hope our main result is of interest and encourages others to study centerpoints in the manifold setting. Targeting optimization procedures, we believe focusing on the spaces SL_n/SO_n would be of greatest interest, both for theory and applications. Computing a centerpoint from a discrete point set would be a notable advancement. It would also be useful to be able to sample from the Riemannian volume restricted to a convex subset.

2.2 Existence of Centerpoints

One might wonder if the centroid of a convex set of a manifold is an adequate centerpoint. Here centroid refers to the center of mass, the point minimizing the average squared-distance. After all, the centroid of a convex subset of $\mathbb{R}^n$ is an approximately optimal centerpoint [35]. However, this is tied closely to the fact that cross-sectional areas of a convex set in $\mathbb{R}^n$ follow a log-concave probability distribution - a consequence of the Brunn-Minkowski inequality. On the otherhand, for a manifold with negative sectional curvature, the distribution of cross-sectional areas is not necessarily even unimodal. This reflects the fact that manifold versions of the Brunn-Minkowski inequality use curvature lower bounds as parameters, and are qualitatively different in negative curvature compared to $\mathbb{R}^n$ [22]. Helly's Theorem is somewhat the opposite, as it holds in situations in which the distance function is convex. Moreover, as cited in Appendix **??**, Hadamard manifolds have convex distance functions. One can find in [45] and [40] proofs that amount to:

Theorem 26. *Let M be an n-dimensional Riemannian manifold of non-positive sectional curvature. Suppose we are given a convex compact set C and a family $\{C_\alpha\} \subset C$ of closed convex sets. Then if for an arbitrary selection of $n+1$ sets $C_{\alpha_1} \cap \cdots \cap C_{\alpha_{n+1}} \neq \emptyset$, it follows that $\cap_\alpha C_\alpha \neq \emptyset$*

The paper [40] actually proves this result for Cat(0) geodesic spaces.

That the halfspace notion of Definition **??** is not typically convex limits the applicability of this generalization of Helly's Theorem. The remainder of this section proves a result that could be a considered a Riemannian variant of the well known corollary of Helly's theorem cited as Theorem **??**. To generalize that result, we rely on a few simple regularity properties of sets of Euclidean centerpoints, which we now collect. In the following lemma, the halfspaces are Euclidean halfspaces, and D is the Hausdorff distance. That is,

$$D(A,B) := \max\{\sup_{b\in B} \inf_{a\in A} |a-b|\,,\, \sup_{a\in A} \inf_{b\in B} |a-b|\},$$

where $|\cdot|$ denotes the Euclidean norm. In other words, $D(A,B)$ is the minimal value ϵ so that A is contained in the ϵ-fattened version of B and vice versa. Also recall the total

variation distance between probability distributions is

$$\sup_{A\in\mathcal{F}} |\mu_1(A) - \mu_2(A)|,$$

where $|\cdot|$ denotes the absolute value, as there will be no confusion. Here A can be any measurable set, the collection of which is labeled $\mathcal{F}$.

Lemma 27. *Let $\{\mu_x(\cdot)\}$ be a family of probability measures on $\mathbb{R}^n$ that share a compact support Y. Assume the measures are indexed by members x of a compact metric space X with metric d, and the measures $\mu_x(\cdot)$ vary continuously with respect to total variation distance. Define the Euclidean centrality function $G : X \times Y \to \mathbb{R}$ by*

$$G(x, y) := \sup_{\hat{v}\in S^{n-1}} \mu_x(H_y(\hat{v})),$$

in order to measure how good of a centerpoint y is for distribution μ_x. Then G is continuous under the product topology and $G(x, \cdot)$ is a quasi-convex function for a fixed x. Fixing an arbitrary $\alpha > 0$, also define the sets

$$U_x := \left\{ y \in \mathbb{R}^n \,|\, G(x, y) \in \left(0, 1 - \frac{1}{n+1} + \alpha\right] \right\}$$

in order to explicitly propose the set of centerpoints for distribution μ_x. For any x, these sets have a non-empty interior. Moreover fix $x \in X$ and suppose $\operatorname{supp}(\mu_x)$ is a connected set. Then $x_i \to x$, $D(U_{x_i}, U_x) \to 0$.

Proof. Each $\{y \,|\, \mu_x(H_y(\hat{v})) < a\}$ is a halfspace. Indeed, there is a unique halfspace with normal $\hat{v}$ of mass a, and the previous set is precisely the points contained in this halfspace. Therefore the intersection over all $\hat{v}$ is a convex set. This shows that preimages under $G(x, \cdot)$ of sets $(-\infty, a)$ are convex, which is the definition of quasi-convex.

As we are using the product topology, the domain of G, which we denote by $K = X \times Y$, is compact. Because $g(x, y, \hat{v}) : (x, y, \hat{v}) \mapsto \mu(H_x(\hat{v}))$ is continuous and K is compact, g is uniformly continuous on $K \times S^{n-1}$. Thus given $\epsilon > 0$, one can choose δ so that when $d(x, x') < \delta$ and $|y - y'| < \delta$, then

$$|g(x, y, \hat{v}_0) - g(x', y', \hat{v}_0)| < \epsilon$$

holds for any $\hat{v}_0$. By compactness in the last argument, we may let $G(x, y) = g(x, y, \hat{v}_{x,y})$. Therefore

$$G(x, y) - G(x', y') = g(x, y, \hat{v}_{x,y}) - g(x', y', \hat{v}_{x',y'}) > g(x, y, \hat{v}_{x,y}) - (g(x, y, \hat{v}_{x',y'}) + \epsilon) \geq -\epsilon$$

holds. Switching roles gives the reverse inequality, $G(x, y) - G(x', y') < \epsilon$, which proves continuity.

Recall the Hausdorff distance is the maximum distance it might require to travel from a point in one of the sets to the other set. We argue by contradiction that $U_{(\cdot)}$ converges

to U_x in the Hausdorff distance metric. Assume the contrary, then either (i) there exists a sequence of points $y_{n_i} \in U_{x_{n_i}}$ such that y_{n_i} are bounded away from U_x or (ii) there exists a sequence of points $y_{n_i} \in U_x$ bounded away from $U_{x_{n_i}}$.

In situation (i), compactness implies an accumulation point y for the sequence y_{n_i}. However, continuity of G requires $y \in U_x$, because $x_{n_i} \to x$ and each $G(x_{n_i}, y_{n_i}) \in (0, 1 - \frac{1}{n+1} + \alpha]$. This contradicts the premise that y_{n_i} are bound away from U_x. In particular, this shows that the maximum distance from a point in U_{x_i} to the set U_x is going to 0.

In situation (ii), again by compactness there is an accumulation point $y \in U_x$ that is bounded away from infinitely many of the $U_{x_{n_i}}$. Because we have assumed $\alpha > 0$, Theorem **??** and the continuity of $G(x, \cdot)$ imply U_x has an interior. Note that in proving continuity of $G(x, \cdot)$, we used the absolute continuity of μ_x with respect to Lebesgue measure.

As a first subcase of (ii), we assume y is in the interior of U_x. We show by contradiction that $G(x, y) < 1 - \frac{1}{n+1} + \alpha$. Supposing to the contrary, there would be $\hat{v}$ such that $G(x, y) = 1 - \frac{1}{n+1} + \alpha = \mu_x(H_y(\hat{v}))$, and we may select $p \in H_y(-\hat{v}) \cap U_x$ because y is in the interior of U_x. As $p \in U_x$ and $\mu_x(H_y(\hat{v})) = 1 - \frac{1}{n+1}$, it must be the case that $\mu_x\left(H_y(-\hat{v}) \cap H_p(\hat{v})\right) = 0$, hence

$$H_y(-\hat{v}) \cap H_p(\hat{v}) \cap \operatorname{supp}(\mu_x) = \emptyset,$$

and this implies the boundary of $H_{\frac{y+p}{2}}(\hat{v})$ is disjoint from $\operatorname{supp}(\mu_x)$. Then $\operatorname{supp}(\mu_x) \cap H_{\frac{y+p}{2}}(-\hat{v})$ and $\operatorname{supp}(\mu_x) \cap H_{\frac{y+p}{2}}(\hat{v})$ are non-empty sets whose union is $\operatorname{supp}(\mu_x)$. As these sets are open in the induced topology on $\operatorname{supp}(\mu_x)$, this contradicts the assumption that $\operatorname{supp}(\mu_x)$ is connected. Therefore we have shown by contradiction that $G(x, y) < 1 - \frac{1}{n+1} + \alpha$. We immediately conclude from the continuity of G that $y \in U_{x_i}$ for large enough i.

In the event that y is not in the interior of U_x, we can still select an interior point $y' \in U_x$ that is arbitrarily close to y, because the set is open and convex. For any such y', the prior argument establishes that $y' \in U_{x_i}$ for large enough i. Therefore, we conclude that the distance between y and U_{x_i} is going to 0, contradicting our assumption. This completes the proof showing $D(U_{x_i}, U_x) \to 0$.

□

As a comment on the proof, the assumption that $\sup(\mu_x)$ is connected was essential in the final conclusion of the proof. This assumption, along with a few others like the use of α, are bootstrapped out of the eventual theorem we prove. It would be nice to eliminate the absolute continuity assumption on μ_x by using a more general convergence tool like Wasserstein distance. We necessarily lose G's continuity, but it remains lower semi-continuous. However, these topological properties alone were insufficient for proving U_x had an interior point or analogous "deep" point, which we found necessary in proving Hausdorff convergence for U_x.

We are now ready for the key step in proving the main result.

Proposition 28. *Let μ be a probability measure whose support lies within a compact set S of a Hadamard manifold M. Further assume μ is absolutely continuous with respect to the Riemannian volume measure vol_g and has connected support. Then there exists a $\frac{1}{n+1}$-centerpoint for μ.*

Before going into the proof details, here is conceptual overview of the proof. We will define a continuous function F from S to itself, and an application of Brouwer's theorem will show there is a fixed point. We design F so that the fixed point is a $(\frac{1}{n+1} - \alpha)$-centerpoint. The α is inherited from the definition of U_x in Lemma **??**, and is removed at the end of the proof. In designing F, we adopt normal coordinates at x and pull back the measure μ from M (i.e. the measure of $U \subset \mathbb{R}^n$ is $\mu(\exp_x(U))$. In these coordinates, there is a Euclidean-convex set of Euclidean centerpoints U_x provided by the previous lemma, for the pulled-back measure. We select the closest of these centerpoints to x and denote this point by u_x. Finally, $F(x)$ is then defined by projecting u_x onto S. As stated precisely in the appendix, it is the Hadamard assumption that implies a strictly convex distance function, making this projection possible.

The technical part of the proof mostly involves showing continuity of $F(x)$, as it is not hard to show that fixed points are $(\frac{1}{n+1} - \alpha)$-centerpoints. The main obstacle is to show that u_x varies continuously. To establish this, we note that the pulled back measures vary continuously with respect to total variation. Then Lemma **??** shows that the Euclidean centerpoint sets U_x, $U_{x'}$ are close in Hausdorff distance, provided x, x' are close. Combining this with convexity of the centerpoint sets, we are able to make $|u_x - u'_x|$ small.

We now provide the details.

Proof. We may WLOG assume S is a closed Riemannian ball of radius R. By parallel transport we may fix a smooth orthonormal frame $V = (\vec{e}_1, \ldots \vec{e}_n)$ on S, thereby determining normal coordinate charts at each $x \in S$ defined by

$$\psi_x : y \mapsto \exp_x(y^i \vec{e}_i(x)).$$

Note that $\psi_x(y)$ varies smoothly both in x and y. We may pull back the measure μ by ψ_x to give the measures $\mu_x(y)dy$. The absolute continuity of these pull back measures with respect to Lebesgue measure is due to μ being absolutely continuous with respect to the Riemannian volume measure. Smoothness of parallel transport ensures that the coordinate charts vary smoothly and therefore the $\mu_x(y)$ vary continuously with respect to total variation distance. Finally, since ψ_x are diffeomorphisms, we see that for each $x \in S$ the measure $\mu_x(y)$ has connected support. The set S is of radius R. Therefore in applying Lemma **??**, we may choose Y to be the closed ball of radius $2R$.

First fix $\alpha > 0$. For all $x \in S$, Lemma **??** then establishes the existence of non-empty compact convex sets $U_x \subset \mathbb{R}^n$ of Euclidean $(\frac{1}{n+1} - \alpha)$-centerpoints. There is a unique point $u_x \in U_x$ that is closest to x. However, it is not necessarily the case that u_x is inside $\psi_x^{-1}(S)$,

because $\psi_x^{-1}(S)$ is not convex with respect to the Euclidean metric. To work around this, project u_x onto S. That is,

$$F(x) := \pi(u_x) := \arg\min_{s \in S} d(s, \psi_x(u_x)) \,,$$

where by $d(\cdot, \cdot)$ we mean the Riemannian distance. The projection is well-defined and continuous by [8, Corollary 5.6]. Therefore F is well-defined. In the following we show

- If $F(x) = x$, then x is a $(\frac{1}{n+1} - \alpha)$-centerpoint contained in S.
- $F(x)$ is continuous.

Then since S is a closed ball, an application of Brouwer's fixed point theorem yields the desired result.

We first show that fixed points are centerpoints. We argue by contradiction and assume x is a fixed point of F which is not a $(\frac{1}{n+1} - \alpha)$-centerpoint. Observe that $u_x \neq \vec{0}$, because this would imply x is a $(\frac{1}{n+1} - \alpha)$-centerpoint. Further, we see that $\psi_x(H_0(-u_x)) \cap S \neq \emptyset$, since $H_{u_x}(-u_x) \subset H_0(-u_x)$ and $\mu(\psi_x(H_{u_x}(-u_x)) > \frac{1}{n+1} - \alpha > 0$ by the centerpoint property. Next choose $s \in \psi_x(H_{\vec{0}}(-u_x)) \cap S$, and consider the geodesic between x, s. This geodesic is contained in S by the assumption that S is convex. Triangle inequalities in the form of Toponogov's Theorem [16] (or convexity of the distance function as a simple alternative) show that, initially, moving from x to s along the geodesic decreases the distance to u_x. This means it is not the case that $\pi(\psi_x(u_x)) = x$.

Next we consider the continuity claim. Once we show $u_x \in \mathbb{R}^n$ varies continuously with respect to $x \in S$, then the continuity of $F(x)$ follows because, as noted in Appendix **??**, the projection is also continuous. As a first step, we remark that the pull-back probability densities $\mu_x(y)$ vary continuously with respect x, because they are defined by smoothing varying diffeomorphisms ψ_x. Moreover, as S is compact and μ is supported on S, we may assume the y are taken from a compact set. Then we may apply uniform continuity to show there is δ so that $d(x, x') < \delta$ implies $|\mu_x(y) - \mu_{x'}(y)| < \epsilon$ for any y. This establishes continuity for the family of measures $\mu_x(y)dy$, with respect to total variation distance. We can now make use of the regularity properties provided by Lemma **??**.

From the lemma's last part, by requiring $d(x, x') < \delta$ for small enough δ, one can ensure $D(U_x, U_{x'}) < \epsilon$. Let $h_x \in U_x$ be the point closest to $u_{x'}$; this ensures $|u_{x'} - h_x| < \epsilon$. It is also not difficult to see that $|u_{x'}| - |u_x| < \epsilon$. Therefore $|h_x| - |u_x| < 2\epsilon$. Critically, the Euclidean distance to the origin is strongly convex and u_x minimizes it on the Euclidean convex set U_x, which also includes h_x. Therefore, qualitatively, since $|h_x|$ and $|u_x|$ are similar in value, we know that $|h_x - u_x|$ is small. Making this quantitative through the Euclidean law of cosines,

$$|u_x - h_x|^2 \leq |h_x|^2 - |u_x|^2 = (|h_x| - |u_x|)(|h_x| + |u_x|) < \epsilon R$$

where a sufficiently large R can be taken to be twice the diameter of S. We conclude

$$|u_x - u_{x'}| \leq |u_x - h_x| + |h_x - u_{x'}| < \sqrt{\epsilon R} + \epsilon\,,$$

which establishes continuity for $F(x)$. This essentially completes the proof, but recall that we have used a small α parameter to define U_x, and this resulted in our proving only the existence of $(\frac{1}{n+1} - \alpha)$-centerpoints for $\alpha > 0$. However, the continuity of the centerpoint function $\sup_{\hat{v} \in S^{n-1}} H_x(\hat{v})$ on S follows from the same argument for proving continuity of G in Lemma **??**. From this and compactness of S, may conclude the existence of $\frac{1}{n+1}$-centerpoints. □

This nearly proves the main part of Theorem **??**. The main difference is the absence of a few simplifying assumptions, namely compactness and connected support. The fact that $x \in S$ provided S is convex was also postponed. We complete the proof here.

Proof for Theorem **??**. We first remove the connected support assumption. For measures μ supported on compact S, we may WLOG assume S is a ball and therefore connected. Then we may define the probability measures $(1-\epsilon)\mu + \epsilon \frac{\mathbb{1}_S}{\mathrm{vol}_g(S)} \cdot \mathrm{vol}_g$. Proposition **??** applies to these measures. Thus it is clear that we may construct $(\frac{1}{n+1} - \epsilon)$-centerpoints for μ. Again using the continuity of the centerpoint function and compactness of S as at the end of Proposition **??**'s proof, it follows that a $\frac{1}{n+1}$-centerpoint exists for μ.

Next we extend the result by removing the assumption that S is compact. Fixing some point $x \in M$, we may define the family of compact sets $S_i = S \cap \bar{B}_g(x, i)$, where i ranges over the positive integers and $\bar{B}_g(x, i)$ denotes the closed ball of radius i around x. These sets satisfy $\lim_{i \to \infty} \mu(S_i) = 1$. Applying Proposition **??** to these S_i, we get points s_i that are at least $\left(\frac{1}{n+1} - \mu(S_i^C)\right)$-centerpoints for μ. Moreover, these s_i must all lie in some compact set $C \subset S$. Indeed, by the analog of the separating hyperplane theorem proven in Lemma **??**, any point $p \notin B_g(x, i)$ will have a halfspace $H_p(\hat{v}) \cap S_i = \emptyset$, and thus be at most a $\mu(S_i^C)$-centerpoint. As in Lemma **??**, the function $G : C \to \mathbb{R}$ defined by $G(y) := \sup_{\hat{v} \in S^{n-1}} \mu(H_y(\hat{v}))$ is continuous. Because $\lim_{i \to \infty} G(s_i) \leq 1 - \frac{1}{n+1}$, C is compact, and G is continuous, it follows that there is some $c \in C \subset S$ such that $G(c) \leq 1 - \frac{1}{n+1}$. Therefore this c is a $\frac{1}{n+1}$-centerpoint.

If we additionally assume S is convex, then the separating hyperplane theorem again gives that the centerpoint satisfies $c \in S$.

Finally, we remind the reader that Proposition **??** established the second part of the theorem, concerning sharpness. □

2.3 Upper Bound on Needed Subgradient Calls

We require one final lemma for the application to convex optimization. In this lemma, as in the past, vol_g will denote the Riemmanian volume measure and $B_g(x, r)$ the open ball

of radius r around x.

Lemma 29. *Suppose f is convex and L-Lipschitz on its convex domain $S \subset M$, where M is a Hadamard manifold. Additionally assume the minimum x_* is in the ϵ-interior of S, meaning $B_g(x_*, \epsilon) \subset S$. Now suppose we are given a sequence of cutting planes $H_{c_i}(v_i)$, $i = 1, 2, \cdots, N$ with $c_i \in S$ and $v_i \in \partial f_{c_i}$ such that the remaining feasible set $S' := S \cap_i H_{c_i}(v_i)$ satisfies the volume bound*

$$vol_g(S' := S \cap_i H_{c_i}(v_i)) < \frac{(\epsilon/L)^n}{n^n}.$$

Then one of the c_i satisfies $f(c_i) - f(x_*) \leq \epsilon$.

Proof. The main fact to be established is that $\mathrm{vol}_g(B_g(x_*, \frac{\epsilon}{L})) > \mathrm{vol}_g(S')$, as this implies that one of the complements of the halfspaces $H_{c_i}(v_i)$ must have intersected $B_g(x_*, \frac{\epsilon}{L})$. By volume comparison methods (see [16]), the volume of this geodesic ball is greater or equal to the volume of a Euclidean ball of equal radius. A reference justifying the exact version required is included in Appendix **??** as Theorem **??**. Hence we obtain

$$\mathrm{vol}_g(B_g(x_*, \frac{\epsilon}{L})) > \frac{(\epsilon/L)^n}{n^n} > \mathrm{vol}_g(S'),$$

by using $\frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)} \frac{\epsilon^n}{L^n} > \frac{1}{n^n} \frac{\epsilon^n}{L^n}$ in the first inequality.

It follows that there exists a point $x' \in B(x_*, \frac{\epsilon}{L})$ that lies in the complement of one of the halfspaces $H_{c_i}(v_i)$. From Lemma **??**, $f(c_i) \leq f(x')$. The Lipschitz bound on f then gives

$$f(c_i) - f(x_*) \leq f(x') - f(x_*) \leq L \cdot d(x', x_*) \leq \epsilon.$$

□

The proof of Theorem **??** is now a rather straightforward consequence.

Proof for Theorem **??**. Lemma **??** shows that one of the origins of the cuts is ϵ from optimal for the function f as soon as the remaining set, denoted by S', has volume $O(\frac{\epsilon^n}{n^n L^n})$.

We must only bound the number of halfspaces needed to reduce the volume of S' to this amount. Proceeding iteratively, apply Theorem **??** with μ being the Riemannian volume measure restricted to S' (i.e. $\mu = \frac{\mathbb{1}_{S'}}{\mathrm{vol}_g(S')} \cdot \mathrm{vol}_g$). As the support of μ is contained in the convex set S, Theorem **??** shows that we may choose the cut centers to be $\frac{1}{n+1}$-centerpoints $c_i \in S$ for the remaining set $S' \subset S$, so that the volume is reduced by a factor $(1 - \frac{1}{n+1})$ each cut. This means the number of iterations needed is $O(n^2 \log(nL \,\mathrm{vol}_g(S)\epsilon^{-1}))$. □

Chapter 3

A Generalized Randomized Rank-Revealing Factorization[1,2]

3.1 Introduction

Rank-revealing factorizations have been around for a long time, including [13], which introduced the world to **QR** with pivoting to solve least-squares problems. Since then, many other algorithms have been proposed, among which we mention [15], [54]; for a more complete list the reader can refer to [36]. While they all perform very well most of them time, the only one that stably produces a strong rank-revealing factorization (in the [36] sense, defined in the next section) in an arithmetic complexity comparable to **QR** belongs to Ming Gu and Stanley Eisenstat [36].

Recently, the idea of using randomized algorithms for rank approximation (or more generally for low-rank approximations of matrices) has received a lot of attention due to the applications in signal processing and information technology, for example [43] and [48]. For a good overview of the types of algorithms involved, see [37].

We provide here an analysis of the ("Randomized URV") factorization, or **RURV**, which will allow us to prove that it has the following three properties.

- It is strong (in the Gu-Eisenstat sense, which will be explained in Section **??**). In particular, it is almost as strong as the best existing deterministic rank-revealing factorization of Gu and Eisenstat [36];
- It is communication-optimal. It uses only **QR** and and matrix multiplication, and thus both its arithmetic complexity and its communication complexity are asymptotically the same as **QR** and matrix multiplication.

[1] Joint work with Grey Ballard, James Demmel, Ioana Dumitriu

[2] Preprint [5], submitted for publication to SIAM J. Matrix Analysis and Application.

- If the information desired is related to the invariant subspaces, it can be applied to a product of matrices and inverses of matrices *without the need to explicitly calculate any products or inverses.*

To place these three properties in context, we compare with recent trends in the randomized numerical linear algebra literature. In work focused on sketching, such as the approaches in the overview [37], it is customary to make the assumption that the rank is small, generally much smaller than the size of the matrix. In such cases, the speedups achieved by the algorithms in [37] and others like them over **QR** is significant, both in terms of arithmetic complexity and other features, like parallelization; naturally, the results can be achieved only with (arbitrarily) high probability. The downside is that this literature largely focuses on taking advantage of matrices with low numerical rank, and quickly producing low-rank approximations of such matrices. Such developments are insufficient for effective use as a numerical subroutine within the communication optimal generalized eigenvalue algorithm, which at minimum require parts of the first and third properties mentioned above.

Other recent approaches using randomization include [29] and [49]. These works recognize that QR with column pivoting tends to have good rank-revealing properties, but that column pivoting induces extra communication. They build on this realization by using randomization to guide pivot selection during the QR algorithm and introduce block pivoting strategies. However, in contrast to our work and earlier work such as [36], theoretical bounds are not provided. We therefore emphasize that our **RURV** is communication optimal while maintaining the key theoretical properties of strong rank-revealing QR-factorizations. Moreover, **RURV** is conceptually simple, as it depends only on the existence of communication-avoiding **QR** algorithms, which were popularized in [27].

A subset of the authors introduced **RURV** in [24], for the purpose of using it as a building block for a divide-and-conquer eigenvalue computation algorithm whose arithmetic complexity was shown to be the same as that of matrix multiplication. The analysis of **RURV** performed at the time was not optimal, and the authors of [24] had not realized that **RURV** has the third property listed above. This property makes **RURV** unique among rank-revealing factorizations, as far as we can tell, and it is crucial in the complexity analysis of the aforementioned divide-and-conquer algorithm for nonsymmetric eigenvalue computations [3].

The rest of the paper is structured as follows: in Section **??** we give the necessary definitions and the algorithms we will use, and Section **??** proves some necessary probability results; Section **??** deals with the analysis of **RURV**, the short Section **??** generalizes the algorithm to work for a product of matrices and inverses, and Section **??** presents some numerical experiments validating the correctness and tightness of the results of Sections **??** and **??**.

3.2 Randomized Rank-Revealing Decompositions

Let A be an $n \times n$ matrix with singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$, and assume that there is a "gap" in the singular values at level r, that is, $\sigma_r/\sigma_{r+1} \gg 1$.

Informally speaking, a decomposition of the form $A = URV$ is called *rank revealing* if the following conditions are fulfilled:

1) U and V are orthogonal/unitary and $R = \begin{bmatrix} R_{11} & R_{12} \\ & R_{22} \end{bmatrix}$ is upper triangular, with R_{11} $r \times r$ and R_{22} $(n-r) \times (n-r)$;

2) $\sigma_{min}(R_{11})$ is a "good" approximation to σ_r (at most a factor of a low-degree polynomial in n away from it),

(3) $\sigma_{max}(R_{22})$ is a "good" approximation to σ_{r+1} (at most a factor of a low-degree polynomial in n away from it);

(4) In addition, if $||R_{11}^{-1}R_{12}||_2$ is small (at most a low-degree polynomial in n), then the rank-revealing factorization is called *strong* (as per [36]).

Rank revealing decompositions are used in rank determination [61], least squares computations [14], condition estimation [11], etc., as well as in divide-and-conquer algorithms for eigenproblems. For a good survey paper, we recommend [36].

In the paper [24], the authors proposed a *randomized* rank revealing factorization algorithm **RURV**. Given a matrix A, the routine computes a decomposition $A = URV$ with the property that R is a rank-revealing matrix; the way it does it is by "scrambling" the columns of A via right multiplication by a uniformly random orthogonal (or unitary) matrix V^H (the uniform distribution over the manifold of unitary/orthogonal matrices is known as Haar). One way to obtain a Haar-distributed random matrix is to start from a matrix of independent, identically distributed normal variables of mean 0 and variance 1 (denoted, here and throughout the paper, by $N(0,1)$), and to perform the **QR** algorithm on it. The orthogonal/unitary matrix V obtained through this procedure is Haar distributed.

It is worth noting that there are in the literature other ways of obtaining Haar-distributed matrices, and some involve using fewer random bits and/or fewer arithmetic operations; we have chosen to use this one because it is simple and communication-optimal (as it only involves one **QR** operation, which can be performed optimally from a communication perspective both sequentially and in parallel [27, 6, 4]. On a practical side, we note that using less randomness would not incur any significant overall savings, as the total arithmetic cost is much higher than the cost of generating n^2 normal random variables.

Performing **QR** on the resulting matrix $AV^H =: \hat{A} = UR$ yields two matrices, U (orthogonal or unitary) and R (upper triangular), and it is immediate to check that $A = URV$.

Algorithm 3 Function $[U, R, V]$ =**RURV**(A), computes a randomized rank revealing decomposition $A = URV$, with V a Haar matrix.

1: Generate a random matrix B with i.i.d. $N(0, 1)$ entries.
2: $[V, \hat{R}]$ =**QR**(B).
3: $\hat{A} = A \cdot V^H$.
4: $[U, R]$ = **QR**$(\hat{A})$.
5: Output R, $[U, R]$, or $[U, R, V]$.

We also define the routine **RULV**, nearly identical to **RURV**, which performs the same kind of computation (and obtains a rank revealing decomposition of A), but uses **QL** instead of **QR**, and thus obtains a lower triangular matrix in the middle, rather than an upper triangular one. **red**(Note one can think of the decomposition **RULV** as being the transpose of the decomposition **RURV** performed on A^H.)

Given **RURV** and **RULV**, we now can give a method to find a randomized rank-revealing factorization for a product of matrices and inverses of matrices, *without actually computing any of the inverses or matrix products*. This is a very interesting and useful procedure in itself, and at the same time it is crucial in the analysis of a communication-optimal Divide-and-Conquer algorithm for the non-symmetric eigenvalue problem presented in [3].

Suppose we wish to stably find a randomized rank-revealing factorization $M_k = URV$ for the matrix $M_k = A_1^{m_1} \cdot A_2^{m_2} \cdot \ldots A_k^{m_k}$, where $A_1, \ldots, A_k$ are given matrices, and $m_1, \ldots, m_k \in \{-1, 1\}$, without actually computing M_k or any of the inverses.

Essentially, the method performs **RURV** or, depending on the power, **RULV**, on the last matrix of the product, and then uses a series of **QR**/**RQ** to "propagate" an orthogonal/unitary matrix to the front of the product, while computing factor matrices from which (if desired) the upper triangular R matrix can be obtained. A similar idea was explored by G.W. Stewart in [62] to perform graded **QR**; although it was suggested that such techniques can be also applied to algorithms like **URV**, no randomization was used.

The algorithm is presented in pseudocode below. For the proof of correctness, see Lemma ??.

3.3 Smallest singular value bounds

The estimates for our main theorem are based on the following result, a more general case of which can be found in [30]; in particular, the following is a consequence of Theorem 3.2 and Lemma 3.5.

