THE UNIVERSITY OF CHICAGO

ALGEBRAIC AND DIFFERENTIAL GEOMETRY IN MODERN OPTIMIZATION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON COMPUTATIONAL AND APPLIED MATHEMATICS

BY
ZEHUA LAI

CHICAGO, ILLINOIS
JUNE 2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am very grateful to my advisor, Lek-Heng Lim, for his suggestions, discussions and support. I met a lot of great teachers in University of Chicago, including but not limit to Daniel Sanz-Alonso, Guillaume Bal, Mihai Anitescu, Mary Silber, Jeremy Hoskins, Gregory Lawler, Shmuel Weinberger. I am supported by the generous scholarship of University of Chicago and Neubauer fellowship in the last 5 years as a PhD candicate, which I am very thankful for. I also learn a lot from my collaborators, Shmuel Friedland, Alex Wu, Minda Zhao, Jose Israel Rodriguez, and Bradley Nelson, and my fellow PhD Yian Chen, Zhen Dai, and Adela DePavia.

I want to thank my teachers and classmates from Tsinghua University. The friendship with He Li, Ziheng Huang, Tianyang Pan is invaluable. My experience in S.-T. Yau College Student Mathematics Contest helped me gain the confidence for the Ph.D. application. My teachers, Michael R. Powers, Lijun Yang, Zuoyang Shi, encouraged and supported me.

Chapter 2 is based on a joint work with Lek-Heng Lim, which was my first published paper. It was a wonderful experience. I learned everything from it. Chapter 3 and 4 are based on joint works with Lek-Heng Lim and Ke Ye. Ke Ye invited me to Chinese Academy of Sciences several times and I am grateful for his hospitality. Chapter 5 and 6 are based on joint works with He Li, Yichen Zhang, and Xi Chen. Chapter 7 is based on a joint work with Shmuel Friedland and Lek-Heng Lim.

I thank my family for their love. I would not have survived my PhD and two years of lockdown without the support of my friends, Chen Lin, Xinyu Wang, Zhiyao Xu, Congzhe Fang, Fanyi Su.

ABSTRACT

Stochastic optimization algorithms have become indispensable in modern machine learning. The developments of theories and algorithms of modern optimization also requires the application of tools from different methematical branches, such as algebraic and differential geometry. In this dissertation, we answer several problems in stochastic optimization by a wide range of tools. We disprove the noncommutative arithmetic and geometric mean inequality using results from noncommutative polynomial optimization. We propose new, simpler and efficient models and algorithms for optimization over Grassmannian and flag manifolds. We study the problem of statistical inference in gradient-free optimization and contextual bandit optimization, and prove central limit theorems to construct confidence intervals. We present several versions of the Grothendieck inequality over the skew field of quaternions.

CHAPTER 1
INTRODUCTION

This dissertation consists of six chapters. In Chapter 2, we study the noncommutative arithmetic and geometric mean inequality. An unresolved foundational question in stochastic optimization is the difference between with-replacement sampling and without-replacement sampling — does the latter have superior convergence rate compared to the former? A groundbreaking result of Recht and Ré reduces the problem to a noncommutative analogue of the arithmetic-geometric mean inequality where n positive numbers are replaced by n positive definite matrices. If this inequality holds for all n, then without-replacement sampling (also known as random reshuffling) indeed outperforms with-replacement sampling in some important optimization problems. The conjectured Recht–Ré inequality has so far only been established for $n = 2$ and a special case of $n = 3$. We will show that the Recht–Ré conjecture is false for general n. Our approach relies on the noncommutative Positivstellensatz, which allows us to reduce the conjectured inequality to a semidefinite program and the validity of the conjecture to certain bounds for the optimum values, which we show are false as soon as $n = 5$.

In Chapter 3, we study Riemannian optimization algorithms on Grassmannian manifolds. There are two widely used models for the Grassmannian $\mathrm{Gr}(k, n)$, as the set of equivalence classes of orthogonal matrices $\mathrm{O}(n)/\big(\mathrm{O}(k) \times \mathrm{O}(n-k)\big)$, and as the set of trace-k projection matrices $\{P \in \mathbb{R}^{n\times n} : P^\top = P = P^2,\ \mathrm{tr}(P) = k\}$. The former, standard in manifold optimization, has the downside of relying on equivalence classes but working with orthogonal matrices is generally good numerical practice. The latter, widely adopted in coding theory and probability, uses actual matrices (as opposed to equivalence classes) but working with projection matrices is numerically unstable. We present an alternative that has both advantages and suffers from neither of the disadvantages; by representing k-dimensional subspaces as symmetric orthogonal matrices of trace $2k-n$, we obtain

$$\mathrm{Gr}(k,n) \cong \{Q \in \mathrm{O}(n) : Q^{\top} = Q,\ \mathrm{tr}(Q) = 2k-n\}.$$

As with the other two models, we show that differential geometric objects and operations — tangent vector, metric, normal vector, exponential map, geodesic, parallel transport, gradient, Hessian, etc — have closed-form analytic expressions that are computable with standard numerical linear algebra. In the proposed model, these expressions are considerably simpler, a result of representing $\mathrm{Gr}(k,n)$ as a linear section of a compact matrix Lie group $\mathrm{O}(n)$, and can be computed with at most one QR decomposition and one exponential of a special skew-symmetric matrix that takes only $O\bigl(nk(n-k)\bigr)$ time. In particular, we completely avoid eigen- and singular value decompositions in our steepest descent, conjugate gradient, quasi-Newton, and Newton methods for the Grassmannian. Another important feature of these algorithms, particularly evident in steepest descent and Newton method, is that they exhibit clear signs of numerical stability; various measures of errors consistently reduce to the order of machine precision throughout extensive numerical experiments.

In Chapter 4, we study the geometry of flag manifolds under different embeddings into a product of Grassmannians. We show that differential geometric objects and operations — tangent vector, metric, normal vector, exponential map, geodesic, parallel transport, gradient, Hessian, etc — have closed-form analytic expressions that are computable with standard numerical linear algebra. Furthermore, we are able to derive a coordinate minimization method in the flag manifold that performs well compared to other gradient descent methods.

In Chapter 5, we investigates the problem of online statistical inference of model parameters in stochastic optimization problems via the Kiefer-Wolfowitz algorithm with random search directions. We first present the asymptotic distribution for the Polyak-Ruppert-averaging type Kiefer-Wolfowitz (AKW) estimators, whose asymptotic covariance matrices depend on the distribution of search directions and the function-value query complexity. The distributional result reflects the trade-off between statistical efficiency and function query complexity. We further analyze the choice of random search directions to minimize certain summary statistics of the asymptotic covariance matrix. Based on the asymptotic distribution, we conduct online statistical inference by providing two construction procedures of valid confidence intervals.

In Chapter 6, we study the online statistical inference of model parameters in a contextual bandit framework of sequential decision-making. With the fast development of big data, it has been easier than before to learn the optimal decision rule by updating the decision rule recursively and making online decisions. We propose a general framework for online and adaptive data collection environment that can update decision rules via weighted stochastic gradient descent. We allow different weighting schemes of the stochastic gradient and establish the asymptotic normality of the parameter estimator. Our proposed estimator significantly improves the asymptotic efficiency over the previous averaged stochastic gradient descent approach via inverse probability weights. We also conduct an optimality analysis on the weights in a linear regression setting. We provide a Bahadur representation of the proposed estimator and show that the remainder term in the Bahadur representation entails a slower convergence rate compared to classical stochastic gradient descent due to the adaptive data collection.

In Chapter 7, we present several versions of the Grothendieck inequality over the skew field of quaternions: The first one is the standard Grothendieck inequality for rectangular matrices, and two additional inequalities for self-adjoint matrices, as introduced by the first and the last authors in a recent paper. We give several results on “conic Grothendieck inequality”: as Nesterov $\pi/2$-Theorem, which corresponds to the cones of positive semidefinite matrices; the Goemans–Williamson inequality, which corresponds to the cones of weighted Laplacians; the diagonally dominant matrices. The most challenging technical part is the proof of the analog of Haagerup result that the inverse of the hypergeometric function $x_2F_1(\frac{1}{2}, \frac{1}{2}; \ell; x^2)$ has first positive Taylor coefficient and all other Taylor coefficients are nonpositive.

CHAPTER 2

RECHT–RÉ NONCOMMUTATIVE ARITHMETIC-GEOMETRIC MEAN CONJECTURE IS FALSE

This is a joint work with Lek-Heng Lim.

2.1 Introduction

The breathtaking reach of deep learning, permeating every area of science and technology, has led to an outsize role for randomized optimization algorithms. It is probably fair to say that in the absence of randomized algorithms, deep learning would not have achieved its spectacular level of success. Fitting an exceedingly high-dimensional model with an exceedingly large training set would have been prohibitively expensive without some form of *random sampling*, which in addition provides other crucial benefits such as saddle-point avoidance [70, 112]. As such, in machine learning computations, stochastic variants of gradient descent [25, 113, 150], alternating projections [187], coordinate descent [153], and other algorithms have largely overtaken their classical deterministic counterparts in relevance and utility.

There are numerous random sampling strategies but the most fundamental question, before all other considerations, is deciding between *sampling with replacement* or *sampling without replacement*. In the vast majority of randomized algorithms, a random sample is selected or a random action is performed *with replacement* from a pool, making the randomness in each iteration independent and thus easier (often much easier) to analyze. However, when it comes to practical realizations of these algorithms, one invariably samples *without replacement*, since they are easier (often much easier) to implement. Take the ubiquitous stochastic gradient descent for example, many if not most implementations would pass through each item exactly once in a random order — this is sampling without replacement, also called *random reshuffling*. Likewise, in implementations of randomized coordinate descent, coordinates are usually just chosen in a random order — again sampling without replacement.

Apart from its ease of implementation, there are other reasons for favoring without-replacement sampling. Empirical evidence [24, 16, 202] suggests that in stochastic gradient descent, without-replacement sampling regularly outperforms with-replacement sampling. Theoretical results also point towards without-replacement sampling: Under standard convexity assumptions, the convergence rate of a without-replacement sampling algorithm typically beats a with-replacement sampling one by a factor of $O(n^{-1})$ [174, 149, 88].

Recht and Ré [165] proposed a matrix theoretic approach to compare the efficacy of with- and without-replacement sampling methods. Since nearly every common optimization algorithm, deterministic or randomized, works with a linear or quadratic approximation of the objective function locally, it suffices to examine the two sampling strategies on linear or quadratic functions to understand their local convergence behaviors. In this case, the iteration reduces to matrix multiplication and both sampling procedures are linearly convergent (often called "exponentially convergent" in machine learning). The question of which is better then reduces to comparing their linear convergence rates. In this context, Recht and Ré showed that without-replacement sampling outperforms with-replacement sampling provided the following noncommutative version of the arithmetic-geometric mean inequality holds.

Conjecture 2.1.1 ([165]). *Let n be a positive integer, $A_1, \dots, A_n$ be symmetric positive semidefinite matrices, and $\|\cdot\|$ be the spectral norm. Then for any $m \leq n$,*

$$\frac{1}{n^m}\left\| \sum_{1\leq j_1,\dots,j_m\leq n} A_{j_1}\cdots A_{j_m}\right\| \geq \frac{(n-m)!}{n!}\left\| \sum_{\substack{1\leq j_1,\dots,j_m\leq n,\\ j_1,\dots,j_m \text{ distinct}}} A_{j_1}\cdots A_{j_m}\right\|. \tag{2.1}$$

While one may also ask if (2.1) holds for other norms, the most natural and basic choice is the spectral norm, i.e., the operator 2-norm. Unless specified otherwise, $\|\cdot\|$ will always denote the spectral norm in this article.

To give an inkling of how (2.1) arises, consider the Kaczmarz algorithm [187] where we attempt to solve an overdetermined linear system $Cx = b$, $C \in \mathbb{R}^{p \times d}$, $p > d$, with ith row vector[1] $c_i^\top$ where $c_i \in \mathbb{R}^d$. For $k = 1, 2, \ldots$, the $(k+1)$th iterate is formed with a randomly chosen i and

$$x^{(k+1)} = x^{(k)} + \frac{b_i - \langle c_i, x^{(k)} \rangle}{\|c_i\|^2} c_i.$$

The kth error $e^{(k)} = x^{(k)} - x^*$ is then

$$e^{(k+1)} = \left(I - \frac{c_i c_i^\top}{\|c_i\|^2} \right) e^{(k)} =: P_{c_i} e^{(k)},$$

where $P_c \in \mathbb{R}^{d \times d}$, the orthogonal projector onto $\operatorname{span}\{c\}^\perp$, is clearly symmetric positive semidefinite. A careful analysis would show that the relative efficacy of with- and without-replacement sampling depends on a multitude of inequalities like $\|A^4 + B^4 + AB^2A + BA^2B\| \geq 2\|AB^2A + BA^2B\|$, which are difficult to analyze on a case-by-case basis. Nevertheless, more general heuristics would lead to (2.1) — if it holds, then without-replacement sampling is expected to outperform with-replacement sampling. In fact, the gap can be significant — for random Wishart matrices, the ratio between the two sides of (2.1) increases exponentially with m [165].

To date, extensive numerical simulations have produced no counterexample. Conjecture 2.1.1 has been rigorously established only in very special cases, notably for $(m, n) = (2, 2)$ [165] and $(m, n) = (3, 3k)$ [207].

Our contributions: We show how to transform Conjecture 2.1.1 into a form where the *noncommutative Positivstellensatz* applies, which implies in particular that for any specific values of m and n, the conjecture can be checked via two semidefinite programs. This allows us to show in Section 2.3 that the conjecture is false as soon as $m = n = 5$. We also establish

[1] We adopt standard convention that any vector $x \in \mathbb{R}^d$ is a column vector; a row vector will always be denoted $x^\top$.

in Section 2.2 that the conjecture holds for $m = 2$ and 3 with arbitrary n by extending the approach in [207]. While the conjectured inequality (2.1) is clearly sharp (as we may choose all A_i's to be equal) whenever it is true, we show in Section 2.5 that the $m = 2$ case may nonetheless be improved in a different sense, and we do likewise for $m = 3$ in Section 2.2. The $m = 4$ case remains open but our noncommutative Positivstellensatz approach permits us to at least check that it holds for $n = 4$ and 5 in Section 2.3.

Over the next two sections, we will transform Recht and Ré's Conjecture 2.1.1 into a "Loewner form" (Conjecture 2.2.2), a "sum-of-squares form" (Conjecture 2.3.2), and finally a "semidefinite program form" (Conjecture 2.3.3). All four conjectures are equivalent but the correctness of the last one for any m, n can be readily checked as a semidefinite program.

After the main results of this work are published [130], a concrete counter-example for $n = 5$ is constructed in [53].

2.2 Recht–Ré inequality for $m = 2$ and 3

Our goal here is to establish (2.1) for a pair and a triple of matrices. In so doing, we take Conjecture 2.1.1 a step closer to a form where noncommutative Positivstellensatz applies. There is independent value in establishing these two special cases given that the classical noncommutative arithmetic-geometric-harmonic mean inequality [18] is only known for a pair of matrices but nonetheless attracted a lot of interests from linear algebraists. These special cases also have implications on randomized algorithms — take the Kaczmarz algorithm for example, the fact that Conjecture 2.1.1 holds for $m = 2$ and 3 implies that if we randomly choose two or three distinct samples, perform the iterations, and sample again, then this "replacing after every two or three samples" strategy will converge faster than a "replacing after every sample" strategy.

We begin by providing some context for the inequality (2.1). The usual arithmetic-geometric mean inequality for n nonnegative real numbers $a_1, \dots, a_n$, i.e.,

$$(a_1 + \cdots + a_n)/n \geq (a_1 \cdots a_n)^{1/n},$$

is a special case of Maclaurin's inequality [96]. If we define

$$s_m := \frac{1}{\binom{n}{m}} \sum_{1 \leq j_1 < \cdots < j_m \leq n} a_{j_1} \cdots a_{j_m},$$

then $s_1 \geq \sqrt{s_2} \geq \cdots \geq \sqrt[n]{s_n}$. So $s_1 \geq \sqrt[m]{s_m}$ gives us

$$\frac{1}{n^m}(a_1 + \cdots + a_n)^m \geq \frac{(n-m)!}{n!} \sum_{\substack{1 \leq j_1, \dots, j_m \leq n, \\ j_1, \dots, j_m \text{ distinct}}} a_{j_1} \cdots a_{j_m},$$

which is just (2.1) for 1-by-1 positive semidefinite matrices.

For real symmetric or complex Hermitian matrices A, B, the Loewner order is defined by $A \succeq B$ iff $A - B$ is positive semidefinite. The Maclaurin's inequality has several noncommutative extensions but we regard the following as the starting point for all noncommutative arithmetic-geometric mean inequalities.

Proposition 2.2.1. *For any unitarily invariant norm* $\|\cdot\|$ *and Hermitian matrices* A, B, $\|AB + BA\| \leq \|A^2 + B^2\|$ *and* $2\|AB + BA\| \leq \|(A + B)^2\|$.

Proof. Since $-A^2 - B^2 \preceq AB + BA \preceq A^2 + B^2$, by Lemma 2.1 in [21], the desired inequalities hold for any unitarily invariant norm. □

The result was extended to compact operators on a separable Hilbert space and strengthened to $2\|A^*B\| \leq \|A^*A + B^*B\|$ in [19], with yet other extensions in [21, 20]. In [165], Conjecture 2.1.1 was also formulated as an extension of Proposition 2.2.1, with the second inequality corresponding to the $m = n = 2$ case.

Straightforward counterexamples for $n = 3$ show that we cannot simply drop the norm in (2.1) and replace the inequality $\geq$ with the Loewner order $\succeq$. Nevertheless Conjecture 2.1.1 may be written as *two* Loewner inequalities, as demonstrated by [207].