Definition 30. *Let $s_{r,n}$ be a random variable denoting the smallest singular value of an $r \times r$ corner of an $n \times n$ real Haar matrix.*

Proposition 31. *The probability density function (pdf) of $s_{r,n}$, with $r < n/2$, is given by*

$$f_{r,n}(x) = c_{r,n} \frac{1}{\sqrt{x}} (1-x)^{\frac{1}{2}r(n-r)-1} {}_2F_1\left(\frac{1}{2}(n-r-1), \frac{1}{2}(r-1); \frac{1}{2}(n-1)+1; 1-x\right) ,$$

Algorithm 4 Function U =**GRURV**$(k; A_1, \ldots, A_k; m_1, \ldots, m_k)$, computes a randomized rank-revealing decomposition $UR_1^{m_1} \cdots R_i^{m_i} V = A_1^{m_1} \cdot A_2^{m_2} \cdot \cdots A_k^{m_k}$, where $m_1, \ldots, m_k \in \{-1, 1\}$.

if $m_k = 1$, **then**
  $[U, R_k, V] =$ **RURV**(A_k)
else
  $[U, L_k, V] =$ **RULV**(A_k^H)
  $R_k = L_k^H$
end if
$U_{current} = U$
for $i = k - 1$ downto 1 **do**
  if $m_i = 1$, **then**
    $[U, R_i] =$ **QR**$(A_i \cdot U_{current})$
    $U_{current} = U$
  else
    $[U, R_i] =$ **RQ**$(U_{current}^H \cdot A_i)$
    $U_{current} = U^H$
  end if
end for
return $U_{current}$, optionally $V, R_1, \ldots, R_k$

where ${}_2F_1$ *is the ordinary hypergeometric function [1], and*

$$c_{r,n} = \frac{\frac{1}{2}r(n-r)\ \Gamma\left(\frac{1}{2}(n-r+1)\right)\ \Gamma\left(\frac{1}{2}(r+1)\right)}{\Gamma\left(\frac{1}{2}\right)\ \Gamma\left(\frac{1}{2}(n+1)\right)}\ .$$

This Proposition allows us to estimate very closely the probability that $s_{r,n}$ is small. In particular, the correct scaling for the asymptotics of $s_{r,n}$ under r and/or $n \to \infty$ was proved in [31] to be $\sqrt{r(n-r)}$ (that is, $s_{r,n}\sqrt{r(n-r)} = O(1)$ almost surely), which means that the kind of upper bounds one should search for $s_{r,n}$ are of the form "$s_{r,n} \leq a/\sqrt{r(n-r)}$" for some constant a. This constant a will depend on how confident we want to be that the inequality holds; if we wish to say that the inequality fails with probability δ, then we will have a as a function of δ.

Lemma 32. *Let* $\delta > 0$, $r, (n-r) > 30$*; then the probability that* $s_{r,n} \leq \frac{\delta}{\sqrt{r(n-r)}}$ *is* $\mathbb{P}\left[s_{r,n} \leq \frac{\delta}{\sqrt{r(n-r)}}\right] \leq 2.02\delta$.

Proof. What we essentially need to do here is find an upper bound on $f_{r,n}$ which, when integrated over small intervals next to 0, yields the bound in the Lemma.

We will first upper bound the term $(1-x)^{\frac{1}{2}r(n-r)-1}$ in the expression of $f_{r,n}(x)$ by 1.

Secondly, we note that the hypergeometric function has all positive arguments, and hence from its definition, it is monotonically decreasing from 0 to 1, and so we bound it by its value at $x = 0$. As per [1, Formula 15.1.20],

$$ {}_2F_1\left(\frac{1}{2}(n-r-1), \frac{1}{2}(r-1); \frac{1}{2}(n-1)+1; 1\right) = \frac{\Gamma\left(\frac{1}{2}(n+1)\right)\Gamma\left(\frac{3}{2}\right)}{\Gamma\left(\frac{1}{2}(n-r+2)\right)\Gamma\left(\frac{1}{2}(r+2)\right)} \ , $$

and after some obvious cancellation we obtain that

$$ f_{r,n}(x) \leq \frac{1}{2} r(n-r) \cdot \frac{\Gamma\left(\frac{1}{2}(n-r+1)\right)}{\Gamma\left(\frac{1}{2}(n-r+2)\right)} \cdot \frac{\Gamma\left(\frac{1}{2}(r+1)\right)}{\Gamma\left(\frac{1}{2}(r+2)\right)} \cdot \frac{1}{\sqrt{x}}. $$

The following expansion can be derived from Stirling's formula and is given as a particular case of [1, Formula 6.1.47] (with $a = 0,\ b = 1/2$):

$$ z^{1/2}\frac{\Gamma(z)}{\Gamma(z+1/2)} = 1 + \frac{1}{8z} + \frac{1}{128z^2} + o\left(\frac{1}{z^2}\right) \ , $$

as z real and $z \to \infty$. In particular, $z > 30$ means that $z^{1/2}\frac{\Gamma(z)}{\Gamma(z+1/2)} < 1.01$.

Provided that $r, n-r > 30$, we thus have

$$ f_{r,n}(x) \leq 1.01 \cdot \sqrt{r(n-r)}\sqrt{\frac{r(n-r)}{(r+1)(n-r+1)}}\frac{1}{\sqrt{x}} \ , $$

and so

$$ f_{r,n}(x) \leq 1.01 \cdot \sqrt{r(n-r)} \cdot \frac{1}{\sqrt{x}} \ . $$

Note that this last inequality allows us to conclude the following:

$$ \begin{aligned} \mathbb{P}\left[s_{r,n} \leq \frac{\delta}{\sqrt{r(n-r)}}\right] &= \mathbb{P}\left[s_{r,n}^2 \leq \frac{\delta^2}{r(n-r)}\right] \\ &\leq 1.01 \int_0^{\frac{\delta^2}{r(n-r)}} \sqrt{r(n-r)}\frac{1}{\sqrt{t}}dt \\ &= 2.02\delta. \end{aligned} $$

□

As an immediate corollary to Lemma **??** we obtain the following result, which is what we will actually use in our calculations.

Corollary 33. *Let $\delta > 0$, $r, n-r > 30$. Then*

$$ \mathbb{P}\left[\frac{1}{s_{r,n}} \leq \frac{2.02}{\delta}\sqrt{r(n-r)}\right] \geq 1 - \delta \ . $$

3.4 Analysis for RURV

Bounding the probability of failure for RURV

It was proven in [24] that, with high probability, **RURV** computes a good rank revealing decomposition of A in the case of A real. Specifically, the quality of the rank-revealing decomposition depends on computing the asymptotics of $s_{r,n}$, the smallest singular value of an $r \times r$ submatrix of a Haar-distributed orthogonal $n \times n$ matrix. All the results of [24] can be extended verbatim to Haar-distributed unitary matrices; however, the analysis employed in [24] is not optimal. Using the bounds obtained for $s_{r,n}$ in the previous section, we can improve them here.

We will tighten the argument to obtain one of the upper bounds for $\sigma_{max}(R_{22})$. In addition, the result of [24] states only that **RURV** is, with high probability, a rank-revealing factorization. Here we strengthen these results to argue that it is actually a *strong* rank-revealing factorization (as defined in the Introduction), since with high probability $||R_{11}^{-1}R_{12}||$ will be small.

In proving Theorem **??**, we require two lemmas that we state here and prove afterwards.

Lemma 34. *Let A be an $n \times n$ matrix whose SVD is $A = P\Sigma Q^H$, and with singular values $\sigma_1, \ldots, \sigma_n$.*

*Let R be the matrix produced by the **RURV** algorithm on A,* in exact arithmetic, *so that $UR = AV^H$. Then defining $X = Q^H V^H$,*

$$\sigma_{\max}(R_{22}) \leq \sigma_{\min}(X_{11})^{-1}\sigma_{r+1} \ , \tag{3.1}$$

$$\tag{3.2}$$

where X_{11} is the upper-left $r \times r$ submatrix.

Lemma 35. *Carrying over the notation of Lemma* **??**,

$$\|R_{11}^{-1}R_{12}\|_2 \leq 3\sigma_{\min}^{-1}(X_{11}) + 6\frac{\sigma_{r+1}}{\sigma_r}\sigma_{\min}^{-3}(X_{11}) \ .$$

We are ready for the main result.

Theorem 36. *Let A be an $n \times n$ matrix with singular values $\sigma_1, \ldots, \sigma_r, \sigma_{r+1}, \ldots, \sigma_n$. Let $1 > \delta > 0$. Let R be the matrix produced by the **RURV** algorithm on A,* in exact arithmetic. *Assume that $r, n - r > 30$.*

Then with probability $1 - \delta$, the following three events occur:

$$\frac{\delta}{2.02}\frac{\sigma_r}{\sqrt{r(n-r)}} \leq \sigma_{\min}(R_{11}) \leq \sigma_r \ , \tag{3.3}$$

$$\sigma_{r+1} \leq \sigma_{\max}(R_{22}) \leq 2.02\frac{\sqrt{r(n-r)}}{\delta}\sigma_{r+1} \ , \tag{3.4}$$

$$||R_{11}^{-1}R_{12}||_2 \leq \frac{6.1\sqrt{r(n-r)}}{\delta} + \frac{\sigma_{r+1}}{\sigma_r}\frac{50\sqrt{r^3(n-r)^3}}{\delta^3}. \tag{3.5}$$

We note the upper bound in (??) *and lower bound in* (??) *always hold. Moreover, if we additionally assume* $\delta > \sqrt{2} \cdot 1.01 \cdot n \cdot \frac{\sigma_{r+1}}{\sigma_r}$*, then we can strengthen* (??) *to*

$$||R_{11}^{-1} R_{12}||_2 \leq \frac{4.04}{\delta} \cdot \sqrt{r(n-r)} + 1 \tag{3.6}$$

Remark 37. *The factor* $\sqrt{r(n-r)}$ *in the equations* (??), (??), *matches the best deterministic algorithms up to a constant. When the gap is large enough so that* $\frac{\sigma_{r+1}}{\sigma_r}$ *is* $O(1/n)$ *with some small constant, so that the additional hypothesis applies,* (??) *also matches the best deterministic algorithms up to a constant. Even when the gap is small,* (??) *shows the factorization is strong with high probability.*

Proof. To prove this theorem, we will rely on Lemma ?? and Lemma ??. For the sake of argument flow, we have moved the proofs of these lemmas to the end of the section.

There are two cases of the problem, $r \leq n/2$ and $r > n/2$. Let V be the Haar matrix used by the algorithm. From [33, Theorem 2.4-1], the singular values of $V[1:r, 1:r]$ when $r > n/2$ consist of $(2r-n)$ 1's and the singular values of $V[(r+1):n, (r+1):n]$. Thus, the case $r > n/2$ reduces to the case $r \leq n/2$.

The upper bound in inequality (??) and the lower bound in inequality (??) follow from the Cauchy interlace theorem (see [38, Theorem 7.3.9]). The lower bound in inequality (??) follows immediately from [24, Theorem 5.2] and Corollary ??. We provide proofs of the upper bounds of inequalities (??) and (??), below.

Theorem 5.2 from [24] states that

$$\sigma_{\max}(R_{22}) \leq 3\sigma_{r+1} \cdot \frac{s_{r,n}^{-4} \cdot \left(\frac{\sigma_1}{\sigma_r}\right)^3}{1 - \frac{\sigma_{r+1}^2}{s_{r,n}^2 \sigma_r^2}} \; ;$$

provided that $\sigma_{r+1} < \sigma_r s_{r,n}$. This upper bound is lax, and we tighten it here. Note that Lemma ?? establishes that

$$\sigma_{\max}(R_{22}) \leq \sigma_{r+1} \sigma_{\min}^{-1}(X_{11}) \, ,$$

where X is Haar distributed and X_{11} is its upper-left $r \times r$ submatrix. We conclude $\sigma_{\max}(R_{22}) \leq 2.02 \frac{\sqrt{r(n-r)}}{\delta} \sigma_{r+1}$ with probability $1 - \delta$, by using Corollary ??. This completes the proof of (2).

To prove (??), we use Lemma ??, which establishes that

$$\|R_{11}^{-1} R_{12}\|_2 \leq 3\sigma_{\min}^{-1}(X_{11}) + 6 \frac{\sigma_{r+1}}{\sigma_r} \sigma_{\min}^{-3}(X_{11}) \, ,$$

where again X is Haar distributed. We conclude

$$\|R_{11}^{-1} R_{12}\|_2 \leq 3 s_{r,n}^{-1} + 6 \frac{\sigma_{r+1}}{\sigma_r} s_{r,n}^{-3} \, ,$$

and apply Corollary **??** to get the result (**??**).

It remains to show the strengthened bound (**??**) on $\|R_{11}^{-1}R_{12}\|_2$ when $\delta > \sqrt{2}\cdot 1.01\cdot n\cdot\frac{\sigma_{r+1}}{\sigma_r}$. We use the following notation. Let $A = P\Sigma Q^H = P\cdot\mathrm{diag}(\Sigma_1,\Sigma_2)\cdot Q^H$ be the singular value decomposition of A, where $\Sigma_1 = \mathrm{diag}(\sigma_1,\ldots,\sigma_r)$ and $\Sigma_2 = \mathrm{diag}(\sigma_{r+1},\ldots,\sigma_n)$. Let V^H be the random unitary matrix in **RURV**, so that $A = URV$. Then $X = Q^HV^H$ has the same distribution as V^H, by virtue of the fact that V's distribution is uniform over unitary matrices.

Write

$$X = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} ,$$

where X_{11} is $r\times r$ and X_{22} is $(n-r)\times(n-r)$. Then

$$U^HP\cdot\Sigma X = R .$$

Denote $\Sigma\cdot X = [Y_1, Y_2]$ where Y_1 is an $n\times r$ matrix and Y_2 is $n\times(n-r)$. Since U^HP is unitary, it is not hard to check that

$$R_{11}^{-1}R_{12} = Y_1^+Y_2 ,$$

where Y_1^+ is the pseudoinverse of Y_1, i.e. $Y_1^+ = (Y_1^HY_1)^{-1}Y_1^H$.

There are two crucial facts that we need to check here: one is that R_{11}^{-1} actually exists, and the other is that the pseudoinverse (as defined above) is well-defined, that is, that Y_1 is full rank. We start with the second one of these facts.

The matrix Y_1 is full-rank with probability 1. This is true due to two reasons: the first one is that the first r singular values of A, ordered decreasingly on the diagonal of Σ, are strictly positive. The second one is that X is Haar distributed, and hence Lemma **??** shows that X_{11} is invertible with probability 1. It follows that Y_1^+ is well-defined.

To argue that R_{11}^{-1} exists, note that $Y_1 = P^HU[R_{11};0]$ so $\mathrm{rank}(Y_1)$=$\mathrm{rank}(R_{11})$ as P^HU is unitary. Since Y_1 is full-rank, it follows that R_{11} is invertible.

Having made sure that the equation relating $R_{11}^{-1}R_{12}$ and Y_1 is correct, we proceed to study the right hand side. From the definition of Y, we obtain that

$$Y_1^HY_1 = X_{11}^H\Sigma_1^2X_{11} + X_{21}^H\Sigma_2^2X_{21} \quad , \quad \text{and} \quad Y_1^HY_2 = X_{11}^H\Sigma_1^2X_{12} + X_{21}^H\Sigma_2^2X_{22} \quad .$$

Hence

$$R_{11}^{-1}R_{12} = \left(X_{11}^H\Sigma_1^2X_{11} + X_{21}^H\Sigma_2^2X_{21}\right)^{-1}\left(X_{11}^H\Sigma_1^2X_{12} + X_{21}^H\Sigma_2^2X_{22}\right) .$$

We split this into two terms. Let T_1 be defined as follows:

$$\begin{aligned} T_1 &:= \left(X_{11}^H\Sigma_1^2X_{11} + X_{21}^H\Sigma_2^2X_{21}\right)^{-1}X_{11}^H\Sigma_1^2X_{12} \\ &= X_{11}^{-1}\left(\Sigma_1^2 + (X_{21}X_{11}^{-1})^H\Sigma_2^2(X_{21}X_{11}^{-1})\right)^{-1}\Sigma_1^2X_{12} , \end{aligned}$$

where the last equality reflects the factoring out of X_{11}^H to the left and of X_{11} to the right inside the first parenthesis, followed by cancellation. Since X_{12} is a submatrix of a unitary matrix, $||X_{12}|| \leq 1$, and thus

$$||T_1||_2 \leq ||X_{11}^{-1}||_2 \cdot || \left(I_r + \Sigma_1^{-2}(X_{21}X_{11}^{-1})^H \Sigma_2^2 (X_{21}X_{11}^{-1})\right)^{-1} ||_2 \leq \frac{1}{s_{r,n}} \cdot \frac{1}{1 - \frac{\sigma_{r+1}^2}{s_{r,n}^2 \sigma_r^2}} \ ,$$

where the last inequality follows from the fact that for a matrix A with $||A|| < 1$, $||(I - A)^{-1}|| \leq \frac{1}{1-||A||}$. The right hand side has been obtained by applying norm inequalities and using the fact that $||X_{11}^{-1}|| = s_{r,n}^{-1}$. The assumption on δ can be rearranged to $\frac{\delta}{\sqrt{2} \cdot 1.01 \cdot n} > \frac{\sigma_{r+1}}{\sigma_r}$. Combine this with $\frac{1}{s_{r,n}} < \frac{2.02}{\delta} \cdot \sqrt{r(n-r)}$ to get

$$\frac{\sigma_{r+1}}{s_{r,n}\sigma_r} < \frac{\delta}{\sqrt{2} \cdot 1.01 \cdot n} \frac{2.02}{\delta} \cdot \sqrt{r(n-r)} = \sqrt{2} \frac{\sqrt{r(n-r)}}{n} \leq \frac{1}{\sqrt{2}} \tag{3.7}$$

We conclude that

$$\|T_1\|_2 \leq \frac{4.04}{\delta} \cdot \sqrt{r(n-r)} \tag{3.8}$$

We now apply similar reasoning to the second (remaining) term

$$T_2 := \left(X_{11}^H \Sigma_1^2 X_{11} + X_{21}^H \Sigma_2^2 X_{21}\right)^{-1} X_{21}^H \Sigma_2^2 X_{22} \ ;$$

to yield that

$$\begin{aligned} ||T_2||_2 &\leq ||X_{11}^{-1}||_2^2 \cdot || \left(I_r + \Sigma_1^{-2}(X_{21}X_{11}^{-1})^H \Sigma_2^2 (X_{21}X_{11}^{-1})\right)^{-1} ||_2 \cdot ||\Sigma_1^{-2}||_2 \cdot ||\Sigma_2^2||_2 \\ &\leq \frac{\sigma_{r+1}^2}{s_{r,n}^2 \sigma_r^2} \cdot \frac{1}{1 - \frac{\sigma_{r+1}^2}{s_{r,n}^2 \sigma_r^2}} \ , \end{aligned}$$

because $||X_{21}||$ and $||X_{22}|| \leq 1$. Finally, note that (**??**) together with the fact that the function $x^2/(1-x^2)$ is increasing on $(0, \infty)$ give $\|T_2\|_2 \leq 1$. Combining this with (**??**), the conclusion follows.

□

We return now for the proofs of the two lemmas.

Proof of Lemma **??**. Let $Y = \Sigma X = P^H U R$. We begin by introducing a few block notations naturally suggested by the singular value gap, separating the first r coordinates from the final $n - r$ coordinates:

$$X = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \ , Y = \begin{bmatrix} Y_1 & Y_2 \end{bmatrix} \ , \Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \ ,$$

Note that R is the upper-triangular factor resulting from QR-factorization of Y. From this and understanding that the QR-factorization records the Gram-Schmidt orthogonalization process, we see that R_{22} has the same singular values as $\text{Proj}_{Y_{1\perp}} Y_2$, where $Y_{1\perp}$ is any matrix whose columns are a basis for the orthogonal complement of Y_1. In particular, $\sigma_{\max}(R_{22}) = \|\text{Proj}_{Y_{1\perp}} Y_2\|_2$.

It is also clear from $Y = \Sigma X$ that if Σ contains 0 elements on the diagonal, then the corresponding rows of $\text{Proj}_{Y_{1\perp}} Y_2$ are 0. Therefore we make the assumption that there are no 0 singular values.

We next relate Y_1 to Y_2 in order to analyze this matrix. Using a common matrix identity in first equality and the orthogonality of X in the third equality, we see that

$$\text{im}(Y_1)_\perp = \ker(Y_1^H) = \ker(X_1^H \Sigma) = \text{im}(\Sigma^{-1} X_2)$$

Thus, we seek to bound $\|\text{Proj}_{\Sigma^{-1}X_2} \Sigma X_2\|_2$.

We will present a coordinate-based bound. First, it is true for any invertible U that $\text{Proj}_Y X = \text{Proj}_{YU} X$. We set U to be the $(n-r) \times (n-r)$ orthogonal matrix L of right-singular vectors of $\text{Proj}_{\Sigma^{-1}X_2} \Sigma X_2$. Thus,

$$\text{Proj}_{\Sigma^{-1}X_2} \Sigma X_2 = \text{Proj}_{\Sigma^{-1}X_2 L} \Sigma X_2 \,,$$

and by the definition of L,

$$\|\text{Proj}_{\Sigma^{-1}X_2} \Sigma X_2\|_2 = \|\text{Proj}_{\Sigma^{-1}X_2 L} \Sigma X_2 L[1 : n-r, 1]\|_2 \ .$$

To keep notation as simple as possible, we denote $Z = X_2 L$ and partition $Z = (z, Z')$. Selecting any column z' from Z', observe that $(\Sigma z)^H (\Sigma^{-1} z') = z^H z' = 0$. Thus we can compute the needed projection by performing Gram-Schmidt on $\Sigma^{-1} z$ with respect to $\Sigma^{-1} Z'$, which we observe to produce $\Sigma^{-1} z - \text{Proj}_{\Sigma^{-1}Z'} \Sigma^{-1} z$.

$$\|\text{Proj}_{\Sigma^{-1}Z} \Sigma z\|_2 = \frac{(\Sigma z)^H \left(\Sigma^{-1} z - \text{Proj}_{\Sigma^{-1}Z'} \Sigma^{-1} z\right)}{\|\Sigma^{-1} z - \text{Proj}_{\Sigma^{-1}Z'} \Sigma^{-1} z\|_2} = \|\Sigma^{-1} z - \text{Proj}_{\Sigma^{-1}Z'} \Sigma^{-1} z\|_2^{-1}$$

We are looking for an upper bound for the latter, so it suffices to bound the quantity $\|\Sigma^{-1} z - \text{Proj}_{\Sigma^{-1}Z'} \Sigma^{-1} z\|_2$ away from 0. Note that if we can restrict the problem to the last $n - r$ coordinates and get a non-trivial lower bound, we have achieved our goal.

For simplicity, we denote the last $(n-r)$ coordinates of Z by B, of Z' by B', and of z by b. The quantity $\|\Sigma_2^{-1} b - \text{Proj}_{\Sigma_2^{-1}B'} \Sigma_2^{-1} b\|_2$ is a least squares error, specifically,

$$\|\Sigma_2^{-1} b - \text{Proj}_{\Sigma_2^{-1}B'} \Sigma_2^{-1} b\|_2^{-1} = \left(\min_{x \in \mathbb{R}^{n-r-1}} \|\Sigma_2^{-1} B' x - \Sigma_2^{-1} b\|_2 \right)^{-1} \,,$$

and on the other hand, trivially,

$$\left(\min_{x \in \mathbb{R}^{n-r-1}} \|\Sigma_2^{-1} B' x - \Sigma_2^{-1} b\|_2 \right)^{-1} \leq \sigma_{r+1} \left(\min_{x \in \mathbb{R}^{n-r-1}} \|B' x - b\|_2 \right)^{-1} \,.$$

To complete the proof, we bound the quantity $\min_{x \in \mathbb{R}^{n-r-1}} \|B'x - b\|_2$ away from 0 in terms of the smallest singular value of the lower right $(n-r) \times (n-r)$ submatrix of the original random matrix X. Indeed,

$$\min_{x \in \mathbb{R}^{n-r-1}} \|B'x - b\|_2 = \min_{x \in \mathbb{R}^{n-r-1}} \|B \begin{pmatrix} 1 \\ x \end{pmatrix} \|_2 \geq \sigma_{\min}(B)$$

However, also recall that B is exactly the lower $(n-r) \times (n-r)$ block of $X_2 L$. As L is unitary and applied on the right, we see $\sigma_{\min}(B) = \sigma_{\min}(X_{22})$. Finally, it is not difficult to check that the orthogonality of X implies that $\sigma_{\min}(X_{22}) = \sigma_{\min}(X_{11})$. Therefore, in total, we have shown

$$\|R_{22}\|_2 \leq \sigma_{r+1} \sigma_{\min}(X_{11})^{-1} .$$

□

Proof of Lemma **??**. Let R' be the upper-triangular result of $\mathbf{QR}(A\tilde{V}^H)$, where $\tilde{V}^H$ is formed by swapping column $i \leq r$ of V^H with column $j + r$. This is equivalent to saying that R' is the upper-triangular result of $\mathbf{QR}(\Sigma \tilde{X})$, where $\tilde{X}$ swaps column i with column $j + r$. This point of view will be more helpful in the proof. From Lemma 3.1 of [36],

$$\left|(R_{11}^{-1} R_{12})[i, j]\right| \leq \frac{|\det(R'_{11})|}{|\det(R_{11})|} . \tag{3.9}$$

We are particularly interested in the case when $i = r$ and $j = n - r$, as will become apparent below.

Since our bound is going to use the coordinate-based inequality (**??**), much as in Lemma **??**, it is again useful to change coordinates to an optimal choice. One can check that $R_{11}^{-1} R_{12} = (\Sigma X_1)^+ \Sigma X_2$, since (**??**) can be viewed as a generalization of Cramer's Rule. Therefore for orthogonal matrices of appropriate sizes $\bar{U}$, $\bar{V}$,

$$\|R_{11}^{-1} R_{12}\|_2 = \|\bar{U}^H R_{11}^{-1} R_{12} \bar{V}\|_2 = \|(\Sigma X_1 \bar{U})^+ (\Sigma X_2 \bar{V})\|_2 .$$

Choosing now $\bar{U}$ and $\bar{V}$ to be given by the SVD of $R_{11}^{-1} R_{12}$ in appropriate column order, we can ensure that the norm of $R_{11}^{-1} R_{12}$ is the lower right entry of $\bar{U} R_{11}^{-1} R_{12} \bar{V}^H = (R_{11} \bar{U}^H)^{-1} R_{12} \bar{V}^H$.

Suppose now that we bounded the entries of $R_{11}^{-1} R_{12}$ in terms of a function of the matrix X that is invariant under right multiplication of X by a block-orthogonal matrix $\begin{pmatrix} \bar{U}^H & 0 \\ 0 & \bar{V} \end{pmatrix}$. Then the bound on the bottom right entry $(r, n-r)$ of $R_{11}^{-1} R_{12}$ would apply to the operator norm. Hence, using Equation (**??**), our the task is to bound

$$R_{11}^{-1} R_{12}[r, n-r] \leq \frac{|\det(R'_{11})|}{|\det(R_{11})|} = \frac{|R'[r, r]|}{|R[r, r]|}.$$

where R' has resulted from swapping column r in X with column n in X.

With the preliminaries over, we introduce notation to assist in the proof. Using Matlab notation, let $X_{11} = X[1:r, 1:r]$, $X_1' = X[1:n, 1:r-1]$, $X_{11}' = X[1:r, 1:r-1]$, $X_{21}' = X[r+1:n, 1:r-1]$; x_r will denote the r-th column of X, x_n will denote the last column of X_2 (also of X). We make use of one projection extensively, and therefore denote it by letter $\Pi := \mathrm{Proj}_{(\Sigma X_1')_\perp}\Sigma(\cdot)$. Finally, projection onto the first r coordinates is used, and we denote this by ζ.

We are ready for the main part of the lemma's proof. First upper bound $|R'[r,r]| = \|\Pi\Sigma x_n\|_2$. Let $\begin{pmatrix} w \\ 0 \end{pmatrix}$ be the maximal right singular unit-vector of the operator $\Pi\Sigma\zeta$. We have

$$|R'[r,r]| = \|\Pi\Sigma x_n\|_2 \leq \|\Pi\Sigma \begin{pmatrix} w \\ 0 \end{pmatrix}\|_2 + \max_{\|t\|=1} \|\Pi\Sigma \begin{pmatrix} 0 \\ t \end{pmatrix}\|_2 \leq \|\Pi\Sigma \begin{pmatrix} w \\ 0 \end{pmatrix}\|_2 + \sigma_{r+1}$$

To analyze this further, decompose $w = c_1 v + c_2 u$ where $v \in X_{11\perp}'$ and $u \in \mathrm{im}(X_{11}')$ are unit-vectors orthogonal to each other. We will have to control this u term later in the proof. To this end, note that we have the following least squares interpretation:

$$\|\Pi\Sigma \begin{pmatrix} u \\ 0 \end{pmatrix}\|_2 = \min_{y\in\mathbb{R}^{r-1}} \|\Sigma X_1' y - \Sigma \begin{pmatrix} u \\ 0 \end{pmatrix}\|_2 \ ,$$

and by making $y = X_{11}'^{+}u$ and taking advantage of the fact that u is in the image of X_{11}' so that $X_{11}'X_{11}'^{+}u = u$, we obtain

$$\|\Pi\Sigma \begin{pmatrix} u \\ 0 \end{pmatrix}\|_2 \leq \|\Sigma_2 X_{21}' X_{11}'^{+} u\|_2 \leq \sigma_{r+1}\sigma_{\min}^{-1}(X_{11}) \ . \tag{3.10}$$

In the second inequality above we have used that $\sigma_{\min}(X_{11}') \geq \sigma_{\min}(X_{11})$.

Our main upper bound on $|R'[r,r]|$ is then

$$|R'[r,r]| \leq \|\Pi\Sigma \begin{pmatrix} v \\ 0 \end{pmatrix}\|_2 + \sigma_{r+1} + \sigma_{r+1}\sigma_{\min}^{-1}(X_{11}) \tag{3.11}$$

Due to the presence of $\|\Pi\Sigma \begin{pmatrix} v \\ 0 \end{pmatrix}\|_2$ in the bound above, we will need to consider two different cases depending on whether this term is small or large; as such, we will need to develop two different lower bounds on $|R[r,r]|$. We now build the the first lower bound on $|R[r,r]| = \|\Pi\Sigma x_r\|_2$.

$$\begin{aligned} |R[r,r]| = \|\Pi\Sigma x_r\|_2 = \min_{y\in\mathbb{R}^{r-1}} \|\Sigma X_1' y - \Sigma x_r\|_2 &\geq \min_{y_1\in\mathbb{R}^{r-1}} \|\Sigma_1 X_{11}' y_1 - \Sigma_1 (x_r)_1\|_2 \\ &= \min_{y_1\in\mathbb{R}^{r-1}} \|\Sigma_1 X_{11} \begin{pmatrix} y_1 \\ 1 \end{pmatrix}\|_2 \geq \sigma_r \sigma_{\min}(X_{11}) \, , \end{aligned}$$

where $(x_r)_1 = x_r[:r]$. This proves the first lower bound we need,

$$|R[r,r]| \geq \sigma_r \sigma_{\min}(X_{11}) \tag{3.12}$$

The second lower bound on $|R[r,r]|$ is similar in spirit, and is introduced to take care of the case when $\|\Pi\Sigma \begin{pmatrix} v \\ 0 \end{pmatrix}\|_2$ is large. Recall v is the unique direction orthogonal to the the columns of X'_{11}, and additionally let c be the magnitude of the projection of x_r onto $\begin{pmatrix} v \\ 0 \end{pmatrix}$. Again interpreting projection through least squares,

$$c = \min_{y \in \mathbb{R}^{r-1}} \|X_{11} \begin{pmatrix} y \\ 1 \end{pmatrix}\|_2 \geq \sigma_{\min}(X_{11})$$

Now use the reverse triangle inequality,

$$|R[r,r]| = \|\Pi\Sigma x_r\|_2 \geq \|\Pi\zeta\Sigma x_r\|_2 - \|\Pi(I-\zeta)\Sigma x_r\|_2 \,. \tag{3.13}$$

We need to lower bound the term $\|\Pi\zeta\Sigma x_r\|_2$ and to upper bound the term $\|\Pi(I-\zeta)\Sigma x_r\|_2$. Decompose $\zeta x_r = \pm cv + u'$ with u' in the image of X'_{11}. Use the same algebra as in (**??**) to control the u' term, establishing that

$$\|\Pi\zeta\Sigma x_r\|_2 \geq c\|\Pi\Sigma \begin{pmatrix} v \\ 0 \end{pmatrix}\|_2 - \|\Pi\Sigma \begin{pmatrix} u' \\ 0 \end{pmatrix}\|_2 \geq c\|\Pi\Sigma \begin{pmatrix} v \\ 0 \end{pmatrix}\|_2 - \sigma_{r+1}\sigma_{\min}^{-1}(X_{11}) \,.$$

To upper bound the second term of (**??**), note that

$$\|(I-\zeta)\Sigma x_r\|_2 \leq \|(I-\zeta)\Sigma\|_2 = \sigma_{r+1} \,.$$

Combining all of these observations and using the lower bound on c, the second lower bound on $|R[r,r]|$ is

$$|R[r,r]| \geq \sigma_{\min}(X_{11})\|\Pi\Sigma \begin{pmatrix} v \\ 0 \end{pmatrix}\|_2 - 2\sigma_{r+1}\sigma_{\min}^{-1}(X_{11}) \tag{3.14}$$

To conclude the proof, we present two cases of the ratio between $|R'[r,r]|$, bounded in (**??**), and $|R[r,r]|$, bounded in (**??**) and (**??**). For the first case, we make the assumption

$\|\Pi\Sigma \begin{pmatrix} v \\ 0 \end{pmatrix}\|_2 \geq 4\sigma_{r+1}\sigma_{\min}^{-2}(X_{11})$. In this situation, (**??**) is superior. We get

$$\begin{aligned}
\frac{|R'[r,r]|}{|R[r,r]|} &\leq \frac{\|\Pi\Sigma \begin{pmatrix} v \\ 0 \end{pmatrix}\|_2 + \sigma_{r+1} + \sigma_{r+1}\sigma_{\min}^{-1}(X_{11})}{\sigma_{\min}(X_{11})\|\Pi\Sigma \begin{pmatrix} v \\ 0 \end{pmatrix}\|_2 - 2\sigma_{r+1}\sigma_{\min}^{-1}(X_{11})} \\
&\leq \frac{1.5\|\Pi\Sigma \begin{pmatrix} v \\ 0 \end{pmatrix}\|_2\sigma_{\min}^{-1}(X_{11})}{\|\Pi\Sigma \begin{pmatrix} v \\ 0 \end{pmatrix}\|_2 - 2\sigma_{r+1}\sigma_{\min}^{-2}(X_{11})} \\
&\leq \frac{1.5\|\Pi\Sigma \begin{pmatrix} v \\ 0 \end{pmatrix}\|_2\sigma_{\min}^{-1}(X_{11})}{.5\|\Pi\Sigma \begin{pmatrix} v \\ 0 \end{pmatrix}\|_2} = 3\sigma_{\min}^{-1}(X_{11})
\end{aligned}$$

The second case is $\|\Pi\Sigma \begin{pmatrix} v \\ 0 \end{pmatrix}\|_2 < 4\sigma_{r+1}\sigma_{\min}^{-2}(X_{11})$. In this situation we must use (**??**),

$$\frac{|R'[r,r]|}{|R[r,r]|} \leq \frac{\|\Pi\Sigma \begin{pmatrix} v \\ 0 \end{pmatrix}\|_2 + \sigma_{r+1} + \sigma_{r+1}\sigma_{\min}^{-1}(X_{11})}{\sigma_r\sigma_{\min}(X_{11})} \tag{3.15}$$

$$\leq 6\frac{\sigma_{r+1}}{\sigma_r}\sigma_{\min}^{-3}(X_{11}) \tag{3.16}$$

The bound we have given depends on $\sigma_{\min}(X_{11})$, therefore satisfying the unitary invariance property we required. To make the statement simple we have added the two bounds together in the lemma statement.

□

Stability of RURV

The following appeared in [24] as Lemma 5.4. We include a short proof for completeness, and also because it is used to prove Theorem **??**.

Theorem 38. RURV *is backward stable.*

Proof. We need two facts: that **QR** is backward stable (e.g., implemented via Householder reflectors) and that, while multiplication by a square matrix is not, in general, backward stable, multiplication by a square unitary matrix is (this is a simple exercise appearing in many Numerical Linear Algebra books, which we leave for the reader).

We input a matrix A and output two matrices, U and R, such that (in the absence of floating point error) we should have $UR = AV^H$, for some V unitary, with U unitary and R

upper triangular. For backward stability, we would like to show that in practice the outputs U and R satisfy $(U + dU)R = (A + dA)(V + dV)^H$, with $V + dV$ unitary, $U + dU$ unitary, and $||dA||/||A|| = O(\epsilon_{mach})$. We know $||dV||$ and $||dU||$ are $O(\epsilon)$ because **QR** is stable.

We start with the fact that matrix multiplication by $(V + dV)^H$ is backward stable; recalling the definition $AV^H =: \hat{A}$ with round-off, then

$$\hat{A} = (A + (dA)_1)(V + dV)^H \ ;$$

with $||(dA)_1||/||A|| = O(\epsilon_{mach})$; and since **QR** is also stable, the output $[U, R]$ will have the property that

$$(U + dU)R = \hat{A} + (dA)_2 \ ,$$

with $||(dA)_2||/||\hat{A}|| = O(\epsilon_{mach})$. Combining these,

$$(U+dU)R = A(V+dV)^H+(dA)_1(V+dV)^H+(dA)_2 = (A+(dA)_1+(dA)_2(V+dV))(V+dV)^H \ .$$

As $(V + dV)^H$ is orthogonal, this means that $||(dA)_2(V + dV)||/||A|| = O(\epsilon_{mach})$, and we conclude that

$$(U + dU)R = (A + dA)(V + dV)^H \ ,$$

where $dA = (dA)_1 + (dA)_2(V + dV)$ has the property that $||dA||/||A|| = O(\epsilon_{mach})$. □

3.5 Analysis of GRURV

In this short section we prove that, given a matrix $M_k = A_1^{m_1} \cdot A_2^{m_2} \cdot \ldots A_k^{m_k}$, where $m_1, \ldots, m_k \in \{-1, 1\}$, and such that only the matrices A_i may be available, **GRURV** can be applied to get the same rank-revealing factorization we would obtain in the case of applying **RURV** to the explicitly formed product M_k.

Lemma 39. ***GRURV*** *(*Generalized Randomized URV*) computes the rank-revealing decomposition* $M_k = U_{current} R_1^{m_1} \ldots R_k^{m_k} V$.

Proof. Let us examine the case when $k = 2$ ($k > 2$ results immediately through simple induction).

Let us examine the cases:

In the first case, $m_2 = 1$. In this case, $M_2 = A_1^{m_1} A_2$; the first **RURV** yields $M_2 = A_1^{m_1} U R_2 V$.

- if $m_1 = 1$, $M_2 = A_1 U R_2 V$; performing **QR** on $A_1 U$ yields $M_2 = U_{current} R_1 R_2 V$.
- if $m_1 = -1$, $M_2 = A_1^{-1} U R_2 V$; performing **RQ** on $U^H A_1$ yields $M_2 = U_{current} R_1^{-1} R_2 V$.

In the second case, $m_2 = -1$. In this case, $M_2 = A_1^{m_1} A_2^{-1}$; the first **RULV** yields $M_2 = A_1^{m_1} U L_2^{-H} V = A_1^{m_1} U R_2^{-1} V$.

- if $m_1 = 1$, $M_2 = A_1 U L_2^{-H} V = A_1 U R_2^{-1} V$; performing **QR** on $A_1 U$ yields $M_2 = U_{current} R_1 R_2^{-1} V$.
- finally, if $m_2 = -1$, $M_2 = A_1^{-1} U L_2^{-H} V = A_1^{-1} U R_2^{-1} V$; performing **RQ** on $U^H A_1$ yields $M_2 = U_{current} R_1^{-1} R_2^{-1} V$.

Note now that in all cases $M_k = U_{current} R_1^{m_1} \dots R_k^{m_k} V$. Since the inverse of an upper triangular matrix is upper triangular, and since the product of two upper triangular matrices is upper triangular, it follows that $R := R_1^{m_1} \dots R_k^{m_k}$ is upper triangular. Thus, we have obtained a rank-revealing decomposition of M_k; the same rank-revealing decomposition as if we have performed QR on $M_k V^H$. □

This allows us to conclude the following important stability result for **GRURV**. We note that it was claimed without proof as Theorem 2.3 in the technical report [3].

Theorem 40. *In the absence of floating point error, the result of the algorithm* ***GRURV*** *is essentially the same as the result of* ***RURV*** *on the (explicitly formed) matrix M_k. This means that all results of Theorem* **??** *apply for the product matrix $R = R_1^{m_1} \dots R_k^{m_k}$.*

Note that we may also return the matrices $R_1, \dots, R_k$, from which the factor R can later be reassembled, if desired.

Theorem 41. GRURV *is backward stable.*

Proof. Simple induction once again shows that it is sufficient to consider the case $k = 2$.

For the case $k = 2$, we will do the calculations for $m_1 = 1$ and $m_2 = \pm 1$; the other two cases can be dealt with in the same fashion.

Note **QR**, **RQ**, and **QL** can be performed in a backward-stable manner, all multiplications performed are multiplications by unitary matrices (and thus backward stable), and we have shown that **RURV** (respectively **RULV**, as the only difference is the application of **QL** instead of **QR**) is also backward stable.

Let now $m_1 = 1$ and let $[U_2, R_2]$ be the outputs of the first **QR** operation performed by **GRURV**, and $[U_1, R_1]$ be the outputs of the second. Then stability of **QR** gives $(U_1 + dU_1)R_1 = A_1 \cdot U_2 + d(A_1 \cdot U_2)$ for some $d(A_1 \cdot U_2)$ satisfying $||d(A_1 \cdot U_2)||/||A_1 \cdot U_2|| = O(\epsilon_{mach})$. However, the stability of **RURV** ensures the existence of dU_2' satisfying $||dU_2'||/||U_2|| = O(\epsilon_{mach})$ such that $U_2 + dU_2'$ is orthogonal. As this establishes U_2 has condition number $1 - O(\epsilon)$, we therefore conclude $d(A_1 \cdot U_2) = (dA_1) \cdot U_2$ for some $||dA_1||/||A_1|| = O(\epsilon_{mach})$. More specifically, this is for $dA_1 = d(A_1 U_2) U_2^{-1}$. Combining these steps, we have shown

$$(U_1 + dU_1)R_1 = (A_1 + dA_1) \cdot U_2 \,.$$

Now we break into cases:

- $m_2 = 1$. Then a simple consequence of Theorem **??** is $U_2 R_2 (V + dV) = A_2 + dA_2$, with $||dA_2||/||A_2|| = O(\epsilon_{mach})$.

  Putting it all together,

  $$(U_1 + dU_1) R_1 R_2 (V + dV) = (A_1 + dA_1) \cdot U_2 R_2 (V + dV) = (A_1 + dA_1)(A_2 + dA_2) \ ,$$

  with $||dA_1||/||A_1|| = O(\epsilon_{mach})$ and $||dA_2||/||A_2|| = O(\epsilon_{mach})$, so in this case **GRURV** is backward stable.

- $m_2 = -1$. Then $U_2 R_2^H (V + dV) = (A_2 + dA_2)^H$, with $||dA_2||/||A_2|| = O(\epsilon_{mach})$, again because **RULV** is backward stable. By Hermitian transposing and inverting, we obtain that $U_2 R_2^{-1} (V + dV) = (A_2 + dA_2)^{-1}$.

  Putting it all together,

  $$(U_1 + dU_1) R_1 R_2^{-1} (V + dV) = (A_1 + dA_1) \cdot U_2 R_2^{-1} (V + dV) = (A_1 + dA_1)(A_2 + dA_2)^{-1} \ ,$$

  with $||dA_1||/||A_1|| = O(\epsilon_{mach})$ and $||dA_2||/||A_2|| = O(\epsilon_{mach})$, so again **GRURV** is backward stable.

The other two cases are dealt with in the same fashion, and simple induction on k shows that **GRURV** is backward stable for any $k \geq 2$. □

3.6 Numerical Experiments

In this section, we present numerical experiments to test the four bounds of Theorem **??**. Since Theorem **??** utilizes the asymptotic behavior of singular values of submatrices of Haar matrices, becoming more accurate as the dimensions of the matrix and submatrix increase, we will perform tests over a range of matrix dimensions.

To test the effects of dimension and gap on the effectiveness of **RURV**, we set up problems with specified singular value distributions and Haar distributed left and right singular vectors. For our experiments, we consider two types of singular value distributions:

- stair step distribution: $\sigma_1 = \sigma_r$, $\sigma_{r+1} = \sigma_n = 1$, with a specified gap σ_r/σ_{r+1};
- logarithmically spaced distribution: $\sigma_1 = 10^{13}$, $\sigma_n = 1$, $\sigma_i/\sigma_{i+1} = \sigma_j/\sigma_{j+1}$ for all $i, j \neq r$, with a specified gap σ_r/σ_{r+1}.

We also perform two types of experiments

- fix the matrix dimension $n = 1500$, and vary the gap σ_r/σ_{r+1} ranging from 10^1 to 10^{10};
- fix the gap $\sigma_r/\sigma_{r+1} = 10^7$, and vary the matrix dimension n ranging from 250 to 2000.

Across all experiments, we fix $r = n/2$ and repeat each experiment 1000 times, constructing matrices with the fixed singular value distribution and Haar-distributed random singular vectors.

We compare the results of these four experiments against the theoretical bounds of Theorem ?? in Figures ??, ??, ??, and ??. In each figure, the top left plot shows the ratio $\sigma_1/\sigma_{\min}(R_{11})$ and Inequality (??), the top right plot shows the ratio $\sigma_{\max}(R_{22})/\sigma_{r+1}$ and Inequality (??), the bottom left plot shows $\|R_{11}^{-1}R_{22}\|_2$ and Inequalities (??) and (??), and the bottom right plot shows an example singular value distribution to illustrate the distribution as a stair step or logarithmically spaced. The experimental results are presented as box plots, with boxes corresponding to the inter-quartile range (middle 50%) and horizontal lines corresponding to min, median, and max. In order to compare against the probabilistic bounds, across all experiments, we specify a fixed failure probability of 3% ($\delta = 0.03$) to obtain the theoretical bounds, and we plot the 97th percentile of the experimental data as a black line. The theoretical bounds (??), (??), and (??) appear as blue lines with asterisk markers; they have values

$$\begin{aligned}
\frac{\sigma_r}{\sigma_{\min}(R_{11})} &\leq \frac{2.02}{\delta}\sqrt{r(n-r)}\ , \\
\frac{\sigma_{\max}(R_{22})}{\sigma_{r+1}} &\leq \frac{2.02}{\delta}\sqrt{r(n-r)}\ , \\
||R_{11}^{-1}R_{12}||_2 &\leq \frac{4.04}{\delta}\cdot\sqrt{r(n-r)}+1\ .
\end{aligned}$$

For (??) from Theorem ?? to be valid, $\sigma_r/\sigma_{r+1} > \sqrt{2}\cdot 1.01\cdot\frac{n}{\delta}$; we do not plot the bound when it does not apply (see Figures ?? and ??). Theoretical bound (??) appears as a red line with circle markers.

Overall, we observe that the probabilistic bounds (??), (??), (??) hold up empirically and are very tight given that the black line never exceeds the blue line but remains close in all experiments.

In the experiment varying the gap for the stair step singular value distribution (Figure ??), a deterministic bound governs the behavior of the algorithm for small gaps. The blue dotted line corresponds to the deterministic upper bounds

$$\frac{\sigma_r}{\sigma_{\min}(R_{11})} \leq \frac{\sigma_r}{\sigma_n}\ , \tag{3.17}$$

$$\frac{\sigma_{\max}(R_{22})}{\sigma_{r+1}} \leq \frac{\sigma_1}{\sigma_{r+1}}\ , \tag{3.18}$$

$$\|R_{11}^{-1}R_{12}\|_2 \leq \frac{\sigma_1}{\sigma_n}\ . \tag{3.19}$$

We also plot these bounds in the same experiment for the logarithmically spaced distribution (Figure ??), but in that case they are very loose.

To get a more detailed view of the distributions of these quantities than the box plots, we provide histograms for the empirical data for two examples from the experiments, one

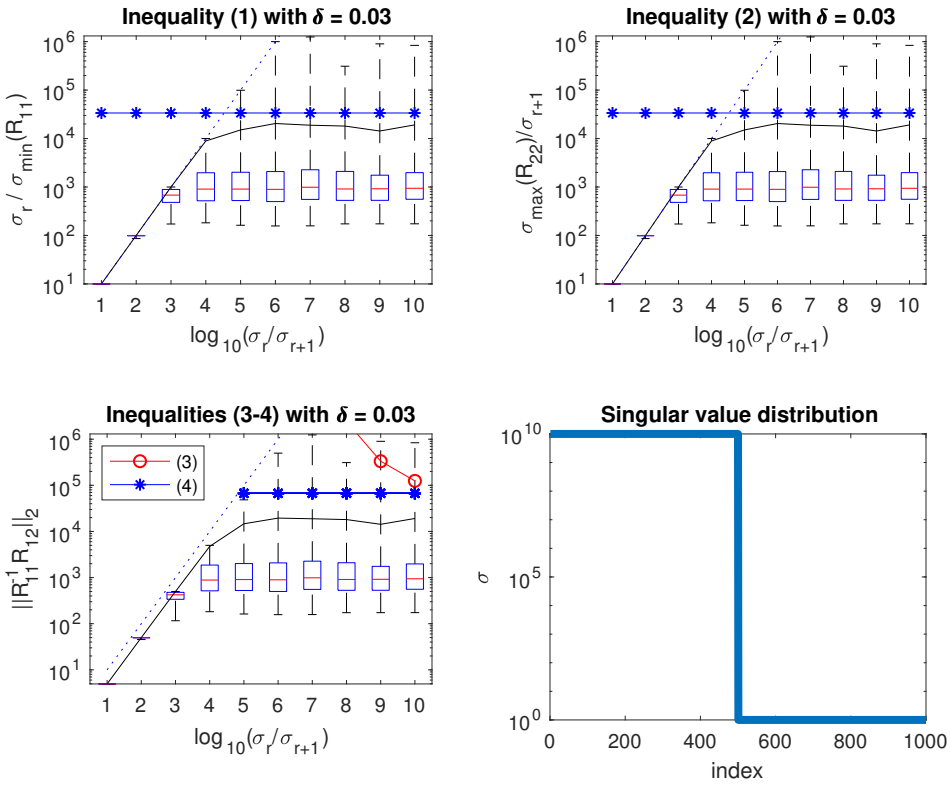

Figure 3.1: Experimental results for test matrices with stair step distribution with dimension $n = 1500$ and varying gap σ_r/σ_{r+1} given on the x-axis. The black line is the 97th percentile. The solid blue line with markers is the corresponding error bound. The dotted blue line is a deterministic bound given by (??–??). The example singular value distribution corresponds to $\sigma_r/\sigma_{r+1} = 10^{10}$.

for each singular value distribution. In the histogram plots, we also include vertical bars to show the 97th percentile (black line) and the theoretical bound (blue line with asterisk marker). Figure ?? shows the logarithms of the quantities $\sigma_1/\sigma_{\min}(R_{11})$, $\sigma_{\max}(R_{22})/\sigma_{r+1}$, and $\|R_{11}^{-1}R_{22}\|_2$ for dimension $n = 1500$ and gap $\sigma_r/\sigma_{r+1} = 10^7$. Figure ?? shows the same quantities for the logarithmically space distribution with the same dimension ($n = 1500$) and gap ($\sigma_r/\sigma_{r+1} = 10^7$).

3.7 Conclusion

We have introduced an algorithm for finding the QR-factorization of products of matrices and their inverses, without explicitly computing the products or inverses. This algorithm is notable for its simplicity, strong theoretical underpinnings, and usefulness as a subroutine within the divide-and-conquer approach to the generalized eigenvalue problem. Among other

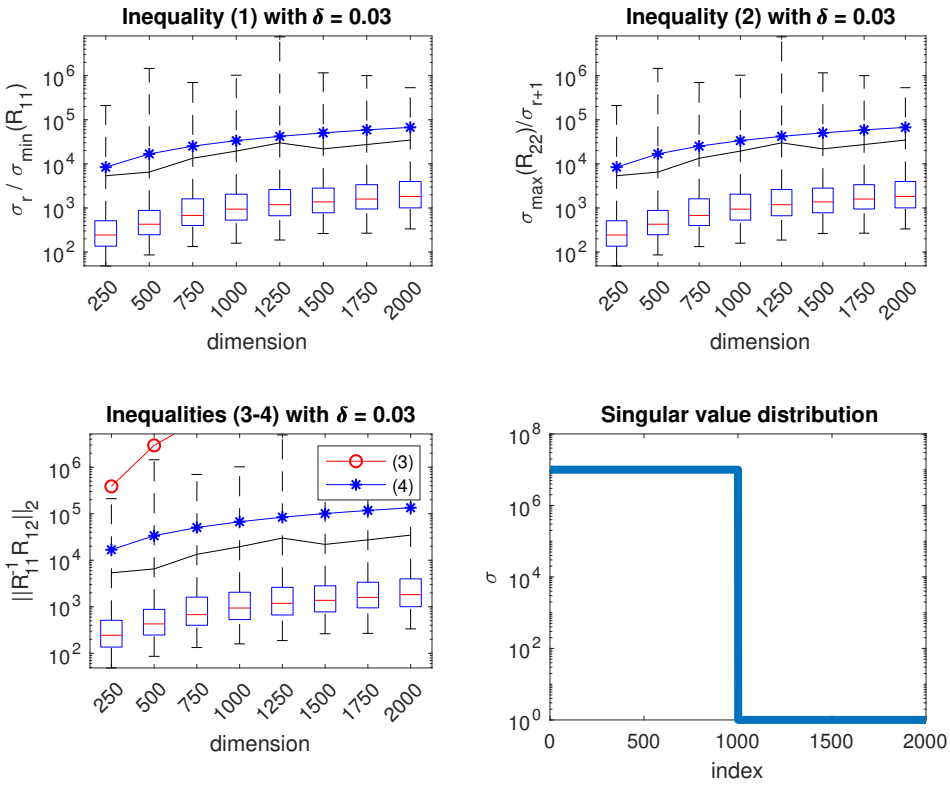

Figure 3.2: Experimental results for test matrices with stair step distribution with gap $\sigma_r/\sigma_{r+1} = 10^7$ and varying dimension given on the x-axis. The black line is the 97th percentile. The solid blue line with markers is the corresponding error bound. The example singular value distribution corresponds to $n = 2000$.

important properties, **GURV** was shown to be strongly rank-revealing, backward stable, and communication-optimal. Moreover, extensive numerical experiments demonstrate that the bounds we presented are essentially tight.

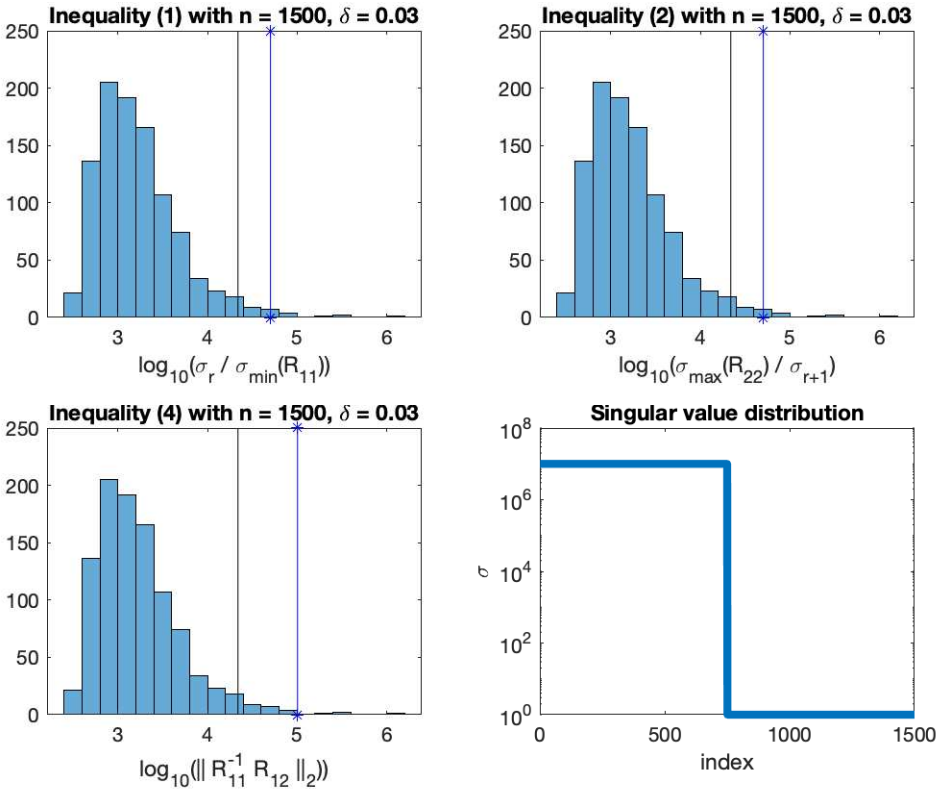

Figure 3.3: Histograms of 1000 trials of the stair step distribution from $n = 1500$ and $\sigma_r/\sigma_{r+1} = 10^7$. The black line indicates the 97th percentile. The blue line indicates where our theoretical bounds from (??), (??), and (??) predict the 97% confidence interval to be.

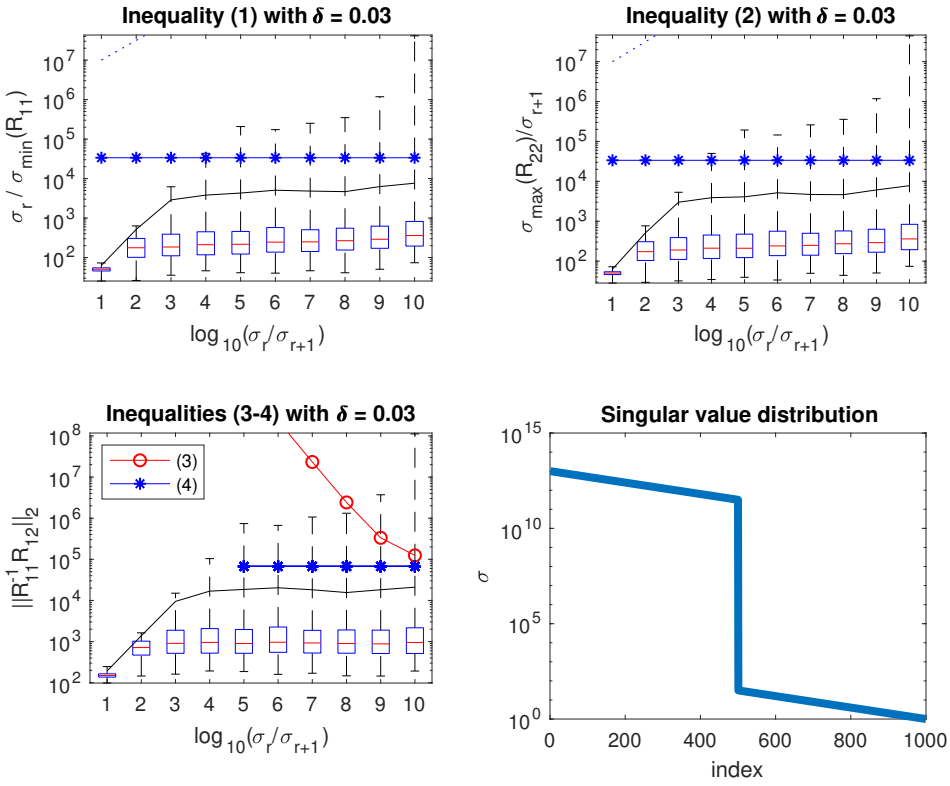

Figure 3.4: Experimental results for test matrices with logarithmically spaced distribution with dimension $n = 1500$ and varying gap σ_r/σ_{r+1} given on the x-axis. The black line is the 97th percentile. The solid blue line with markers is the corresponding error bound. The dotted blue line is a deterministic bound given by (??–??). The example singular value distribution corresponds to $\sigma_r/\sigma_{r+1} = 10^{10}$.