Conjecture 2.2.2 (Loewner form)**.** *Let $A_1, \ldots, A_n$ be symmetric positive semidefinite and $A_1 + \cdots + A_n \preceq nI$. Then for any $m \leq n$,*

$$-\frac{n!}{(n-m)!}I \preceq \sum_{\substack{1 \leq j_1, \ldots, j_m \leq n, \\ j_1, \ldots, j_m \text{ distinct}}} A_{j_1} \cdots A_{j_m} \preceq \frac{n!}{(n-m)!}I. \tag{2.2}$$

We prefer this equivalent formulation (2.2) as the original formulation (2.1) hides an asymmetry — note that there is an upper bound and a lower bound in (2.2) and there is no reason to expect that they should have the same magnitude. In fact, as we will see in the later sections, the best upper and lower bounds have different magnitudes in every case that we examined.

We will next prove Conjecture 2.1.1 in its equivalent form Conjecture 2.2.2 for $m = 2$ and 3. Our proofs rely on techniques introduced by [207] in his proof for the case $m = 3$, $n = 3k$, but our two additional contributions are that (i) we will obtain better lower bounds (deferred to Section 2.5), and (ii) our proof will work for arbitrary n (not necessarily a multiple of 3).

Theorem 2.2.3 (Recht–Ré for $m = 2$)**.** *Let $A_1, \ldots, A_n$ be symmetric positive semidefinite and $A_1 + \cdots + A_n \preceq nI$. Then*

$$-n(n-1)I \preceq \sum_{i \neq j} A_i A_j \preceq n(n-1)I. \tag{2.3}$$

Proof. The right inequality in (2.3) follows from

$$\begin{aligned}(n-1)\sum_{i,j} A_iA_j - n\sum_{i\neq j} A_iA_j &= (n-1)\sum_i A_i^2 - \sum_{i\neq j} A_iA_j \\ &= \sum_{i<j}(A_i - A_j)^2 \succeq 0,\end{aligned}$$

and so

$$\sum_{i\neq j} A_iA_j \preceq \frac{n-1}{n}\sum_{i,j} A_iA_j \preceq n(n-1)I.$$

For the left inequality in (2.3), expand $(\sum_i A_i)^2 \succeq 0$ to get

$$\sum_i A_i^2 \succeq -\sum_{i\neq j} A_iA_j.$$

Let $B \coloneqq nI - \sum_i A_i \succeq 0$ and $B_i \coloneqq A_i + \frac{1}{n}B \succeq 0$. So $\sum_i B_i = nI$. Then

$$\begin{aligned}-n\sum_{i\neq j} A_iA_j &= -(n-1)\sum_{i\neq j} A_iA_j - \sum_{i\neq j} A_iA_j \\ &\preceq (n-1)\sum_i A_i^2 - \sum_{i\neq j} A_iA_j \\ &= \sum_{i<j}(A_i - A_j)^2 = \sum_{i<j}(B_i - B_j)^2 \\ &= (n-1)\sum_i B_i^2 - \sum_{i\neq j} B_iB_j \\ &= n\sum_i B_i^2 - \Big(\sum_i B_i\Big)^2 \\ &= n\sum_i B_i^2 - n^2 I.\end{aligned}$$

Therefore

$$-\sum_{i\neq j} A_iA_j - (n-1)nI \preceq \sum_i B_i^2 - n^2 I = \sum_i (B_i^2 - nB_i).$$

The eigenvalues of B_i fall between 0 and n, so the eigenvalues of $B_i^2 - nB_i$ are all nonpositive, i.e., $B_i^2 - nB_i \preceq 0$. Hence $-\sum_{i \neq j} A_i A_j - (n-1)nI \preceq 0$. □

The right inequality of (2.3) is clearly sharp. In Section 2.5, we will prove a stronger result, improving the constant in the left inequality of (2.3) to $n(n-1)/4$.

Following [207], we write $\mathbb{E}_{i_1,\ldots,i_k}$ for expectation or average over all indices $1 \leq i_1, \ldots, i_k \leq n$, and $\widetilde{\mathbb{E}}_{i_1,\ldots,i_k}$ for that over distinct indices $1 \leq i_1, \ldots, i_k \leq n$.

Theorem 2.2.4 (Recht–Ré for $m = 3$)**.** *Let $A_1, \ldots, A_n$ be symmetric positive semidefinite and $A_1 + \cdots + A_n \preceq nI$. Then*

$$-I \preceq \widetilde{\mathbb{E}}_{i,j,k} A_i A_j A_k \preceq I. \tag{2.4}$$

Proof. Let A, B, C be positive semidefinite. Then $ABC + CBA \preceq ABA + CBC$. If $B \preceq C$, then $ABA \preceq ACA$.

We start with the right inequality of (2.4),

$$\begin{aligned}
\widetilde{\mathbb{E}}_{i,j,k} A_i A_j A_k &= \frac{1}{2} \widetilde{\mathbb{E}}_{i,j,k} (A_i A_j A_k + A_k A_j A_i) \\
&\preceq \frac{1}{2} \widetilde{\mathbb{E}}_{i,j,k} (A_i A_j A_i + A_k A_j A_k) \\
&= \widetilde{\mathbb{E}}_{i,j,k} A_i A_j A_i.
\end{aligned}$$

Fix a positive integer $l < n$ whose value we decide later, and deduce from last inequality that

$$\begin{aligned}
\widetilde{\mathbb{E}}_{i,j,k} A_i A_j A_k &\preceq \widetilde{\mathbb{E}}_{i,j,k} \Big[\Big(1 - \frac{1}{l}\Big) A_i A_j A_k + \frac{1}{l} A_i A_j A_i \Big] \\
&= \frac{1}{l^2(n-l)} \widetilde{\mathbb{E}}_{i_1,\ldots,i_n} \big[(A_{i_1} + \cdots + A_{i_l}) \\
&\qquad \cdot (A_{i_{l+1}} + \cdots + A_{i_n})(A_{i_1} + \cdots + A_{i_l}) \big].
\end{aligned}$$

Since $A_{i_{l+1}} + \cdots + A_{i_n} \preceq nI - (A_{i_1} + \cdots + A_{i_l})$,

$$\widetilde{\mathbb{E}}_{i,j,k} A_i A_j A_k \preceq \frac{1}{l^2(n-l)} \widetilde{\mathbb{E}}_{i_1,\ldots,i_l} \big[(A_{i_1} + \cdots + A_{i_l}) \\ \cdot \big(nI - (A_{i_1} + \cdots + A_{i_l})\big)(A_{i_1} + \cdots + A_{i_l}) \big].$$

Consider the function $f(x) = x^2(n-x)$. Let the line $y = cx + d$ be tangent to f at $x = l$. We require that $c \geq 0$ and $f(x) \leq cx + d$ for $0 \leq x \leq n$. Elementary calculation shows that such a line exists as long as $1/2 \leq l/n \leq 2/3$. Let $A = A_{i_1} + \cdots + A_{i_l}$. As $cA + dI$ and $A(nI - A)A$ are simultaneous diagonalizable, each eigenvalue of $cA + dI - A(nI - A)A$ can be obtained by applying the function $g(x) = cx + d - x^2(n-x)$ to an eigenvalue of A. Hence

$$\begin{aligned} \widetilde{\mathbb{E}}_{i,j,k} A_i A_j A_k &\preceq \frac{1}{l^2(n-l)} \widetilde{\mathbb{E}}_{i_1,\ldots,i_l} \big[c(A_{i_1} + \cdots + A_{i_l}) + dI \big] \\ &\preceq \frac{cl + d}{l^2(n-l)} I, \end{aligned}$$

where the first inequality follows from the fact that it holds for each eigenvalue. Note that if we choose $A_1 = \cdots = A_n = I$, all inequalities above as well as the right inequality of (2.4) hold with equality. So as long as $1/2 \leq l/n \leq 2/3$, l, c, d will give us

$$\frac{1}{l^2(n-l)}(cl + d) = 1$$

and thus the right inequality of (2.4).

For the left inequality of (2.4), we start by noting

$$(A_1 + \cdots + A_{n-1}) A_n (A_1 + \cdots + A_{n-1}) \succeq 0.$$

Taking expectation, we have $-(n-2)\widetilde{\mathbb{E}}_{i,j,k}A_iA_jA_k \preceq \widetilde{\mathbb{E}}_{i,j,k}A_iA_jA_i$ and thus

$$\begin{aligned}
-\widetilde{\mathbb{E}}_{i,j,k}A_iA_jA_k &= -\frac{n-2}{n-1}\widetilde{\mathbb{E}}_{i,j,k}A_iA_jA_k - \frac{1}{n-1}\widetilde{\mathbb{E}}_{i,j,k}A_iA_jA_k \\
&\preceq \frac{1}{n-1}\widetilde{\mathbb{E}}_{i,j,k}(A_iA_jA_i - A_iA_jA_k) \\
&\preceq \frac{1}{2(n-1)}\widetilde{\mathbb{E}}_{i,j,k}\big[(A_i-A_j)A_k(A_i-A_j)\big].
\end{aligned}$$

As in the proof of Theorem 2.2.3, set $B \coloneqq nI - \sum_i A_i \succeq 0$ and $B_i \coloneqq A_i + \frac{1}{n}B \succeq 0$. Then

$$\begin{aligned}
-\widetilde{\mathbb{E}}_{i,j,k}A_iA_jA_k &\preceq \frac{1}{2(n-1)}\widetilde{\mathbb{E}}_{i,j,k}\big[(B_i-B_j)B_k(B_i-B_j)\big] \\
&= \frac{1}{n-1}\widetilde{\mathbb{E}}_{i,j,k}B_iB_jB_i - \frac{1}{n-1}\widetilde{\mathbb{E}}_{i,j,k}B_iB_jB_k.
\end{aligned}$$

Let $X_i \coloneqq B_i(nI-B_i)B_i$ and $Y_i \coloneqq (nI-B_i)B_i(nI-B_i)$. Routine calculations give

$$\begin{aligned}
\widetilde{\mathbb{E}}_iX_i &= (n-1)\widetilde{\mathbb{E}}_{i,j,k}B_iB_jB_i, \\
\widetilde{\mathbb{E}}_iY_i &= (n-1)\widetilde{\mathbb{E}}_{i,j,k}B_iB_jB_i + (n-1)(n-2)\widetilde{\mathbb{E}}_{i,j,k}B_iB_jB_k,
\end{aligned}$$

which allows us to express $\widetilde{\mathbb{E}}_{i,j,k}B_iB_jB_i$ and $\widetilde{\mathbb{E}}_{i,j,k}B_iB_jB_k$ in terms of $\widetilde{\mathbb{E}}_iX_i$ and $\widetilde{\mathbb{E}}_iY_i$. Then

$$\begin{aligned}
-\widetilde{\mathbb{E}}_{i,j,k}A_iA_jA_k &\preceq \frac{1}{n-1}\widetilde{\mathbb{E}}_{i,j,k}B_iB_jB_i - \frac{1}{n-1}\widetilde{\mathbb{E}}_{i,j,k}B_iB_jB_k \\
&= \frac{1}{(n-1)^2(n-2)}\widetilde{\mathbb{E}}_i[(n-1)X_i - Y_i] \\
&= \frac{n}{(n-1)^2(n-2)}\widetilde{\mathbb{E}}_i[-B_i(B_i-nI)(B_i-I)].
\end{aligned}$$

As $-x(x-n)(x-1) \le (n-1)^2x/4$ for $0 \le x \le n$,

$$-\widetilde{\mathbb{E}}_{i,j,k}A_iA_jA_k \preceq \frac{n}{4(n-2)}\widetilde{\mathbb{E}}_iB_i = \frac{n}{4(n-2)}I.$$

When $n \geq 3$, we have $\frac{n}{4(n-2)} \leq 1$. □

Our proof in fact shows that the constant in the left inequality of (2.4) can be improved to $\frac{n}{4(n-2)}$. Nevertheless, we will see in the next section (Table 2.1) that this is not sharp.

2.3 Noncommutative Positivstellensatz

In the seminal paper [101], Helton proved an astounding result: Every positive polynomial in noncommutative variables can be written as a sum of squares of polynomials. The corresponding statement for usual polynomials, i.e., in commutative variables, is well-known to be false and is the subject of Hilbert's 17th Problem. Subsequent developments ultimately led to a noncommutative version of the Positivstellensatz for semialgebraic sets. We refer interested readers to [159] for an overview of this topic.

Stating noncommutative Positivstellensatz will require that we introduce some terminologies. Let $X_1, \dots, X_n$ be n noncommutative variables, i.e., $X_i X_j \neq X_j X_i$ whenever $i \neq j$. A *monomial* of degree d or a *word* of length d is an expression of the form $X_{i_1} \cdots X_{i_d}$. The monomials span a real infinite-dimensional vector space $\mathbb{R}\langle X_1, \dots, X_n \rangle$, called the space of *noncommutative polynomials*. For any $d \in \mathbb{N}$, the finite-dimensional subspace of noncommutative polynomials of degree $\leq d$ will be denoted $\mathbb{R}\langle X_1, \dots, X_n \rangle_d$. The *transpose* of $f \in \mathbb{R}\langle X_1, \dots, X_n \rangle$ is denoted $f^\top$ and is defined on monomials by reversing the order of variables $(X_{i_1} \cdots X_{i_d})^\top = X_{i_d} \cdots X_{i_1}$ and extended linearly to all of $\mathbb{R}\langle X_1, \dots, X_n \rangle$. If $f^\top = f$, then f is called *symmetric*.

The bottom line is that noncommutative polynomials may be evaluated on square matrices of the same dimensions, i.e., they define matrix-valued functions of matrix variables. For our purpose, if $A_1, \dots, A_n$ are real symmetric matrices, then $f(A_1, \dots, A_n)$ is also a matrix, but it may not be a symmetric matrix unless f is a symmetric polynomial.

Let $L = \{\ell_1, \dots, \ell_k\} \subseteq \mathbb{R}\langle X_1, \dots, X_n \rangle_1$ be a set of k linear polynomials, i.e., $d = 1$. We will refer to $\ell_1, \dots, \ell_k$ as *linear constraints* and

$$\mathcal{B}_L \coloneqq \{(A_1, \dots, A_n) \mid \ell_1(A_1, \dots, A_n) \succeq 0, \dots, \ell_k(A_1, \dots, A_n) \succeq 0\}$$

as the *feasible set.* Note that elements of $\mathcal{B}_L$ are n tuples of symmetric matrices. We say that $\mathcal{B}_L$ is *bounded* if there exists $r > 0$ such that all $(A_1, \dots, A_n) \in \mathcal{B}_L$ satisfy $\|A_1\| \leq r, \dots, \|A_n\| \leq r$. Let $d \in \mathbb{N}$. We write

$$\Sigma_d(L) \coloneqq \left\{ \sum_{i=1}^{k} \sum_{j=1}^{p_i} f_{ij}^\top \ell_i f_{ij} \;\middle|\; f_{ij} \in \mathbb{R}\langle X_1, \dots, X_n \rangle_d, k, p_1, \dots, p_k \in \mathbb{N} \right\}$$

for the set of *noncommutative sum-of-squares* generated by L. The following theorem is a simplified version of the noncommutative Positivstellensatz, i.e., Theorem 1.1 in [102], that will be enough for our purpose.

Theorem 2.3.1 (Noncommutative Positivstellensatz)**.** *Let f be a symmetric polynomial with* $\deg(f) \leq 2d + 1$ *and the feasible set $\mathcal{B}_L$ be bounded with nonempty interior. Then*

$$f(A_1, \dots, A_n) \succeq 0 \;\; \textit{for all} \;\; (A_1, \dots, A_n) \in \mathcal{B}_L$$

if and only if $f \in \Sigma_d(L)$.

Readers familiar with the commutative Positivstellsatz [134] would see that the noncommutative version is, surprisingly, much simpler and neater.

To avoid notational clutter, we introduce the shorthand

$$\sum_{j_i \neq j_k} \coloneqq \sum_{\substack{1 \leq j_1, \dots, j_m \leq n, \\ j_1, \dots, j_m \text{ distinct}}}$$

for sum over distinct indices. Applying Theorem 2.3.1 with linear constraints $X_1 \succeq 0, \dots, X_n \succeq 0$, $X_1 + \cdots + X_n \preceq nI$, Conjecture 2.2.2 becomes the following.

Conjecture 2.3.2 (Sum-of-squares form)**.** *Let $m \le n \in \mathbb{N}$ and $d = \lfloor m/2 \rfloor$. For the linear constraints $\ell_1 = X_1, \dots, \ell_n = X_n, \ell_{n+1} = n - X_1 - \cdots - X_n$, let*

$$\lambda_1 = \operatorname{argmin}\left\{ \lambda \in \mathbb{R} \;\middle|\; \lambda - \sum_{j_i \neq j_k} X_{j_1} \cdots X_{j_m} \in \Sigma_d(L) \right\},$$
$$\lambda_2 = \operatorname{argmin}\left\{ \lambda \in \mathbb{R} \;\middle|\; \lambda + \sum_{j_i \neq j_k} X_{j_1} \cdots X_{j_m} \in \Sigma_d(L) \right\}.$$

Then both λ_1 and $\lambda_2 \le n!/(n-m)!$.

In polynomial optimization [134], the commutative Positivstellsatz is used to transform a constrained optimization problem into a sum-of-squares problem that can in turn be transformed into a semidefinite programming (SDP) problem. This also applies to noncommutative polynomial optimization problems, i.e., we may further transform Conjecture 2.3.2 into an SDP form.

The vector space $\mathbb{R}\langle X_1, \dots, X_n \rangle_d$ has dimension $q \coloneqq 1 + n + n^2 + \cdots + n^d$ and a basis comprising all q monomials of degree $\le d$. We will assemble all basis elements into a q-tuple of monomials that we denote by β. With respect to this basis, any $f \in \mathbb{R}\langle X_1, \dots, X_n \rangle_d$ may be represented uniquely as $f = \beta^\top u$ for some $u \in \mathbb{R}^q$. Therefore a *noncommutative square* may be expressed as

$$\sum_{j=1}^{p} f_j^\top \ell f_j = \sum_{j=1}^{p} u_j^\top \beta \ell \beta^\top u_j = \operatorname{tr}\left[\beta \ell \beta^\top \left(\sum_{j=1}^{p} u_j u_j^\top \right) \right]$$

by simply writing $f_j = \beta^\top u_j$, $u_j \in \mathbb{R}^q$, $j = 1, \dots, p$. Since a symmetric matrix Y is positive semidefinite if and only if it can be written as $Y = \sum_{j=1}^{p} u_j u_j^\top$, we obtain the following one-to-one correspondence between noncommutative squares and positive semidefinite matrices:

$$\sum_{j=1}^{p} f_j^\top \ell f_j \in \mathbb{R}\langle X_1, \ldots, X_n\rangle_{2d+1},\ f_j \in \mathbb{R}\langle X_1, \ldots, X_n\rangle_d$$

$$\Updownarrow$$

$$\sum_{j=1}^{p} u_j u_j^\top \in \mathbb{R}^{q\times q},\ u_j \in \mathbb{R}^q.$$

With this correspondence, the two minimization problems in Conjecture 2.3.2 become two SDPs.