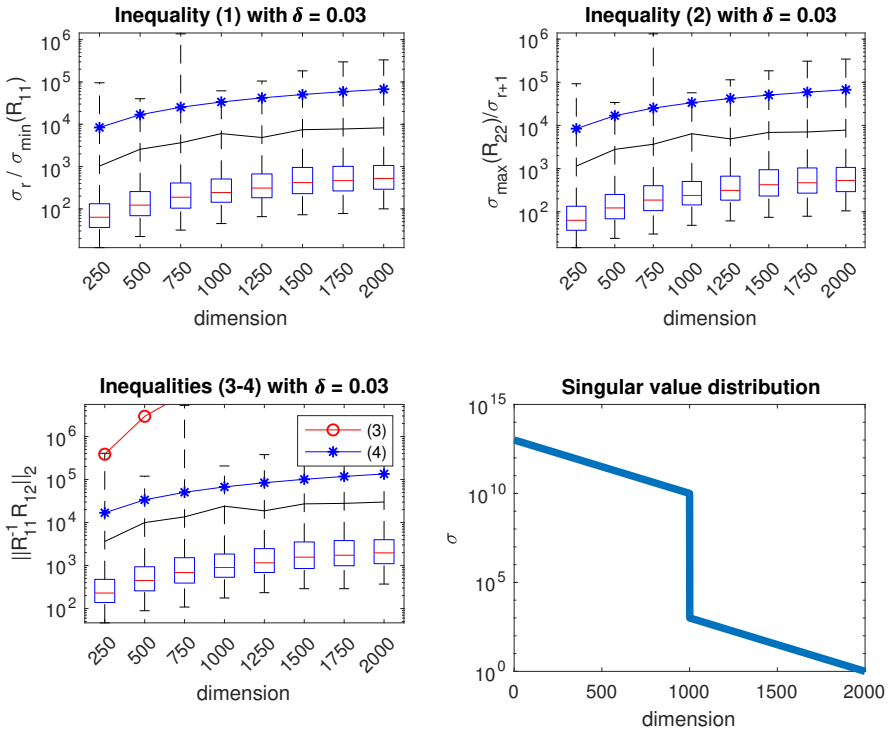

Figure 3.5: Experimental results for test matrices with logarithmically spaced distribution with gap $\sigma_r/\sigma_{r+1} = 10^7$ and varying dimension given on the x-axis. The black line is the 97th percentile. The solid blue line with markers is the corresponding error bound. The example singular value distribution corresponds to $n = 2000$.

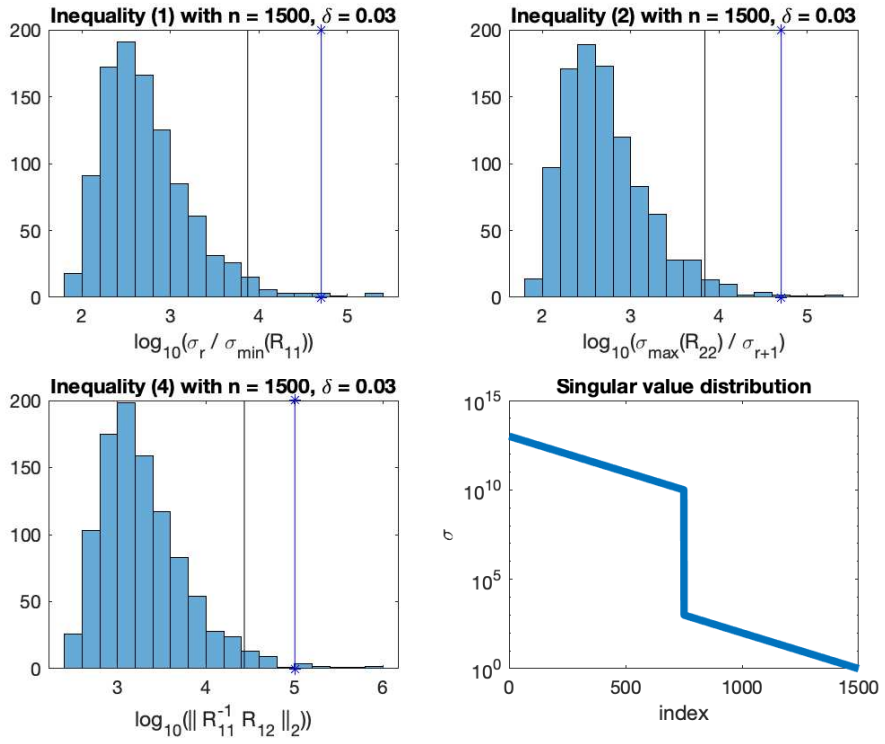

Figure 3.6: Histograms of 1000 trials of the logarithmically spaced distribution from $n = 1500$ and $\sigma_r/\sigma_{r+1} = 10^7$. The black line indicates the 97th percentile. The blue line indicates where our theoretical bounds from (??), (??), and (??) predict the 97% confidence interval to be.

Chapter 4

An improved analysis of low rank matrix approximations[1,2]

4.1 Introduction

Many different problem domains produce matrices that can be approximated by a low-rank matrix. In some cases such as a divide-and-conquer approach to eigenproblems [2], there may be many large and small singular values separated by a gap. In other cases such as identifying a low rank subspace from noisy data, we might expect there to be relatively few large singular values. Perhaps most generically in applied problems, there is no pronounced gap, but the spectrum still decays fairly quickly, and one might prefer to work with a more compact representation when computing quantities such as matrix-vector products.

We next define some related properties which can be of interest to these problems. The following definitions have appeared in the rank-revealing literature, such as in [50, 35, 24, 36] in similar forms. Here and later the singular values are sorted in descending order.

Definition 42. *[low-rank approximation] A matrix A_k satisfying $\|A - A_k\|_2 \leq \gamma\sigma_{k+1}(A)$ for some $\gamma \geq 1$ will be said to be a (k, γ) low-rank approximation of A.*

Definition 43. *[spectrum preserving] If A_k satisfies $\sigma_j(A) \geq \sigma_j(A_k) \geq \gamma^{-1}\sigma_j(A)$ for $j \leq k$ and some $\gamma \geq 1$, it is (k, γ) spectrum preserving.*

Many results in the rank-revealing literature use a strengthening of Definition **??**,

Definition 44. *[kernel approximation] If A_k satisfies $\sigma_{j+k}(A) \leq \sigma_j(A - A_k) \leq \gamma\sigma_{k+j}(A)$ for $1 \leq j \leq n - k$ and some $\gamma \geq 1$, it is a (k, γ) kernel approximation of A.*

In all of these definitions, if we assume A_k is rank k, then $\gamma = 1$ is optimal from the truncated-SVD, so all methods can be compared with this standard. Though we made

[1] Joint work with James Demmel, Laura Grigori

[2] Preprint [25]

the above definitions quite strong, we will not prove our results satisfy them exactly. In particular, we drop the upper bound in Definition **??** and the lower bound in Definition **??** from our considerations. One can derive analogs for these dropped quantities using techniques developed in this paper. For example, one could replace $\sigma_j(A)$ with $\delta \cdot \sigma_j(A)$ in Definition **??** and it is not generally difficult to give a better bound on δ than on γ by using Definition **??** and Weyl's inequality. However, the stated complementary bounds in Defs. **??** and **??** do not hold for all the algorithmic variations we consider, and we choose not to complicate the results with these considerations.

Different algorithms may end up representing A_k in different ways, but generally A_k is represented as a product of matrices which have at least one dimension much smaller than those of the original A. Note in this work we do not require A_k to be rank k. Nevertheless, the dimensions of A_k will be chosen as a function of k in order to compete with the truncated SVD of rank k, and this motivates the choice of notation. For the choices made in this paper, it is always the case that $\text{rank}(A_k) = O\left(k \cdot \text{polylog}(n)\right)$.

This paper has two main goals, both motivated by the history of low-rank factorizations. First, we show that many important low-rank factorizations can be viewed as an LU-factorization followed by deleting the Schur-complement. We call this prototype algorithm **GLU**. Second, older research into low-rank factorizations bounded more quantities than recent results on randomized algorithms. In particular, Definitions **??** and **??** do not receive much discussion in randomized algorithms. We will provide bounds on all of Definitions **??**, **??**, **??** for **GLU**. In doing this, we first derive sharp deterministic bounds for approximate LU and QR factorizations in Sections **??** and **??**, and then in section **??** we complete the bounds by using properties of random matrix ensembles.

GLU is essentially an LU-factorization that allows the leading block to be rectangular instead of square. Allowing the leading block to be rectangular enables much better low-rank approximation properties. Let A be an $m \times n$ matrix, A_{11} be the leading $l' \times l$ block which is assumed to have full column rank so that $l' \geq l$, and U and V be invertible matrices. First we have an exact factorization of matrix A that is the natural generalization of a full LU-factorization,

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} I & \\ A_{21}A_{11}^{+} & I \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ & \mathcal{S}(A_{11}) \end{pmatrix},$$

where $\mathcal{S}(A_{11}) = A_{22} - A_{21}A_{11}^{+}A_{12}$ denotes what we call the generalized Schur-complement. By applying the sketching matrices U and V and deleting the Schur complement, we get a low-rank factorization that can have remarkably good properties. Defining $\bar{A} = UAV$,

$$A_k := U^{-1} \begin{pmatrix} I \\ \bar{A}_{21}\bar{A}_{11}^{+} \end{pmatrix} \begin{pmatrix} \bar{A}_{11} & \bar{A}_{12} \end{pmatrix} V^{-1} \tag{4.1}$$

is a complete description of our proposed **GLU** approximation. The inverses may look daunting at first because they are large matrices, but we will see that they are only tools to

facilitate the analysis; actually the leading rows of U and leading columns of V are the only parts required.

We have emphasized that **GLU** factorization unifies many factorizations through appropriate choices of the settings of U,V. We believe it is also important that other choices are novel and practical, as we illustrate in main results Theorem **??** and Theorem **??**. That said, this paper will not argue that these novel instantiations of **GLU** should necessarily be adopted over similar methods like the low-rank factorization described in [21]. On the contrary, we find that the factorization underlying [21], which we term **CW**, can be viewed as an abridged version of **GLU**. Thus while our bounds are tighter in the case of Definition **??**, we briefly sketch in Remark **??** that the improved bounds on Definitions **??** and **??** under the SRHT ensemble as in Theorem **??** also apply to **CW**.

The remainder of the introduction is divided into four sections for clarity. The first and second aim to highlight our contributions. The third provides references to related work. The fourth gives notation we adopt.

Unifying Approach

GLU generalizes past low-rank LU factorizations in two ways. First, it allows pre- and post-multiplication by matrices other than permutations. Second, it allows for rectangular Schur complements. Even without generalizing to rectangular Schur complements, GLU encompasses several well-known procedures. We provide examples to illustrate this in section **??**. We refer the reader to Table 1 in the preprint version of this paper for a summary of several deterministic and randomized approximation algorithms. It displays separately the case when $k \leq l = l'$ and the more general case when $k \leq l \leq l'$, and cites existing as well as new bounds on the spectral and kernel approximation provided by these algorithms. We discuss a novel and practical instance when $l < l'$ in section **??**. Here we focus on $k \leq l = l'$ and identify the equivalence between existing deterministic and randomized algorithms. In this case, the rank-k approximation A_k can be written as

$$\begin{aligned} A_k &= U^{-1} \begin{pmatrix} I_l \\ \bar{A}_{21}\bar{A}_{11}^{-1} \end{pmatrix} \begin{pmatrix} \bar{A}_{11} & \bar{A}_{12} \end{pmatrix} V^{-1} \\ &= AV_1(U_1AV_1)^{-1}U_1A, \end{aligned} \tag{4.2}$$

where V_1 contains the leading l columns of V, U_1 contains the leading l rows of U, and $\bar{A} = UAV$. See (**??**) for more details. Now we define some notation we will use later. Let Q_1 be the orthogonal factor obtained from the thin QR-decomposition of AV_1, so Q_1 is of dimensions $m \times l$. In the case when U_1 contains the leading l rows of a permutation matrix U, we denote $UQ_1 = \begin{pmatrix} \bar{Q}_{11} \\ \bar{Q}_{21} \end{pmatrix}$, where $\bar{Q}_{11}$ is $l \times l$. While $l \geq k$ is always the case, in applications l varies from being exactly k, as for deterministic algorithms, to being a polylog-factor larger than k for randomized algorithms.

Deterministic algorithms are typically based on rank revealing QR and LU factorizations.

Both factorizations select k columns from the matrix A, that is V_1 represents a column permutation and AV_1 are the selected columns. In the case of a rank revealing QR factorization, $U_1 = Q_1^T$ and the approximation becomes $A_k = Q_1 Q_1^T A$. See (**??**) for a detailed derivation. Let $Q_1^T A = (R_{11}\ R_{12})$. The strong rank-revealing QR-factorization [36] chooses the column permutation V_1 such that $||R_{11}^{-1}R_{12}||_{max}$ is bounded by a small constant and the approximation A_k is spectrum preserving and a kernel approximation of A: γ in Definitions **??** and **??** is a low degree polynomial in n and k. The rank revealing LU factorization selects k columns and k rows from the matrix A, that is both U_1 and V_1 are permutation matrices. For example in [34] the columns are selected by using a pivoting strategy referred to as tournament pivoting and based on rank revealing QR, while the rows are selected such that $||\bar{Q}_{21}\bar{Q}_{11}^{-1}||_{max}$ is bounded. The obtained approximation

$$A_k = AV_1(U_1AV_1)^{-1}U_1A$$

is again spectrum preserving and a kernel approximation of A, with γ in Definitions **??** and **??** being a low degree polynomial in n and k.

For randomized algorithms, V_1 is a random matrix, typically based on JL transforms or fast JL transforms, such as the sub-sampled randomized Hadamard transform (SRHT) of Definition **??** introduced originally in [58]. The randomized SVD (see e.g. [37]) is obtained by choosing $U_1 = Q_1^T$ and corresponds to computing l steps of the QR factorization of UAV. We refer to this factorization as a randomized QR factorization. The randomized SVD via row extraction is obtained by choosing U_1 a row permutation such that $||\bar{Q}_{21}\bar{Q}_{11}^{-1}||_{max}$ is bounded. In other words, this factorization corresponds to computing l steps of the LU factorization of UAV, and we refer to this as randomized LU with row selection. Notably in the case of the approximations based on LU factorization, both deterministic and randomized algorithms bound $||\bar{Q}_{21}\bar{Q}_{11}^{-1}||_{max}$ [36, 34] to obtain guarantees on the approximation.

Detailed Bounds

In the context we consider, **GLU** satisfies bounds at least as sharp as in the literature, and many are new.

Given $k \leq l \leq l'$, the clean formulation of A_k described in eq. (**??**) becomes a bit more complicated,

$$A_k = [U_1^+(I - (U_1AV_1)(U_1AV_1)^+) + (AV_1)(U_1AV_1)^+][U_1A], \tag{4.3}$$

where U_1 and (U_1AV_1) are of dimensions $l' \times m$ and $l' \times l$ respectively. However, the algorithmic implementation is still straightforward and inexpensive. See (**??**) for a detailed derivation. Proposition **??** gives the precursor bounds for the spectral and the kernel approximation provided by A_k for general U_1 and V_1, and as in Section **??** properties of U_1 and V_1 specific to the algorithm are used to complete the bound. Both Propositions **??** and **??** provide new deterministic bounds not found in the literature. For example, Proposition **??** generalizes Theorem 9.1 of [37] to include values $j > 1$. This generalization proves

useful when analyzing Definition ?? for randomized algorithms, which we observe to be an advantage of **GLU** over **CW**.

Section ?? contains our new results after suitable random ensembles are chosen, that is when V_1 and U_1 are random matrices. Extra attention is given to the SRHT ensemble of Definition ??, because the especially good bounds it can provide were not fully exploited in past literature. Using this ensemble, from Algorithm ?? for computing **GLU** we can see the number of arithmetic operations is $O(nm\log(l') + mll')$. Plugging in l' and l from Theorem ??, we can produce a low-rank approximation in $\tilde{O}(nm + k^2m\epsilon^{-3})$ time that relative to the squared error of the truncated SVD of rank k, $A_{\text{opt},k}$,

- approximates A with only $1 + O(\epsilon)$ times the squared Frobenius norm error .
- approximates A with only $O\left(1 + \frac{\log(m/\delta)\epsilon}{k\log(k/\delta)}\frac{\|A-A_{\text{opt},k}\|_F^2}{\|A-A_{\text{opt},k}\|_2^2}\right)$ times the squared spectral norm error.

This holds with probability $1 - 5\delta$, and l, l' grow poly-logarithmically with δ, as in Remark ??. In other words, the algorithm we propose attains $\gamma = O(1)$ in Definition ?? for many families of A matrices encountered in practice with modest spectral decay (which makes the Frobenius norm not too much larger than the spectral norm). The same Theorem ?? shows this $\gamma = O(1)$ bound carries over to Definition ??. Further, Theorem ?? shows that Definition ?? is satisfied with $\gamma = O(\frac{k}{n})$. To our knowledge, no other work has found such a representation of A in time less than $\Omega(nmk)$ satisfying any of these properties. Instead, randomized low-rank approximation literature on algorithms running in $\Omega(nmk)$ time do not typically discuss the spectral norm of the residual (Definition ??), choosing to focus on the Frobenius norm. Moreover, the fast linear algebra community has typically not considered properties like spectrum preserving and kernel approximation (Definitions ??, ??).

Our bounds have interesting implications for the growth factor of pre/post-conditioned Gaussian Elimination. Corollary ?? is a step towards a theoretical understanding of conditioning Gaussian Elimination to avoid pivoting. Besides this, it expands the classes of distributions for which pivoting is provably unnecessary, to a class including Gaussian-distributed matrices. We pose an open question at the end, motivated by this analysis.

Related Work

Low-rank matrix approximations have been extensively studied, hence this work is related to a large body of literature. Because of our emphasis on the LU-factorization viewpoint, we should mention some work related to LU factorizations. Such papers providing information regarding Definitions ??, ??, ?? are few, notably including perhaps the first [50], as well as later more efficient versions like [34]. These papers do not exploit randomness, however.

Exploiting randomness for low-rank factorizations has led to major speedups. Some literature in recent years has exploited this for LU factorizations, including perhaps most

relevantly [59]. Their work has somewhat different goals, in that it seeks to find left and right permutation matrices, which makes it in some ways more like [34]. Also, their paper only discusses spectral norm bounds on the residual. Interestingly, the fast version of their procedure (their Algorithm 4.4) uses an ensemble equivalent to the SRHT ensemble. The bounds we have in Theorem **??** are better for the spectral norm of the residual. Comparing our Theorem **??** with their Theorem 4.12, our approximation is always a factor on the order of $\sqrt{n}$ more accurate, and a factor n more accurate when the spectrum decays sufficiently quickly. Our results utilizing the SRHT ensemble build on [12], which proved the SRHT ensemble has geometry preserving properties beyond those of the Johnson-Lindenstrauss transform properties. They used this fact to provide sharper spectral norm bounds on the residual for the randomized QR decomposition approach to low-rank matrix approximation.

Outside of research into LU factorizations, many papers have focused on studying JL embeddings. This has culminated in algorithms considered to run in nnz(A) time for many problems related to and including low-rank approximations. Notable such papers include [21] and [52]. This body of literature has focused more on the properties of the random ensemble, and little on the properties of the factorization itself. For example, [52] uses the same factorization as [21], whose technical report we believe to be the first paper to use sketching from the left and right to speed up the algorithm. Few of these papers for nnz(A) algorithms study any error bounds beyond the Frobenius norm of the residual.

To date, procedures for the residual being within an ϵ factor as accurrate as the truncated SVD with respect to the spectral norm do not gain any speed advantage by using fast Johnson-Lindenstrauss ensembles. This is because a repeated-squaring must be used, and therefore structured sketching matrices have no advantage. Important work in this area includes [35] and [51].

The list is far from complete, and many different takes on the problem have been proposed which tangentially touch this paper, [37] and [65] are useful for finding more pointers into the literature.

Notations

As this paper is notation heavy, we first take a moment to collect some conventions we will use.

- A is $m \times n$.
- $A_{\text{opt},k}$ will be the truncated rank-k SVD.
- Assume $m \geq n$. $[Q, R] = \mathbf{tQR}(A)$ will be the thin QR-decomposition of A, so Q is $m \times n$
- $[Q, R] = \mathbf{QR}(A)$ will be the square QR-decomposition of A, so Q is $m \times m$.

- Assume $m \leq n$. $[L, Q] = \mathbf{tLQ}(A)$ will be the thin LQ-decomposition of A, so Q is $m \times n$
- $[L, Q] = \mathbf{LQ}(A)$ will be the square LQ-decomposition of A, so Q is $n \times n$.
- $[U, \Sigma, V] = \mathbf{tSVD}(A)$ will be the thin variant (square Σ) and with decreasing singular values. So given $m \geq n$, $A = U\Sigma V^T$, singular values are $\sigma_1 \geq \ldots, \geq \sigma_n$ and U is $m \times n$.
- A^+ is the $n \times m$ Moore-Penrose pseudo-inverse.
- $[U, \Sigma, V] = \mathbf{SVD}(A)$ will be the full variant ($m \times n$ Σ) and with decreasing singular values. So U is $m \times m$, V is $n \times n$.
- $\mathcal{S}(A_{11}) = A_{22} - A_{21}A_{11}^+A_{12}$ is the Schur complement of A_{11}; if the dimension of A_{11} is $' \times l$, then $\mathcal{S}(A_{11})$ is $(m - l') \times (n - l)$. Here $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$.
- Matlab-like notation to select submatrices, e.g. $A[: k, : k]$ is the leading $k \times k$ minor of A.
- To simplify notation, we denote $(X_{11})^+$ as X_{11}^+.

4.2 Generalized LU-factorization

Classically as in [50] and [34], the rank-revealing LU factorization finds permutations P_r, P_c (usually iteratively over the procedure), forming $\bar{A} = P_r A P_c$, and LU-factors $\bar{A}$ but deletes the Schur-complement after k-steps. Thus,

$$\bar{A} = \begin{pmatrix} I & 0 \\ \bar{A}_{21}\bar{A}_{11}^{-1} & I \end{pmatrix} \begin{pmatrix} \bar{A}_{11} & \bar{A}_{12} \\ 0 & \mathcal{S}(\bar{A}_{11}) \end{pmatrix} \approx \begin{pmatrix} I \\ \bar{A}_{21}\bar{A}_{11}^{-1} \end{pmatrix} \begin{pmatrix} \bar{A}_{11} & \bar{A}_{12} \end{pmatrix} =: \bar{A}_k.$$

This naturally suggests the approximation $A \approx A_k := P_r^T \bar{A}_k P_c^T$. Letting P_{c1} be the first k columns of P_c and P_{r1} be the first k rows of P_r, some algebra (see Remark **??** for the more general case) shows the approximation to be $A \approx AP_{c1}(P_{r1}AP_{c1})^{-1}P_{r1}A$.

This paper generalizes the rank-revealing LU-factorization in two directions. First, we include other matrices on the left and right besides permutations. This allows for speedups through matrix sketching. Second, we generalize one step further by using rectangular Schur complements. This can greatly improve the quality of the low-rank approximation, as we will see in Proposition **??** and Theorem **??**.

We describe this second modification in greater detail now. For the sake of analysis it will be convenient to let U, V be square matrices in the following discussion and subsequent

Proposition **??**. The relevant matrices are the $m \times n$ matrix A which we wish to approximate, the invertible $m \times m$ matrix U, and the invertible $n \times n$ matrix V. Now define

$$\bar{A} := UAV = \begin{pmatrix} I_{l'} & 0 \\ \bar{A}_{21}\bar{A}_{11}^{+} & I_{m-l'} \end{pmatrix} \begin{pmatrix} \bar{A}_{11} & \bar{A}_{12} \\ 0 & \mathcal{S}(\bar{A}_{11}) \end{pmatrix},$$

where this is valid when the $l' \times l$ block $\bar{A}_{11}$ has full column rank so that $\bar{A}_{11}^{+}\bar{A}_{11} = I$. In particular we are assuming $l' \geq l$. To help visualize the construction, the following depicts the block sizes.

$$\begin{aligned} \bar{A} &= \begin{pmatrix} l', l & l', n-l \\ m-l', l & m-l', n-l \end{pmatrix} \\ &= \begin{pmatrix} l', l' & l', m-l' \\ m-l', l' & m-l', m-l' \end{pmatrix} \begin{pmatrix} l', l & l', n-l \\ m-l', l & m-l', n-l \end{pmatrix}. \end{aligned}$$

Deleting the $(m-l') \times (n-l)$ Schur complement and undoing the U, V factors gives the approximation we use as a definition,

$$A \approx A_k := U^{-1} \begin{pmatrix} I_{l'} \\ \bar{A}_{21}\bar{A}_{11}^{+} \end{pmatrix} \begin{pmatrix} \bar{A}_{11} & \bar{A}_{12} \end{pmatrix} V^{-1}. \tag{4.4}$$

In (**??**), U and V are square, but for low-rank approximations this would be expensive. Only the leading l' rows and l columns respectively of U and V respectively are actually required, but we find the square form helpful for the analysis. Accordingly for U, we assume that we may express

$$U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = \begin{pmatrix} L'_{11}U'_1 \\ U'_2 \end{pmatrix} = \begin{pmatrix} L'_{11} & 0 \\ 0 & I_{m-l'} \end{pmatrix} \begin{pmatrix} U'_1 \\ U'_2 \end{pmatrix} = L'U', \tag{4.5}$$

where $U' = \begin{pmatrix} U'_1 \\ U'_2 \end{pmatrix}$ is an orthogonal matrix, U_1 and U'_1 are $l' \times m$, and L_{11} is $l' \times l'$ lower-triangular. Note by assumption, L'_{21} and L'_{12} are 0 matrices, and $L'_{22} = I_{m-l'}$. Conceptually this means the first l' rows of U are arbitrary full-rank and the other rows are the orthogonal complement. We also assume L' is invertible, so that U is invertible as well. Any reasonable sketching matrix U_1 satisfies this property with probability 1. Similarly, we assume V may be expressed as

$$V = \begin{pmatrix} V_1 & V_2 \end{pmatrix} = \begin{pmatrix} V'_1R'_{11} & V'_2 \end{pmatrix} = V' \begin{pmatrix} R'_{11} & 0 \\ 0 & I_{n-l} \end{pmatrix} = V'R', \tag{4.6}$$

where V' is orthogonal, V_1 and V'_1 are $n \times l$, and R'_{11} is $l \times l$ upper-triangular. Again note the assumption R'_{21} and R'_{12} and both 0 matrices, and $R'_{22} = I_{n-l}$. We will again assume R' is invertible so that V is as well. That R'_{22} and L'_{22} are identity matrices will be used several times in our algebra, so we emphasize this fact.

Schur complements of rectangular blocks do not appear to be commonly used. The following derives a few useful identities for them in the context of LU-factorization.

Lemma 45. *We continue to assume $l' \geq l$ and that $\bar{A}_{11}$ has full column rank so that $\bar{A}_{11}^{+}\bar{A}_{11} = I$. Further introduce matrices $U = L'U'$ and $V = V'R'$ structured as explained in (??) and (??). Set $[Q, R] = \boldsymbol{QR}(AV)$ so that R is $m \times n$. Block R so that R_{11} is $l \times l$ and $X := UQ$ so so that X_{11} is $l' \times l$. Then the following identities hold for $\bar{A} = UAV$,*

$$\mathcal{S}(\bar{A}_{11}) = \mathcal{S}(X_{11})R_{22} \tag{4.7}$$

$$\bar{A}_{11} = X_{11}R_{11} \tag{4.8}$$

$$\bar{A}_{21}\bar{A}_{11}^{+} = X_{21}X_{11}^{+}. \tag{4.9}$$

Proof. There is a factorization through a generalized LU-factorization of $\bar{A}$, in which the lower-triangular factor is the identity on the diagonal and the lower left factor is $\bar{A}_{21}\bar{A}_{11}^{+}$,

$$\bar{A} = \begin{pmatrix} I_{l'} & \\ \bar{A}_{21}\bar{A}_{11}^{+} & I_{m-l'} \end{pmatrix} \begin{pmatrix} \bar{A}_{11} & \bar{A}_{12} \\ & \mathcal{S}(\bar{A}_{11}) \end{pmatrix}. \tag{4.10}$$

However we could alternatively first use a QR-factorization of AV followed by a generalized LU-factorization of X (so that X_{11} is $l' \times l$),

$$\begin{aligned} \bar{A} &= UAV \\ &= XR \\ &= \begin{pmatrix} I_{l'} & \\ X_{21}X_{11}^{+} & I_{m-l'} \end{pmatrix} \begin{pmatrix} X_{11} & X_{12} \\ & \mathcal{S}(X_{11}) \end{pmatrix} \begin{pmatrix} R_{11} & R_{12} \\ & R_{22} \end{pmatrix} \\ &= \begin{pmatrix} I_{l'} & \\ X_{21}X_{11}^{+} & I_{m-l'} \end{pmatrix} \begin{pmatrix} X_{11}R_{11} & \cdots \\ & \mathcal{S}(X_{11})R_{22} \end{pmatrix}. \end{aligned} \tag{4.11}$$

The proof amounts to equating the blocks now between (??) and (??), but we provide a justification which essentially argues that the lower left block of the generalized LU factorization makes it unique. (??) follows first because $I_{l'}X_{11}R_{11} = \bar{A}_{11}$.

Second, by definition $\begin{pmatrix} \bar{A}_{11} \\ \bar{A}_{21} \end{pmatrix} = \begin{pmatrix} X_{11}R_{11} \\ X_{21}R_{11} \end{pmatrix}$. We assumed $\bar{A}_{11}$ has full column rank and we continually assume U, V are invertible; therefore X_{11} has full column rank and R_{11} has full row rank (it is invertible). Consequently we may compute the pseudo-inverse $\bar{A}_{21}\bar{A}_{11}^{+} = X_{21}R_{11}(X_{11}R_{11})^{+} = X_{21}X_{11}^{+}$. This gives (??).

Finally, (??) follows by equating the corresponding lower-right block of the upper-triangular factors, and is justified because we have shown the left-triangular factors in (??) and (??) are identical (and invertible). □

Singular values of a matrix product obey a well-known bound called the multiplicative Weyl inequality. We make use of this and its less known reverse version. Therefore we state the inequality with a reference, and prove its reverse version.

Lemma 46. *Say A is $m \times n$, B is $n \times p$. For $1 \leq k \leq j$,*

$$\sigma_j(AB) \leq \sigma_{j-k+1}(A)\sigma_k(B). \tag{4.12}$$

Now assume for simplicity that $n \geq m \geq p$, both A, B are full rank, and $im(B) \subset ker(A)_\perp$. In other words, A is short-wide and B is tall-skinny, and the image of B is orthogonal to the kernel of A. Then for $1 \leq k \leq m - j$ and $j \leq p$, an inequality in the other direction is

$$\sigma_{m-k+1}(A)\sigma_{j+k-1}(B) \leq \sigma_j(AB). \tag{4.13}$$

Besides these multiplicative inequalities, the additive Weyl inequality holds for any matrices A, B and $1 \leq k, j \leq n$ where n is the smaller of the row and column numbers, and says

$$\sigma_j(A + B) \leq \sigma_{j-k+1}(A) + \sigma_k(B). \tag{4.14}$$

Proof. (**??**) and (**??**) are well-known. For example, section 7.3, exercise 18 from [38].