Conjecture 2.3.3 (Semidefinite program form)**.** *Let $m \leq n \in \mathbb{N}$ and $d = \lfloor m/2 \rfloor$. Let β be a monomial basis of $\mathbb{R}\langle X_1, \ldots, X_n\rangle_d$ and let $X_{n+1} = n - X_1 - \cdots - X_n$. Let λ_1 be the minimum value of the SDP:*

$$\begin{aligned} \text{minimize} \quad & \lambda \\ \text{subject to} \quad & \lambda - \sum_{j_i \neq j_k} X_{j_1} \cdots X_{j_m} = \sum_{i=1}^{n+1} \operatorname{tr}(\beta X_i \beta^\top Y_i), \\ & Y_1 \succeq 0, \ldots, Y_{n+1} \succeq 0; \end{aligned} \tag{2.5}$$

and λ_2 be that of the SDP:

$$\begin{aligned} \text{minimize} \quad & \lambda \\ \text{subject to} \quad & \lambda + \sum_{j_i \neq j_k} X_{j_1} \cdots X_{j_m} = \sum_{i=1}^{n+1} \operatorname{tr}(\beta X_i \beta^\top Y_i), \\ & Y_1 \succeq 0, \ldots, Y_{n+1} \succeq 0. \end{aligned} \tag{2.6}$$

Then both λ_1 and $\lambda_2 \leq n!/(n-m)!$.

Note that the minimization is over the scalar variable λ and the matrix variables $Y_1, \ldots, Y_{n+1}$; the equality constraint equating two noncommutative polynomials is simply saying that the coefficients on both sides are equal, i.e., for each monomial, we get a *linear constraint* involving $\lambda, Y_1, \ldots, Y_{n+1}$ — the X_i's play no role other than to serve as placeholders for these

	λ_1	λ_2	$n!/(n-m)!$
$m = 2$, $n = 2$	2.0000	0.5000	2
$m = 2$, $n = 3$	6.0000	1.5000	6
$m = 2$, $n = 4$	12.0000	3.0000	12
$m = 2$, $n = 5$	20.0000	5.0000	20
$m = 3$, $n = 3$	6.0000	3.4113	6
$m = 3$, $n = 4$	24.0000	8.5367	24
$m = 3$, $n = 5$	60.0000	17.3611	60
$m = 4$, $n = 4$	24.0000	22.4746	24
$m = 4$, $n = 5$	120.0000	80.2349	120
$m = 5$, $n = 5$	120.0000	**144.6488**	120

Table 2.1: Results from SDPs.

linear constraints. We may express (2.5) and (2.6) as SDPs in standard form with a single matrix variable $Y \coloneqq \operatorname{diag}(\lambda, Y_1, \dots, Y_{n+1})$, see (2.7) for example.

Readers acquainted with (commutative) polynomial optimization [134] would be familiar with the above discussions. In fact, the only difference between the commutative and noncommutative cases is that $\sum_{i=0}^{d} n^i$, the size of a noncommutative monomial basis, is much larger than $\binom{d+n}{d}$, the size of a commutative monomial basis.

For any fixed values of m and n, Conjecture 2.3.3 is in a form that can be checked by standard SDP solvers. The dimension of the SDP grows exponentially with m, and without access to significant computing resources, only small values of m, n are within reach. Fortuitously, $m = n = 5$ already yields the required violation $144.6488 \nleq 120$, showing that Conjecture 2.3.3 and thus Conjecture 2.1.1 is false in general. We tabulate our results for $m \le n \le 5$ in Table 2.1.

The fact that the SDP in (2.6) for $m = n = 5$ has a minimum $\lambda_2 > 144 > 120 = 5!$ shows that there are uncountably many instances with $A_1 \succeq 0$, $A_2 \succeq 0$, $A_3 \succeq 0$, $A_4 \succeq 0$, $A_5 \succeq 0$, and $A_1 + A_2 + A_3 + A_4 + A_5 \preceq 5I$ such that the matrix

$$\sum_{\sigma \in \mathfrak{S}_5} A_{\sigma(1)} A_{\sigma(2)} A_{\sigma(3)} A_{\sigma(4)} A_{\sigma(5)}$$

has an eigenvalue that is less than $-144 < -120 = -5!$. Here $\mathfrak{S}_n$ is the symmetric group on n elements. We emphasize that neither (2.6) nor its dual would give us five such matrices explicitly, although the dual does provide another way to verify our result, as we will see in Section 2.4.

Indeed, the beauty of the noncommutative Positivstellensatz approach is that it allows us to show that Conjecture 2.1.1 is false for $m = n = 5$ without actually having to produce five positive semidefinite matrices $A_1, \dots, A_5$ that violates the inequality (2.1). It would be difficult to find $A_1, \dots, A_5$ explicitly as one does not even know the smallest dimensions required for these matrices to give a counterexample to (2.1). Our approach essentially circumvents the issue by replacing them with noncommutative variables $X_1, \dots, X_5$ — the reader may have observed that the dimensions of the matrices $A_1, \dots, A_5$ did not make an appearance anywhere in this article.

2.4 Verification via Farkas

We take a closer look at the $m = n = 5$ case that provided a refutation to the Recht–Ré conjecture. In this case, the basis β has $1 + 5 + 5^2 = 31$ monomials; the SDP in (2.6) has $1 + 5 + 5^2 + 5^3 + 5^4 + 5^5 = 3906$ linear constraints, $31^2 \times 6 + 1 = 5767$ variables, and takes the form:

$$\begin{aligned} \text{minimize} \quad & \operatorname{tr}(C_0 Y) \\ \text{subject to} \quad & \operatorname{tr}(C_i Y) = b_i, \quad i = 1, \dots, 3906, \\ & Y = \operatorname{diag}(\lambda, Y_1, \dots, Y_6) \succeq 0. \end{aligned} \tag{2.7}$$

Here $C_0, C_1, \dots, C_{3906} \in \mathbb{S}^{187}_{++}$, $b \in \mathbb{R}^{3096}$, λ is a scalar variable, and $Y_1, \dots, Y_6$ are 31-by-31 symmetric matrix variables. To put (2.7) into standard form, the block diagonal structure of Y may be further encoded as linear constraints requiring that off-diagonal blocks be zero. The output of our program gives a minimizer of the form $Y^* = \operatorname{diag}(\lambda^*, Y_1^*, \dots, Y_6^*) \in \mathbb{S}^{187}_{++}$

with

$$\lambda^* = 144.6488, \quad Y_1^*, \dots, Y_6^* \in \mathbb{S}_{++}^{31}. \tag{2.8}$$

The actual numerical entries of the matrices appearing in (2.7) and (2.8) are omitted due to space constraints; but they can be found in the output of our program.

The values in (2.8) are of course approximate because of the inherent errors in numerical computations. In our opinion, the gap between the computed 144.6488 and the conjectured 120 is large enough to override any concerns of a mistaken conclusion resulting from numerical errors. Nevertheless, to put to rest any lingering doubts, we will directly show that the conjectured value $\lambda = 120$ is infeasible by producing a Farkas certificate. Consider the feasibility problem:

$$\begin{aligned} \text{minimize} \quad & 0 \\ \text{subject to} \quad & \operatorname{tr}(C_i Y) = b_i, \quad i = 1, \dots, 3906, \\ & \operatorname{tr}(C_0 Y) = 120, \\ & Y \succeq 0, \end{aligned} \tag{2.9}$$

with $C_0, C_1, \dots, C_{3906} \in \mathbb{S}_{++}^{187}$ and $b \in \mathbb{R}^{3096}$ as in (2.7). Note that $C_0 = e_1 e_1^\top$ is the matrix with one in the $(1,1)$th entry and zero everywhere else. So (2.9) is the feasibility problem of the optimization problem (2.7) with the additional linear constraint $y_{11} = 120$ and where we have disregarded the block diagonal constraints[2] on Y. The dual of (2.9) is

$$\begin{aligned} \text{maximize} \quad & 120 y_0 + b^\top y \\ \text{subject to} \quad & y_0 C_0 + y_1 C_1 + \cdots + y_{3906} C_{3906} \preceq 0. \end{aligned}$$

Our program produces a Farkas certificate $y \in \mathbb{R}^{3096}$ with $120 y_0 + b^\top y \approx 47.3 > 0$, implying that (2.9) is infeasible. While this is a consequence of Farkas Lemma for SDP [133], all we

[2] If (2.9) is already infeasible, then adding these block diagonal constraints just makes it even more infeasible.

need is the following trivial version.

Lemma 2.4.1. *Let $m, n \in \mathbb{N}$. Let $C_0, C_1, \dots, C_m \in \mathbb{S}^n$ and $b \in \mathbb{R}^{m+1}$. If there exists a $y \in \mathbb{R}^{m+1}$ with*

$$y_0 C_0 + \cdots + y_m C_m \preceq 0, \quad b^\top y > 0,$$

then there does not exist a $Y \in \mathbb{S}^n$ with

$$\operatorname{tr}(C_0 Y) = b_0, \dots, \operatorname{tr}(C_m Y) = b_m, \quad Y \succeq 0.$$

Proof. If such a Y exists, then

$$0 \geq \operatorname{tr}\big((y_0 C_0 + \cdots + y_m C_m) Y\big) = y_0 b_0 + \cdots + y_m b_m > 0,$$

a contradiction. □

Hence a matrix of the form

$$Y = \operatorname{diag}(120, Y_1, \dots, Y_6) \in \mathbb{S}^{187}$$

is *infeasible* for (2.7), providing another refutation of Conjecture 2.3.3 and thus Conjecture 2.1.1. In particular, showing that $\lambda = 120$ is infeasible for (2.7) does not require any of the values computed in (2.8). Of course, aside from being the conjectured value of λ_2, there is nothing special about $\lambda = 120$ — for any $\lambda < 144.6488$, we may similarly compute a Farkas certificate y to show that such a value of λ is infeasible for (2.7).

We conclude with a few words on the computational costs of the SDPs in this and the last section. Our resulting dense linear system for $m = n = 5$ requires $3906 \times 5767 \approx 22$ million floating point storage. Using a personal computer with an Intel Core i7-9700k processor and 16GB of RAM, our SeDuMi [188] program in Matlab takes 150 seconds. For $m = n = 6$, storage alone would have taken 26 billion floating numbers, beyond our modest computing resources.

2.5 Improving the Recht–Ré inequality

An unexpected benefit of the noncommutative Positivstellensatz approach is that it leads to better bounds for the $m = 2$ and 3 cases that we know are true. Observe that the values for λ_2 in Table 2.1 for $m = 2$ are exactly smaller than the values for $n!/(n-m)!$ by a factor of 1/4. This suggests that the Recht–Ré inequality (2.3) for $m = 2$ in Theorem 2.2.3 may be improved to

$$-\frac{1}{4}n(n-1)I \preceq \sum_{i\neq j} A_iA_j \preceq n(n-1)I.$$

Table 2.1 only shows this for $n = 2, 3, 4, 5$ but in this section, we will give a proof for arbitrary $n \geq 2$. Although our proof below does not depend on the SDP formulation in (2.6), the correct coefficients in (2.11) for arbitrary n would have been impossible to guess without solving (2.6) for $m = 2$ and some small values of n.

So far we have not explored the symmetry evident in our formulations of the Recht–Ré inequality: In Conjecture 2.2.2, the matrix expression

$$\lambda I \pm \sum_{j_i\neq j_k} A_{j_1}\cdots A_{j_m}$$

and the constraints $A_1 \succeq 0, \ldots, A_n \succeq 0$, $A_1 + \cdots + A_n \preceq nI$ are clearly invariant under any permutation $\sigma \in \mathfrak{S}_n$. In Conjecture 2.3.3, the noncommutative sum-of-squares

$$\lambda \pm \sum_{j_i\neq j_k} X_{j_1}\cdots X_{j_m} = \sum_{i=1}^{n+1} \operatorname{tr}(\beta X_i \beta^\top Y_i), \tag{2.10}$$

where $X_{n+1} = n - X_1 - \cdots - X_n$, is also invariant under $\mathfrak{S}_n$ and so we may average over all permutations to get a *symmetrized sum-of-squares*. For commutative polynomials, results from classical invariant theory are often used to take advantage of symmetry [81]. We will see next that such symmetry may also be exploited for noncommutative polynomials.

Consider the case $m = 2$, $n = 3$. The monomial basis of $\mathbb{R}\langle X_1, X_2, X_3\rangle_1$ is $\beta = (1, X_1, X_2, X_3)$. The symmetry imposes linear constraints on the matrix variables in (2.6), requiring them to take the following forms:

$$Y_1 = \begin{bmatrix} a & b & c & c \\ b & d & e & e \\ c & e & f & g \\ c & e & g & f \end{bmatrix}, \qquad Y_2 = \begin{bmatrix} a & c & b & c \\ c & f & e & g \\ b & e & d & e \\ c & g & e & f \end{bmatrix},$$

$$Y_3 = \begin{bmatrix} a & c & c & b \\ c & f & g & e \\ c & g & f & e \\ b & e & e & d \end{bmatrix}, \qquad Y_4 = \begin{bmatrix} x & y & y & y \\ y & z & w & w \\ y & w & z & w \\ y & w & w & z \end{bmatrix}.$$

These symmetries allow us to drastically reduce the degree of freedom in our SDP: For any $m = 2, n \geq 2$, the matrices $Y_1, \dots, Y_n$ are always determined by precisely 11 variables that we label $a, b, c, d, e, f, g, x, y, z, w$. We computed their values explicitly for $n = 2, 3, 4$. For $n = 2$,

$$Y_1 = \begin{bmatrix} \frac{5}{4} & -\frac{3}{4} & \frac{1}{4} \\ -\frac{3}{4} & \frac{1}{2} & 0 \\ \frac{1}{4} & 0 & \frac{1}{2} \end{bmatrix}, \quad Y_3 = \begin{bmatrix} \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} & 0 \\ -\frac{1}{4} & 0 & \frac{1}{2} \end{bmatrix},$$

and Y_2 can be determined from Y_1. For $n = 3$,

$$Y_1 = \begin{bmatrix} \frac{5}{2} & -1 & 0 & 0 \\ -1 & \frac{4}{9} & \frac{1}{9} & \frac{1}{9} \\ 0 & \frac{1}{9} & \frac{4}{9} & \frac{1}{9} \\ 0 & \frac{1}{9} & \frac{1}{2} & \frac{4}{9} \end{bmatrix}, Y_4 = \begin{bmatrix} \frac{1}{2} & -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{4}{9} & \frac{1}{9} & \frac{1}{9} \\ -\frac{1}{3} & \frac{1}{9} & \frac{4}{9} & \frac{1}{9} \\ -\frac{1}{3} & \frac{1}{9} & \frac{1}{2} & \frac{4}{9} \end{bmatrix},$$

and Y_2, Y_3 can be determined from Y_1. For $n = 4$,

$$Y_1 = \begin{bmatrix} \frac{15}{4} & -\frac{9}{8} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} \\ -\frac{9}{8} & \frac{3}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ -\frac{1}{8} & \frac{1}{8} & \frac{3}{8} & \frac{1}{8} & \frac{1}{8} \\ -\frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{3}{8} & \frac{1}{8} \\ -\frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{3}{8} \end{bmatrix}, Y_5 = \begin{bmatrix} \frac{3}{4} & -\frac{3}{8} & -\frac{3}{8} & -\frac{3}{8} & -\frac{3}{8} \\ -\frac{3}{8} & \frac{3}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ -\frac{3}{8} & \frac{1}{8} & \frac{3}{8} & \frac{1}{8} & \frac{1}{8} \\ -\frac{3}{8} & \frac{1}{8} & \frac{1}{8} & \frac{3}{8} & \frac{1}{8} \\ -\frac{3}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{3}{8} \end{bmatrix},$$

and Y_2, Y_3, Y_4 can be determined from Y_1. The rational numbers above are all chosen by observing the floating numbers output of the SDP (2.6).

The values of the matrices Y_i's for $n = 2, 3, 4$ allow us to guess that the variables $a, b, c, d, e, f, g, x, y, z, w$ are:

$$\begin{gathered} a = \frac{5(n-1)}{4}, \quad b = -\frac{3(n-1)}{2n}, \quad c = \frac{3-n}{2n}, \\ d = f = z = \frac{2(n-1)}{n^2}, \quad e = g = w = \frac{n-2}{n^2}, \\ x = \frac{n-1}{4}, \quad y = -\frac{n-1}{2n}. \end{gathered} \tag{2.11}$$

The proof of our next theorem will ascertain that these choices are indeed correct — they yield the sum-of-squares decomposition in (2.10) for $m = 2$.