We next prove (**??**). Let Σ_1, Σ_2 be the square singular value matrices of A,B respectively. Then AB is spectrally equivalent to $\Sigma_1 U \Sigma_2$ for some $m \times p$ orthogonal matrix $U = V_1^T U_2$, with U_2 being the left singular matrix of B and V_1 being the right singular matrix of A. This U has orthonormal columns because it is norm preserving; $\mathrm{im}(U_2) \subset \mathrm{im}(V_1) = \ker(V_1^T)_\perp$ so if we let V extend V_1 to a square orthogonal matrix, then $\|V_1^T U_2 x\|_2 = \|V^T U_2 x\|_2 = \|x\|_2$. Σ_1 is invertible based on the full rank assumption, and $U\Sigma_2$ is $m \times p$ with full column rank. Note that $(U\Sigma_2)^+\Sigma_1^{-1}$ is a left inverse for $\Sigma_1 U \Sigma_2$. Therefore $(\Sigma_1 U \Sigma_2)^+ = (U\Sigma_2)^+\Sigma_1^{-1}P$ where P orthogonally projects onto $\mathrm{im}(\Sigma_1 U \Sigma_2)$. Apply (**??**) to conclude $\sigma_j((\Sigma_1 U \Sigma_2)^+) \leq \sigma_j((U\Sigma_2)^+\Sigma_1^{-1})$. Combine this with another application of (**??**),

$$\begin{aligned}
\sigma_{p-j+1}^{-1}(AB) = \sigma_j((AB)^+) \leq \sigma_j((U\Sigma_2)^+\Sigma_1^{-1}) &\leq \sigma_{j-k+1}((U\Sigma_2)^+)\sigma_k(\Sigma_1^{-1}) \\
&= \sigma_{p-(j-k+1)+1}^{-1}(V_1^T U_2 \Sigma_2)\sigma_{m-k+1}^{-1}(A) \\
&= \sigma_{p-(j-k+1)+1}^{-1}(\Sigma_2)\sigma_{m-k+1}^{-1}(A) \\
&= \sigma_{p-(j-k+1)+1}^{-1}(B)\sigma_{m-k+1}^{-1}(A).
\end{aligned} \tag{4.15}$$

We used that $V_1^T U_2$ is an orthogonal matrix to advance to line (**??**). Finally, reassign $j = p - j + 1$ to get the claimed (**??**).

□

The next Proposition is critical for understanding the rank-revealing properties for **GLU**. It will combine with Proposition **??** to culminate in Theorem **??**.

Proposition 47. *Let A be an $m \times n$ matrix, $U = L'U'$ and $V = V'R'$ as in (**??**) and (**??**), $[Q, R] = \boldsymbol{QR}(AV)$, and finally $\bar{A} = UAV$. Block $Q, R, A, \bar{A}$ as in Lemma **??**; in particular Q_{11} is $l' \times l$ and R_{11} is $l \times l$. Then the low-rank approximation suggested in (**??**), namely*

$$A_k := U^{-1} \begin{pmatrix} I \\ \bar{A}_{21}\bar{A}_{11}^+ \end{pmatrix} \begin{pmatrix} \bar{A}_{11} & \bar{A}_{12} \end{pmatrix} V^{-1},$$

satisfies

$$\|A - A_k\|_F^2 \leq \|R_{22}\|_F^2 + \|(UQ)_{11}^{+}(UQ)_{12}R_{22}\|_F^2 \tag{4.16}$$

$$\|(A - A_k) - (A - A_k)_{opt,j-1}\|_F^2 \leq \|R_{22} - R_{22\,opt,j-1}\|_F^2 + \|(UQ)_{11}^{+}(UQ)_{12}(R_{22} - R_{22\,opt,j-1})\|_F^2 \tag{4.17}$$

$$\|A - A_k\|_2^2 \leq \|R_{22}\|_2^2 + \|(UQ)_{11}^{+}(UQ)_{12}R_{22}\|_2^2 \tag{4.18}$$

$$\sigma_j^2(A - A_k) \leq \|R_{22} - R_{22\,opt,j-1}\|_2^2 + \|(UQ)_{11}^{+}(UQ)_{12}(R_{22} - R_{22\,opt,j-1})\|_2^2 \tag{4.19}$$

$$\sigma_i(A_k) \geq \sigma_i(A_k[:, : l']) = \sigma_i(R_{11}R_{11}'^{-1}). \tag{4.20}$$

In the above, the relations for σ_j hold for $1 \leq j \leq \min(m,n) - k$. The relation for σ_i holds for $1 \leq i \leq k$. Also note that $\sigma_j(R_{11}R_{11}'^{-1})$ could be thought of as the singular values of A restricted to $im(AV_1)$.

Proof. The approximation loss in A_k is exactly the Schur complement $\mathcal{S}(\bar{A}_{11})$. To establish this, we first do some matrix algebra. In this algebra we will again recall the simplifying notation $X := UQ$ from Lemma **??**. Now to start, we have

$$\begin{aligned} A_k &= U^{-1} \begin{pmatrix} I \\ \bar{A}_{21}\bar{A}_{11}^{+} \end{pmatrix} \begin{pmatrix} \bar{A}_{11} & \bar{A}_{12} \end{pmatrix} V^{-1} \\ &= U^{-1} \begin{pmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{21}\bar{A}_{11}^{+}\bar{A}_{12} \end{pmatrix} V^{-1}. \end{aligned} \tag{4.21}$$

Next apply (**??**) from Lemma **??** to get $\mathcal{S}(\bar{A}_{11}) = \mathcal{S}(X_{11})R_{22}$. From this and the fact that $U^{-1}\bar{A}V^{-1} = A$,

$$\begin{aligned} A - A_k &= U^{-1} \begin{pmatrix} 0 & \\ & \mathcal{S}(\bar{A}_{11}) \end{pmatrix} V^{-1} \\ &= U'^{-1}L'^{-1} \begin{pmatrix} 0 & \\ & \mathcal{S}(X_{11})R_{22} \end{pmatrix} R'^{-1}V'^{-1} \\ &= U'^{T} \begin{pmatrix} 0 & \\ & \mathcal{S}(X_{11})R_{22} \end{pmatrix} V'^{T}. \end{aligned} \tag{4.22}$$

Now to get (**??**), recalling $U = L'U'$ with $L_{22}' = I$,

$$\begin{aligned} \|A_k - A\|_F^2 &= \|\mathcal{S}(X_{11})R_{22}\|_F^2 \\ &= \| \left[(UQ)_{22} - (UQ)_{21}((UQ)_{11})^{+}(UQ)_{12}\right] R_{22}\|_F^2 \\ &= \| \begin{pmatrix} (U'Q)_{21} & (U'Q)_{22} \end{pmatrix} \begin{pmatrix} -((UQ)_{11})^{+}((UQ)_{12})R_{22} \\ R_{22} \end{pmatrix} \|_F^2 \\ &\leq \|R_{22}\|_F^2 + \|X_{11}^{+}X_{12}R_{22}\|_F^2. \end{aligned}$$

And for (**??**), similar steps produce

$$\|A_k - A\|_2^2 \leq \| \begin{pmatrix} X_{11}^+ X_{12} R_{22} \\ R_{22} \end{pmatrix} \|_2^2 \leq \|X_{11}^+ X_{12} R_{22}\|_2^2 + \|R_{22}\|_2^2.$$

Even more generally, from the multiplicative Weyl inequality,

$$\sigma_j(A_k - A) \leq \sigma_j(\begin{pmatrix} -((UQ)_{11})^+((UQ)_{12})R_{22} \\ R_{22} \end{pmatrix}).$$

Using this, and the additive Weyl inequality [38] in the second inequality,

$$\begin{aligned}\sigma_{j+s-1}^2(A - A_k) &\leq \sigma_{j+s-1}^2(\begin{pmatrix} -((UQ)_{11})^+((UQ)_{12})(R_{22} - R_{22\text{opt},j-1} + R_{22\text{opt},j-1}) \\ R_{22} - R_{22\text{opt},j-1} + R_{22\text{opt},j-1} \end{pmatrix}) \\ &\leq \sigma_s^2(\begin{pmatrix} X_{11}^+ X_{12}(R_{22} - R_{22\text{opt},j-1}) \\ R_{22} - R_{22\text{opt},j-1} \end{pmatrix}).\end{aligned}$$

In particular, this establishes

$$\begin{aligned}\sigma_j^2(A - A_k) &\leq \sigma_1^2(\begin{pmatrix} X_{11}^+ X_{12}(R_{22} - R_{22\text{opt},j-1}) \\ R_{22} - R_{22\text{opt},j-1} \end{pmatrix}) \\ &\leq \|X_{11}^+ X_{12}(R_{22} - R_{22\text{opt},j-1})\|_2^2 + \|R_{22} - R_{22\text{opt},j-1}\|_2^2,\end{aligned}$$

and also by noting that the trailing $\min(m,n) - j$ singular values of $A - A_k$ are bound in this manner,

$$\begin{aligned}\|A - A_k - (A - A_k)_{\text{opt},j-1}\|_F^2 &\leq \| \begin{pmatrix} X_{11}^+ X_{12}(R_{22} - R_{22\text{opt},j-1}) \\ R_{22} - R_{22\text{opt},j-1} \end{pmatrix} \|_F^2 \\ &= \|X_{11}^+ X_{12}(R_{22} - R_{22\text{opt},j-1})\|_F^2 + \|R_{22} - R_{22\text{opt},j-1}\|_F^2.\end{aligned}$$

This completes (**??**)- (**??**). We proceed to the lower bound on $\sigma_i(A_k)$ claimed in (**??**). If we let $\bar{A}_k$ for the moment denote the middle matrix in (**??**), then

$$\begin{aligned}\sigma_i(A_k) = \sigma_i(L'^{-1}\bar{A}_k R'^{-1}) &\geq \sigma_i(\begin{pmatrix} (L'^{-1}\bar{A}_k R'^{-1})_{11} \\ (L'^{-1}\bar{A}_k R'^{-1})_{21} \end{pmatrix}) \\ &= \sigma_i(\begin{pmatrix} L_{11}'^{-1}\bar{A}_{11}R_{11}'^{-1} \\ \bar{A}_{21}R_{11}'^{-1} \end{pmatrix}) \\ &= \sigma_i(\begin{pmatrix} L_{11}'^{-1}[L_{11}'(U'Q)_{11}R_{11}]R_{11}'^{-1} \\ \bar{A}_{21}R_{11}'^{-1} \end{pmatrix}) \\ &= \sigma_i(\begin{pmatrix} (U'Q)_{11}R_{11}R_{11}'^{-1} \\ [X_{21}X_{11}^+X_{11}R_{11}]R_{11}'^{-1} \end{pmatrix}) \\ &= \sigma_i(\begin{pmatrix} (U'Q)_{11}R_{11}R_{11}'^{-1} \\ [(U'Q)_{21}R_{11}]R_{11}'^{-1} \end{pmatrix}) \\ &= \sigma_i(R_{11}R_{11}'^{-1}).\end{aligned} \tag{4.23}$$

Here we have used the identities of Lemma **??**. For (**??**) in particular, we used $\bar{A}_{21} = (\bar{A}_{21}\bar{A}_{11}^{+})(\bar{A}_{11})$ and then used (**??**) and (**??**) on the quantities in parentheses. □

Remark 48. *Recall the sizes $V_1 = V[:, :l]$, $U_1 = U[:l', :]$. When $l' = l$, the factorization in (**??**) can readily be rewritten in the more elegant form*

$$A_k = AV_1(U_1AV_1)^{-1}U_1A \tag{4.24}$$

*One nice feature of this is that only U_1, V_1 are actually needed to compute A_k. We will later see that the residual bounds in Proposition **??** can be computed with only U_1, V_1 so it makes sense that we can find an analog of (**??**) for $l' > l$. However, we actually need to set the rows $U_2 = U[l'+1:, :]$ to be a basis for the orthogonal complement of the rows of U_1 in order to achieve this. Then $U^{-1} = [U_1^{+}, U_2^{+}]$, and we get a different form of (**??**) that is often faster to compute,*

$$\begin{aligned}
A_k &= U^{-1}\begin{pmatrix} I \\ \bar{A}_{21}\bar{A}_{11}^{+} \end{pmatrix}\begin{pmatrix} \bar{A}_{11} & \bar{A}_{12} \end{pmatrix}V^{-1} \\
&= \begin{pmatrix} U_1^{+} & U_2^{+} \end{pmatrix}\begin{pmatrix} I \\ \bar{A}_{21}\bar{A}_{11}^{+} \end{pmatrix}U_1A \\
&= (U_1^{+} + U_2^{+}U_2AV_1(U_1AV_1)^{+})U_1A \\
&= \left[U_1^{+} + (I - U_1^{+}U_1)AV_1(U_1AV_1)^{+}\right][U_1A] \\
&= [U_1^{+}(I - (U_1AV_1)(U_1AV_1)^{+}) + (AV_1)(U_1AV_1)^{+}][U_1A]
\end{aligned} \tag{4.25}$$

This final form should be viewed as a generalized LU-factorization. The left factor is $m \times l'$ and the right factor (U_1A) is $l' \times n$. Also recall U_1 is $l' \times m$ so the pseudo-inverse can be cheaply computed.

We summarize the factorization discussed above in (only partially specified because of U, V and the oversampling parameters l, l') Algorithm **GLU** and Algorithm **RLU**. Recall that using square U, V was only to help with the theoretical guarantees. Therefore, in order to simplify notation, we now let U, V denote what were up until now labeled as U_1, V_1. We also emphasize Algorithm **RLU** is the special case of Algorithm **GLU** when the latter sets $l = l'$.

The bounds in Proposition **??** are not fully developed in that neither R_{22} nor $R_{11}R_{11}'^{-1}$ have been examined yet, and also the choice of U greatly influences X. In Section **??** the R_{22} factor will be studied; it will be bound in terms of S^TV where S is the right singular matrix of A. See Proposition **??** and the resulting Theorem **??**. Section **??** describes how choosing suitable random ensembles for U, V allows for the Frobenius norm of the residual to be arbitrarily close to that of the truncated SVD, as well as many other bounds. We therefore present what we consider to be our main results in Section **??**.

Algorithm 5 $[T, S] = \mathbf{GLU}(A, k)$. Generalized LU approximation computes a low-rank approximation $A \approx A_k = TS$, where T is a tall-skinny matrix and S is a short-wide matrix.

Input: target rank k, matrix $A \in \mathbb{R}^{m \times n}$
Output: $T \in \mathbb{R}^{m \times l'}$, $S \in \mathbb{R}^{l' \times n}$, where
Ensure: $T = U^+(I - (UAV)(UAV)^+) + (AV)(UAV)^+$, $S = UA$
Select oversampling parameters $l' \geq l \geq k$.
Generate full-rank $n \times l$ matrix V and full-rank $l' \times m$ matrix U.
$\hat{A} = UAV$
$T_1 = U^+(I - \hat{A}\hat{A}^+)$
$T_2 = AV$
$T_2 = T_2\hat{A}^+$
$T = T_1 + T_2$
$S = UA$

Algorithm 6 $[T, \hat{A}, S] = \mathbf{RLU}(A)$. Rank-revealing LU computes a low-rank approximation $A \approx A_k = T\hat{A}^{-1}S$, where T is a tall-skinny matrix, S is a short-wide matrix, and $\hat{A}$ is a small dense matrix.

Input: target rank k, matrix $A \in \mathbb{R}^{m \times n}$
Output: $T \in \mathbb{R}^{m \times l}$, $S \in \mathbb{R}^{l \times n}$, $\hat{A} \in \mathbb{R}^{l \times l}$
Ensure: $T = AV$, $S = UA$, $\hat{A} = UAV$
Select oversampling parameter $l \geq k$.
Generate a full-rank $n \times l$ matrix V and a full-rank $l \times m$ matrix U.
$T = AV$
$S = UA$
$\hat{A} = UT$

4.3 Relationship to other Approaches

In this section we illustrate how **GLU** provides a general framework by proving the equivalence with Algorithm **PRR_RLU** and Algorithm **RQR** below. We will also see a close connection to the popular approach Algorithm **CW**, from Clarkson and Woodruff [21]. We show show our approach is strictly more accurate when $l' > l$, and the same when $l' = l$.

This version of the randomized SVD is described in section 5.2 of [37]. It additionally extracts rows from A based on the product AV, leading to a speedup in many settings but at the cost of approximation quality. See discussion in [37] around (5.3).

Algorithm 7 $[T, S] = \mathbf{RQR}(A, k)$. Randomized QR approximation computes a low-rank approximation $A \approx A_k = TS$ where T is a tall-skinny matrix with orthonormal columns, and S is a short-wide matrix

1: **Input:** target rank k, matrix $A \in \mathbb{R}^{m\times n}$
2: **Output:** orthogonal matrix $T \in \mathbb{R}^{m\times l}$, matrix $S \in l \times n$
3: **Ensure:** T has orthonormal columns, $S = T^T A$
4: Select the oversampling parameter $l \geq k$.
5: Generate a full rank $n \times l$ matrix V.
6: $\hat{A} = AV$.
7: $[T, \_] = \mathbf{tQR}(\hat{A})$.
8: $S = T^T A$

Algorithm 8 $[T, S] = \mathbf{PRR\_RLU}(A, k)$. Randomized LU with row selection approximation based on Panel Rank-Revealing computes a randomized LU-factorization $A_k = TS$, performing sketching on the columns and a panel rank-revealing QR to select rows

1: **Input:** target rank k, matrix $A \in \mathbb{R}^{m\times n}$
2: **Output:** $T \in \mathbb{R}^{m\times l}$ and $S \in \mathbb{R}^{l\times n}$
3: **Ensure:** $T = AV(P_1AV)^{-1} = Q(P_1Q)^{-1}$, $S = P_1A \in \mathbb{R}^{l\times n}$ where P is a permutation matrix and $P_1 = P[:l, :]$ and $Q \in \mathbb{R}^{m\times l}$ has orthonormal columns
4: Select oversampling parameter $l \geq k$
5: Generate a full-rank $n \times l$ random matrix V
6: $[Q, R] = \mathbf{tQR}(AV)$.
7: Permutation P is selected so that $PQ = \bar{Q} = \begin{pmatrix} \bar{Q}_{11} \\ \bar{Q}_{21} \end{pmatrix}$ results in $||\bar{Q}_{21}\bar{Q}_{11}^{-1}||_{max}$ being bounded by a small constant (see [47]). Here $P_1 = P[:l, :]$ and $\bar{Q}_{11} = P_1Q$.
8: $T = P^T \begin{pmatrix} I \\ \bar{Q}_{21}\bar{Q}_{11}^{-1} \end{pmatrix}$, also note then $T = AV(P_1AV)^{-1}$.
9: $S = P_1A$

We show how these algorithms fit into the LU-framework. The fact is simple, but it appears to have been overlooked in the literature. Therefore it has its own proposition:

Proposition 49. ***PRR_RLU*** *is equivalent to* ***RLU*** *when the latter chooses the same V and $U := P_1$.*

RQR *is equivalent to* ***RLU*** *when the latter chooses the same V and $U := T^T$.*

Proof. The proof is mainly to recall the various definitions. First, Algorithm **PRR_RLU** produces $A_k = Q(P_1Q)^{-1}P_1A$. As claimed within the algorithm, because $QR = AV$ it follows that $AV(P_1AV)^{-1}P_1A = QR(P_1QR)^{-1}P_1A = A_k$. As $AV(P_1AV)^{-1}P_1A$ is the output

factorization of Algorithm **RLU**, the factorizations agree.

We move on to the other equivalence. Recall $[T, R] = \mathbf{tQR}(AV)$. Selecting the same V and $U = T^T$, the random LU-approximation given by Algorithm **RLU** would be

$$AV(T^TAV)^{-1}T^TA = TR(T^TTR)^{-1}T^TA = TT^TA \tag{4.26}$$

which agrees with Algorithm **RQR**. □

The most popular approach involving sketching from the left and right is perhaps the method first introduced in [21], and also described in the overview [65]. It is not equivalent to **GLU**, but it is still closely related as we point out now. As seen in Theorem 47 of [65], the output of what we call Algorithm **CW** after Clarkson, Woodruff is

$$A \approx A'_k = AV_1(U_1AV_1)^+U_1A \tag{4.27}$$

where we take U, V in the expanded form as below (**??**). We now show that this procedure is strictly less accurate than **GLU** when $l' > l$, and the same when $l' = l$.

Proposition 50. *Let $\bar{A} = UAV$ with $U = L'U'$ and $V = V'R'$ as in Proposition* **??**. *Additionally set $\tilde{A} = U_1A$, and let B be the projection of $\tilde{A}$ onto the orthogonal complement of $\tilde{A}V_1$. Finally let A_k be the output of Algorithm* ***GLU*** *and A'_k be the output (**??**) of Algorithm* ***CW***. *Then*

$$\begin{aligned}
\|A - A'_k\|_F^2 &= \|A - A_k\|_F^2 + \|A_k - A'_k\|_F^2 \\
\|A_k - A'_k\|_F^2 &= \|U_1^+B\|_F^2 \\
\|A - A'_k\|_2^2 &\leq \|A - A_k\|_2^2 + \|U_1^+B\|_2^2
\end{aligned}$$

Proof. Similar to Remark **??**,

$$\begin{aligned}
A'_k = AV_1(U_1AV_1)^+U_1A &= U^{-1}\begin{pmatrix} \bar{A}_{11} \\ \bar{A}_{21} \end{pmatrix}\bar{A}_{11}^+\begin{pmatrix} \bar{A}_{11} & \bar{A}_{12} \end{pmatrix}V^{-1} \\
&= U^{-1}\begin{pmatrix} \bar{A}_{11} & \bar{A}_{11}\bar{A}_{11}^+\bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} - \mathcal{S}(\bar{A}_{11}) \end{pmatrix}V^{-1}
\end{aligned}$$

From this calculation and the calculation leading to (**??**), it follows that

$$\begin{aligned}
\|A - A'_k\|_F^2 &= \|U^{-1}\left[\bar{A} - \begin{pmatrix} \bar{A}_{11} & \bar{A}_{11}\bar{A}_{11}^+\bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} - \mathcal{S}(\bar{A}_{11}) \end{pmatrix}\right]V^{-1}\|_F^2 \\
&= \|\begin{pmatrix} 0 & U_1^+(I - \bar{A}_{11}\bar{A}_{11}^+)\bar{A}_{12} \\ 0 & \mathcal{S}(\bar{A}_{11}) \end{pmatrix}\|_F^2 \\
&= \|A - A_k\|_F^2 + \|U_1^+(I - \bar{A}_{11}\bar{A}_{11}^+)\bar{A}_{12}\|_F^2 \\
&= \|A - A_k\|_F^2 + \|U_1^+(I - \tilde{A}V_1(\tilde{A}V_1)^+)\tilde{A}V_2\|_F^2 \\
&= \|A - A_k\|_F^2 + \|U_1^+(I - \tilde{A}V_1(\tilde{A}V_1)^+)\begin{pmatrix} \tilde{A}V'_1 & \tilde{A}V'_2 \end{pmatrix}\|_F^2 \\
&= \|A - A_k\|_F^2 + \|U_1^+(I - \tilde{A}V_1(\tilde{A}V_1)^+)\tilde{A}\|_F^2 \\
&= \|A - A_k\|_F^2 + \|U_1^+B\|_F^2
\end{aligned}$$

This gives the Frobenius norm claims. Compare this with the similar (**??**). Repeating the similar steps gives the spectral norm claim. The only difference comes from an inequality instead of an equality in one step; for any unit vector x,

$$\begin{aligned}\| \begin{pmatrix} U_1^+(I - \bar{A}_{11}\bar{A}_{11}^+)\bar{A}_{12} \\ \mathcal{S}(\bar{A}_{11}) \end{pmatrix} x\|_2^2 &= \| \begin{pmatrix} U_1^+(I - \bar{A}_{11}\bar{A}_{11}^+)\bar{A}_{12}x \\ \mathcal{S}(\bar{A}_{11})x \end{pmatrix} \|_2^2 \\ &= \|U_1^+(I - \bar{A}_{11}\bar{A}_{11}^+)\bar{A}_{12}x\|_2^2 + \|\mathcal{S}(\bar{A}_{11})x\|_2^2 \\ &\leq \|\mathcal{S}(\bar{A}_{11})\|_2^2 + \|U_1^+(I - \bar{A}_{11}\bar{A}_{11}^+)\bar{A}_{12}\|_2^2\end{aligned}$$

□

The work in [21] only considered the properties of the factorization (**??**) in the context of Johnson-Lindenstrauss transforms, specifically when l' is a poly-log factor larger than l, and not focusing on deterministic bounds. If we compare the factorizations directly, perhaps the most obvious difference is that the output of Algorithm **GLU** is typically rank l' whereas the output of **CW** is rank l. Related to this, the factorization in **CW** may be slightly cheaper to perform, although in typical settings (k is relatively small) the same term dominants the cost of both algorithms. Specializing to the SRHT ensemble for which our results are strongest, in Remark **??** we will see that the same bounds in Definition **??** and Definition **??** apply to both **CW** and to **GLU**. However, it does not appear the case that Theorem **??** 's strong bound on Def. **??** can be carried over.

4.4 QR Deterministic Bounds

The following lemma is important in randomized low rank approximation results. Our proof is novel, and (**??**), (**??**) significantly generalize past versions.

Proposition 51. *Let A be an $m \times n$ matrix with $[P, \Sigma, S] = \boldsymbol{SVD}(A)$. As with Proposition* **??**, *it is again convenient to suppose V is $n \times n$ with $V = V'R'$ as described in (***??***). Also let $[Q, R] = \boldsymbol{QR}(AV)$. Then block Q, R, S^TV, Σ using $Q_1 := Q[:, : l]$, $R_{11} = R[: l, : l]$, $(S^TV)_{11} = (S^TV)[: k, : l]$, and $\Sigma_1 = \Sigma[: k, : k]$, $\Sigma_2 = \Sigma[k+1 :, k+1 :]$. Then the singular values of $Q_1Q_1^TA - A$ are identical to those of R_{22}, i.e. for any $1 \leq j \leq \min(m,n) - l$*

$$\sigma_j(Q_1Q_1^TA - A) = \sigma_j(R_{22}).$$

Moreover, assuming $(S^TV)_{11}$ has full row-rank (and therefore $k \leq l$), we have that

$$\|Q_1Q_1^TA - A\|_F^2 \leq \|\Sigma_2\|_F^2 + \|\Sigma_2(S^TV)_{21}(S^TV)_{11}^+\|_F^2 \tag{4.28}$$

$$\|Q_1Q_1^TA - A\|_2^2 \leq \sigma_{k+1}^2 + \|\Sigma_2(S^TV)_{21}(S^TV)_{11}^+\|_2^2. \tag{4.29}$$

We may generalize this last equation with the goal of covering Definition **??**. *For any $1 \leq j \leq \min(m,n) - l$, there exists an $n \times (n - j + 1)$ orthogonal matrix $\tilde{S}$ independent of V,*

satisfying

$$\sigma_j(Q_1Q_1^TA - A)_2^2 \leq \sigma_{j+k}^2 + \|\Sigma_{j,2}(\tilde{S}^TV)_{21}(\tilde{S}^TV)_{11}^+\|_2^2 \tag{4.30}$$

$$\|(Q_1Q_1^TA - A) - (Q_1Q_1^TA - A)_{opt,j-1}\|_F^2 \leq \|\Sigma_{j,2}\|_F^2 + \|\Sigma_{j,2}(\tilde{S}^TV)_{21}(\tilde{S}^TV)_{11}^+\|_F^2 \tag{4.31}$$

with $(\tilde{S}^TV)_{11}$ *being* $k \times l$ *as before, and* $\Sigma_{j,2} := \boldsymbol{diag}(\sigma_{k+j}, \ldots, \sigma_{\min(m,n)}, 0, \ldots, 0)$ *is of dimension* $(m-k) \times (n-k)$*, where* ***diag*** *denotes the diagonal matrix.*

Proof. We first observe by direct computation,

$$\sigma_j(Q_1Q_1^TA - A) = \sigma_j(Q_2Q_2^TA) = \sigma_j(Q_2 \begin{pmatrix} 0 & R_{22} \end{pmatrix} R'^{-1}V'^T) = \sigma_j(R_{22}),$$

to establish the first claim. Next we invoke the common fact that for the spectral and Frobenius norms, $Q_1Q_1^TA$ is the best approximation to A whose columns are in $\mathrm{im}(Q_1)$. For example, one can check that $Q_1Q_1^T$ satisfies the orthogonal projection properties with respect to these norms. Set $\bar{A} = P^TAV$. Then we explicitly propose an approximation $\tilde{A}_k$ whose columns are contained in $\mathrm{im}(Q_1) = \mathrm{im}(AV_1)$, namely

$$\tilde{A}_k := P \begin{pmatrix} \bar{A}_{11} \\ \bar{A}_{12} \end{pmatrix} \begin{pmatrix} I & \bar{A}_{11}^+\bar{A}_{12} \end{pmatrix} V^{-1} = AV_1 \begin{pmatrix} I & \bar{A}_{11}^+\bar{A}_{12} \end{pmatrix} V^{-1}$$

In contrast to before, $\bar{A}_{11}$ is $k \times l$, making it short and wide. Repeating the algebra around (**??**) in the first step,

$$\begin{aligned} \|A - \tilde{A}_k\|_F^2 = \|\mathcal{S}((P^TAV)_{11})\|_F^2 &= \|\mathcal{S}(\Sigma_1(S^TV)_{11})\|_F^2 \\ &= \|\Sigma_2(S^TV)_{22} - \Sigma_2(S^TV)_{21}(\Sigma_1(S^TV)_{11})^+\Sigma_1(S^TV)_{12}\|_F^2 \\ &= \|\Sigma_2(S^TV)_{22} - \Sigma_2(S^TV)_{21}(S^TV)_{11}^+(S^TV)_{12}\|_F^2 \qquad (4.32) \\ &\leq \| \begin{pmatrix} \Sigma_2 & -\Sigma_2(S^TV)_{21}(S^TV)_{11}^+ \end{pmatrix} \|_F^2 \\ &= \|\Sigma_2\|_F^2 + \|\Sigma_2(S^TV)_{21}(S^TV)_{11}^+\|_F^2. \end{aligned}$$

To be clear, S^TV was blocked so that $(S^TV)_{11}$ is $k \times l$. Note we were able to distribute the pseudo-inverse in $(\Sigma_1(S^TV)_{11})^+$. In the generic case this follows from $(S^TV)_{11}$ having full row rank (this will be with probability 1 for suitably random V) and Σ_1 being invertible. If Σ_1 has trailing 0 values, the assumption of full row rank of $(S^TV)_{11}$ ensures $\mathrm{im}(AV_1) = \mathrm{im}(A)$ and therefore we can instead use $\tilde{A}_k := A$ to get the bound of 0.