Theorem 2.5.1 (Better Recht–Ré for $m = 2$)**.** *Let $A_1, \dots, A_n$ be positive semidefinite* *matrices. If* $A_1 + \cdots + A_n \preceq nI$*, then*

$$-\frac{1}{4}n(n-1)I \preceq \sum_{i \neq j} A_i A_j \preceq n(n-1)I.$$

Proof. The upper bound has already been established in Theorem 2.2.3. It remains to establish the lower bound. We start from the following readily verifiable inequalities

$$\frac{1}{2n(n-1)}(A_j - A_k)A_i(A_j - A_k) \succeq 0, \quad (2.12)$$

$$\frac{5(n-1)}{4}\Big(I - \frac{6}{5n}A_i + \frac{2(3-n)}{5n(n-1)}\sum_{j \neq i} A_j\Big)A_i\Big(I - \frac{6}{5n}A_i + \frac{2(3-n)}{5n(n-1)}\sum_{j \neq i} A_j\Big) \succeq 0 \quad (2.13)$$

$$\frac{n-1}{5n^2}\Big(A_i + \frac{2n-1}{n-1}\sum_{j \neq i} A_j\Big)A_i\Big(A_i + \frac{2n-1}{n-1}\sum_{j \neq i} A_j\Big) \succeq 0 \quad (2.14)$$

$$\frac{1}{2n^2}(A_j - A_k)\Big(n - \sum_i A_i\Big)(A_j - A_k) \succeq 0, \quad (2.15)$$

$$\frac{n-1}{4}\Big(I - \frac{2}{n}\sum_i A_i\Big)\Big(n - \sum_i A_i\Big)\Big(I - \frac{2}{n}\sum_i A_i\Big) \succeq 0. \quad (2.16)$$

Sum (2.12) over all distinct i, j, k; sum (2.13) over all i; sum (2.14) over all i; sum (2.15) over all distinct j, k; add all results to (2.16). The final inequality is our required lower bound. □

For $n = m = 2$, the new lower bound is sharp. Take

$$A_1 = \begin{bmatrix} \frac{3}{2} & 0 \\ 0 & 0 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} \frac{1}{6} & \frac{\sqrt{2}}{3} \\ \frac{\sqrt{2}}{3} & \frac{4}{3} \end{bmatrix},$$

then $\|A_1 + A_2\| = 2$ and the smallest eigenvalue of $A_1A_2 + A_2A_1$ is $-1/2$. We conjecture that this bound is sharp for all $m = 2$, $n \geq 2$.

The method in this section also extends to higher m. For example, we may impose symmetry constraints for $m = n = 3$ and see if the Y_1, Y_2, Y_3, Y_4 obtained have rational values, and if so write down a sums-of-squares proof by factoring the Y_i's.

2.6 Conclusion and open problems

We conclude our article with a discussion of some open problems and why we think the Recht–Ré conjecture, while false as it is currently stated, only needs to be refined.

An immediate open question is whether the conjecture is true for $m = 4$: Table 2.1 shows that it holds for $(m, n) = (4, 4)$ and $(4, 5)$; we suspect that it is true for all $n \geq 4$.

As we pointed out after Conjecture 2.2.2, the Recht–Ré inequality as stated in (2.1) conceals an asymmetry — it actually contains two inequalities, as shown in (2.2). What we have seen is that the lower bound is never attained in any of the cases we have examined. For $m = 2$ and 3, the lower bound is too large, and we improved it in Theorem 2.5.1 and the proof of Theorem 2.2.4 respectively. For $m = 5$, the lower bound is too small, which is why the Recht–Ré inequality is false. A natural follow-up question is then: "What is the correct lower bound?" On the other hand, we conjecture that the remaining half of the Recht–Ré inequality, i.e., the upper bound in (2.2), holds true for all $m \leq n \in \mathbb{N}$.

[60] has another conjecture similar to Conjecture 2.1.1 but where the norms appear after the summation.

Conjecture 2.6.1 ([60])**.** *Let $A_1, \dots, A_n$ be positive semidefinite matrices. Then*

$$\frac{1}{n^m} \sum_{1 \leq j_1, \dots, j_m \leq n} \|A_{j_1} \cdots A_{j_m}\| \geq \frac{(n-m)!}{n!} \sum_{\substack{1 \leq j_1, \dots, j_m \leq n, \\ j_1, \dots, j_m \text{ distinct}}} \|A_{j_1} \cdots A_{j_m}\|.$$

For $m = 2$ and 3, the conjecture has been proved for any unitarily invariant norm [109]. It is not clear to us if the noncommutative Positivstellensatz might perhaps also shed light on Conjecture 2.6.1.

Lastly, if our intention is to analyze the relative efficacies of with-replacement and without-replacement sampling strategies in randomized algorithms, then it is more pertinent to study these inequalities for random matrices, i.e., we do not just assume that the indices are random variables but also the entries of the matrices. For example, if we want to analyze the Kaczmarz algorithm, then we ought to take expectation not only with respect to all permutations but also with respect to how we generate the entries of the matrices. This would provide a more realistic platform for comparing different sampling strategies.

CHAPTER 3

SIMPLER GRASSMANNIAN OPTIMIZATION

This is a joint work with Lek-Heng Lim and Ke Ye.

3.1 Introduction

As a manifold, the Grassmannian $\operatorname{Gr}(k,n)$ is just the set of k-planes in n-space with its usual differential structure; this is an abstract description that cannot be employed in algorithms and applications. In order to optimize functions $f : \operatorname{Gr}(k,n) \to \mathbb{R}$ using currently available technology, one needs to put a coordinate system on $\operatorname{Gr}(k,n)$. The best known way, as discovered by Edelman, Arias, and Smith in their classic work [65], is to realize $\operatorname{Gr}(k,n)$ as a *matrix manifold* [3], where every point on $\operatorname{Gr}(k,n)$ is represented by a matrix or an equivalence class of matrices and from which one may derive closed-form analytic expressions for other differential geometric objects (e.g., tangent, metric, geodesic) and differential geometric operations (e.g., exponential map, parallel transport) that in turn provide the necessary ingredients (e.g., Riemannian gradient and Hessian, conjugate direction, Newton step) for optimization algorithms. The biggest advantage afforded by the approach in [65] is that a judiciously chosen system of extrinsic *matrix coordinates* for points on $\operatorname{Gr}(k,n)$ allows all aforementioned objects, operations, and algorithms to be computed solely in terms of standard *numerical linear algebra*, which provides a ready supply of stable and accurate algorithms [86] with high-quality software implementations [9]. In particular, one does not need to solve any differential equations numerically when doing optimization on matrix manifolds à la [65].

3.1.1 Existing models

There are two well-known models for $\operatorname{Gr}(k, n)$ supplying such matrix coordinates — one uses orthogonal matrices and the other projection matrices. In optimization, the by-now standard model (see, for example, [59, 111, 145, 173, 198]) is the one introduced in [65], namely,

$$\operatorname{Gr}(k,n) \cong \mathrm{O}(n)/\big(\mathrm{O}(k) \times \mathrm{O}(n-k)\big) \cong \mathrm{V}(k,n)/\,\mathrm{O}(k), \tag{3.1}$$

where $\mathrm{V}(k,n) := \{V \in \mathbb{R}^{n \times k} : V^{\top}V = I\} \cong \mathrm{O}(n)/\,\mathrm{O}(n-k)$ is the Stiefel manifold. In this homogeneous space model, which is also widely used in areas other than optimization [11, 14, 84, 95, 143, 144, 209], a point $\mathbb{V} \in \operatorname{Gr}(k,n)$, i.e., a k-dimensional subspace $\mathbb{V} \subseteq \mathbb{R}^n$, is represented by its orthonormal basis, written as columns of a matrix $V = [v_1, \dots, v_k] \in \mathrm{V}(k,n)$. Since any two orthonormal bases $V_1, V_2 \in \mathrm{V}(k,n)$ of $\mathbb{V}$ must be related by $V_1 = V_2 Q$ for some $Q \in \mathrm{O}(k)$, such a representation is not unique and so this model requires that we represent $\mathbb{V}$ not as a single $n \times k$ orthonormal matrix but as a whole equivalence class $[V] := \{VQ \in \mathrm{V}(k,n) : Q \in \mathrm{O}(k)\}$ of orthonormal bases of $\mathbb{V}$. A brief word about our notations: Throughout this article, we adopt the convention that a vector space $\mathbb{V} \in \operatorname{Gr}(k,n)$ will be typeset in blackboard bold, with the corresponding letter in normal typeface $V \in \mathrm{V}(k,n)$ denoting an (ordered) orthonormal basis. Equivalence classes will be denoted in double brackets, so $[V] = \mathbb{V}$. Diffeomorphism of two smooth manifolds will be denoted by $\cong$.

It is straightforward to represent a point $\mathbb{V} \in \operatorname{Gr}(k,n)$ by an actual matrix as opposed to an equivalence class of matrices. Since any subspace $\mathbb{V}$ has a unique orthogonal projection matrix $P_{\mathbb{V}}$, this gives us an alternative model for the Grassmannian that is also widely used (notably in linear programming [181, 208] but also many other areas [46, 31, 49, 66, 142, 157]):

$$\operatorname{Gr}(k,n) \cong \{P \in \mathbb{R}^{n \times n} : P^{\top} = P = P^2,\ \operatorname{tr}(P) = k\}. \tag{3.2}$$

Note that $\operatorname{rank}(P) = \operatorname{tr}(P) = \dim(\mathbb{V})$ for orthogonal projection matrices. The reader is reminded that an *orthogonal* projection matrix is not an orthogonal matrix — the 'orthogonal' describes the projection, not the matrix. To avoid confusion, we drop 'orthogonal' from future descriptions — all projection matrices in our article will be orthogonal projection matrices.

As demonstrated in [100], it is also possible to derive closed-form analytic expressions for various differential geometric objects and present various optimization algorithms in terms of the matrix coordinates in (3.2). Nevertheless, the problem with the model (3.2) is that algorithms based on projection matrices are almost always *numerically unstable*, especially in comparison with algorithms based on orthogonal matrices. This is likely the reason why there are no numerical experiments in [100]. Roughly speaking an orthogonal matrix preserves (Euclidean) norms and therefore rounding errors do not get magnified through a sequence of orthogonal transformations [55, Section 3.4.4] and consequently algorithms based on orthogonal matrices tend to be numerically stable (details are more subtle, see [199, pp. 124–166] and [104]). Projection matrices not only do not preserve norms but are singular and give notoriously unstable algorithms — possibly the best known illustration of numerical instability [192, 197] is one that contrasts Gram–Schmidt, which uses projection matrices, with Householder QR, which uses orthogonal matrices.[1] In fact, the proper way to compute projections is to do so via a sequence of orthogonal matrices [185, pp. 260–261], as a straightforward computation is numerically unstable [47, pp. 849–851].

The alternative (3.1) is currently universally adopted for optimization over a Grassmannian. One issue with the model (3.1) is that a point on $\mathrm{Gr}(k, n)$ is not a single matrix but an *equivalence class* of uncountably many matrices. Equivalence classes are tricky to implement in numerical algorithms and standard algorithms in numerical linear algebra [9] do not work with equivalence classes of matrices. Given a function $f : \mathrm{Gr}(k, n) \to \mathbb{R}$ to be

[1] For example, computing the QR decomposition of a Hilbert matrix $A = [1/(i + j - 1)]_{i,j=1}^{15}$, we get $\|Q^*Q - I\| \approx 8.0 \times 10^0$ with Gram–Schmidt, 1.7×10^0 with modified Gram–Schmidt, 2.4×10^{-15} with Householder QR.

optimized, any optimization algorithm [59, 65, 111, 145, 173, 198] that rely on the model (3.1) side steps the issue by lifting f to an $\operatorname{O}(k)$-invariant function $\tilde{f} : \operatorname{V}(k,n) \to \mathbb{R}$, i.e., where $\tilde{f}(VQ) = \tilde{f}(V)$ for all $Q \in \operatorname{O}(k)$. This incurs additional costs in two ways: : (a) whenever a point $\mathbb{V} \in \operatorname{Gr}(k,n)$ needs to be lifted to a point $V \in \operatorname{V}(k,n)$, this incurs the cost of finding an orthonormal basis V for $\mathbb{V}$; (b) whenever one needs to check equality of points $\operatorname{im}(V_1) \overset{?}{=} \operatorname{im}(V_2)$, this incurs the cost of one matrix product $V_1^{\top} V_2$ and its norm.[2] These additional costs cannot be avoided in the model (3.1) as we represent a linear subspace by an equivalence class. For comparison, (a) and (b) are immaterial when points are represented as actual matrices, like in model (3.2) or our proposed model. Moreover it is impossible to continuously choose such 'Stiefel coordinates' $V \in \operatorname{V}(k,n)$ for every point $\mathbb{V} \in \operatorname{Gr}(k,n)$, as we will discuss in Section 3.7.5. A second and more serious issue with the model (3.1) is that its associated optimization algorithms in [65] are still significantly less stable than those for our proposed model. As we will see in Section 3.8, and for reasons explained therein, loss-of-orthogonality remains very much a problem when we use (3.1) to represent a Grassmannian. This is likely the reason why the numerical experiments in [65] had used extended precision arithmetic.

We would like to mention a noncompact analogue of (3.1) that is popular in combinatorics [2, 73, 80, 119, 135]:

$$\operatorname{Gr}(k,n) \cong \mathbb{R}^{n\times k}_{k} / \operatorname{GL}(k), \tag{3.3}$$

where $\mathbb{R}^{n\times k}_{k} := \{A \in \mathbb{R}^{n\times k} : \operatorname{rank}(A) = k\}$ and $\operatorname{GL}(k) := \{X \in \mathbb{R}^{k\times k} : \det(X) \neq 0\}$. It has also been shown [3] that one may obtain closed-form analytic expressions for differential geometric quantities with the model (3.3) and so in principle one may use it for optimization purposes. Nevertheless, from the perspective of numerical algorithms, the model (3.3) suffers from the same problem as (3.2) — by working with rank-k matrices, i.e., whose condition

[2] Note that $\operatorname{im}(V_1) = \operatorname{im}(V_2)$ is equivalent to $V_1V_1^{\top} = V_2V_2^{\top}$, which is equivalent to $\|V_1^{\top}V_2\| = k$ as $\|V_1V_1^{\top} - V_2V_2^{\top}\|^2 = 2k^2 - 2\|V_1^{\top}V_2\|^2$.

number can be arbitrarily large, algorithms based on (3.3) are inherently numerically unstable. In fact, since the model (3.3) also represents points as equivalence classes, it has both shortcomings of (3.1) and (3.2) but neither of their good features. The natural redress of imposing orthogonal constraints on (3.3) to get a well-conditioned representative for each equivalence class would just lead one back to the model (3.1).

Looking beyond optimization, we stress that each of the aforementioned models has its own (sometimes unique) strengths. For example, (3.3) is the only model we know in which one may naturally define the positive Grassmannian [73], an important construction in combinatorics [119] and physics [80]. The model (3.2) is indispensable in probability and statistics as probability measures [142, Section 3.9] and probability densities [46, Section 2.3.2] on $\mathrm{Gr}(k,n)$ are invariably expressed in terms of projection matrices.

3.1.2 Proposed model

We propose to use a model for the Grassmannian that combines the best features, suffers from none of the defects of the aforementioned models, and, somewhat surprisingly, is also simpler:

$$\mathrm{Gr}(k,n) \cong \{Q \in \mathrm{O}(n) : Q^{\top} = Q,\ \mathrm{tr}(Q) = 2k - n\}. \tag{3.4}$$

This model, which represents k-dimensional subspace as a symmetric orthogonal matrix of trace $2k-n$, is known but obscure. It was mentioned in passing in [17, p. 305] and was used in [116] to derive geodesics for the *oriented Grassmannian*, a different but related manifold. Note that (3.4) merely provides an expression for *points*, our main contribution is to derive expressions for other differential geometric objects and operations, as well as the corresponding optimization algorithms, thereby fully realizing (3.4) as a model for optimization. A summary of these objects, operations, and algorithms is given in Table 3.1. From a differential geometric perspective, Sections 3.2–3.5 may be regarded as an investigation into the embedded geometry of $\mathrm{Gr}(k,n)$ as a submanifold of $\mathrm{O}(n)$.

OBJECTS/OPERATIONS	RESULTS
point	Proposition 3.2.1
change-of-coordinates	Proposition 3.2.2, 3.2.3, 3.2.4, 3.2.5
tangent vector	Proposition 3.3.1, 3.3.2, Corollary 3.3.3
metric	Proposition 3.3.4, 3.3.5
normal vector	Proposition 3.3.6, Corollary 3.3.7
curve	Proposition 3.4.2
geodesic	Theorem 3.4.3, Proposition 3.4.5
geodesic distance	Corollary 3.4.6
exponential map	Corollary 3.4.4
logarithmic map	Corollary 3.4.7
parallel transport	Proposition 3.4.8
gradient	Proposition 3.5.1, Corollary 3.5.3
Hessian	Proposition 3.5.2
retraction and vector transport	Proposition 3.6.4, 3.6.5, 3.6.6
steepest descent	Algorithm 1, 2
Newton method	Algorithm 3
conjugate gradient	Algorithm 4
quasi-Newton	Algorithm 5

Table 3.1: Guide to results.

The two key advantages of the model (3.4) in computations are that: (i) we represent points on $\mathrm{Gr}(k, n)$ as actual matrices, not equivalence classes; (ii) we work only with orthogonal matrices and in numerical stable ways.

Numerical stability is an important feature of the algorithms for model (3.4); as we will see in Section 3.8, the errors and gradients in our steepest descent and Newton algorithms consistently reduce to the order of machine precision. Moreover, another bonus with (3.4) is that the expressions and algorithms in Table 3.1 are considerably simpler compared to those in [3, 65, 100]. We will not need to solve quadratic eigenvalue problems, nor compute exp/cos/sin/sinc of nonnormal matrices, nor even EVD or SVD except in cases when they can be trivially obtained. Aside from standard matrix arithmetic, our optimization algorithms require just two operations. In fact all differential geometric objects and operations can be computed with at most a QR decomposition and an exponentiation of a skew-symmetric matrix,

$$\exp\left(\begin{bmatrix} 0 & B \\ -B^{\top} & 0 \end{bmatrix}\right), \quad B \in \mathbb{R}^{k\times(n-k)}. \tag{3.5}$$

The problem of computing matrix exponential has been thoroughly studied and there is a plethora of algorithms [106, 147], certainly more so than other transcendental matrix functions like cosine, sine, or sinc [106]. For normal matrices, matrix exponentiation is a well-conditioned problem — the numerical issues described in [147] only occur with nonnormal matrices. For us, $\left[\begin{smallmatrix} 0 & B \\ -B^{\top} & 0 \end{smallmatrix}\right]$ is skew-symmetric and thus normal; in fact its exponential will always be an orthogonal matrix.