For the spectral norm bound, the steps are the same, except as in the proof in Proposition **??**, the final equality becomes an inequality.

We actually are interested in the lower singular values as well though, so we extend the proof. In the following, P_Y and P_{AV_1} project onto the complements of the images of Y and AV_1 respectively, Y is rank $j-1$, and AV_1 is rank l. The same projection notation applies to the other projections. Using the additive Weyl inequality in the inequality step, similar to the use within Prop **??**, for $s \leq \min(m,n) - j$,

Additionally, in the third equality, $Y + AV_1$ is used to refer to direct sum of the images of Y and AV_1, and the equality holds under the assumption these spaces are orthogonal. The fourth (last) equality holds when $\mathrm{im}(Y') \oplus \mathrm{im}(\tilde{Y}) = \mathrm{im}(Y) \oplus \mathrm{im}(AV_1)$, and $\mathrm{im}(Y')$ is orthogonal to $\mathrm{im}(\tilde{Y})$.

Now we make our choice of $Y + AV_1$. First let $P_1 = P[:, : j-1]$, the leading $j-1$ left singular vectors of A. Noting that $\mathrm{im}(P_1) \oplus \mathrm{im}((P_1 P_1^T A - A)V_1)$ is rank $l+j-1$ and contains $\mathrm{im}(Q_1) = \mathrm{im}(AV_1)$, and that P_1 is orthogonal to $(A - P_1 P_1^T A)V_1$, these are valid choices of Y' and $\tilde{Y}$ respectively. In summary,

$$\begin{aligned} \mathrm{im}(Y) \oplus \mathrm{im}(AV_1) = \mathrm{im}(P_1) \oplus \mathrm{im}((P_1 P_1^T A - A)V_1) &= \mathrm{im}(P_1) \oplus \mathrm{im}(AV_1) \qquad (4.33) \\ Y &= \text{ trailing j-1 columns of Q factor of tQR}(\begin{pmatrix} AV_1 & P_1 \end{pmatrix}) \\ Y' &= P_1 \\ \tilde{Y} &= P_{P_1} AV_1 \end{aligned}$$

We emphasize in (**??**) that the first two are orthogonal direct sums. This puts us essentially back into the situation surrounding (**??**). Indeed,

$$\sigma_{j+s-1}(Q_1 Q_1^T A - A) \leq \sigma_s(P_{\tilde{Y}} P_{Y'} A) = \sigma_s(P_{BV_1} B) = \sigma_s(B - \tilde{Q}_1 \tilde{Q}_1^T B),$$

where $B = P_{P_1} A = A - P_1 P_1^T A = A - A_{\mathrm{opt},j-1}$, and $\tilde{Q}_1$ is an orthogonal matrix such that $\mathrm{im}(\tilde{Q}_1) = \mathrm{im}(BV_1) = \mathrm{im}((A - A_{\mathrm{opt},j-1})V_1)$. In particular with $s = 1$,

$$\sigma_j(Q_1 Q_1^T A - A) \leq \sigma_1(\tilde{Q}_1 \tilde{Q}_1^T (A - A_{\mathrm{opt},j-1}) - (A - A_{\mathrm{opt},j-1})),$$

as well as by comparing the singular values individually by varying s,

$$\|(Q_1 Q_1^T A - A) - (Q_1 Q_1^T A - A)_{\mathrm{opt},j-1}\|_F^2 \leq \|\tilde{Q}_1 \tilde{Q}_1^T (A - A_{\mathrm{opt},j-1}) - (A - A_{\mathrm{opt},j-1})\|_F^2.$$

As a result, the RHS's we need to bound are the same as those bound when we established (**??**), and we may carry out the same steps as those around (**??**). The only change is A is replaced with $B = A - A_{\mathrm{opt},j-1}$, and accordingly Q_1 changes to have the same image as BV_1. The effect of this is the order of right singular vector matrix S changes; the leading $j-1$ singular values and singular vectors removed. To capture this change, we may notationally let $\tilde{S}$ be the permutation of the columns of S, moving the leading $j-1$ columns to the end. Then in the spectral case with $s = 1$,

$$\sigma_j^2(Q_1 Q_1^T A - A) \leq \sigma_{j+k}^2 + \|\Sigma_{j,2}(\tilde{S}^T V)_{21}(\tilde{S}^T V)_{11}^+\|_2^2.$$

The Frobenius norm version follows similarly. □

In (**??**) and (**??**), one could factor out the σ part to make the equations immediately take the form of Definitions **??** and **??**. However, as in Theorem **??**, the unfactored form can have advantages.

Lemma 52. *Continue in the situation of Proposition ??. For $j \leq k$,*

$$\sigma_j(Q_1 Q_1^T A) \geq \sigma_j(R_{11} R_{11}'^{-1}) \geq \sigma_j(A)\sigma_{min}((S^T V')_{11}))$$

Proof. As in the previous proof, we see that the result is the same as if we right multiplied by V' rather than V. That is, we seek to bound from below the j-th singular value of

$$Q_1^T A = \begin{pmatrix} R_{11} & R_{12} \end{pmatrix} R'^{-1} V'^T.$$

Using this expression, we see that

$$\sigma_j(Q_1 Q_1^T A) \geq \sigma_j(R_{11} R_{11}'^{-1}) = \sigma_j(\Sigma(S^T V')[:, : l]) \geq \sigma_j(A)\sigma_{min}((S^T V')_{11}),$$

where the reversed Weyl inequality was used in the last step. □

We state the following mainly to collect the results of the section into a single statement which resembles the definitions of strong QR factorizations from the literature.

Proposition 53. *Let A be an $m \times n$ matrix with SVD $A = P\Sigma S^T$. Set $[Q, R] = \boldsymbol{QR}(AV)$, where $V = V'R'$ is an $n \times n$ matrix. Then the singular values of $Q_1 Q_1^T A - A$ are identical to those of $(m-l) \times (n-l)$ matrix R_{22}. Moreover,*

$$\|R_{22}\|_F^2 \leq \|\Sigma_2\|_F^2 + \|\Sigma_2 (S^T V)_{21} (S^T V)_{11}^+\|_F^2$$

Also for $j \leq k$,

$$\sigma_j(A) \geq \sigma_j(Q_1 Q_1^T A) \geq \sigma_j(R_{11} R_{11}'^{-1}) \geq \sigma_j(A)\sigma_{\min}((S^T V')_{11}) \tag{4.34}$$

as well as for any given $j \leq \min(m, n) - k$, there is an orthogonal $n \times (n-j)$ matrix $\tilde{S}$ independent of V such that

$$\sigma_j^2(R_{22}) \leq \sigma_{j+k}^2 + \|\Sigma_{j,2} (\tilde{S}^T V)_{21} (\tilde{S}^T V)_{11}^+\|_2^2 \tag{4.35}$$

$$\|(R_{22}) - (R_{22})_{opt,j-1}\|_F^2 \leq \|\Sigma_{j,2}\|_F^2 + \|\Sigma_{j,2} (\tilde{S}^T V)_{21} (\tilde{S}^T V)_{11}^+\|_F^2, \tag{4.36}$$

*with $(\tilde{S}^T V)_{11}$ being $k \times l$ as before, and $\Sigma_{j,2} := \boldsymbol{diag}(\sigma_{k+j}, \ldots, \sigma_n, 0, \ldots, 0)$ is of dimension $(m-k) \times (n-k)$, where **diag** denotes the diagonal matrix.*

Proof. Excluding the upper bound in (??), the bounds are restatements of facts in Proposition ?? and Lemma ??. The upper bound is a consequence of the Weyl inequality,

$$\sigma_j(Q_1 Q_1^T A) \leq \sigma_1(Q_1 Q_1^T)\sigma_j(A) = \sigma_j(A)$$

□

4.5 Application of Randomness

In this section, we combine our deterministic bounds with the past literature on sketching matrices. There are three applications. We first note that ensembles U and V used in Algorithm **GLU**'s guarantees in Proposition **??** can be viewed through the oblivious subspace embedding property commonly used in literature. Second, we specialize the random ensemble to the subsampled randomized Hadamard transform (SRHT) introduced in [58] but whose analysis was strengthened in [12]. Our approach fits nicely with their work to give particularly strong operator norm bounds, but in asymptotically less time. Third, we specialize to the Gaussian ensemble to see an application to analyzing the growth factor in Gaussian Elimination.

We begin by recalling a property associated with Johnson-Lindenstrauss embeddings. Different authors establish it in different ways, as in [58], [12], [21], but they all have found it necessary in providing sharp Frobenius bounds.

Definition 54. *We say U from $\mathbb{R}^n$ to $\mathbb{R}^s$ is (ϵ, δ, n) multiplication approximating if for any A, B having n rows, then*

$$\|A^T U^T U B - A^T B\|_F^2 \leq \epsilon \|A\|_F^2 \|B\|_F^2,$$

with probability $1 - \delta$.

We also include a definition used consistently in the literature,

Definition 55. *An (k, ϵ, δ) oblivious subspace embedding (OSE) from $\mathbb{R}^n$ to $\mathbb{R}^s$ is a distribution $U \sim \mathbb{D}$ over $s \times n$ matrices. It must with probability $1 - \delta$ succeed in making*

$$1 - \epsilon \leq \sigma_{\min}^2(UQ) \leq \sigma_{\max}^2(UQ) \leq 1 + \epsilon$$

hold for any given orthogonal $n \times k$ matrix Q. We will assume $l \geq k$ and $\epsilon < 1/6$.

For Definition **??**, there is a consequence we require. The first part is essentially Lemma 4.1 of [12] but we need to state it more generally.

Lemma 56. *Let U be an $s \times n$ matrix that is a (k, ϵ, δ) OSE from $\mathbb{R}^n$ to $\mathbb{R}^s$, and Q be an $(n \times k)$ orthogonal matrix. Provided $\epsilon < 1/6$, then with probability $1 - \delta$ both of the following hold,*

$$\|(UQ)^+ - (UQ)^T\|_2 \leq 3\epsilon$$
$$\|U\|_2^2 = O\left(\frac{n}{k}\right),$$

where in the second of these we require the additional assumption $\delta > 2e^{-k/5}$.

Proof. Let $A = UQ$. Then from Definition **??** and power series expansion, the singular values of A lie within $[\sqrt{1-\epsilon}, \sqrt{1+\epsilon}]$ and hence for simplicity we may say they lie within $[1-\epsilon, 1+\epsilon]$ with probability $1-\delta$. Let $l \times k$ diagonal matrix Σ contain these singular values. Therefore

$$\begin{aligned}\|A^+ - A^T\|_2 = \|\Sigma^T - \Sigma^+\|_2 &= \max_{i \leq k} |\lambda_i - \lambda_i^{-1}| \\ &\leq |1 - \epsilon - \frac{1}{1-\epsilon}| \leq 3\epsilon,\end{aligned}$$

where we have chosen to write the small extreme of σ_i; the large extreme is identical.

For the second fact, let $V \leq \mathbb{R}^n$ be a uniformly distributed k-dimensional subspace with $\dim(V) = k$ independent of U, i.e. V is spanned by the first k columns of a Haar distributed matrix on $\mathbb{R}^n$ independent of U. A consequence of Definition **??** is that $\|Uv\|_2 \leq 2$ with probability $1-\delta$ holding uniformly for unit vectors v contained in V. Otherwise some fixed subspace V_0 would also fail to have this property with probability δ, violating Definition **??**.

Now let x be the maximal right singular vector of U. The subsequent Lemma **??** gives $\sup_{v \in V, \|v\|_2 = 1} |\langle x, v \rangle| = \Omega(\sqrt{\frac{k}{n}})$ with probability $1-\delta$. Next choose $v \in \text{argmax}_{v \in V, \|v\|_2 = 1} |\langle x, v \rangle|$ to be a unit-vector with smallest angle with respect to x, and observe $\|Uv\|_2 = \Omega(\sqrt{\frac{k}{n}})\|Ux\|_2$. We conclude $\|U\|_2^2 = O\left(\frac{n}{k}\right)$ with probability $1-\delta$. Otherwise this would contradict $\|Uv\|_2 \leq 2$ holding with probability $1-\delta$. □

Lemma 57. *Let V be a k-dimensional uniformly distributed subspace of $\mathbb{R}^n$, and $x \in \mathbb{R}^n$ be a unit vector drawn from a distribution independent of V. Then $\sup_{v \in V, \|v\|_2 = 1} |\langle x, v \rangle| = \Omega(\sqrt{\frac{k}{n}})$ with probability $1 - 2e^{-k/5}$.*

Proof. We may assume $V = \text{span}(e_1, \ldots, e_k)$, and represent x as $\frac{(X_1, \ldots, X_n)^T}{\sqrt{X_1^2 + \cdots + X_n^2}}$ where X_i are i.i.d. variance $\frac{1}{n}$ Gaussians. Indeed, V can be taken to be the first k columns of Haar distributed orthogonal matrix $\tilde{V}$, and the WLOG assumption is equivalent to changing to the coordinates of $\tilde{V}$. As a result, we are interested in

$$\sup_{v \in V, \|v\|_2 = 1} |\langle x, v \rangle| = \frac{(X_1, \ldots, X_n)}{\sqrt{X_1^2 + \cdots + X_n^2}} \cdot \frac{(X_1, \ldots X_k, 0, \ldots)^T}{\sqrt{X_1^2 + \cdots + X_k^2}} = \frac{\sqrt{X_1^2 + \cdots + X_k^2}}{\sqrt{X_1^2 + \cdots + X_n^2}}.$$

Standard large-deviation bounds for chi-squared distribution, which is a sub-exponential random variable, can be used to lower bound this. We take bounds from [44] (4.3), (4.4). The right tail bound is

$$\mathbb{P}[X_1^2 + \cdots + X_n^2 > 1 + 2\frac{\sqrt{\delta}}{\sqrt{n}} + 2\frac{\delta}{n}] \leq e^{-\delta},$$

and the left tail bound is

$$\mathbb{P}[X_1^2 + \cdots + X_k^2 < \frac{k}{n} - 2\frac{\sqrt{k\delta}}{n}] \leq e^{-\delta}.$$

From these and setting $\delta = k/5$, we conclude

$$\sup_{v \in V, \|v\|_2 = 1} |\langle x, v \rangle| \geq \left(\frac{\frac{k}{n} - 2\frac{k}{\sqrt{5}n}}{1 + 2\frac{\sqrt{k}}{\sqrt{5n}} + 2\frac{k}{5n}} \right)^{.5} \geq \frac{1}{25}\sqrt{\frac{k}{n}}$$

holds with probability $1 - 2e^{-k/5}$. □

The following lemma largely follows the steps of [12] but we have tried to abstract out the key probabilistic properties responsible in order to be more general. Another difference is that we also treat the spectral norm. It is a natural consequence of the prior lemmas, and will bridge the gap between deterministic Proposition **??** and randomized Theorem **??**. We do not attempt to tightly bound the constant coefficients.

Lemma 58. *Assume $l \times m$ matrix U is drawn from a distribution that is a $(k, \sqrt{\epsilon}, \delta)$ OSE from $\mathbb{R}^m$ to $\mathbb{R}^l$. Let B be a fixed $(m-k) \times n$ matrix, and $Q = [Q_1, Q_2]$ be a fixed orthogonal $m \times m$ matrix blocked so that Q_1 is $m \times k$. Then provided $\delta > 2e^{-k/5}$, with probability $1 - \delta$*

$$\|(UQ_1)^+(UQ_2)A\|_2 = O\left(\frac{m}{k}\right).$$

Further assume U is $(\frac{\epsilon}{k}, \delta, n)$ multiplication approximating, then with probability at least $1 - 2\delta$,

$$\|(UQ_1)^+(UQ_2)B\|_F^2 = O(\epsilon)\|B\|_F^2.$$

Proof. For the Frobenius bound, apply Lemma **??** in (**??**), and Definition **??** in (**??**) by noting $Q_2^T Q_1 = 0$,

$$\begin{aligned}
\|(UQ_1)^+(UQ_2)B\|_F^2 &\leq 2\|(UQ_1)^T(UQ_2)B\|_F^2 + 2\|((UQ_1)^+ - (UQ_1)^T)(UQ_2)B\|_F^2 \\
&\leq 2\|Q_1^T U^T U Q_2 B\|_F^2 + 6\epsilon\|UQ_2B\|_F^2 \qquad (4.37)\\
&\leq 2\|Q_1^T U^T U Q_2 B\|_F^2 + 12\epsilon\|Q_2B\|_F^2 \qquad (4.38)\\
&\leq 2\|Q_1^T U^T U Q_2 B\|_F^2 + 12\epsilon\|B\|_F^2 \\
&\leq 2\frac{\epsilon}{k}\|Q_2B\|_F^2\|Q_1^T\|_F^2 + 12\epsilon\|B\|_F^2 \qquad (4.39)\\
&\leq 2\epsilon\|B\|_F^2 + 12\epsilon\|B\|_F^2 = 14\epsilon\|B\|_F^2.
\end{aligned}$$

In the above, the step to (**??**) used Definition **??**, noting

$$\|Q_2BU\|_F^2 = \|Q_2BU(Q_2BU)^T\|_F \leq (1+\epsilon)\|Q_2B\|_F^2 \leq\leq 2\|Q_2B\|_F^2,$$

with probability $1-\delta$.

For the spectral bound, we may argue

$$\begin{aligned}
\|(UQ_1)^+(UQ_2)A\|_2^2 &\leq \|(UQ_1)^+(UQ_2)\|_2^2 \cdot \|A\|_2^2 \\
&\leq \|(UQ_1)^T(UQ_2)\|_2^2 \cdot \|A\|_2^2 + \|((UQ_1)^+ - (UQ_1)^T)(UQ_2)\|_2^2 \cdot \|A\|_2^2 \\
&\leq \frac{7}{6}\|UQ_2\|_2^2 \cdot \|A\|_2^2 + 3\epsilon\|(UQ_2)\|_2^2 \cdot \|A\|_2^2 \qquad (4.40) \\
&= \frac{7}{6}\|U\|_2^2 \cdot \|Q_2\|_2^2 \cdot \|A\|_2^2 \qquad (4.41) \\
&= O(\frac{m}{k})\|A\|_2^2 \,.
\end{aligned}$$

In the former steps, we note in particular that (**??**) follows from Definition **??**, and **??** from Lemma **??**. □

In the following, one of our main results, we continue with the notation of Propositions **??** and **??**. We provide a bound on Definitions **??** and **??**. While these bounds do appear quite weak (often weaker than a naive Frobenius norm adaptation), we note that they match the guarantees of past literature for algorithms running in $o(nmk)$ time, e.g. [34], [37], [59]. On the other hand, in Theorem **??** we notably achieve very sharp bounds on Definitions **??** and **??**, by exploiting a special property of the SRHT ensemble.

Theorem 59. *Assume U_1 is drawn from a distribution that is an (l, ϵ, δ) OSE from $\mathbb{R}^m$ into $\mathbb{R}^{l'}$. Similarly assume V_1^T is drawn from a distribution that is a (k, ϵ, δ) OSE from $\mathbb{R}^n$ into $\mathbb{R}^l$. Then provided $\delta > 2e^{-k/5}$, with probability $1-2\delta$ for $j \leq k$,*

$$\sigma_j(A_k) = \Omega\left(\sqrt{\frac{k}{n}}\right)\sigma_j(A)\,.$$

Fixing a given $1 \leq j \leq \min(m,n) - k$, with probability $1-4\delta$ we also have

$$\sigma_j(A - A_k) = O\left(\sqrt{\frac{mn}{kl}}\right)\sigma_{k+j}(A).$$

If we additionally assume U_1 is drawn from a $(\sqrt{\frac{\epsilon}{l}}, \delta, m)$ multiplication approximating and similarly V_1^T is drawn from a $(\sqrt{\frac{\epsilon}{k}}, \delta, n)$ multiplication approximating, then for a given $1 \leq j \leq \min(m,n) - k$,

$$\|A - A_k\|_F^2 = (1 + O(\epsilon))\,\|A - A_{opt,k}\|_F^2$$

holds with probability $1-4\delta$.

Proof. We start with the Frobenius norm bound. The starting point is Proposition **??**, which includes

$$\|R_{22} - R_{22\text{opt},j-1}\|_F^2 \leq \|\Sigma_{j,2}\|_F^2 + \|\Sigma_{j,2}(\tilde{S}^T V)_{21}(\tilde{S}^T V)_{11}^+\|_F^2.$$

Then as V_1^T satisfies the JL properties, apply Lemma ?? with $B = \Sigma_{j,2}^T$, $Q_1 = \tilde{S}_1$, $Q_2 = \tilde{S}_2$, and $U = V_1^T$, to conclude that for a given $1 \leq j \leq \min(m,n) - k$, $\|R_{22} - R_{22\text{opt},j-1}\|_F^2 = (1 + O(\epsilon))\|\Sigma_{j,2}\|_F^2$ with probability $1 - 2\delta$. To complete the Frobenius bound, recall from Proposition ?? that

$$\|(A - A_k) - (A - A_k)_{\text{opt},j-1}\|_F^2 \leq \|R_{22} - R_{22\text{opt},j-1}\|_F^2 + \|(UQ)_{11}^+(UQ)_{12}(R_{22} - R_{22\text{opt},j-1})\|_F^2,$$

and again apply Lemma ??, this time with $B = R_{22} - R_{22\text{opt},j-1}$, to get $\|(A - A_k) - (A - A_k)_{\text{opt},j-1}\|_F^2 = (1 + O(\epsilon))\|\Sigma_{j,2}\|_F^2$ with probability $1 - 4\delta$.

The spectral bound proceeds similarly, but using the spectral bounds of Proposition ??, Proposition ??, and ?? instead. Thus

$$\|R_{22} - R_{22\text{opt},j-1}\|_2^2 \leq \|\Sigma_{j,2}\|_2^2 + \|\Sigma_{j,2}(\tilde{S}^T V)_{21}(\tilde{S}^T V)_{11}^+\|_2^2 = O\left(\sqrt{\frac{m}{l'}}\right)\sigma_{j+k}^2.$$

And then using (??) of Proposition ??,

$$\sigma_j^2(A - A_k) \leq \|R_{22} - R_{22\text{opt},j-1}\|_2^2 + \|(UQ)_{11}^+(UQ)_{12}(R_{22} - R_{22\text{opt},j-1})\|_2^2 = O\left(\frac{mn}{l'l}\right)\sigma_{j+k}^2,$$

which proves the spectral claim.

For the multiplicative lower bound on the singular values of A_{11}, from (??) and (??) in Proposition ?? and Proposition ?? respectively, it follows that for $j \leq k$,

$$\sigma_j(A_k) \geq \sigma_j(R_{11}R_{11}'^{-1}) \geq \sigma_{\min}((S_1^T V_1'))\sigma_j(A) = \Omega\left(\sqrt{\frac{k}{n}}\right)\sigma_j(A).$$

This last step requires additional explanation. First,

$$\sigma_{\min}(S_1^T V_1') = \sigma_{\min}(S_1^T V_1 R_{11}'^{-1}) \geq \sigma_{\min}(S_1^T V_1)\sigma_{\min}(R_{11}'^{-1})) \geq \frac{5}{6}\sigma_{\min}(R_{11}'^{-1}) = \frac{5}{6}\frac{1}{\|R'_{11}\|_2},$$

where we used $\sigma_{\min}(S_1^T V_1) \geq \frac{5}{6}$ holds by Definition ?? with probability $1 - \delta$. It remains to upper bound $\|R_{11}'\|_2$. We know $V_1 = V'\begin{pmatrix} R_{11}' \\ 0 \end{pmatrix}$, so $\|R_{11}'\|_2 = \|V_1\|_2$. But $\|V_1\|_2 = O\left(\sqrt{\frac{n}{k}}\right)$ due to Lemma ??, with probability $1 - \delta$. This completes the proof of the lower bound. □

Next, we specialize to the SRHT ensemble in order to see a case where the bounds of Definition ?? and Definition ?? are stronger than in ??.

Definition 60. *The SRHT ensemble embedding $\mathbb{R}^n$ into $\mathbb{R}^s$ is defined by generating*

$$\sqrt{\frac{n}{s}}PHD,$$

where P is $s \times n$ selecting s rows, H is the normalized Hadamard transform, and D is a $n \times n$ diagonal matrix of uniformly random signs.

The key special additional property of the SRHT ensemble is from Lemma 4.8 of [12].

Lemma 61. *Let V^T be drawn from an SRHT of dimension $l \times n$. Then for $m \times n$ matrix A with rank ρ, with probability $1 - 2\delta$,*

$$\|AV\|_2^2 \leq 5\|A\|_2^2 + \frac{\log(\rho/\delta)}{l}(\|A\|_F + \sqrt{8\log(n/\delta)}\|A\|_2)^2$$

The SRHT with $10\epsilon^{-1}(\sqrt{k} + \sqrt{8\log(m/\delta)})^2)\log(k/\delta)$ rows is a $(k, \sqrt{\epsilon}, \delta)$ OSE (by substituting $\delta = \delta/3$ into Lemma 4.1 of [12]). For the multiplication approximating property, it is straightforward to plug this setting of r into Lemma 4.11 of [12] (along with setting $R = \sqrt{4\ln(3\delta^{-1})}$ and $\delta = \delta/3$). Thus it satisfies the multiplication approximating property with parameters $(\frac{\epsilon}{k}, \delta, n)$. We may substitute these parameters into Theorem **??**, but numerous other ensembles could also be used. We have singled out the SRHT because it enjoys a remarkably good bound for the spectral norm approximation quality due to the prior lemma, but past work has not exploited this property fully. In particular, when the spectral norm and Frobenius norm are comparable (i.e. quickly decaying singular values), the quality is constant in the dimension rather than polynomial. Loosely speaking, as long as $\frac{\|A-A_k\|_F}{\|A-A_k\|_2} = O(\sqrt{k})$, then $\|A - A_k\|_2$ is around a constant factor from that of the k-truncated SVD. The theorem further strengthens this by proving the generalization to the lower singular values of $A - A_k$.

Theorem 62. *Let U_1, V_1^T be drawn from SRHT ensembles with dimensions $l' \times m$, $n \times l$. We set*

$$l \geq 10\epsilon^{-1}(\sqrt{k} + \sqrt{8\log(n/\delta)})^2\log(k/\delta),$$

as well as

$$l' \geq 10\epsilon^{-1}(\sqrt{l} + \sqrt{8\log(m/\delta)})^2\log(k/\delta).$$

*Letting ρ be the rank of A, for simplicity assume $l' \geq \log(m/\delta)\log(\rho/\delta)$ and $l \geq \log(n/\delta)\log(\rho/\delta)$. Then for any fixed $1 \leq j \leq \min(m,n) - k$, with probability $1 - 5\delta$ the approximation of A using **GLU**, A_k, satisfies*

$$\begin{aligned}\sigma_j^2(A - A_k) &= O(1)\sigma_{k+j}^2 + O\left(\frac{\log(\rho/\delta)}{l}\right)\|A - A_{opt,k+j-1}\|_F^2 \\ &= O\left(1 + \frac{\epsilon\log(\min(m,n)/\delta)}{k\log(k/\delta)}\frac{\|A - A_{opt,k+j-1}\|_F^2}{\sigma_{k+j}^2})\right)\sigma_{k+j}^2\,.\end{aligned}$$

Proof. It suffices to prove the first claim. Begin by using Proposition **??** and Lemma **??**,

$$\sigma_j^2(R_{22}) \leq \|\Sigma_{j,2}\|_2^2 + \|\Sigma_{j,2}(\tilde{S}^T V)_{21}(\tilde{S}^T V)_{11}^+\|_2^2 \leq \|\Sigma_{j,2}\|_2^2 + 2\|\Sigma_{j,2}(\tilde{S}^T V)_{21}\|_2^2,$$

with probability $1 - \delta$. Next apply Lemma **??** to the second term to get

$$\begin{aligned}\sigma_j^2(R_{22}) &= O\left(1 + \frac{\log(\rho/\delta)\log(n/\delta)}{l}\right)\|\Sigma_{j,2}\|_2^2 + O\left(\frac{\log(\rho/\delta)}{l}\right)\|\Sigma_{j,2}\|_F^2 \\ &= O(1)\|\Sigma_{j,2}\|_2^2 + O\left(\frac{\log(\rho/\delta)}{l}\right)\|\Sigma_{j,2}\|_F^2, \qquad (4.42)\end{aligned}$$

where ρ is the rank of A, with probability $1-2\delta$. Continue from the result of Proposition **??**,

$$\begin{aligned}\sigma_j^2(A-A_k) &\leq \|R_{22}-R_{22\text{opt},j-1}\|_2^2 + \|(U_1Q_1)^+(U_1Q_2)(R_{22}-R_{22\text{opt},j-1})\|_2^2 \\ &\leq \|R_{22}-R_{22\text{opt},j-1}\|_2^2 + 2\|(U_1Q_2)(R_{22}-R_{22\text{opt},j-1})\|_2^2.\end{aligned}$$

From Theorem **??** we also know $\|(R_{22}-R_{22\text{opt},j-1})\|_F^2 \leq (1+O(\epsilon))\|\Sigma_{j,2}\|_F^2 \leq 2\|\Sigma_{j,2}\|_F^2$ because the SRHT with the parameter settings specified for l and l' satisfies the multiplication approximatin and OSE properties. Thus repeating the same steps using Lemma **??** and Lemma **??** to complete the proof for the first bound,

$$\begin{aligned}\sigma_j^2(A-A_k) &\leq \|R_{22}-R_{22\text{opt},j-1}\|_2^2 + 2\|(U_1Q_2)(R_{22}-R_{22\text{opt},j-1})\|_2^2 \\ &\leq C_1\frac{\log(\rho/\delta)\log(m/\delta)}{l'}\|R_{22}-R_{22\text{opt},j-1}\|_2^2 + C_2\frac{\log(\rho/\delta)}{l'}\|R_{22}-R_{22\text{opt},j-1}\|_F^2 \\ &= O(1)\|R_{22}-R_{22\text{opt},j-1}\|_2^2 + O\left(\frac{\log(\rho/\delta)}{l'}\right)\|R_{22}-R_{22\text{opt},j-1}\|_F^2.\end{aligned}$$

By using the bounds on $\sigma_j(R_{22})$ from (**??**) and the fact that $\|R_{22}-R_{22\text{opt},j-1}\|_2 = \sigma_j(R_{22})$, we further obtain

$$\sigma_j^2(A-A_k) \leq C_1\sigma_{k+j}^2 + C_2\frac{\log(\rho/\delta)}{l}\|\Sigma_{j,2}\|_F^2.$$

□

A few remarks are in order.