There are other algorithmic advantages afforded by (3.4) that are difficult to explain without context and will be discussed alongside the algorithms in Section 3.7 and numerical results in Section 3.8. In particular, our algorithms will work with what we call "effective gradients," "effective Newton steps," "effective conjugate directions," etc — these are all matrices of size $k \times (n-k)$ like the matrix B in (3.5), i.e., they have the intrinsic dimension of $\mathrm{Gr}(k, n)$. With this we would also like to add a note of caution. One cannot infer an accurate estimate of computational complexity based on a simple dimension count of the models in Table 3.2. There are many differential geometric objects and operations involved in an algorithm, such as those in Table 3.1, and not just points. The matrices arising in actual computations are highly structured and the computational cost depends heavily on the specific problem.

3.1.3 Nomenclatures and notations

For easy reference, we will introduce names for the models (3.1)–(3.4) based on the type of matrices used as coordinates for points.

We note that there are actually two homogeneous space models for $\mathrm{Gr}(k, n)$ in (3.1), one as a quotient of $\mathrm{O}(n)$ and the other as a quotient of $\mathrm{V}(k, n)$. While they are used somewhat

NAME	MODEL
orthogonal model	$\mathrm{O}(n)/\bigl(\mathrm{O}(k)\times\mathrm{O}(n-k)\bigr)$
Stiefel model	$\mathrm{V}(k,n)/\,\mathrm{O}(k)$
full-rank model	$\mathbb{R}^{n\times k}_{k}/\,\mathrm{GL}(k)$
projection model	$\{P\in\mathbb{R}^{n\times n} : P^{\top}=P=P^{2},\ \mathrm{tr}(P)=k\}$
involution model	$\{Q\in\mathrm{O}(n) : Q^{\top}=Q,\ \mathrm{tr}(Q)=2k-n\}$

Table 3.2: Matrix manifold models for the Grassmannian $\mathrm{Gr}(k,n)$.

interchangeably in [65], we distinguish them in Table 3.2 as their change-of-coordinates maps to the involution model are different (see Section 3.2).

The name *involution model* is warranted for (3.4) because for any $Q\in\mathbb{R}^{n\times n}$, any two of the following conditions clearly imply the third:

$$Q^{\top}Q=I,\qquad Q^{\top}=Q,\qquad Q^{2}=I.$$

Thus a symmetric orthogonal matrix may also be viewed as a symmetric involution or an orthogonal involution matrix. We will need the eigendecomposition of a matrix in the involution model for *all* of our subsequent calculations; for easy reference we state this as a lemma. Such an eigendecomposition is trivial to compute, requiring only a single QR decomposition (of the matrix $\frac{1}{2}(I+Q)$; see Lemma 3.7.1).

Lemma 3.1.1. *Let $k=1,\dots,n$ and $Q\in\mathbb{R}^{n\times n}$ be such that*

$$Q^{\top}Q=I,\qquad Q^{\top}=Q,\qquad \mathrm{tr}(Q)=2k-n.$$

Then Q has an eigenvalue decomposition

$$Q = VI_{k,n-k}V^\top = [y_1, \dots, y_k, z_1, \dots, z_{n-k}] \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & -1 & & \\ & & & & \ddots & \\ & & & & & -1 \end{bmatrix} \begin{bmatrix} y_1^\top \\ \vdots \\ y_k^\top \\ z_1^\top \\ \vdots \\ z_{n-k}^\top \end{bmatrix},$$

where $V \in \mathrm{O}(n)$ and $I_{k,n-k} \coloneqq \operatorname{diag}(I_k, -I_{n-k}) = \operatorname{diag}(1, \dots, 1, -1, \dots, -1)$.

Proof. Existence of an eigendecomposition follows from the symmetry of Q. A symmetric involution has all eigenvalues ±1 and the multiplicity of 1 must be k since $\operatorname{tr}(Q) = 2k-n$. □

Henceforth, for a matrix Q in the involution model, we write

$$\begin{aligned} &Y_Q \coloneqq [y_1, \dots, y_k] \in \mathrm{V}(k, n), \qquad Z_Q \coloneqq [z_1, \dots, z_{n-k}] \in \mathrm{V}(n-k, n), \\ &V_Q = [Y_Q, Z_Q] = V \in \mathrm{O}(n) \end{aligned} \tag{3.6}$$

for its matrix of 1-eigenvectors, its matrix of -1-eigenvectors, and its matrix of all eigenvectors respectively. While these matrices are not unique, the 1-eigenspace and -1-eigenspace

$$\operatorname{im}(Y_Q) = \operatorname{span}\{y_1, \dots, y_k\} \in \mathrm{Gr}(k, n), \qquad \operatorname{im}(Z_Q) = \operatorname{span}\{z_1, \dots, z_{n-k}\} \in \mathrm{Gr}(n-k, n)$$

are uniquely determined by Q.

3.2 Points and change-of-coordinates

We begin by exhibiting a diffeomorphism to justify the involution model, showing that as *smooth manifolds*, $\mathrm{Gr}(k, n)$ and $\{Q \in \mathrm{O}(n) : Q^\top = Q,\ \operatorname{tr}(Q) = 2k - n\}$ are the same.

In the next section, we will show that if we equip the latter with appropriate Riemannian metrics, then as *Riemannian manifolds*, they are also the same, i.e., the diffeomorphism is an isometry. The practically minded may simply take this as establishing a system of matrix coordinates for points on $\operatorname{Gr}(k,n)$.

Proposition 3.2.1 (Points)**.** *Let $k = 1, \dots, n$. Then the map*

$$\begin{aligned} \varphi : \operatorname{Gr}(k,n) &\to \{Q \in \mathrm{O}(n) : Q^\top = Q,\ \operatorname{tr}(Q) = 2k - n\}, \\ \varphi(\mathbb{W}) &= P_{\mathbb{W}} - P_{\mathbb{W}^\perp}, \end{aligned} \tag{3.7}$$

is a diffeomorphism with $\varphi^{-1}(Q) = \operatorname{im}(Y_Q)$ where $Y_Q \in \mathrm{V}(k,n)$ is as in (3.6)*. Here $P_{\mathbb{W}}$ denotes the orthogonal projection matrix to the subspace $\mathbb{W}$.*

Proof. One can check that $Q = P_{\mathbb{W}} - P_{\mathbb{W}^\perp}$ is symmetric, orthogonal, and has trace $2k - n$. So the map φ is well-defined. If we write $\psi(Q) = \operatorname{im}(Y_Q)$, then $\varphi(\psi(Q)) = Q$ and $\psi(\varphi(\mathbb{W})) = \mathbb{W}$, so $\psi = \varphi^{-1}$. To see that φ is smooth, we may choose any local coordinates, say, represent $\mathbb{W} \in \operatorname{Gr}(k,n)$ in terms of any orthonormal basis $W = [w_1, \dots, w_k] \in \mathrm{V}(k,n)$ and observe that

$$\varphi(\mathbb{W}) = 2WW^\top - I,$$

which is smooth. With a linear change-of-coordinates, we may assume that

$$W = \begin{bmatrix} I_k \\ 0 \end{bmatrix}.$$

The differential $(d\varphi)_{\mathbb{W}}$ is given by the (clearly invertible) linear map

$$(d\varphi)_{\mathbb{W}} \left(\begin{bmatrix} 0 \\ X \end{bmatrix} \right) = 2 \left(\begin{bmatrix} I_k \\ 0 \end{bmatrix} \begin{bmatrix} 0 & X^\top \end{bmatrix} + \begin{bmatrix} 0 \\ X \end{bmatrix} \begin{bmatrix} I_k & 0 \end{bmatrix} \right) = 2 \begin{bmatrix} 0 & X^\top \\ X & 0 \end{bmatrix}$$

for all $X \in \mathbb{R}^{(n-k)\times k}$. So φ is a diffeomorphism. □

Since the manifolds in Table 3.2 are all diffeomorphic to $\operatorname{Gr}(k,n)$, they are diffeomorphic to each other. Our next results are not intended to establish that they are diffeomorphic but to construct these diffeomorphisms and their inverses explicitly, so that we may switch to and from the other systems of coordinates easily.

In the next proposition, $[V] = \Bigl\{V\bigl[\begin{smallmatrix} Q_1 & 0 \\ 0 & Q_2 \end{smallmatrix}\bigr] : Q_1 \in \mathrm{O}(k),\ Q_2 \in \mathrm{O}(n-k)\Bigr\}$ denotes equivalence class in $\mathrm{O}(n)/\bigl(\mathrm{O}(k)\times\mathrm{O}(n-k)\bigr)$.

Proposition 3.2.2 (Change-of-coordinates I)**.** *Let $k = 1,\dots,n$. Then*

$$\varphi_1 : \mathrm{O}(n)/\bigl(\mathrm{O}(k)\times\mathrm{O}(n-k)\bigr) \to \{Q \in \mathrm{O}(n) : Q^\top = Q,\ \operatorname{tr}(Q) = 2k-n\},$$
$$\varphi_1([V]) = V^\top I_{k,n-k} V$$

is a diffeomorphism with $\varphi_1^{-1}(Q) = [V_Q]$ with $V_Q \in \mathrm{O}(n)$ as in (3.6).

Proof. Note that $Q = V_1 I_{k,n-k} V_1^\top = V_2 I_{k,n-k} V_2^\top$ iff

$$V_2 = V_1 \begin{bmatrix} Q_1 & 0 \\ 0 & Q_2 \end{bmatrix}$$

for some $(Q_1, Q_2) \in \mathrm{O}(k)\times\mathrm{O}(n-k)$ iff $[V_1] = [V_2]$. Hence both φ_1 and φ_1^{-1} are well-defined and are inverses of each other. Observe that φ_1 is induced from the map

$$\widetilde{\varphi}_1 : \mathrm{O}(n) \to \{Q \in \mathrm{O}(n) : Q^\top = Q,\ \operatorname{tr}(Q) = 2k-n\}, \quad \widetilde{\varphi}_1(V) = V^\top I_{k,n-k} V,$$

which is a surjective submersion. The proof that φ_1^{-1} is well-defined shows that the fibers of $\widetilde{\varphi}_1$ are exactly the $\mathrm{O}(k)\times\mathrm{O}(n-k)$-orbits in $\mathrm{O}(n)$. Hence φ_1, as the composition of $\widetilde{\varphi}_1$ and the quotient map $\mathrm{O}(n) \to \mathrm{O}(n)/\bigl(\mathrm{O}(k)\times\mathrm{O}(n-k)\bigr)$, is a diffeomorphism. □

The next result explains the resemblance between the projection and involution models — each is a scaled and translated copy of the other. The scaling and translation are judiciously chosen so that orthogonal projections become symmetric involutions, and this seemingly innocuous difference will have a significant impact on the numerical stability of Grassmannian optimization algorithms.

Proposition 3.2.3 (Change-of-coordinates II)**.** *Let $k = 1, \dots, n$. Then*

$$\varphi_2 : \{P \in \mathbb{R}^{n\times n} : P^\top = P = P^2,\ \operatorname{tr}(P) = k\} \to \{Q \in \mathrm{O}(n) : Q^\top = Q,\ \operatorname{tr}(Q) = 2k - n\},$$
$$\varphi_2(P) = 2P - I$$

is a diffeomorphism with $\varphi_2^{-1}(Q) = \frac{1}{2}(I + Q)$.

Proof. Note that $2P - I = P - P^\perp$ where $P^\perp$ is the projection onto the orthogonal complement of $\operatorname{im}(P)$, so both φ_2 and φ_2^{-1} are well-defined. They are clearly diffeomorphisms and are inverses to each other. □

In the next proposition, $[Y] = \{YQ : Q \in \mathrm{O}(k)\}$ denotes a equivalence class in $\mathrm{V}(k, n)/\mathrm{O}(k)$.

Proposition 3.2.4 (Change-of-coordinates III)**.** *Let $k = 1, \dots, n$. Then*

$$\varphi_3 : \mathrm{V}(k, n)/\mathrm{O}(k) \to \{Q \in \mathrm{O}(n) : Q^\top = Q,\ \operatorname{tr}(Q) = 2k - n\},$$
$$\varphi_3([Y]) = 2YY^\top - I$$

is a diffeomorphism with $\varphi_3^{-1}(Q) = [Y_Q]$ with $Y_Q \in \mathrm{V}(k, n)$ as in (3.6).

Proof. Given $[Y] \in \mathrm{V}(k, n)/\mathrm{O}(k)$, the matrix $YY^\top$ is the projection matrix onto the k-dimensional subspace $\operatorname{im}(Y) \in \mathrm{Gr}(k, n)$. Hence φ_3 is a well-defined map by Proposition 3.2.3. To show that its inverse is given by $\psi_3(Q) = [Y_Q]$, observe that any $Y \in \mathrm{V}(k, n)$ can be extended to a full orthogonal matrix $V := [Y, Y^{\perp}] \in \mathrm{O}(n)$ and we have

$$V^{\top} Y = \begin{bmatrix} I_k \\ 0 \end{bmatrix}, \qquad Q = 2YY^{\top} - I = V \begin{bmatrix} 2I_k & 0 \\ 0 & 0 \end{bmatrix} V^{\top} - I = V I_{k,n-k} V^{\top}.$$

This implies that $\psi_3 \circ \varphi_3([Y]) = [Y_Q] = [Y]$. That φ_3 is a diffeomorphism follows from the same argument in the proof of Proposition 3.2.1. □

In the next proposition, $[A] = \{AX : X \in \mathrm{GL}(k)\}$ denotes equivalence class in $\mathbb{R}^{n \times k}_k / \mathrm{GL}(k)$. Also, we write $A = Y_A R_A$ for the QR factorization of $A \in \mathbb{R}^{n \times k}_k$, i.e., $Y_A \in \mathrm{V}(k, n)$ and $R_A \in \mathbb{R}^{k \times k}$ is upper triangular.

Proposition 3.2.5 (Change-of-coordinates IV)**.** *Let $k = 1, \dots, n$. Then*

$$\begin{aligned} \varphi_4 : \mathbb{R}^{n \times k}_k / \mathrm{GL}(k) &\to \{Q \in \mathrm{O}(n) : Q^{\top} = Q,\ \mathrm{tr}(Q) = 2k - n\}, \\ \varphi_4([A]) &= 2Y_A Y_A^{\top} - I \end{aligned}$$

is a diffeomorphism with $\varphi_4^{-1}(Q) = [Y_Q]$ with Y_Q is as in (3.6)*.*

Proof. First observe that $\mathrm{V}(k, n) \subseteq \mathbb{R}^{n \times k}_k$ and the inclusion map $\mathrm{V}(k, n) \hookrightarrow \mathbb{R}^{n \times k}_k$ induces a diffeomorphism $\mathrm{V}(k, n)/\mathrm{O}(k) \cong \mathbb{R}^{n \times k}_k / \mathrm{GL}(k)$ — if we identify them, then φ_4^{-1} becomes φ_3^{-1} in Proposition 3.2.4 and is thus a diffeomorphism. It follows that φ_4 is a diffeomorphism. That the maps are inverses to each other follows from the same argument in the proof of Proposition 3.2.4. □

The maps $\varphi, \varphi_1, \varphi_2, \varphi_3, \varphi_4$ allow one to transform an optimization problem formulated in terms of abstract k-dimensional subspaces or in terms of one of the first four models in Table 3.2 into a mathematically (but not computationally) equivalent problem in terms of the involution model. Note that these are change-of-coordinate maps for *points* — they are good for translating expressions that involve only points on $\mathrm{Gr}(k, n)$. In particular, one cannot simply apply these maps to the analytic expressions for other differential geometric objects and operations in [3, 65, 100] and obtain corresponding expressions for the involution model. Deriving these requires considerable effort and would take up the next three sections.

Henceforth we will identify the Grassmannian with the involution model:

$$\operatorname{Gr}(k,n) \coloneqq \{Q \in \mathrm{O}(n) : Q^{\top} = Q,\ \operatorname{tr}(Q) = 2k - n\},$$

i.e., in the rest of our article, points on $\operatorname{Gr}(k,n)$ are symmetric orthogonal matrices of trace $2k-n$. With this, the well-known isomorphism

$$\operatorname{Gr}(k,n) \cong \operatorname{Gr}(n-k,n), \tag{3.8}$$

which we will need later, is simply given by the map $Q \mapsto -Q$.

3.3 Metric, tangents, and normals

The simple observation in Lemma 3.1.1 implies that a neighborhood of any point $Q \in \operatorname{Gr}(k,n)$ is just like a neighborhood of the special point $I_{k,n-k} = \operatorname{diag}(I_k, -I_{n-k}) \in \operatorname{Gr}(k,n)$. Consequently, objects like tangent spaces and curves at Q can be determined by simply determining them at $I_{k,n-k}$. Although $\operatorname{Gr}(k,n)$ is not a Lie group, the involution model, which models it as a linear section of $\mathrm{O}(n)$, allows certain characteristics of a Lie group to be retained. Here $I_{k,n-k}$ has a role similar to that of the identity element in a Lie group.

We will provide three different expressions for vectors in the *tangent space* $\mathbb{T}_Q \operatorname{Gr}(k,n)$ at a point $Q \in \operatorname{Gr}(k,n)$: an implicit form (3.9) as traceless symmetric matrices that anticommutes with Q and two explicit forms (3.10), (3.11) parameterized by $k \times (n-k)$ matrices. Recall from Lemma 3.1.1 that any $Q \in \operatorname{Gr}(k,n)$ has an eigendecomposition of the form $Q = VI_{k,n-k}V^{\top}$ for some $V \in \mathrm{O}(n)$.

Proposition 3.3.1 (Tangent space I)**.** *Let* $Q \in \operatorname{Gr}(k,n)$ *with eigendecomposition* $Q = VI_{k,n-k}V^\top$*. The tangent space of* $\operatorname{Gr}(k,n)$ *at* Q *is given by*

$$\begin{aligned}
\mathbb{T}_Q \operatorname{Gr}(k,n) &= \left[X \in \mathbb{R}^{n\times n} : X^\top = X,\ XQ + QX = 0,\ \operatorname{tr}(X) = 0\right] && (3.9)\\
&= \left[V \begin{bmatrix} 0 & B \\ B^\top & 0 \end{bmatrix} V^\top \in \mathbb{R}^{n\times n} : B \in \mathbb{R}^{k\times(n-k)}\right] && (3.10)\\
&= \left[QV \begin{bmatrix} 0 & B \\ -B^\top & 0 \end{bmatrix} V^\top \in \mathbb{R}^{n\times n} : B \in \mathbb{R}^{k\times(n-k)}\right]. && (3.11)
\end{aligned}$$

Proof. By definition, a curve γ in $\operatorname{Gr}(k,n)$ passing through Q satisfies

$$\gamma(t)^\top - \gamma(t) = 0, \quad \gamma(t)^\top\gamma(t) = I_n, \quad \operatorname{tr}(\gamma(t)) = 2k - n, \quad t \in (-\varepsilon, \varepsilon),$$

together with the initial condition $\gamma(0) = Q$. Differentiating these equations at $t = 0$, we get

$$\dot\gamma(0)^\top - \dot\gamma(0) = 0, \quad \dot\gamma(0)^\top Q + Q^\top\dot\gamma(0) = 0, \quad \operatorname{tr}(\dot\gamma(0)) = 0,$$

from which (3.9) follows. Now take $X \in \mathbb{T}_Q \operatorname{Gr}(k,n)$. By (3.9), $V^\top XVI_{k,n-k} = V^\top(XQ)V$ is skew-symmetric and $V^\top XV$ is symmetric. Partition

$$V^\top XV = \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix}, \qquad A \in \mathbb{R}^{k\times k},\ B \in \mathbb{R}^{k\times(n-k)},\ C \in \mathbb{R}^{(n-k)\times(n-k)}.$$

Note that A and C are symmetric matrices since X is. So if

$$V^\top XVI_{k,n-k} = \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} = \begin{bmatrix} A & -B \\ B^\top & -C \end{bmatrix}$$

is skew-symmetric, then we must have $A = 0$ and $C = 0$ and we obtain (3.10). Since

$Q = VI_{k,n-k}V^\top$ and $Q = Q^\top$, (3.11) follows from (3.10) by writing $V = QVI_{k,n-k}$. □

The implicit form in (3.9) is inconvenient in algorithms. Of the two explicit forms (3.10) and (3.11), the description in (3.10) is evidently more economical, involving only V, as opposed to both Q and V as in (3.11). Henceforth, (3.10) will be our preferred choice and we will assume that a tangent vector at $Q \in \operatorname{Gr}(k,n)$ always takes the form

$$X = V \begin{bmatrix} 0 & B \\ B^\top & 0 \end{bmatrix} V^\top, \tag{3.12}$$

for some $B \in \mathbb{R}^{k\times(n-k)}$. This description appears to depend on the eigenbasis V, which is not unique, as Q has many repeated eigenvalues. The next proposition, which relates two representations of the same tangent vector with respect to two different V's, guarantees that the tangent space obtained will nonetheless be the same regardless of the choice of V.

Proposition 3.3.2 (Tangent vectors). *If $V_1 I_{k,n-k} V_1^\top = Q = V_2 I_{k,n-k} V_2^\top$, then any $X \in \mathbb{T}_Q \operatorname{Gr}(k,n)$ can be written as*

$$X = V_2 \begin{bmatrix} 0 & B \\ B^\top & 0 \end{bmatrix} V_2^\top = V_1 \begin{bmatrix} 0 & Q_1 B Q_2^\top \\ Q_2 B^\top Q_1^\top & 0 \end{bmatrix} V_1^\top,$$

for some $Q_1 \in \mathrm{O}(k)$ and $Q_2 \in \mathrm{O}(n-k)$ such that

$$V_2 = V_1 \begin{bmatrix} Q_1 & 0 \\ 0 & Q_2 \end{bmatrix}. \tag{3.13}$$

Proof. This is a consequence of the fact that $V_1 I_{k,n-k} V_1^\top = Q = V_2 I_{k,n-k} V_2^\top$ iff there exist $Q_1 \in \mathrm{O}(k)$ and $Q_2 \in \mathrm{O}(n-k)$ such that (3.13) holds. □

Another consequence of using (3.10) is that the tangent space at any point Q is a copy of the tangent space at $I_{k,n-k}$, conjugated by any eigenbasis V of Q; by Proposition 3.3.2, this is independent of the choice of V.

Corollary 3.3.3 (Tangent space II). *The tangent space at $I_{k,n-k}$ is*

$$\mathbb{T}_{I_{k,n-k}} \operatorname{Gr}(k,n) = \left[\begin{bmatrix} 0 & B \\ B^\top & 0 \end{bmatrix} : B \in \mathbb{R}^{k\times(n-k)} \right].$$

For any $Q \in \operatorname{Gr}(k,n)$ with eigendecomposition $Q = VI_{k,n-k}V^\top$,

$$\mathbb{T}_Q \operatorname{Gr}(k,n) = V\big(\mathbb{T}_{I_{k,n-k}} \operatorname{Gr}(k,n)\big)V^\top.$$

With the tangent spaces characterized, we may now define an inner product $\langle\cdot,\cdot\rangle_Q$ on each $\mathbb{T}_Q \operatorname{Gr}(k,n)$ that varies smoothly over all $Q \in \operatorname{Gr}(k,n)$, i.e., a *Riemannian metric*. With the involution model, $\operatorname{Gr}(k,n)$ is a submanifold of $\operatorname{O}(n)$ and there is a natural choice, namely, the Riemannian metric inherited from that on $\operatorname{O}(n)$.

Proposition 3.3.4 (Riemannian metric). *Let $Q \in \operatorname{Gr}(k,n)$ with $Q = VI_{k,n-k}V^\top$ and*

$$X = V \begin{bmatrix} 0 & B \\ B^\top & 0 \end{bmatrix} V^\top, \quad Y = V \begin{bmatrix} 0 & C \\ C^\top & 0 \end{bmatrix} V^\top \in \mathbb{T}_Q \operatorname{Gr}(k,n).$$

Then

$$\langle X, Y\rangle_Q \coloneqq \operatorname{tr}(XY) = 2\operatorname{tr}(B^\top C) \tag{3.14}$$

defines a Riemannian metric. The corresponding Riemannian norm is

$$\|X\|_Q \coloneqq \sqrt{\langle X, X\rangle_Q} = \|X\|_{\text{Fro}} = \sqrt{2}\|B\|_{\text{Fro}}. \tag{3.15}$$

The Riemannian metric in (3.14) is induced by the unique (up to a positive constant multiple) bi-invariant Riemannian metric on $\mathrm{O}(n)$:

$$g_Q(X,Y) \coloneqq \operatorname{tr}(X^\top Y), \quad Q \in \mathrm{O}(n), \quad X,Y \in \mathbb{T}_Q \mathrm{O}(n).$$

Here bi-invariance may be taken to mean

$$g_{V_1 Q V_2^\top}(V_1 X V_2^\top, V_1 Y V_2^\top) = g_Q(X,Y)$$

for all $Q, V_1, V_2 \in \mathrm{O}(n)$ and $X, Y \in \mathbb{T}_Q \mathrm{O}(n)$.

There are also natural Riemannian metrics [3, 65, 100] on the other four models in Table 3.2 but they differ from each other by a constant. As such, it is not possible for us to choose our metric (3.14) so that the diffeomorphisms in Propositions 3.2.2–3.2.5 are all isometry but we do have the next best thing.

Proposition 3.3.5 (Isometry)**.** *All models in Table 3.2 are, up to a constant factor, isometric as Riemannian manifolds.*

Proof. We verify that the diffeomorphism φ_1 in Proposition 3.2.2 gives an isometry between the orthogonal model and the involution model up a constant factor of 8. A tangent vector [65, Equation 2.30] at a point $[V] \in \mathrm{O}(n)/\big(\mathrm{O}(k) \times \mathrm{O}(n-k)\big)$ takes the form

$$V \begin{bmatrix} 0 & B \\ -B^\top & 0 \end{bmatrix} \in \mathbb{T}_{[V]} \mathrm{O}(n)/\big(\mathrm{O}(k) \times \mathrm{O}(n-k)\big), \quad B \in \mathbb{R}^{k \times (n-k)};$$

and the Riemannian metric [65, Equation 2.31] on $\mathrm{O}(n)/\big(\mathrm{O}(k) \times \mathrm{O}(n-k)\big)$ is given by

$$g_{[V]}\left(V \begin{bmatrix} 0 & B_1 \\ -B_1^\top & 0 \end{bmatrix}, V \begin{bmatrix} 0 & B_2 \\ -B_2^\top & 0 \end{bmatrix} \right) = \operatorname{tr}(B_1^\top B_2).$$

At I_n, the differential can be computed by

$$(d\varphi_1)_{[I_n]}\Bigl(I_n \begin{bmatrix} 0 & B \\ -B^\top & 0 \end{bmatrix}\Bigr) = 2I_{k,n-k} \begin{bmatrix} 0 & B \\ -B^\top & 0 \end{bmatrix} = 2 \begin{bmatrix} 0 & B \\ B^\top & 0 \end{bmatrix}.$$

Since both g and $\langle \cdot, \cdot \rangle$ are invariant under left multiplication by $\mathrm{O}(n)$, we have

$$\Bigl\langle (d\varphi_1)_{[V]}\Bigl(V \begin{bmatrix} 0 & B_1 \\ -B_1^\top & 0 \end{bmatrix}\Bigr), (d\varphi_1)_{[V]}\Bigl(V \begin{bmatrix} 0 & B_2 \\ -B_2^\top & 0 \end{bmatrix}\Bigr)\Bigr\rangle_{\varphi_1([V])} = 8\operatorname{tr}(B_1^\top B_2).$$

The proofs for $\varphi_2, \varphi_3, \varphi_4$ are similar and thus omitted. □

As the above proof shows, the diffeomorphism φ_1 may be easily made an isometry of the orthogonal and involution models by simply changing our metric in (3.14) to "$\langle X, Y\rangle_Q := \frac{1}{8}\operatorname{tr}(XY)$." Had we wanted to make φ_2 into an isometry of the projection and involution models, we would have to choose "$\langle X, Y\rangle_Q := \frac{1}{2}\operatorname{tr}(XY)$" instead. We see no reason to favor any single existing model and we stick to our choice of metric in (3.14).

In the involution model, $\mathrm{Gr}(k, n) \subseteq \mathrm{O}(n)$ as a smoothly embedded submanifold and every point $Q \in \mathrm{Gr}(k, n)$ has a *normal space* $\mathbb{N}_Q \mathrm{Gr}(k, n)$. We will next determine the expressions for normal vectors.

Proposition 3.3.6 (Normal space)**.** *Let $Q \in \mathrm{Gr}(k, n)$ with $Q = VI_{k,n-k}V^\top$. The normal space of $\mathrm{Gr}(k, n)$ at Q is given by*

$$\mathbb{N}_Q \mathrm{Gr}(k, n) = \left[V \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} V^\top \in \mathbb{R}^{n\times n} : \begin{matrix} \Lambda_1 \in \mathbb{R}^{k\times k}, & \Lambda_2 \in \mathbb{R}^{(n-k)\times(n-k)} \\ \Lambda_1^\top = -\Lambda_1, & \Lambda_2^\top = -\Lambda_2 \end{matrix} \right].$$

Proof. The tangent space of a point $Q \in \mathrm{O}(n)$ is given by

$$\mathbb{T}_Q \mathrm{O}(n) = \{Q\Lambda \in \mathbb{R}^{n\times n} : \Lambda^\top = -\Lambda\}.$$

A tangent vector $Q\Lambda \in \mathbb{T}_Q \operatorname{O}(n)$ is normal to $\operatorname{Gr}(k, n)$ at Q iff

$$0 = \langle X, Q\Lambda \rangle_Q = \operatorname{tr}(X^\top Q\Lambda),$$

for all $X \in \mathbb{T}_Q \operatorname{Gr}(k, n)$. By (3.12), $X = V \left[\begin{smallmatrix} 0 & B \\ B^\top & 0 \end{smallmatrix}\right] V^\top$ where $Q = V I_{k,n-k} V^\top$. Thus

$$\operatorname{tr}\left(V^\top \Lambda V \begin{bmatrix} 0 & -B \\ B^\top & 0 \end{bmatrix} \right) = 0 \tag{3.16}$$

for all $B \in \mathbb{R}^{k\times(n-k)}$. Since (3.16) must hold for all $B \in \mathbb{R}^{k\times(n-k)}$, we must have

$$\Lambda = V \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} V^\top, \tag{3.17}$$

for some skew-symmetric matrices $\Lambda_1 \in \mathbb{R}^{k\times k}$, $\Lambda_2 \in \mathbb{R}^{(n-k)\times(n-k)}$, and therefore,

$$Q\Lambda = V I_{k,n-k} V^\top \Lambda = V \begin{bmatrix} \Lambda_1 & 0 \\ 0 & -\Lambda_2 \end{bmatrix} V^\top.$$

Conversely, any Λ of the form in (3.17) must satisfy (3.16). □

Propositions 3.3.1 and 3.3.6 allow us to explicitly decompose the tangent space of $\operatorname{O}(n)$ at a point $Q \in \operatorname{Gr}(k, n)$ into

$$\begin{gathered} \mathbb{T}_Q \operatorname{O}(n) = \mathbb{T}_Q \operatorname{Gr}(k, n) \oplus \mathbb{N}_Q \operatorname{Gr}(k, n), \\ Q\Lambda = QV \begin{bmatrix} 0 & B \\ -B^\top & 0 \end{bmatrix} V^\top + V \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} V^\top. \end{gathered}$$

For later purposes, it will be useful to give explicit expressions for the two projection maps.

Corollary 3.3.7 (Projection maps)**.** *Let* $Q \in \operatorname{Gr}(k,n)$ *with* $Q = VI_{k,n-k}V^{\top}$ *and*

$$\operatorname{proj}_Q^{\mathbb{T}} : \mathbb{T}_Q \operatorname{O}(n) \to \mathbb{T}_Q \operatorname{Gr}(k,n), \qquad \operatorname{proj}_Q^{\mathbb{N}} : \mathbb{T}_Q \operatorname{O}(n) \to \mathbb{N}_Q \operatorname{Gr}(k,n)$$

be the projection maps onto the tangent and normal spaces of $\operatorname{Gr}(k,n)$ *respectively. Then*

$$\begin{aligned} \operatorname{proj}_Q^{\mathbb{T}}(Q\Lambda) &= \frac{1}{2}(Q\Lambda - \Lambda Q) = \frac{1}{2}V(S + S^{\top})V^{\top}, \\ \operatorname{proj}_Q^{\mathbb{N}}(Q\Lambda) &= \frac{1}{2}(Q\Lambda + \Lambda Q) = \frac{1}{2}V(S - S^{\top})V^{\top}, \end{aligned} \tag{3.18}$$

for any decomposition $Q\Lambda = VSV^{\top}$*, where* $S \in \mathbb{R}^{n \times n}$ *and* $I_{k,n-k}S$ *is skew-symmetric.*

Proof. We see from Propositions 3.3.1 and 3.3.6 that the maps are well defined, i.e., $\frac{1}{2}(Q\Lambda - \Lambda Q) \in \mathbb{T}_Q \operatorname{Gr}(k,n)$ and $\frac{1}{2}(Q\Lambda + \Lambda Q) \in \mathbb{N}_Q \operatorname{Gr}(k,n)$, and the images are orthogonal as

$$\langle Q\Lambda - \Lambda Q, Q\Lambda + \Lambda Q \rangle_Q = 0.$$

The alternative expressions follow from taking $S = I_{k,n-k}V^{\top}\Lambda V$. □

3.4 Exponential map, geodesic, and parallel transport

An explicit and easily computable formula for a geodesic curve is indispensable in most Riemannian optimization algorithms. By Lemma 3.1.1, any $Q \in \operatorname{Gr}(k,n)$ can be eigendecomposed as $VI_{k,n-k}V^{\top}$ for some $V \in \operatorname{O}(n)$. So a curve γ in $\operatorname{Gr}(k,n)$ takes the form

$$\gamma(t) = V(t)I_{k,n-k}V(t)^{\top}, \tag{3.19}$$

with $V(t)$ a curve in $\operatorname{O}(n)$ that can in turn be written as

$$V(t) = V\exp(\Lambda(t)), \tag{3.20}$$

where $\Lambda(t)$ is a curve in the space of $n \times n$ skew-symmetric matrices, $\Lambda(0) = 0$, and $V(0) = V$. We will show in Proposition 3.4.2 that in the involution model the curve $\Lambda(t)$ takes a particularly simple form. We first prove a useful lemma using the CS decomposition [82, 184].