Remark 63. *First, the SRHT ensemble is only defined for powers of 2. This is not a theoretical issue because matrices can be padded. However, as discussed in [12] there are orthogonal ensembles related to the SRHT, namely the discrete cosine transform and Hartley transform, for which the key probabilistic requirement in Lemma* **??** *carries over, so this corollary also carries over.*

Remark 64. *Second, we consider much of the work in this section as adapting [12] to algorithm* ***GLU*** *which sketches A's columns and rows and proves a spectral norm bound comparable to the above. Their work does not specify how to proceed after finding $A \approx Q_1Q_1^TA$, and therefore follows* ***RQR****. Therefore if one follows their approach, creating a compressed representation of A would still require $O(nmk)$ time because Q_1^TA must be computed. We state the relevant part of their theorem here to provide context:*

Theorem 65 ([12], Thm 2.1)**.** *Let $A \in \mathbb{R}^{m\times n}$ have rank ρ and n a power of 2. Fix an integer k satisfying $2 \leq k < \rho$. Let $0 < \epsilon < 1/3$ and $0 < \delta < 1$. Let $Y = AV^T$ where $V \in \mathbb{R}^{r\times n}$ is drawn from the SRHT ensemble with $r = 6\epsilon^{-1}(\sqrt{k}+\sqrt{8\log(n/\delta)})^2)\log(k/\delta))$. Then with probability $1-5\delta$*

$$\|A-YY^+A\|_2 \leq (4+\sqrt{\frac{3\log(n/\delta)\log(\rho/\delta)}{r}})\|A-A_k\|_2 + \sqrt{\frac{3\log(\rho/\delta)}{r}}\|A-A_k\|_F$$

From this we see our Theorem ?? has qualitatively the same accuracy guarantee on the residual error. For many types matrices A, in particular for those with fast spectral decay, Theorem ?? will be within a constant factor of the rank k truncated-SVD's spectral approximation error.

Remark 66. *In comparing **CW** with the outcomes of Theorems ?? and Theorem ??, many results carry over, and we briefly sketch this here. Given U_1 is from an SRHT ensemble, it is not difficult to see in Proposition ?? that $\frac{l}{\sqrt{m}}\tilde{A}$ has smaller singular values than A. Moreover, the singular values of B are bound through those of $\tilde{A}$, using Proposition ??, as B is the projection of $\tilde{A}$ onto the sketch generated by $\tilde{A}V_1$. Then U_1^+ is orthogonal besides undoing the scaling of $\tilde{A}$, by multiplying the singular values by $\frac{l}{\sqrt{m}}$. This sketch describes why the Frobenius norm and Spectral norm bounds on the residual still apply, i.e. Definition ??.*

The bound on Definition ?? we use is from Theorem ??, as it is not strengthened by using an SRHT ensemble. In particular it gave $\sigma_j(A_k) = \Omega(\sqrt{\frac{k}{n}})\sigma_j(A)$. Recall that the deterministic identity behind the result is from ??, using the relations around (??). Intuitively because only the leading l columns of $\bar{A}$ are used in proving this bound, the same argument applies.

*In contrast to the above, Definition ?? does not fit very naturally with **CW**. This is because, though U_1^+B in Proposition ?? by itself can easily be bound, it does not necessarily interact nicely with $S(\bar{A}_{11})$. Thus we are unable to extend the result of Theorem ?? for $j > 1$.*

Remark 67. *Let us consider the computational cost of computing the GLU approximation of A through Theorem ??, storing the result in the form of (??), following Algorithm ??.*

Simply by following the algorithmic description, we see the largest cost terms are $O(nm\log(l') + mll')$. We present a short table tabulating this.

$\hat{A} = U_1(AV_1)$	$O(nm\log(l))$
$T_1 = U_1^+(I - \hat{A}\hat{A}^+)$	$O(ml'\log(m) + ll'^2)$ because up to a factor U_1 has orthonormal columns, thus $U_1^+ = \sqrt{\frac{l'}{m}}(PHD)^T = \sqrt{\frac{l'}{m}}DHP^T$
$T_2 = AV_1$	Stored from first step
$T_2 = T_2\hat{A}^+$	$O(mll')$
$T = T_1 + T_2$	$O(ml')$
$S = U_1A$	$O(mn\log(l'))$

Specializing as in the theorem, we additionally required
$l \geq 10\epsilon^{-1}(\sqrt{k} + \sqrt{8\log(n/\delta)})^2\log(k/\delta)$
and $l' \geq 10\epsilon^{-1}(\sqrt{l} + \sqrt{8\log(m/\delta)})^2\log(k/\delta)$.

Using these bounds on l and l', we say the runtime is $\tilde{O}(nm + k^2 m\epsilon^{-3})$. Various poly-log factors are hidden here, involving n, m, k, δ. In more detail, plugging in l and l' into the prior complexity bound and assuming $m < n$ so that $l' = O(l\epsilon^{-1}\log(k/\delta))$, we get Big-Oh of

$$nm\log\left(\epsilon^{-2}(k+\log(n/\delta))\log^2(k/\delta)\right) + m\epsilon^{-3}(k^2+\log^2(n/\delta))\log^3(k/\delta)).$$

Note that in the runtime bound, because there is asymmetry between m and n, it turns out to be faster if $m < n$ and thus A is short-wide. If this is not the case for A, then one could simply run the algorithm on A^T.

Remark 68. *As stated, Theorem ?? provides bounds for the **GLU** with sketching from the left and right. We noted in the prior remark how this retains the performance of [12] while increasing the speed. We could stop the analysis at (??), and also borrow the bounds already found in Theorem ?? and Proposition ??. Then we obtain new bounds for the randomized QR factorization,*

Corollary 69. *Let $n \times l$ matrix V_1^T be drawn from an SHRT ensemble, $l \geq 10\epsilon^{-1}(\sqrt{k} + \sqrt{8\log(n/\delta)})^2\log(k/\delta)$, and for simplicity assume $l \geq \log(n/\delta)\log(\rho/\delta)$. Then we have*

$$\|R_{22} - R_{22_{opt,j-1}}\|_F^2 \leq (1+O(\epsilon))\|A - A_{opt,k+j-1}\|_F^2\,,$$

with probability $1-2\delta$, as well as

$$\sigma_j^2(R_{22}) \leq O(\sigma_{k+j}^2) + O(\frac{\log(\rho/\delta)}{l})\|A - A_{opt,k+j-1}\|_F^2\,,$$

for $1 \leq j \leq min(m,n) - k$ with probability $1-3\delta$ for a particular j. We also have upper and lower bounds on the largest singular values, as for $1 \leq j \leq k$,

$$\sigma_j(A) \geq \sigma_j(Q_1 Q_1^T A) = \Omega(\sqrt{\frac{k}{n}})\sigma_j(A)$$

holds with probability $1 - 2\max(\delta, e^{-k/5})$. Actually, borrowing the deterministic bound of [35] found in equation (4.7),

$$\sigma_j(A) \geq \sigma_j(Q_1 Q_1^T A) \geq \frac{\sigma_j}{1 + O(\sqrt{\frac{n}{k}})\frac{\sigma_{k+1}}{\sigma_j}}$$

holds with probability $1-\delta$.

We move on to the third application, controlling the growth factor during Gaussian elimination by right and left multiplication by square random matrices. The theoretical result we establish is that the growth factor is well behaved if we multiply by square Gaussian random matrices. Note the bounds in Propositions ?? and ?? will in this case be the same for Gaussian random matrices as for Haar random matrices, because they differ by lower and upper triangular factors and U_1, V_1 are now square. We make use of bounds proven for the

Haar ensemble. The work [24], which viewed the problem in terms of the Haar ensemble, required a randomized QR-factorization as a subroutine to compute the generalized Schur-decomposition of the matrix by a divide-and-conquer approach. This required a bound on the smallest singular value of the $k \times k$ minors. Eventually a tight bound on these was given in [23] by means of the exact probability distribution, which we will use.

As pointed out in [7], Theorem 3.2 and Lemma 3.5 of [23] give an exact density of the smallest singular value of a Haar minor. Analyzing this formula gives the following bound, which is sharp up to a constant in the primary range of interest, $\sigma_{\min} = O\left(\frac{1}{\sqrt{k(n-k)}}\right)$.

Lemma 70. *Let $\delta > 0$, $k, (n-k) > 30$; then $\mathbb{P}\left[\sigma_{\min} \leq \frac{\delta}{\sqrt{k(n-k)}}\right] \leq 2.02\delta$.*

We will define the ℓ_2 growth factors of $\bar{A}$ as $\rho_U(\bar{A}) := \max_p \|\mathcal{S}_p(\bar{A})\|_2 / \|\bar{A}\|_2$ and $\rho_L(\bar{A}) := \max_p \|\bar{A}_{21}\bar{A}_{11}^{-1}\|_2$ where $\mathcal{S}_p$ is the Schur complement of the top $p \times p$ block. From Proposition **??**,(**??**), and Proposition **??** it is not difficult to see that both are bounded as

$$\begin{aligned}
\rho_U(\bar{A}), \rho_L(\bar{A}) &\leq \max_j \left[\|X[:j,:j]^{-1}\|_2 \|R[j+1:,j+1:]\|_2 / \|\bar{A}\|_2\right] \\
&\leq \max_j \left[\|(UQ)[:j,:j]^{-1}\|_2 \|(S^TV)[j+1:,j+1:]^{-1}\|_2\right] \\
&= \max_j \left[\sigma_{\min}^{-1}((UQ)[:j,:j]) \sigma_{\min}^{-1}((S^TV)[:j,:j])\right].
\end{aligned}$$

Note that ρ_U and ρ_L control what is typically called the growth factor of $\bar{A}$. The growth factor is the largest magnitude entry appearing in the matrices L, U returned by Gaussian Elimination. This is because of norm equivalence, with the operator and max-element norm differing by at most a factor of $\sqrt{n}$. Therefore our ℓ_2 growth factors are equivalent for the purpose of proving stability.

Corollary 71. *Suppose we want to solve $Ax = b$ by Gaussian Elimination, and we precondition, postcondition A by Haar distributed matrices U, V. That is, we solve $UAVx' = Ub$ and output V^Tx'. Then the U and L ℓ_2-growth factors introduced above satisfy*

$$\mathbb{E}[\log(\max(\rho_U(\bar{A}), \rho_L(\bar{A})))] = O(\log(n))$$

Proof. Because U and V are Haar, the matrices UQ and S^TV in Propositions **??** and **??** are Haar distributed. Apply Lemma **??** to the minors (call them generically M) of UQ and S^TV with size in the range $[30, n-30]$,

$$\mathbb{P}[\sigma_{\min}^{-1}(M) > n^{2+a}] < 2.02n^{-1-a}$$

To control all minors in this range, simply perform a union bound over all $< 2n$ minors being considered. Let B_1 be the inverse of the smallest singular value of the minors in range

$[30, n-30]$ of UQ and S^TV. Then $\mathbb{P}[B_1 \geq n^{2+a}] \leq 4.04n^{-a}$. Setting $a = x-2$, this is $\mathbb{P}[\log_n(B_1) \geq x] \leq 4.04n^{2-x}$.

To deal with the minors in range $[0, 30]$, we cite a result in random matrix theory which says that these minors scaled by $\sqrt{n}$ approach a matrix of i.i.d. $N(0,1)$ random variables. The convergence is with respect to total variation distance, see [41]. Let B_2 be the inverse of the smallest singular value of these 60 minors. For the claimed result, what matters is $\mathbb{E}[\log_n(B_2)] = C_1'$ for some constant C_1', due to the $\frac{1}{\sqrt{n}}$ scaling. This is apparent from work similar to [23] but for Gaussian matrices, see for example the bound on the condition number in [17].

Combining the bounds for B_1 and B_2,

$$\begin{aligned}
\mathbb{E}[\log_n(\max(\rho_U(\bar{A}), \rho_L(\bar{A})))] &\leq \mathbb{E}[\log_n(\max_j \left[\sigma_{\min}^{-1}((UQ)[:j,:j])\sigma_{\min}^{-1}((S^TV)[:j,:j])\right])] \\
&\leq \mathbb{E}[\log_n(B_1)] + \mathbb{E}[\log_n(B_2)] \\
&\leq C_1' + \int_0^2 1dx + 4.04\int_2^\infty n^{2-x}dx \\
&= C_1 + 4.04\log(n)\int_0^\infty e^{-x}dx \\
&\leq C\log(n)
\end{aligned}$$

□

Of course, it is impractical to use a Gaussian or Haar matrix to condition a matrix in this context. We might as well then solve the system by means of QR-factorization. However, this sheds light on the strategy of using conditioners to avoid pivoting during Gaussian Elimination. This has been popularized in work such as [8]. The theoretical support of such work has been lacking. Corollary **??** is the first theoretical result we are aware of that shows a random conditioners can be used to provably avoid the need to pivot.

It also could be considered a generalization of the well-known fact that Gaussian random matrices have low pivot growth during Gaussian elimination. Indeed, we have shown that this is the case for any distribution of singular values not just that of the Gaussian random matrix. The most interesting question still remains if faster conditioners can be used to make the approach both theoretically and practically sound for all matrices A. More concretely we pose the question,

Remark 72. *Is there a random matrix ensemble S such that SA can be computed quickly, but also $\sigma_{\min}((SA)[:k,:k]) = O\left(\frac{1}{poly(n)}\right)$ when A is an orthogonal matrix?*

4.6 Conclusion

We have provided a thorough analysis of a new low-rank approximation procedure **GLU**. Along the way, we have seen it is closely related to many different past approaches. Our procedure is as fast as past approaches to within a log factor, and comes with spectral and frobenius norm bounds on the residual, as well as multiplicative bounds for the other singular values.

For future work, Remark **??** seems useful and interesting. Finding applications which particularly benefit from the speed and accuracy guarantees of our procedure is also of interest.

Appendix A

Appendix One

A.1 Lebesgue Case

For simplicity's sake, we will now use $|\cdot|$ for the Lebesgue case, just as we used this notation for the counting measure in the discrete case. The dual is almost the same as in the discrete case, and can be read off from Theorem **??**. The difference is that the dual variable associated to $\mathbb{R}^d$ is allowed to be negative.

$$\begin{aligned} \underset{y}{\text{maximize}} \quad & y^T \dim(\mathbf{E}) \\ \text{subject to} \quad & y^T \dim(\phi_i(\mathbf{E})) \leq \alpha_i, \, \forall \phi_i \\ & y_V \geq 0, \; \forall \, V \neq \mathbb{R}^d \end{aligned} \tag{A.1}$$

We require an analog construction to replace Definition **??** and Proposition **??**.

Definition 73 (Product Parallelepiped)**.** *Consider independent subspaces Y_i, as well as real valued scaling parameters y_i. Also fix unit vectors e_{ij} providing a basis for Y_i. Then define the set*

$$S := \{x \in \mathbb{R}^d \,|\, x = \sum_{i,j} a_{ij} e_{ij} \text{ with } 0 \leq a_{ij} < M^{k_i}\} \tag{A.2}$$

Note the above fits the classical definition of a parallelepiped. We also emphasize that we will only use the construction when the $\oplus Y_i = \mathbb{R}^d$; that is, we only create full dimensional parallelepipeds.

Proposition 74. *Suppose we are given a dual vector y, whose non-zero values are attached to a list of independent subspaces $\mathbf{Y} = (Y_1, \ldots, Y_t)$. If $\oplus_i Y_i \neq \mathbb{R}^d$, then augment $\mathbf{Y}$ with a complementary space Y_{t+1}, leaving $y_{Y_{t+1}} = 0$. Now form the product parallelepiped S of Def.* **??***, which is full dimensional due to the additional complementary space Y_{t+1}.*

Then $|S| = \Theta(M^{y^T rank(\mathbf{Y})})$. If in addition y is dual feasible, then $|\phi_i(S)| = O(M^{\alpha_i})$ holds for each ϕ_j.

Proof. Since S is a parallelepiped, its volume can be calculated by the determinant of its axis. Hence, letting E be the matrix with columns the vectors e_{ij}, we use multilinearity of determinants in the first equality to get

$$|S| = M^{y^T \text{rank}(\mathbf{Y})} \det(E) = \Theta(M^{y^T \text{rank}(\mathbf{Y})})$$

It remains to consider the images of this set under the ϕ_i in the case y is feasible. This requires a bound on $|\phi_i(S)|$. Again using multilinearity of determinants in the first equality below, and feasibility property (**??**) of y in the third equality,

$$|\phi_i(S)| = \prod_j M^{\text{rank}(\phi_i(Y_j))y_{Y_j}} \det(\phi_i(E)) = O(M^{C_i(y)}) = O(M^{\alpha_i})$$

□

The algorithm that moves the support to a flag is identical, as the possibility $y_{\mathbb{R}^d} < 0$ does not affect it. We now can make the assumption that the dual is supported on a flag. The consequence of this is the main technical lemma in the Lebesgue case,

Lemma 75. *Consider independent subspaces $Y_1, \ldots, Y_t$ with corresponding dual values y_{Y_i}. Follow the construction of Proposition* **??**. *Assume the subspaces are ordered so that y_{Y_i} monotonically decreases with i. In keeping with Def.* **??** *have $U_i := Y_1 + \cdots + Y_i$ for $i = 1, \ldots, t$, and for convenience $U_0 := \{0\}$. For any linear map L, set*

$$d_i := dim(L(U_i)) - dim(L(U_{i-1}))$$

Then we have the bound

$$|L(S)| = O\left(\prod_{i=1}^{t} M^{y_{Y_i} \cdot d_i}\right)$$

In particular, this holds for L chosen to be any of the ϕ_j.

The proof is similar to before, so we omit it. So in the Lebesgue case, we have now constructed a solid, full-dimensional parallelepiped. Such a shape clearly tiles. As before, this lemma yields the Lebesgue version of Theorem **??** and subsequently Theorem **??** as consequences. In fact, one may remove the requirement that $\alpha \geq 0$ because the new construction in Proposition **??** makes sense for $y < 0$.

A.2 Exactly Optimal Tilings

Technical Definition

This paper has shown how to construct tilings that provably have the best polynomial dependence on M. For example, the methods have established tiles of volume $O(M^{3/2})$ for the

matrix multiplication problem. However, so far we have not discussed the hidden constant. This section makes a brief attempt to shed light on the hidden constant, while acknowledging that its behavior is complicated in general. In particular, we provide a technical definition of exact optimality and describe two useful settings in which it applies.

Let us fix $\alpha = \vec{1}$ under the discrete case, and also use the more precise requirement for the tile S,

$$\sum |\phi_i(S)| = \sum c_i M \leq M$$

The weights c_i enforce that the total memory we can use is M; but also, it is necessary that $c_i M \geq 1$ for the tile to be nonempty. While $h_*(\vec{1})$ still uniformly upper bounds the asymptotic behavior of inequality (**??**) for every c one might choose, the inequality still differs by a constant factor. Choice of c could be regarded as strategic use of memory. Incorporating the c into inequality (**??**), we are led to consider the following inequality:

$$|S| \leq M^{h_*(\vec{1})} \max_{\substack{\vec{1}^T c=1 \\ c \geq \frac{1}{M}}} \min_{\substack{s \in \mathcal{P} \\ \vec{1}^T s = h_*(\vec{1})}} \prod_i c_i^{s_i}$$

This expression is rather unwieldy. The remainder of the section aims to show that as $M \to \infty$, the complicated max min term reduces to a more elegant function of the polytope $\mathcal{P}$. We will make this more elegant formulation our definition of exact asymptotic optimality.

We will need the following standard result from the reference [56] as Corollary 37.3.2

Proposition 76. *Let C, D be convex, compact subsets of $\mathbb{R}^n, \mathbb{R}^m$ respectively. Also assume that the real function $f(x, y)$ is concave in $x \in C$ and convex in $y \in D$, as well as jointly continuous.*

Then the weak duality of the min, max relation is actually strong duality, meaning

$$\max_{x \in C} \min_{y \in D} f(x, y) = \min_{y \in D} \max_{x \in C} f(x, y)$$

With this tool, we state and then prove the following

Proposition 77.

$$\lim_{M \to \infty} \max_{\substack{\vec{1}^T c=1 \\ c \geq \frac{1}{M}}} \min_{\substack{s \in \mathcal{P} \\ \vec{1}^T s = h_*(\vec{1})}} \prod_i c_i^{s_i} = \frac{1}{h_*(\vec{1})^{h_*(\vec{1})}} \min_{\substack{s \in \mathcal{P} \\ \vec{1}^T s = h_*(\vec{1})}} \prod_i s_i^{s_i}$$

Relaxing to $\vec{1}^T c = 1 + o(1)$ for convenience, the optimal c_i can be taken to be $c_i = \max(\frac{1}{M}, \frac{s_i}{h_*(\vec{1})})$ where s_i solve the optimization problem

$$\min_{\substack{s \in \mathcal{P} \\ \vec{1}^T s = h_*(\vec{1})}} \sum_i s_i \log(s_i)$$

We note that this is a convex optimization problem, after taking the log. Consequently, for the cases in which the polytope $\mathcal{P}$ is computable, this quantity is as well.

Proof. We start by applying the previous duality theorem to the max min relation bounding $|S|$. The theorem applies once we replace $c_i^{s_i}$ with $s_i \log(c_i)$. The relevant quantity becomes

$$\min_{\substack{s \in \mathcal{P} \\ \vec{1}^T s = h_*(\vec{1})}} \max_{\substack{\vec{1}^T c = 1 \\ c \geq \frac{1}{M}}} \sum_i s_i \log(c_i)$$

If we assume each $s_i \geq \frac{h_*(\vec{1})}{M}$, it is straightforward to eliminate c_i in the inside optimization problem by using Lagrange multipliers, giving

$$(1, \ldots, 1) = \lambda(\frac{s_1}{c_1}, \ldots, \frac{s_n}{c_n})$$

The solution to this is $c_i = \frac{s_i}{h_*(\vec{1})}$. We now argue that asymptotically this formula applies when some $s_i = 0$, provided we adopt the continuous extension of $x \log(x)$. To do this we construct s' close to s, setting $s'_i = \frac{h_*(\vec{1})}{M}$ for those values $s_i < \frac{h_*(\vec{1})}{M}$, and $s'_i = s_i$ otherwise. From elementary calculus applied to $x \log(x)$, we see that asymptotically with respect to M,

$$\forall c_i \geq \frac{1}{M},\ s_i \log(c_i) - s'_i \log(c_i) = o(1)$$

which implies

$$\max_{\substack{\vec{1}^T c = 1 \\ c \geq \frac{1}{M}}} \sum_i s_i \log(c_i) - \max_{\substack{\vec{1}^T c = 1 \\ c \geq \frac{1}{M}}} \sum_i s'_i \log(c_i) = o(1)$$

And now use the clean exact solution for $s' \geq \frac{1}{M}$,

$$\max_{\substack{\vec{1}^T c = 1 \\ c \geq \frac{1}{M}}} \sum_i s_i \log(c_i) = s_i \log(s_i) - h_*(\vec{1}) \log(h_*(\vec{1})) + o(1)$$

Critically, the decay of $o(1)$ has no dependence on s; recall it comes from the function $x \log(x)$. Therefore,

$$\lim_{M \to \infty} \max_{\substack{\vec{1}^T c = 1 \\ c \geq \frac{1}{M}}} \min_{\substack{s \in \mathcal{P} \\ \vec{1}^T s = h_*(\vec{1})}} \sum_i s_i \log(c_i) = \lim_{M \to \infty} \min_{\substack{s \in \mathcal{P} \\ \vec{1}^T s = h_*(\vec{1})}} \max_{\substack{\vec{1}^T c = 1 \\ c \geq \frac{1}{M}}} s_i \log(s_i) - h_*(\vec{1}) \log(h_*(\vec{1})) + o(1)$$

This implies the first claim, because taking the log of the original function can now be undone by applying exp, both of which are monotone. The claim concerning the form of the c_i was derived during the proof. □

This leads to a definition of exact optimality:

Definition 78. *Define the value*

$$\gamma := \frac{1}{h_*(1)^{h_*(1)}} \min_{\substack{s \in \mathcal{P} \\ \vec{1}^T s = h_*(\vec{1})}} \prod_i s_i^{s_i}$$

The family of sets $S(M)$, parametrized by integer M, are exactly optimal tilings if translations of $S(M)$ can tile $\mathbb{Z}^d$ and $S(M)$ satisfies

$$|S(M)| = (1 - o(1)) \cdot \gamma M^{h_*(\vec{1})} \tag{A.3}$$

as well as

$$\sum_i |\phi_i(S(M))| = (1 + o(1))M \tag{A.4}$$

The $o(1)$ term in **??** is required for any reasonable goal because one cannot allocate room for fractions of entries from arrays. Conceptually, we are requiring the ratio of $|S(M)|$ and the theoretical optimum to tend to one, i.e. the relative difference is going to 0.

As a way to conclude and summarize this technical section, we wish to note the factor γ can be understood as finding the max entropy $s \in \mathcal{P}$ that additionally lies on the optimal hyperplane $\vec{1}^T s = h_*(\vec{1})$

Rank One Maps

This could be regarded as a generalized n-body problem. Detailed work on the communication patterns and bounds for the n-body problem was examined in [28] and [42]. This corresponds to arrays with single indices. First, we demonstrate what $h_*(\vec{1})$ is for this case and a method for obtaining asymptotic optimality.

Proposition 79. *Assume the maps ϕ_i are rank 1 with $i \in J$ and $|J| = n$, and the lattice is $\mathbb{Z}^d$. If $\cap_i ker(\phi_i) = \{0\}$, then $h_*(\vec{1}) = d$. Then a d dimensional cube with sides $O(M)$ is asymptotically optimal. Otherwise, the Primal LP of Def.* **??** *is infeasible.*

Proof. First, suppose that $\cap_i \ker(\phi_i) \neq \{0\}$. Then take E_1 to be a non-zero element of this intersection; as kernels are subspaces, the $\langle E_1 \rangle$ is also in the kernel. This implies the corresponding constraint in Def. **??** is

$$\vec{0}^T s \geq 1$$

which can't be satisfied. So the LP is infeasible, meaning one could get "infinite" data re-use.

Now suppose $\cap \ker(\phi_i) = \{0\}$. One subgroup you could use is $\mathbb{Z}^d$ itself. By the rank 1 assumption, the inequality constraint in Def. **??** corresponding to this subgroup is

$$\vec{1}^T s \geq d$$

This implies $h_*(\vec{1}) \geq d$. Now we exhibit a feasible primal vector s for which $1^T s = d$ to complete the proof. One may select a subset $J' \subset J$ with $|J'| = d$ such that $\cap_{i \in J'} \ker(\phi_i) = \{0\}$. This follows by induction; start with $H_0 = \mathbb{Z}^d$. Then recurse by $H_i = H_{i-1} \cap \ker(\phi_i)$. If $\text{rank}(H_i) = \text{rank}(H_{i-1}) - 1$ then include i in J'. Because belonging to $\ker(\phi_i)$ amounts to satisfying a single linear equation, the rank may only decrease by 1. Choose the primal variable s to be $1_{J'}$.

It remains to establish the feasibility of this s. The argument may proceed recursively as above. This time label the elements of J' to be $i_1, \ldots, i_d$, and let T denote any subgroup. Set $H_{i_0} = T$ and $H_{i_j} = H_{i_{j-1}} \cap \ker(\phi_i)$ recursively. Again, ranks of the H_{i_j} decrease by 1 or stay the same, compared to the rank of $H_{i_{j-1}}$. The end result is $\{0\}$; this implies T is not a strict subset of at least $\text{rank}(T)$ of the kernels associated with J'. Consequently s satisfies the constraint of Def. **??** for the subgroup T:

$$s^T \text{rank}(\phi(T)) \geq \text{rank}(T)$$

This implies the feasibility of s and establishes $h_*(\vec{1}) = d$. Observe the dual variable indicating the space $\mathbb{Z}^d$ achieves the value $h_*(\vec{1})$ as well. By Lemma **??**, a cube with sides $O(M)$ is asymptotically optimal. □

This establishes that the running Algorithm **??** on $\mathbb{Z}^d$ produces an asymptotically optimal tiling. For exact optimality, we must restrict to the case where there are d rank-one maps with trivial kernel intersection.

Lemma 80 (Basis Lemma)**.** *The subgroup $\cap_{j \neq i} ker(\phi_j)$ is rank 1; take e_i to be a non-zero element of smallest Euclidean norm from this subgroup. Then each subgroup $ker(\phi_i)$ contains the independent elements $e_1, \ldots, e_{i-1}, e_{i+1}, \ldots, e_d$.*

Proof. We must check the e_i are well defined, and that they are linearly independent.

Every time we intersect with one of the kernels, the rank reduces by 1. The trivial intersection property of the d kernels implies this; every intersection adds a linear constraint, and if one of the linear constraints turned out to be redundant then d intersections would not result in the set $\{0\}$.

Lastly, we make sure that the e_i are independent. If not, then some e_i is in the span of the other $e_{i'}$; however, the other $e_{i'}$ are contained in $\ker(\phi_i)$. This means $e_i \in \ker(\phi_i)$ is as well. Then e_i lies in the intersection of all the kernels, which by assumption is the trivial set $\{0\}$. □

This basis is critical in the following;

Proposition 81. *Let e_i be as in Lemma ??. Then the sets $S := \{\sum a_i e_i | a_i \in \mathbb{Z}, 0 \leq a_i \leq \lfloor M/d \rfloor - 1\}$ are exactly optimal. That is, the output of Algorithm 1 on independent elements $e_1, \ldots, e_d$ of $\mathbb{Z}^d$ meets the requirements of Eq. ?? and ??.*

Proof. The first part of this section established that the optimal $h_*(\vec{1})$ is d and comes from each $s_i = 1$. This is in fact the unique solution to the primal LP of Def. ?? so it is by default the minimizer of γ in Eq. ??. Alternatively, evenly distributed values s_i maximize entropy and consequently would minimize γ. Plugging this in, $c_i = \frac{1}{d}$ and $\gamma = \frac{1}{d^d}$.

It remains to confirm that $|S| = (M/d)^d + O(1)$ and $\sum_i |\phi_i(S)| \leq M$. First, by independence of the e_i, there are $\lfloor M/d \rfloor^d$ lattice points enclosed. Now if $M = a \cdot d + r$,

$$(M/d)^d = a^d \cdot (1 + \frac{r}{M})^d = \lfloor M/d \rfloor^d \cdot (1 + \frac{r}{M})^d \leq \lfloor M/d \rfloor^d e^{r/M} = \lfloor M/d \rfloor^d (1 + o(1))$$

This establishes Eq. ??. For the memory bound constraint, consider $\phi_i(S)$. Applied to any point $z \in S$, it outputs $a_i \cdot e_i$. As a_i only varies between $\lfloor M/d \rfloor$ values, the result follows. □

We may summarize the approach to tiling in Proposition ?? in the the following algorithm.

Algorithm 9 Exactly Optimal Tiling, Rank One Maps

1: Input: rank one maps $\{\phi_i\}_{i=1}^d$ with coordinate representations $a_i \in \mathbb{Z}^d$, satisfying $\cap \ker(\phi_i) = \{0\}$, memory size M
2: Output: tile S and translations by T that tile $\mathbb{Z}^d$
3: Initialize $e_1, \ldots, e_d \in \mathbb{Z}^d$
4: **for** i = 1 to d **do**
5: $A \leftarrow (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_d)^T$
6: U, D, $V \leftarrow$ Smith Normal Form(A)
7: $e_i \leftarrow$ column 1 of V
8: **end for**
9: S, $T \leftarrow$ Algorithm ?? on input subgroup $U_1 = \mathbb{Z}^d$, its independent elements $e_1, \ldots, e_d$, memory size M, and scaling $y_{\mathbb{Z}^d} = 1$
10: **return** S and T

The new component of the algorithm is calculating the independent elements e_i. With this in mind, we examine the calculation of the e_i. Recall $e_i \in \cap_{j \neq i} \ker(\phi_j)$ of smallest Euclidean norm are used in Proposition ??. The implies e_i is in the kernel of matrix $A_i := (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_d)^T$. Decompose this matrix by Smith Normal Form, giving the representation UDV^{-1}. Because the rank of A_i is $d-1$, only the first diagonal entry of D is 0. This means the kernel of the matrix is exactly what V^{-1} maps to $(1, 0, \ldots, 0)^T$, meaning multiples of the first column of V. As V is unimodular, this column is also the shortest integer valued multiple of itself.

Rank d-1 Maps

This section follows the rank 1 case very closely, and consequently is kept brief. As an example, this setting includes the case of matrix multiplication and therefore much of linear algebra. We again discuss asymptotic optimality, followed by exact optimality.

Proposition 82. *In the case where all maps are rank $d-1$, the optimal dual vector y can be taken to have the subspace generated by the kernels of all the maps as its only nonzero coordinate. Call this subspace H and let $k = rank(H)$. Then $y_W = 1/(k-1)$ is optimal for the dual LP of Def.* **??**. *In addition, $h_*(\vec{1}) = k/(k-1)$.*

Proof. As H is sent to a rank $k-1$ space by each of the ϕ_i, y is indeed dual feasible with objective value $k/(k-1)$.

We must show that that this matches the HBL lower bound, as then by strong duality y is dual optimal. Propose the primal value $s = 1/(k-1) \cdot 1_A$, where 1_A indicates any k maps whose (one-dimensional) kernels generate the rank k space. Essentially, this is saying the kernels of these maps are independent.

Then consider any rank l subgroup T, and its images under the maps in A. By independence of kernels in the construction of A, only l of the maps might send this to a rank $l-1$ group, the others send it to a l dimensional space. As there are k non-zero s_i, the LHS of the constraint given by subgroup T in the primal LP of Def. **??** is

$$\begin{aligned}
\text{rank}(\phi(T))^T s &= \sum_{\phi_i \in A} s_i \cdot \text{rank}(\phi_i(T)) \\
&= \frac{1}{k-1} \sum_{\phi_i \in A} \text{rank}(\phi_i(T)) \\
&= \frac{1}{k-1} \left[(l-1) |\{\phi \in A | \ker(\phi) \cap T \neq \{0\}\}| + l |\{\phi \in A | \ker(\phi) \cap T = \{0\}\}| \right] \\
&\geq \frac{1}{k-1} \left[(l-1) \cdot l + l \cdot (k-l) \right] \\
&= (l^2 - l + lk - l^2)/(k-1) = l \cdot (k-1)/(k-1) = l
\end{aligned}$$

Meanwhile, the RHS is l, so the constraint is satisfied. □

Similar to the rank 1 case, for exact optimality, restrict to when the kernels of the ϕ_i are independent. Again let e_i denote a non-zero smallest Euclidean norm representative of $\ker(\phi_i)$, and let E be the subgroup they span.

Proposition 83. *Suppose the number of maps is equal to k and the kernels are independent. Form the set $S := \{\sum a_i \cdot e_i | a_i \in \mathbb{Z}, 0 \leq a_i \leq \lfloor \frac{M}{k} \rfloor^{\frac{1}{k-1}} - 1\}$. That is, apply Algorithm* **??** *to the independent elements e_i of subgroup E, with scaling $y_E = 1/(k-1)$ and memory size M/k. Then S meets the criteria of Eq.* **??** *and* **??** *for exact optimality.*

Proof. (Sketch). We established that $s_i = 1/(k-1)$ has $1^T s = h_*(\vec{1})$. Moreover, because the values are evenly distributed, it minimizes γ. Plugging this in, $c_i = 1/k$, $\gamma = (\frac{1}{k})^{k/(k-1)}$.

The remainder follows analagously the argument of rank one maps: show that M/d being rounded induces $1 + o(1)$ relative difference between $\gamma \cdot M^{h_*(\vec{1})}$ and $|S|$ for Eq. **??** to be satisfied, and then quickly confirm Eq. **??** holds.

□

A.3 Example

Consider $\mathbb{Z}^2$ and maps

$$\phi_1(x, y) = 3x - y$$
$$\phi_2(x, y) = x - 2y$$

The kernels are respectively $\langle e_1 + 3e_2 \rangle$ and $\langle 2e_1 + e_2 \rangle$.

Figure **??** depicts a tile shape S that would be produced by Algorithm **??** when $M = 5$.

This example can be used to demonstrate the use of the T_1 and T_3 in Algorithm **??**. T_2 is $\{0\}$ for this example, because $\langle e_1 + 3e_2, 2e_2 + e_3 \rangle$ has the same rank as $\mathbb{Z}^2$. Using T_3 produces images like Figure **??**.

Using an element of T_3 would change Figure **??** to one like Figure **??**.

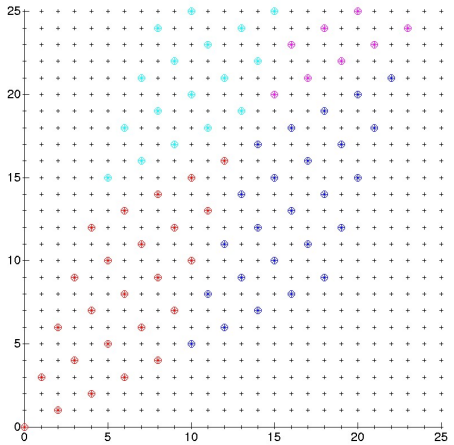

Figure A.1: translations from T_3

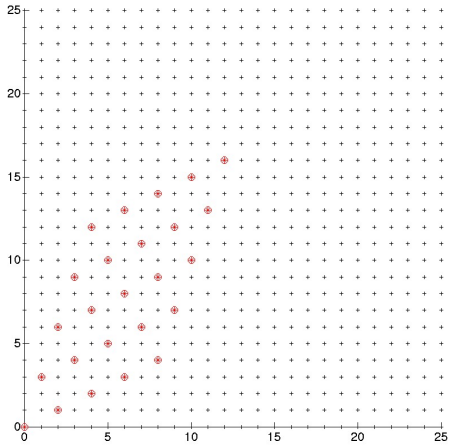

Figure A.2: basic shape

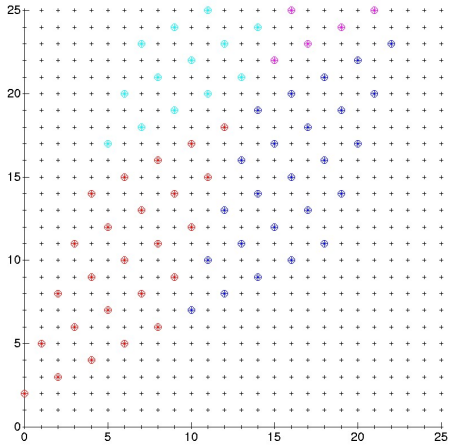

Figure A.3: translations from T_3

Appendix B

Appendix Two

B.1 Riemannian Overview

We will be working in the setting of Riemannian geometry, but will not use much machinery. We provide an informal overview. The definitions we introduce here are generally standard and formalized in introductory texts, one such being [46].

An n-dimensional (smooth) manifold M can be understood as a space that is locally diffeomorphic to $\mathbb{R}^n$, so we identify these subsets of M with coordinates $(x_1, \ldots, x_n)$. This allows us to define smooth curves $\gamma : \mathbb{R} \to M$, by requiring their coordinate representations $(x_1(t), \ldots, x_n(t))$ to be smooth. We may define velocities $\gamma'(t)$ by associating them with $(x_1'(t), \ldots, x_n'(t))$, leading to the notion of the tangent spaces $T_xM \cong \mathbb{R}^n$.

Riemannian manifolds additionally specify a metric for measuring the size of these velocities, by defining an inner product $\langle \cdot, \cdot \rangle_x$ on the tangent space of each $x \in M$. This immediately enables the definition of curve length, as $\int |\gamma'(t)|_{\gamma(t)} dt$. It also gives a method of measuring volume; if g_{ij} is the bilinear form for the metric in a local coordinate choice, then $\sqrt{|g|} dx^1 \wedge \cdots \wedge dx^n$ is the Riemannian volume form.

It also turns out to be helpful to compute directional derivatives for vector fields (or acceleration along curves). Requiring a few natural conditions leads to a unique Riemannian connection $\nabla : T_xM \times T_xM \to T_xM$ determined by the metric. It is known as the Levi-Civita connection. In the coordinates of a local frame $E = (\vec{e}_1, \ldots, \vec{e}_n)$, which provides a basis for the tangent spaces of a neighborhood, the Riemannian connection is given by

$$\nabla_{\vec{e}_i} \vec{e}_j = \Gamma_{ji}^k \vec{e}_k$$

for the Christoffel symbols Γ_{ij}^k. When the acceleration of a curve is 0, i.e. $\nabla_{\gamma'(t)} \gamma'(t) \equiv 0$, we say that curve is a geodesic. This is a second order non-linear ODE system for $\gamma(t) = (x_1(t), \ldots, x_n(t))$,

$$\ddot{x}^k(t) + \dot{x}^i \dot{x}^j \Gamma_{ij}^k(x(t)) = 0.$$

A unique solution will exist locally provided we specify the initial position and velocity. The exponential map is defined by $\exp_p(v) = \gamma(1)$. For Hadamard manifolds, the exponential map is well-defined for any values of p and v.

At any any $x \in M$ we may consider the image of a tangent plane σ_x spanned by $v, w \in T_xM$. Locally around x the image is a surface. The sectional curvature of the 2-plane σ_x is defined to be the Gaussian curvature of the image surface at x. Lower and upper bounds of the sectional curvature enable generalizations of Euclidean tools like ball volume and triangle trigonometry estimates. The Bishop-Gromov volume comparison theorem is one important result along these lines. Although usually stated for its volume upper bound by assuming just a lower bound on curvature, it is understood that the proof also provides a lower volume bound [55]. Here we state a specialization of this theorem that suffices for our application,

Theorem 84 (Bishop-Gromov)**.** *Suppose M is a Hadamard manifold. Let vol_g denote the Riemannian volume and $B_g(x, r)$ denote the open ball of radius r around x. Then*

$$vol_g(B_g(x, r)) \leq \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2} + 1)} r^n.$$

The right side of this inequality is the volume of a Euclidean ball of radius r.

Hadamard manifolds are simply connected manifolds of non-positive sectional curvature. They have been extensively studied in mathematical literature. We collect a few commonly used facts which we made use of or provide intuition. For Hadamard manifolds,

- The exponential maps $\exp_x(\cdot)$ are diffeomorphisms from T_xM to M (Cartan-Hadamard theorem).
- The distance to a point, $d(x, \cdot)$, is strictly convex. The distance to a closed, convex set is convex.
- Geodesics between points are unique and distance minimizing.
- Projection onto closed, convex sets is well defined and continuous.

All of these properties can be found in [8].

In the proof of Theorem **??**, we made use of the separating hyperplane theorem of convex geometry. Here we briefly state and prove a version sufficient for this application.

Lemma 85. *Suppose M is a Hadamard manifold, $S \subset M$ is a closed, convex set, and $p \notin S$. Then there is a halfspace based at p satisfying $H_p(v) \cap S = \emptyset$*

Proof. Consider the function $f(x) = d(x, S)$. By [8] this function is convex and hence for any $p \notin S$ there exists a subgradient $v \in \partial f_p$ (see Definition **??**). Then $H_p(v)$ is such a separating hyperplane, by Lemma **??**. □

In the introduction, we mentioned that the gradient of a differentiable convex function is a subgradient. We provide a short justification for this simple fact, as it is an important source of subgradients. In Riemannian geometry, the gradient is defined by duality using the metric; that is, ∇ satisfies $\langle \nabla f, \cdot \rangle = df(\cdot)$.

Lemma 86. *Let $f(x) : M \to \mathbb{R}$ be convex along geodesics as well as differentiable. Then*

$$f(y) \geq f(x) + \langle \nabla f(x), \exp_x^{-1}(y) \rangle_x$$

Proof. Let $y = \exp_x(t_0 v)$. That f is convex on geodesics means $f(\exp_x(tv))$ is convex in t, so

$$f(y) \geq f(x) + t_0 \frac{d}{dt} f(\exp_x(tv))|_0 \,.$$

But using the chain rule and that $d(\exp_x)|_0 = I$ (see [46]),

$$\frac{d}{dt} f(\exp_x(tv))|_0 = df(d\exp_x|_{\vec{0}}(v)) = df(v) = \langle \nabla f(x), v \rangle_x \,.$$

□

Bibliography

[1] M. Abramowitz and I.A. Stegun, eds. *Handbook of Mathematical Functions*. New York: Dover Publications, 1970.

[2] Zhaojun Bai, James Demmel, and Ming Gu. "An inverse free parallel spectral divide and conquer algorithm for nonsymmetric eigenproblems". In: *Numerische Mathematik* 76.3 (May 1997), pp. 279–308. ISSN: 0945-3245. DOI: 10.1007/s002110050264. URL: https://doi.org/10.1007/s002110050264.

[3] G. Ballard, J. Demmel, and I. Dumitriu. *Communication-optimal parallel and sequential eigenvalue and singular value algorithms*. Tech. rep. EECS-2011-14. UC Berkeley, Feb. 2011. URL: http://www.eecs.berkeley.edu/Pubs/TechRpts/2011/EECS-2011-14.html.

[4] Grey Ballard et al. "A 3D Parallel Algorithm for QR Decomposition". In: *Proceedings of the 30th on Symposium on Parallelism in Algorithms and Architectures*. SPAA '18. Vienna, Austria: ACM, 2018, pp. 55–65. ISBN: 978-1-4503-5799-9. DOI: 10.1145/3210377.3210415. URL: http://doi.acm.org/10.1145/3210377.3210415.

[5] Grey Ballard et al. *A Generalized Randomized Rank-Revealing Factorization*. Version 1. Sept. 2019. URL: https://arxiv.org/abs/1909.06524.

[6] Grey Ballard et al. "Reconstructing Householder Vectors from Tall-Skinny QR". In: *IEEE 28th International Parallel and Distributed Processing Symposium*. May 2014, pp. 1159–1170. DOI: 10.1109/IPDPS.2014.120. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6877344.

[7] G. Ballard et al. *A Generalized Rank-Revealing Factorization*. Version 3. Sept. 17, 2019. arXiv: arXiv:1909.06524v1. URL: https://arxiv.org/abs/1909.06524.

[8] W. Ballmann. *Lectures on Spaces of Nonpositive Curvature*. Berlin, Germany: Springer, 1995.

[9] I. Benjamini and R. Eldan. "Convex Hulls in Hyperbolic Space". In: *Geometriae Dedicata* 160.1 (Oct. 2012), pp. 365–371. DOI: 10.1007/s10711-011-9687-8.

[10] Bennett et al. "Finite Bounds for Holder-Brascamp-Lieb Multilinear Inequalities". In: *Mathematical Research Letters* 55.4 (June 2005), pp. 647–666. DOI: 10.4310/MRL.2010.v17.n4.a6.

[11] Christian H. Bischof. "Incremental Condition Estimation". In: *SIAM Journal on Matrix Analysis and Applications* 11.2 (1990), pp. 312–322. DOI: 10.1137/0611021. eprint: http://dx.doi.org/10.1137/0611021. URL: http://dx.doi.org/10.1137/0611021.

[12] Christos Boutsidis and Alex Gittens. "Improved matrix algorithms via the Subsampled Randomized Hadamard Transform". In: *SIAM J. Matrix Analysis Applications* 34 (2013), pp. 1301–1340.

[13] Peter Businger and Gene H. Golub. "Linear least squares solutions by Householder transformations". English. In: *Numerische Mathematik* 7.3 (1965), pp. 269–276. ISSN: 0029-599X. DOI: 10.1007/BF01436084. URL: http://dx.doi.org/10.1007/BF01436084.

[14] Tony F. Chan and Per Christian Hansen. "Some Applications of the Rank Revealing QR Factorization". In: *SIAM Journal on Scientific and Statistical Computing* 13.3 (1992), pp. 727–741.

[15] Shivkumar Chandrasekaran and Ilse C. F. Ipsen. "On Rank-Revealing Factorizations". In: *SIAM Journal on Matrix Analysis and Applications* 15.2 (1994), pp. 592–622. DOI: 10.1137/S0895479891223781. eprint: http://dx.doi.org/10.1137/S0895479891223781. URL: http://dx.doi.org/10.1137/S0895479891223781.

[16] Jeff. Cheeger and David G. Ebin. *Comparison theorems in riemannian geometry*. Providence, Rhode Island: AMS Chelsea Publishing, 2008. ISBN: 9780821844175.

[17] Zizhong Chen and Jack J. Dongarra. "Condition Numbers of Gaussian Random Matrices". In: *SIAM J. Matrix Anal. Appl.* 27.3 (July 2005), pp. 603–620. ISSN: 0895-4798. DOI: 10.1137/040616413. URL: http://dx.doi.org/10.1137/040616413.

[18] M. Christ, J. Demmel, and N. Knight. *On the Algebraic Structure Underlying Discrete Holder-Brascamp-Lieb Inequalities*. 2018.

[19] M. Christ et al. *Communication Lower Bounds and Optimal Algorithms for Programs That Reference Arrays - Part 1*. Tech. rep. UCB/EECS-2013-61. EECS Department, University of California, Berkeley, May 2013. URL: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-61.html.

[20] M. Christ et al. *On Holder-Brascamp-Lieb Inequalities for Torsion-Free Discrete Abelian Groups*. Oct. 2015. eprint: 1510.04190 (math.CA).

[21] Kenneth L. Clarkson and David P. Woodruff. "Low-Rank Approximation and Regression in Input Sparsity Time". In: *J. ACM* 63.6 (Jan. 2017), 54:1–54:45. ISSN: 0004-5411. DOI: 10.1145/3019134. URL: http://doi.acm.org/10.1145/3019134.

[22] Dario Cordero-Erausquin, Robert J. McCann, and Michael Schmuckenschläger. "A Riemannian interpolation inequality à la Borell, Brascamp and Lieb". In: *Inventiones Mathematicae* (2001). DOI: 10.1007/s002220100.

[23] R. Dedekind. "Uber die von drei Moduln erzeugte Dualgruppe". In: *Mathematische Annalen* 53.4 (1900), pp. 371–403.

[24] James Demmel, Ioana Dumitriu, and Olga Holtz. "Fast linear algebra is stable". In: *Numerische Mathematik* 108.1 (Nov. 2007), pp. 59–91. DOI: 10.1007/s00211-007-0114-x. URL: https://doi.org/10.1007/s00211-007-0114-x.

[25] James Demmel, Laura Grigori, and Alexander Rusciano. *An improved analysis and unified perspective on deterministic and randomized low rank matrix approximations.* Version 1. Oct. 2019. URL: https://arxiv.org/abs/1910.00223.

[26] James Demmel and Alexander Rusciano. *Parallelepipeds obtaining HBL lower bounds.* Version 1. Nov. 2016. URL: https://arxiv.org/abs/1611.05944.

[27] J. Demmel et al. "Communication-optimal Parallel and Sequential QR and LU Factorizations". In: *SIAM Journal on Scientific Computing* 34.1 (2012), A206–A239. DOI: 10.1137/080731992. URL: http://epubs.siam.org/doi/abs/10.1137/080731992.

[28] M. Driscoll et al. "A Communication-Optimal N-Body Algorithm for Direct Interactions". In: *IEEE International Symposium: Parallel and Distributed Processing* 27.4 (May 2013), pp. 1075–1084. DOI: 10.1109/IPDPS.2013.108.

[29] Jed Duersch and Ming Gu. "Randomized QR with Column Pivoting". In: *SIAM Journal on Scientific Computing* 39.4 (2017), pp. C263–C291. DOI: 10.1137/15M1044680. URL: https://doi.org/10.1137/15M1044680.

[30] Ioana Dumitriu. "Smallest eigenvalue distributions for two classes of β-Jacobi ensembles". In: *Journal of Mathematical Physics* 53 (Sept. 2010).

[31] Ioana Dumitriu. "Smallest eigenvalue distributions for two classes of β-Jacobi ensembles". In: *Journal of Mathematical Physics* 53.10, 103301 (2012), pp. 103301.1–103301.15. DOI: http://dx.doi.org/10.1063/1.4748969. URL: http://scitation.aip.org/content/aip/journal/jmp/53/10/10.1063/1.4748969.

[32] A. Garg et al. "Algorithmic and Optimization Aspects of Brascamp-Lieb Inequalities, via Operator Scaling". In: *STOC* 49.4 (June 2017), pp. 397–409. DOI: 10.1145/3055399.3055458.

[33] G.H. Golub and C.F. Van Loan. *Matrix Computations.* Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 2012. ISBN: 9781421407944.

[34] L. Grigori, S. Cayrols, and J. W. Demmel. "Low Rank Approximation of a Sparse Matrix Based on LU Factorization with Column and Row Tournament Pivoting". In: *SIAM J. Sci. Comput.* 40.2 (2018), pp. 181–209.

[35] B. Grunbaum. "Partitions of mass-distributions and of convex bodies by hyperplanes". In: *Pacific J. Math.* 10.4 (1960), pp. 1257–1261. URL: https://projecteuclid.org:443/euclid.pjm/1103038065.

[36] M. Gu and S. Eisenstat. "Efficient Algorithms for Computing a Strong Rank-Revealing QR Factorization". In: *SIAM Journal on Scientific Computing* 17.4 (1996), pp. 848–869. DOI: 10.1137/0917055. eprint: http://epubs.siam.org/doi/pdf/10.1137/0917055. URL: http://epubs.siam.org/doi/abs/10.1137/0917055.

[37] N. Halko, P. G. Martinsson, and J. A. Tropp. "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions". In: *SIAM Rev.* 53.2 (May 2011), pp. 217–288. ISSN: 0036-1445. DOI: 10.1137/090771806. URL: http://dx.doi.org/10.1137/090771806.

[38] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. 2nd. New York, NY, USA: Cambridge University Press, 2012.

[39] D. Irony, S. Toledo, and A. Tiskin. "Communication Lower Bounds for Distributed-Memory Matrix Multiplication". In: *Journal of Parallel and Distributed Computing* 64.9.4 (Sept. 2004), pp. 1017–1026. DOI: 10.1016/j.jpdc.2004.03.021.

[40] Sergei Ivanov. "On Helly's theorem in geodesic spaces". In: *Electronic Research Announcements* 21.1935-9179-2014-0-109 (2014), p. 109. ISSN: 1935-9179. DOI: 10.3934/era.2014.21.109.

[41] Tiefeng Jiang. "Maxima of entries of Haar distributed matrices". In: *Probability Theory and Related Fields* 131 (Jan. 2005), pp. 121–144. DOI: 10.1007/s00440-004-0376-5.

[42] P. Koanantakool and K. Yelick. "A Computation- And Communication-Optimal Parallel Direct 3-Body Algorithm". In: *ACM/IEEE Supercomputing Conference* 24.4 (Nov. 2014). DOI: 10.1109/SC.2014.35.

[43] Rodrigo C. de Lamare and Raimundo Sampaio-Neto. "Adaptive Reduced-Rank Processing Based on Joint and Iterative Interpolation, Decimation, and Filtering". In: *IEEE Transactions on Signal Processing* 57 (2009), pp. 2503–2514.

[44] B. Laurent and P. Massart. "Adaptive estimation of a quadratic functional by model selection". In: *Ann. Statist.* 28.5 (Oct. 2000), pp. 1302–1338. DOI: 10.1214/aos/1015957395. URL: https://doi.org/10.1214/aos/1015957395.

[45] Y. Ledyaev, J. Treiman, and J. Zhu. "Helly's intersection theorem on manifolds of nonpositive curvature". In: *Journal of Convex Analysis* 13.3-4 (2006), pp. 785–798. ISSN: 0944-6532.

[46] J. Lee. *Riemannian Manifolds: An Introduction to Curvature*. Berlin, Germany: Springer, 1997.

[47] Michael W. Mahoney and Petros Drineas. "CUR matrix decompositions for improved data analysis." In: *Proceedings of the National Academy of Sciences of the United States of America* 106 3 (2009), pp. 697–702.

[48] Ivan Markovsky. "Recent progress on variable projection methods for structured low-rank approximation". In: *Signal Processing* 96 (Mar. 2014), pp. 406–419. DOI: 10.1016/j.sigpro.2013.09.021.

[49] Per Gunnar Martinsson et al. "Householder QR Factorization With Randomization for Column Pivoting (HQRRP)". In: *SIAM Journal on Scientific Computing* 39.2 (2017), pp. C96–C115. DOI: 10.1137/16M1081270. URL: https://doi.org/10.1137/16M1081270.

[50] L Miranian and Ming Gu. "Strong rank revealing LU factorizations". In: *Linear Algebra and its Applications* 367 (July 2003), pp. 1–16.

[51] Cameron Musco and Christopher Musco. "Randomized Block Krylov Methods for Stronger and Faster Approximate Singular Value Decomposition". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, 2015, pp. 1396–1404. URL: http://dl.acm.org/citation.cfm?id=2969239.2969395.

[52] J. Nelson and H. L. Nguyen. "OSNAP: Faster Numerical Linear Algebra Algorithms via Sparser Subspace Embeddings". In: *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. Aug. 2013, pp. 117–126. DOI: 10.1109/FOCS.2013.21.

[53] Kevin O'Neill. *A Variation on Holder-Brascamp-Lieb Inequalities*. Oct. 2017. eprint: 1710.06374 (math.CA).

[54] Ching-Tsuan Pan and Ping Tak Peter Tang. "Bounds on Singular Values Revealed by QR Factorizations". In: *BIT Numerical Mathematics* 39.4 (1999), pp. 740–756.

[55] Peter Petersen. *Riemannian Geometry*. New York, New York: Springer, 2006. ISBN: 0387292462.

[56] R. Rockafellar. *Convex Analysis*. Princeton, NJ: Princeton University Press, 1970.

[57] Alexander Rusciano. *A Riemannian Corollary of Helly's Theorem*. Version 2. Apr. 2018. URL: https://arxiv.org/abs/1804.10738.

[58] T. Sarlos. "Improved Approximation Algorithms for Large Matrices via Random Projections". In: *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*. Oct. 2006, pp. 143–152.

[59] Gil Shabat et al. "Randomized LU decomposition". In: *Applied and Computational Harmonic Analysis* 44.2 (2018), pp. 246–272. ISSN: 1063-5203. DOI: https://doi.org/10.1016/j.acha.2016.04.006. URL: http://www.sciencedirect.com/science/article/pii/S1063520316300069.

[60] Suvrit Sra and Reshad Hosseini. "Geometric Optimisation on Positive Definite Matrices with Application to Elliptically Contoured Distributions". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 2562–2570. URL: http://dl.acm.org/citation.cfm?id=2999792.2999898.

[61] G. W. Stewart. "Rank Degeneracy". In: *SIAM Journal on Scientific and Statistical Computing* 5.2 (1984), pp. 403–413. DOI: 10.1137/0905030. eprint: http://dx.doi.org/10.1137/0905030. URL: http://dx.doi.org/10.1137/0905030.

[62] G.W. Stewart. "On graded QR decompositions of products of matrices." eng. In: *ETNA. Electronic Transactions on Numerical Analysis [electronic only]* 3 (1995), pp. 39–49. URL: http://eudml.org/doc/119027.

[63] Constantin Udriste. *Convex Functions and Optimization Methods on Riemannian Manifolds*. Dordrecht: Kluwer Academic Publishers, Jan. 1994. DOI: 10.1007/978-94-015-8390-9.

[64] S. I. Valdimarsson. "The Brascamp-Lieb Polyhedron". In: *Canadian Journal of Mathematics* 62.4 (May 2010), pp. 870–888. DOI: 10.4153/CJM-2010-045-2.

[65] David P. Woodruff. "Sketching As a Tool for Numerical Linear Algebra". In: *Found. Trends Theor. Comput. Sci.* 10.1–2 (Aug. 2014), pp. 1–157. ISSN: 1551-305X. DOI: 10.1561/0400000060. URL: http://dx.doi.org/10.1561/0400000060.

[66] Hongyi Zhang, Sashank J. Reddi, and Suvrit Sra. "Riemannian SVRG: Fast Stochastic Optimization on Riemannian Manifolds". In: *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 2016, pp. 4592–4600. URL: http://papers.nips.cc/paper/6515-riemannian-svrg-fast-stochastic-optimization-on-riemannian-manifolds.pdf.

[67] Hongyi Zhang and Suvrit Sra. "First-order Methods for Geodesically Convex Optimization". In: *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*. 2016, pp. 1617–1638. URL: http://jmlr.org/proceedings/papers/v49/zhang16b.html.

[68] Z. Allen Zhu et al. "Operator Scaling via Geodesically Convex Optimization, Invariant Theory and Polynomial Identity Testing". In: *STOC* 50 (June 2018).