Lemma 3.4.1. *Let $\Lambda \in \mathbb{R}^{n\times n}$ be skew-symmetric. Then there exist $B \in \mathbb{R}^{k\times(n-k)}$ and two skew-symmetric matrices $\Lambda_1 \in \mathbb{R}^{k\times k}$, $\Lambda_2 \in \mathbb{R}^{(n-k)\times(n-k)}$ such that*

$$\exp(\Lambda) = \exp\left(\begin{bmatrix} 0 & B \\ -B^\top & 0 \end{bmatrix}\right) \exp\left(\begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}\right). \tag{3.21}$$

Proof. By (3.8), we may assume $k \le n/2$. Let the CS decomposition of $Q \coloneqq \exp(\Lambda) \in \mathrm{O}(n)$ be

$$Q = \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix} \begin{bmatrix} \cos\Theta & \sin\Theta & 0 \\ -\sin\Theta & \cos\Theta & 0 \\ 0 & 0 & I_{n-2k} \end{bmatrix} \begin{bmatrix} W & 0 \\ 0 & Z \end{bmatrix}^\top, \tag{3.22}$$

where $U, W \in \mathrm{O}(k)$, $V, Z \in \mathrm{O}(n-k)$, $\Theta = \operatorname{diag}(\theta_1, \dots, \theta_k)$ with $\theta_i \in [0, 2\pi]$, $i = 1, \dots, k$. Next we will show that we may always choose U, W, V, Z to have determinant one, i.e., $U, W \in \mathrm{SO}(k)$, $V, Z \in \mathrm{SO}(n-k)$. To see this, note that

$$\det(Q) = \det(\exp(\Lambda)) = 1, \qquad \det \begin{bmatrix} \cos\Theta & \sin\Theta & 0 \\ -\sin\Theta & \cos\Theta & 0 \\ 0 & 0 & I_{n-2k} \end{bmatrix} = 1,$$

and so we must have $\det(U)\det(W)\det(V)\det(Z) = 1$. If $\det(U) = \det(W) = \det(V) = \det(Z) = 1$, then we are done. If either two of them, or all four of them have determinant -1, let $\hat{I}_j \coloneqq \operatorname{diag}(1, \dots, -1, \dots, 1)$, i.e., the identity matrix with its jth diagonal entry replaced by -1, then inserting $\hat{I}_1$ and $\hat{I}_{k+1}$ in (3.22) allows us to change the signs of the determinants at will. For example, if $\det(U) = \det(Z) = -1$, $\det(W) = \det(V) = 1$, then

$$\begin{aligned} Q &= \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix} \hat{I}_1 \hat{I}_1 \begin{bmatrix} \cos\Theta & \sin\Theta & 0 \\ -\sin\Theta & \cos\Theta & 0 \\ 0 & 0 & I_{n-2k} \end{bmatrix} \hat{I}_{k+1} \hat{I}_{k+1} \begin{bmatrix} W & 0 \\ 0 & Z \end{bmatrix}^\top \\ &= \begin{bmatrix} U' & 0 \\ 0 & V' \end{bmatrix} \begin{bmatrix} \cos\Theta' & \sin\Theta' & 0 \\ -\sin\Theta' & \cos\Theta' & 0 \\ 0 & 0 & I_{n-2k} \end{bmatrix} \begin{bmatrix} W' & 0 \\ 0 & Z' \end{bmatrix}^\top \end{aligned}$$

and the new matrices U', V', W', Z' now satisfy $\det(U') = \det(V') = \det(W') = \det(Z') = 1$. It is easy to check that in all cases, we may assume that $U, W \in \mathrm{SO}(k)$, $V, Z \in \mathrm{SO}(n-k)$ without loss of generality. Now

$$\begin{aligned} &\begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix} \begin{bmatrix} \cos\Theta & \sin\Theta & 0 \\ -\sin\Theta & \cos\Theta & 0 \\ 0 & 0 & I_{n-2k} \end{bmatrix} \\ &= \exp\left(\begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix} \begin{bmatrix} 0 & \Theta & 0 \\ -\Theta & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix}^\top \right) \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix} \\ &= \exp\left(\begin{bmatrix} 0 & B \\ -B^\top & 0 \end{bmatrix} \right) \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix}, \end{aligned}$$

where $B := U[\Theta, 0]V^\top \in \mathbb{R}^{k\times(n-k)}$ with $0 \in \mathbb{R}^{k\times(n-2k)}$. As $UW^\top \in \mathrm{SO}(k)$ and $VZ^\top \in \mathrm{SO}(n-k)$, we can find skew symmetric matrices Λ_1, Λ_2 with $\exp(\Lambda_1) = UW^\top$ and $\exp(\Lambda_2) = VZ^\top$ as required. $\square$

Proposition 3.4.2 (Curve)**.** *Let $Q \in \mathrm{Gr}(k, n)$ with eigendecomposition $Q = VI_{k,n-k}V^\top$.*

Then a curve $\gamma(t)$ in $\operatorname{Gr}(k,n)$ through Q may be expressed as

$$\gamma(t) = V \exp\left(\begin{bmatrix} 0 & B(t) \\ -B(t)^\top & 0 \end{bmatrix}\right) I_{k,n-k} \exp\left(\begin{bmatrix} 0 & -B(t) \\ B(t)^\top & 0 \end{bmatrix}\right) V^\top \tag{3.23}$$

for some curve $B(t)$ in $\mathbb{R}^{k\times(n-k)}$ through the zero matrix.

Proof. By (3.19) and (3.20), we have

$$\gamma(t) = V \exp\bigl(\Lambda(t)\bigr) I_{k,n-k} \exp\bigl(-\Lambda(t)\bigr) V^\top.$$

By Lemma 3.4.1, we may write

$$\exp\bigl(\Lambda(t)\bigr) = \exp\left(\begin{bmatrix} 0 & B(t) \\ -B(t)^\top & 0 \end{bmatrix}\right) \exp\left(\begin{bmatrix} \Lambda_1(t) & 0 \\ 0 & \Lambda_2(t) \end{bmatrix}\right),$$

which gives the desired parametrization in (3.23). □

Proposition 3.4.2 yields another way to obtain the expression for tangent vectors in (3.12). Differentiating the curve in (3.23) at $t = 0$, we get

$$\dot\gamma(0) = V \left(\begin{bmatrix} 0 & -2\dot B(0) \\ -2\dot B(0)^\top & 0 \end{bmatrix}\right) V^\top \in \mathbb{T}_Q \operatorname{Gr}(k,n).$$

Choosing $B(t)$ to be any curve in $\mathbb{R}^{k\times(n-k)}$ with $B(0) = 0$ and $\dot B(0) = -B/2$, we obtain (3.12).

The key ingredient in most manifold optimization algorithms is the geodesic at a point in a direction. In [65], the discussion regarding geodesics on the Grassmannian is brief: Essentially, it says that because a geodesic on the Stiefel manifold $\operatorname{V}(k,n)$ takes the form $Q\exp(t\Lambda)$, a geodesic on the Grassmannian $\operatorname{V}(k,n)/\operatorname{O}(k)$ takes the form $[Q\exp(t\Lambda)]$. It is hard to be more specific when one uses the Stiefel model. On the other hand, when we use the involution model, the expression (3.25) in the next theorem describes a geodesic precisely, and any point on γ can be evaluated with a single QR decomposition (to obtain V, see Section 3.7.1) and a single matrix exponentiation (the two exponents are transposes of each other).

Theorem 3.4.3 (Geodesics I). *Let $Q \in \operatorname{Gr}(k,n)$ and $X \in \mathbb{T}_Q \operatorname{Gr}(k,n)$ with*

$$Q = V I_{k,n-k} V^{\top}, \qquad X = V \begin{bmatrix} 0 & B \\ B^{\top} & 0 \end{bmatrix} V^{\top}. \tag{3.24}$$

The geodesic γ emanating from Q in the direction X is given by

$$\gamma(t) = V \exp\left(\frac{t}{2}\begin{bmatrix} 0 & -B \\ B^{\top} & 0 \end{bmatrix}\right) I_{k,n-k} \exp\left(\frac{t}{2}\begin{bmatrix} 0 & B \\ -B^{\top} & 0 \end{bmatrix}\right) V^{\top}. \tag{3.25}$$

The differential equation for γ is

$$\gamma(t)^{\top}\ddot{\gamma}(t) - \ddot{\gamma}(t)^{\top}\gamma(t) = 0, \qquad \gamma(0) = Q, \qquad \dot{\gamma}(0) = X. \tag{3.26}$$

Proof. By Proposition 3.4.2, any curve through Q must take the form

$$\gamma(t) = V \exp\left(\begin{bmatrix} 0 & B(t) \\ -B(t)^{\top} & 0 \end{bmatrix}\right) I_{k,n-k} \exp\left(\begin{bmatrix} 0 & -B(t) \\ B(t)^{\top} & 0 \end{bmatrix}\right) V^{\top},$$

where $B(0) = 0$. Since γ is in the direction X, we have that $\dot{\gamma}(0) = X$, and thus $\dot{B}(0) = -B/2$. It remains to employ the fact that as a geodesic, γ is a critical curve of the length functional

$$L(\gamma) := \int_0^1 \|\dot{\gamma}(t)\|_{\gamma(t)}\, dt$$

where the Riemannian norm is as in (3.15). Let $\varepsilon > 0$. Consider a variation of $\gamma(t)$ with respect to a C^1-curve $C(t)$ in $\mathbb{R}^{k\times(n-k)}$:

$$\gamma_\varepsilon(t) = V \exp\left(\begin{bmatrix} 0 & B(t)+\varepsilon C(t) \\ -B(t)^\top - \varepsilon C(t)^\top & 0 \end{bmatrix}\right) I_{k,n-k} \\ \exp\left(\begin{bmatrix} 0 & -B(t)-\varepsilon C(t) \\ B(t)^\top + \varepsilon C(t)^\top & 0 \end{bmatrix}\right) V^\top.$$

We require $C(0) = C(1) = 0$ so that γ_ε is a variation of γ with fixed end points. The tangent vector of γ_ε at time t is given by

$$V \exp\left(\begin{bmatrix} 0 & B(t)+\varepsilon C(t) \\ -B(t)^\top - \varepsilon C(t)^\top & 0 \end{bmatrix}\right) \left(-2\begin{bmatrix} 0 & \dot{B}(t)+\varepsilon \dot{C}(t) \\ \dot{B}(t) + \varepsilon \dot{C}(t)^\top & 0 \end{bmatrix}\right) \\ \exp\left(\begin{bmatrix} 0 & -B(t)-\varepsilon C(t) \\ B(t)^\top + \varepsilon C(t)^\top & 0 \end{bmatrix}\right) V^\top$$

and so $\|\dot{\gamma_\varepsilon}(t)\|_{\gamma(t)} = 2\sqrt{2}\|\dot{B}(t) + \varepsilon\dot{C}(t)\|_{\text{Fro}}$ where $\|\cdot\|_{\text{Fro}}$ denotes Frobenius norm. Hence,

$$0 = \frac{d}{d\varepsilon} L\bigl(\gamma_\varepsilon(t)\bigr)\Big|_{\varepsilon=0} = 2\sqrt{2}\int_0^1 \frac{\operatorname{tr}\bigl(\dot{B}(t)^\top \dot{C}(t)\bigr)}{\|\dot{B}(t)\|_{\text{Fro}}}\, dt.$$

As $\gamma(t)$ is a geodesic, $\|\dot{\gamma}(t)\|_{\gamma(t)}$ and thus $\|\dot{B}(t)\|_{\text{Fro}}$ must be a constant $K > 0$. Therefore, we have

$$0 = \frac{1}{K}\int_0^1 \operatorname{tr}\bigl(\dot{B}(t)^\top \dot{C}(t)\bigr)\, dt = -\frac{1}{K}\int_0^1 \operatorname{tr}\bigl(\ddot{B}(t)^\top C(t)\bigr)\, dt,$$

implying that $\ddot{B}(t) = 0$ and thus $B(t) = t\dot{B}(0) = -tB/2$. Lastly, since

$$\dot{\gamma}(t) = V \exp\left(\begin{bmatrix} 0 & B(t) \\ -B(t)^{\top} & 0 \end{bmatrix}\right)\left(-2\begin{bmatrix} 0 & \dot{B}(t) \\ \dot{B}(t)^{\top} & 0 \end{bmatrix}\right)\exp\left(\begin{bmatrix} 0 & -B(t) \\ B(t)^{\top} & 0 \end{bmatrix}\right)V^{\top},$$

$$\ddot{\gamma}(t) = V \exp\left(\begin{bmatrix} 0 & B(t) \\ -B(t)^{\top} & 0 \end{bmatrix}\right)\left(-4\begin{bmatrix} \dot{B}(t)\dot{B}(t)^{\top} & 0 \\ 0 & -\dot{B}(t)^{\top}\dot{B}(t) \end{bmatrix} - 2\begin{bmatrix} 0 & \ddot{B}(t) \\ \ddot{B}(t)^{\top} & 0 \end{bmatrix}\right) \tag{3.27}$$

$$\exp\left(\begin{bmatrix} 0 & -B(t) \\ B(t)^{\top} & 0 \end{bmatrix}\right)V^{\top}, \tag{3.28}$$

and the differential equation for a geodesic curve γ is

$$\operatorname{proj}^{\mathbb{T}}_{\gamma(t)}(\ddot{\gamma}) = 0, \qquad \gamma(0) = Q, \qquad \dot{\gamma}(0) = X,$$

we obtain (3.26) from the expression for tangent projection in (3.18). □

Theorem 3.4.3 also gives the *exponential map* of X.

Corollary 3.4.4 (Exponential map)**.** *Let $Q \in \operatorname{Gr}(k, n)$ and $X \in \mathbb{T}_Q \operatorname{Gr}(k, n)$ be as in* (3.24)*. Then*

$$\exp_Q(X) \coloneqq \gamma(1) = V \exp\left(\frac{1}{2}\begin{bmatrix} 0 & -B \\ B^{\top} & 0 \end{bmatrix}\right) I_{k,n-k} \exp\left(\frac{1}{2}\begin{bmatrix} 0 & B \\ -B^{\top} & 0 \end{bmatrix}\right)V^{\top}. \tag{3.29}$$

The length of the geodesic segment from $\gamma(0) = 0$ to $\gamma(1) = \exp_Q(X)$ is

$$L(\gamma) = \|X\|_{\mathrm{Fro}} = \sqrt{2}\|B\|_{\mathrm{Fro}}. \tag{3.30}$$

The Grassmannian is geodesically complete and so any two points can be joined by a length-minimizing geodesic. In the next proposition, we will derive an explicit expression for such a geodesic in the involution model. By (3.8), there will be no loss of generality in assuming that $k \le n/2$ in the following — if $k > n/2$, then we just replace k by $n - k$.

Proposition 3.4.5 (Geodesics II)**.** *Let $k \le n/2$. Let $Q_0, Q_1 \in \mathrm{Gr}(k, n)$ with eigendecompositions $Q_0 = V_0 I_{k,n-k} V_0^\top$ and $Q_1 = V_1 I_{k,n-k} V_1^\top$. Let the* CS *decomposition of $V_0^\top V_1 \in \mathrm{O}(n)$ be*

$$V_0^\top V_1 = \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix} \begin{bmatrix} \cos\Theta & \sin\Theta & 0 \\ -\sin\Theta & \cos\Theta & 0 \\ 0 & 0 & I_{n-2k} \end{bmatrix} \begin{bmatrix} W & 0 \\ 0 & Z \end{bmatrix}^\top \tag{3.31}$$

where $U, W \in \mathrm{O}(k)$, $V, Z \in \mathrm{O}(n-k)$, $\Theta = \operatorname{diag}(\theta_1, \dots, \theta_k) \in \mathbb{R}^{k\times k}$. Then the geodesic γ connecting Q_0 to Q_1 is

$$\gamma(t) = V_0 \exp\left(\frac{t}{2}\begin{bmatrix} 0 & -B \\ B^\top & 0 \end{bmatrix}\right) I_{k,n-k} \exp\left(\frac{t}{2}\begin{bmatrix} 0 & B \\ -B^\top & 0 \end{bmatrix}\right) V_0^\top,$$

where $B = -2U[\Theta, 0]V^\top \in \mathbb{R}^{k\times(n-k)}$ with $0 \in \mathbb{R}^{k\times(n-2k)}$.

Proof. By Theorem 3.4.3, γ is a geodesic curve emanating from $\gamma(0) = V_0 I_{k,n-k} V_0^\top = Q_0$. It remains to verify that

$$\gamma(1) = V_0 \exp\left(\frac{1}{2}\begin{bmatrix} 0 & -B \\ B^\top & 0 \end{bmatrix}\right) I_{k,n-k} \exp\left(\frac{1}{2}\begin{bmatrix} 0 & B \\ -B^\top & 0 \end{bmatrix}\right) V_0^\top = Q_1,$$

when $B = -2U[\Theta, 0]V^\top$. Substituting the expression for B,

$$\gamma(1) = V_0 \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix} \exp\left(\begin{bmatrix} 0 & \Theta & 0 \\ -\Theta & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right) I_{k,n-k} \exp\left(\begin{bmatrix} 0 & -\Theta & 0 \\ \Theta & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right) \begin{bmatrix} U^\top & 0 \\ 0 & V^\top \end{bmatrix} V_0^\top$$

$$= V_0 \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix} \begin{bmatrix} \cos\Theta & \sin\Theta & 0 \\ -\sin\Theta & \cos\Theta & 0 \\ 0 & 0 & I_{n-2k} \end{bmatrix} I_{k,n-k}$$

$$\begin{bmatrix} \cos\Theta & -\sin\Theta & 0 \\ \sin\Theta & \cos\Theta & 0 \\ 0 & 0 & I_{n-k} \end{bmatrix} \begin{bmatrix} U^\top & 0 \\ 0 & V^\top \end{bmatrix} V_0^\top$$

$$= V_0 \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix} \begin{bmatrix} \cos\Theta & \sin\Theta & 0 \\ -\sin\Theta & \cos\Theta & 0 \\ 0 & 0 & I_{n-2k} \end{bmatrix} \begin{bmatrix} W^\top & 0 \\ 0 & Z^\top \end{bmatrix} I_{k,n-k} \begin{bmatrix} W & 0 \\ 0 & Z \end{bmatrix}$$

$$\begin{bmatrix} \cos\Theta & -\sin\Theta & 0 \\ \sin\Theta & \cos\Theta & 0 \\ 0 & 0 & I_{n-k} \end{bmatrix} \begin{bmatrix} U^\top & 0 \\ 0 & V^\top \end{bmatrix} V_0^\top$$

where the last equality holds because we have

$$I_{k,n-k} = \begin{bmatrix} W^\top & 0 \\ 0 & Z^\top \end{bmatrix} I_{k,n-k} \begin{bmatrix} W & 0 \\ 0 & Z \end{bmatrix}$$

whenever $W \in \mathrm{O}(k)$ and $Z \in \mathrm{O}(n-k)$. By (3.31), the last expression of $\gamma(1)$ equals

$$V_0(V_0^\top V_1) I_{k,n-k} (V_0^\top V_1)^\top V_0^\top = V_1 I_{k,n-k} V_1^\top = Q_1. \qquad \square$$

The geodesic expression in Proposition 3.4.5 requires a CS decomposition [82, 184] and is more expensive to evaluate than the one in Theorem 3.4.3. Nevertheless, we do not need Proposition 3.4.5 for our optimization algorithms in Section 3.7, although its next corollary could be useful if one wants to design proximal gradient methods in the involution model.

Corollary 3.4.6 (Geodesic distance). *The geodesic distance between $Q_0, Q_1 \in \operatorname{Gr}(k,n)$ is given by*

$$d(Q_0, Q_1) = 2\sqrt{2}\Bigl(\sum\nolimits_{i=1}^{k} \sigma_i(B)^2\Bigr)^{1/2} = 2\sqrt{2}\Bigl(\sum\nolimits_{i=1}^{k} \theta_i\Bigr)^{1/2} \tag{3.32}$$

where $B \in \mathbb{R}^{k\times(n-k)}$ and $\Theta \in \mathbb{R}^{k\times k}$ are as in Proposition 3.4.5.

Proof. By (3.30), $L(\gamma) = \sqrt{2}\|B\|_{\mathrm{Fro}} = 2\sqrt{2}\|\Theta\|_{\mathrm{Fro}}$ with $B = -2U[\Theta, 0]V^{\top}$ as in Proposition 3.4.5. □

The last expression in (3.32) differs from the expression in [65, Section 4.3] by a factor of $2\sqrt{2}$, which is exactly what we expect since the metrics in the involution and orthogonal models differ by a factor of $(2\sqrt{2})^2 = 8$, as we saw in the proof of Proposition 3.3.5.

The notion of a logarithmic map is somewhat less standard and we remind readers of its definition. Given a Riemannian manifold M and a point $x \in M$, there exists some $r > 0$ such that the exponential map $\exp_x : B_r(0) \to M$ is a diffeomorphism on the ball $B_r(0) \subseteq \mathbb{T}_x M$ of radius r centered at the origin [58, Theorem 3.7]. The *logarithm map*, sometimes called the *inverse exponential map*, is then defined on the diffeomorphic image $\exp_x\bigl(B_r(0)\bigr) \subseteq M$ by

$$\log_x : \exp_x\bigl(B_r(0)\bigr) \to \mathbb{T}_x M, \quad \log_x(v) := \exp_x^{-1}(v)$$

for all $v \in \exp_x\bigl(B_r(0)\bigr)$. The largest r so that $\exp_x$ is a diffeomorphism on $B_r(0)$ is the *injectivity radius* at x and its infimum over all $x \in M$ is the injectivity radius of M.

Corollary 3.4.7 (Logarithmic map). *Let $Q_0, Q_1 \in \operatorname{Gr}(k,n)$ be such that $d(Q_0, Q_1) < \sqrt{2}\pi$. Let $V_0, V_1 \in \mathrm{O}(n)$, and $B \in \mathbb{R}^{k\times(n-k)}$ be as in Proposition 3.4.5. The logarithmic map at*

Q_0 *of* Q_1 *is*

$$\log_{Q_0}(Q_1) = V_0 \begin{bmatrix} 0 & -B \\ B^\top & 0 \end{bmatrix} V_0^\top.$$

Proof. The injectivity radius of $\operatorname{Gr}(k,n)$ is well known to be $\pi/2$ [200]. Write $B_r(0) = \{X \in \mathbb{T}_{Q_0}\operatorname{Gr}(k,n) : \|X\|_Q < r\}$ and $B_r^d(Q_0) = \{Q \in \operatorname{Gr}(k,n) : d(Q_0,Q) < r\}$. By Corollaries 3.4.4 and 3.4.6,

$$\exp_{Q_0}\bigl(B_{\pi/2}(0)\bigr) = B^d_{\sqrt{2}\pi}(Q_0).$$

By Corollary 3.4.4 and Proposition 3.4.5, $\log_{Q_0} : B_{\sqrt{2}\pi}(Q_0) \to \operatorname{Gr}(k,n)$ has the required expression. □

We end this section with the expression for the parallel transport of a vector Y along a geodesic γ at a point Q in the direction X. This will be an essential ingredient for conjugate gradient and Newton methods in the involution model (see Algorithms 3 and 4).

Proposition 3.4.8 (Parallel transport). *Let* $Q \in \operatorname{Gr}(k,n)$ *and* $X, Y \in \mathbb{T}_Q\operatorname{Gr}(k,n)$ *with*

$$Q = VI_{k,n-k}V^\top, \qquad X = V\begin{bmatrix} 0 & B \\ B^\top & 0 \end{bmatrix}V^\top, \qquad Y = V\begin{bmatrix} 0 & C \\ C^\top & 0 \end{bmatrix}V^\top,$$

where $V \in \mathrm{O}(n)$ *and* $B, C \in \mathbb{R}^{k\times(n-k)}$*. Let* γ *be a geodesic curve emanating from* Q *in the direction* X*. Then the parallel transport of* Y *along* γ *is*

$$Y(t) = V\exp\left(\frac{t}{2}\begin{bmatrix} 0 & -B \\ B^\top & 0 \end{bmatrix}\right)\begin{bmatrix} 0 & C \\ C^\top & 0 \end{bmatrix}\exp\left(\frac{t}{2}\begin{bmatrix} 0 & B \\ -B^\top & 0 \end{bmatrix}\right)V^\top. \tag{3.33}$$

Proof. Let γ be parametrized as in (3.25). A vector field $Y(t)$ that is parallel along $\gamma(t)$

may, by (3.12), be written in the form

$$Y(t) = V \exp\left(\frac{t}{2}\begin{bmatrix} 0 & -B \\ B^{\top} & 0 \end{bmatrix}\right)\begin{bmatrix} 0 & C(t) \\ C(t)^{\top} & 0 \end{bmatrix} \exp\left(\frac{t}{2}\begin{bmatrix} 0 & B \\ -B^{\top} & 0 \end{bmatrix}\right) V^{\top}$$

for some curve $C(t)$ in $\mathbb{R}^{k\times(n-k)}$ with $C(0) = C$. Differentiating $Y(t)$ gives

$$\dot{Y}(t) = V \exp\biggl(\frac{t}{2}\begin{bmatrix} 0 & -B \\ B^{\top} & 0 \end{bmatrix}\biggr)\begin{bmatrix} -\frac{1}{2}\bigl(BC(t)^{\top} + C(t)B^{\top}\bigr) & \dot{C}(t) \\ \dot{C}(t)^{\top} & \frac{1}{2}\bigl(B^{\top}C(t) + C(t)^{\top}B\bigr) \end{bmatrix}$$
$$\exp\biggl(\frac{t}{2}\begin{bmatrix} 0 & B \\ -B^{\top} & 0 \end{bmatrix}\biggr)V^{\top}.$$

Since $Y(t)$ is parallel along $\gamma(t)$, we must have

$$\operatorname{proj}^{\mathbb{T}}_{\gamma(t)}\bigl(\dot{Y}(t)\bigr) = 0,$$

which implies that $\dot{C}(t) = 0$ and thus $C(t) = C(0) = C$, giving us (3.33). □

A word about our notation for parallel transport, or rather, the lack of one. Note that $Y(t)$ depends on γ and to indicate this dependence, we may write $Y_{\gamma}(t)$. Other common notations include $\tau_t Y$ [99], $P^{\gamma}_{t} Y$ [115], $\gamma^{t}_{s}(Y)$ [126] ($s = 0$ for us) but there is no single standard notation.

3.5 Gradient and Hessian

We now derive expressions for the Riemannian gradient and Hessian of a C^2 function $f : \operatorname{Gr}(k, n) \to \mathbb{R}$ in the involution model with (3.10) for tangent vectors. As a reminder, this means:

$$
\begin{aligned}
&\operatorname{Gr}(k,n) = \{Q \in \mathbb{R}^{n\times n} : Q^\top Q = I,\ Q^\top = Q,\ \operatorname{tr}(Q) = 2k-n\},\\
&\mathbb{T}_Q \operatorname{Gr}(k,n) = \Bigl[V \begin{bmatrix} 0 & B \\ B^\top & 0 \end{bmatrix} V^\top \in \mathbb{R}^{n\times n} : B \in \mathbb{R}^{k\times(n-k)}\Bigr],
\end{aligned}
\tag{3.34}
$$

where $Q = VI_{k,n-k}V^\top$.

The *Riemannian gradient* ∇f at Q is a tangent vector $\nabla f(Q) \in \mathbb{T}_Q \operatorname{Gr}(k,n)$ and, depending on context, the *Riemannian Hessian* at Q is a bilinear map:

$$
\nabla^2 f(Q) : \mathbb{T}_Q \operatorname{Gr}(k,n) \times \mathbb{T}_Q \operatorname{Gr}(k,n) \to \mathbb{R}.
$$

Proposition 3.5.1 (Riemannian gradient I). *Let $f : \operatorname{Gr}(k,n) \to \mathbb{R}$ be a C^1 function. For any $Q \in \operatorname{Gr}(k,n)$, write*

$$
f_Q \coloneqq \left[\frac{\partial f}{\partial q_{ij}}(Q)\right]_{i,j=1}^n \in \mathbb{R}^{n\times n}. \tag{3.35}
$$

Then

$$
\nabla f(Q) = \frac{1}{4}\bigl[f_Q + f_Q^\top - Q(f_Q + f_Q^\top)Q\bigr]. \tag{3.36}
$$

Proof. The projection of $QX \in \mathbb{T}_Q\mathbb{R}^{n\times n}$ to $\mathbb{T}_Q \operatorname{O}(n)$ is $Q(X - X^\top)/2$. Therefore the projection of $f_Q \in \mathbb{T}_Q\mathbb{R}^{n\times n}$ to $\mathbb{T}_Q \operatorname{O}(n)$ is $(f_Q - Qf_Q^\top Q)/2$. Composing this with the projection of $\mathbb{T}_Q \operatorname{O}(n)$ to $\mathbb{T}_Q \operatorname{Gr}(k,n)$ given in (3.18), we get

$$
\nabla f(Q) = \operatorname{proj}_Q^{\mathbb{T}}\Bigl(\frac{f_Q - Qf_Q^\top Q}{2}\Bigr) = \frac{1}{4}\bigl(f_Q + f_Q^\top - Qf_QQ - Qf_Q^\top Q\bigr)
$$

as required. □

Proposition 3.5.2 (Riemannian Hessian I). *Let $f : \operatorname{Gr}(k,n) \to \mathbb{R}$ be C^2. For any $Q =$*

$VI_{k,n-k}V^\top \in \operatorname{Gr}(k,n)$, let f_Q be as in (3.35) and

$$f_{QQ}(X) \coloneqq \biggl[\sum_{i,j=1}^{n}\Bigl(\frac{\partial^2 f}{\partial q_{ij}\partial q_{kl}}(Q)\Bigr)x_{ij}\biggr]_{k,l=1}^{n}, \quad f_{QQ}(X,Y) \coloneqq \sum_{i,j,k,l=1}^{n}\Bigl(\frac{\partial^2 f}{\partial q_{ij}\partial q_{kl}}(Q)\Bigr)x_{ij}y_{kl}.$$

As a bilinear map, the Hessian of f at Q is given by

$$\nabla^2 f(Q)(X,Y) = f_{QQ}(X,Y) - \frac{1}{2}\operatorname{tr}\bigl(f_Q^\top Q(XY+YX)\bigr) \tag{3.37}$$

for any $X, Y \in \mathbb{T}_Q \operatorname{Gr}(k,n)$.

Proof. Let γ be a geodesic curve emanating from Q in the direction $X \in \mathbb{T}_Q \operatorname{Gr}(k,n)$. Then the Hessian can be computed as

$$\nabla^2 f(Q)(X,X) = \frac{d^2}{dt^2} f\bigl(\gamma(t)\bigr)\Big|_{t=0} = \frac{d}{dt}\operatorname{tr}\bigl(f_{\gamma(t)}^\top \dot\gamma(t)\bigr)\Big|_{t=0} = f_{QQ}(X) + \operatorname{tr}\bigl(f_Q^\top \ddot\gamma(0)\bigr).$$

Since $\gamma(t)$ is given by (3.25),

$$\ddot\gamma(0) = V\begin{bmatrix} -BB^\top & 0 \\ 0 & B^\top B\end{bmatrix}V^\top = -Q\dot\gamma(0)^2$$

and so

$$\nabla^2 f(Q)(X,X) = f_{QQ}(X) - \operatorname{tr}(f_Q^\top Q X^2).$$

To obtain $\nabla^2 f(Q)$ as a bilinear map, we simply polarize the quadratic form above:

$$\begin{aligned}\nabla^2 f(Q)(X,Y) &= \frac{1}{2}\bigl[\nabla^2 f(Q)(X+Y,X+Y) - \nabla^2 f(Q)(X,X) - \nabla^2 f(Q)(Y,Y)\bigr]\\ &= \frac{1}{2}\Bigl[f_{QQ}(X+Y) - f_{QQ}(X) - f_{QQ}(Y) - \operatorname{tr}\bigl(f_Q^\top Q(XY+YX)\bigr)\Bigr]\\ &= f_{QQ}(X,Y) - \frac{1}{2}\operatorname{tr}\bigl(f_Q^\top Q(XY+YX)\bigr).\end{aligned}$$

□

Our optimization algorithms require that we parameterize our tangent space as in (3.34) and we need to express $\nabla f(Q)$ in such a form. This can be easily accomplished. Let $E_{ij} \in \mathbb{R}^{k\times(n-k)}$ be the matrix whose (i, j) entry is zero and other entries are one. Let

$$X_{ij} \coloneqq V \begin{bmatrix} 0 & E_{ij} \\ E_{ij}^{\top} & 0 \end{bmatrix} V^{\top} \in \mathbb{T}_Q \operatorname{Gr}(k, n). \tag{3.38}$$

Then $\mathcal{B}_Q \coloneqq \{X_{ij} : i = 1, \dots, k,\ j = 1, \dots, n-k\}$ is an orthogonal (but not orthonormal since Riemannian norm $\|X_{ij}\|_Q = 1/\sqrt{2}$) basis of $\mathbb{T}_Q \operatorname{Gr}(k, n)$.

Corollary 3.5.3 (Riemannian gradient II). *Let f, Q, f_Q be as in Propositions 3.5.1. If we partition*

$$V^{\top}(f_Q + f_Q^{\top})V = \begin{bmatrix} A & B \\ B^{\top} & C \end{bmatrix}, \tag{3.39}$$

where $A \in \mathbb{R}^{k\times k}$, $B \in \mathbb{R}^{k\times(n-k)}$, $C \in \mathbb{R}^{(n-k)\times(n-k)}$, then

$$\nabla f(Q) = \frac{1}{2} V \begin{bmatrix} 0 & B \\ B^{\top} & 0 \end{bmatrix} V^{\top}. \tag{3.40}$$

Proof. By (3.39), we may rewrite (3.36) as

$$\nabla f(Q) = \frac{1}{4}\left(V \begin{bmatrix} A & B \\ B^{\top} & C \end{bmatrix} V^{\top} - V \begin{bmatrix} A & -B \\ -B^{\top} & C \end{bmatrix} V^{\top} \right) = \frac{1}{2} V \begin{bmatrix} 0 & B \\ B^{\top} & 0 \end{bmatrix} V^{\top}. \qquad \square$$

In our optimization algorithms, (3.39) is how we actually compute Riemannian gradients. Note that in the basis $\mathcal{B}_Q$, the gradient of f is essentially given by the matrix $B/2 \in \mathbb{R}^{k\times(n-k)}$. So in algorithms that rely only on Riemannian gradients, we just need the top right block B, but the other blocks A and C would appear implicitly in the Riemannian Hessians.

We may order the basis $\mathcal{B}_Q$ lexicographically (note that X_{ij}'s are indexed by two indices), then the bilinear form $\nabla^2 f(Q)$ has the matrix representation

$$H_Q := \begin{bmatrix} \nabla^2 f(Q)(X_{11}, X_{11}) & \nabla^2 f(Q)(X_{11}, X_{12}) & \dots & \nabla^2 f(Q)(X_{11}, X_{k,n-k}) \\ \nabla^2 f(Q)(X_{12}, X_{11}) & \nabla^2 f(Q)(X_{12}, X_{12}) & \dots & \nabla^2 f(Q)(X_{12}, X_{k,n-k}) \\ \vdots & \vdots & \ddots & \vdots \\ \nabla^2 f(Q)(X_{k,n-k}, X_{11}) & \nabla^2 f(Q)(X_{k,n-k}, X_{12}) & \dots & \nabla^2 f(Q)(X_{k,n-k}, X_{k,n-k}) \end{bmatrix}. \tag{3.41}$$

In practice, the evaluation of H_Q may be simplified; we will discuss this in Section 3.7.3. To summarize, in the lexicographically ordered basis $\mathcal{B}_Q$,

$$\big[\nabla f(Q)\big]_{\mathcal{B}_Q} = \frac{1}{2}\operatorname{vec}(B) \in \mathbb{R}^{k(n-k)}, \qquad \big[\nabla^2 f(Q)\big]_{\mathcal{B}_Q} = H_Q \in \mathbb{R}^{k(n-k)\times k(n-k)},$$

and the Newton step $S \in \mathbb{R}^{k\times(n-k)}$ is given by the linear system

$$H_Q \operatorname{vec}(S) = -\frac{1}{2}\operatorname{vec}(B). \tag{3.42}$$

3.6 Retraction map and vector transport

Up till this point, everything that we have discussed is authentic Riemannian geometry, even though we have used extrinsic coordinates to obtain expressions in terms of matrices and matrix operations. This section is a departure, we will discuss two notions created for sole use in manifold optimization: *retraction maps* [6, 180] and *vector transports* [4]. They are relaxations of exponential maps and parallel transports respectively and are intended to be pragmatic substitutes in situations where these Riemannian operations are either too difficult to compute (e.g., requiring the exponential of a nonnormal matrix) or unavailable in closed form (e.g., parallel transport on a Stiefel manifold). While the involution model does not suffer from either of these problems, retraction algorithms could still serve as a good option for initializing Riemannian optimization algorithms.

As these definitions are not found in the Riemannian geometry literature, we state a version of [4, Definitions 4.1.1 and 8.1.1] below for easy reference.

Definition 3.6.1. *A map $R : \mathbb{T}M \to M$, $(x, v) \mapsto R_x(v)$ is a* retraction *map if it satisfies the following two conditions:*

- *$R_x(0) = x$ for all $x \in M$;*
- *$dR_x(0) : \mathbb{T}_x M \to \mathbb{T}_x M$ is the identity map for all $x \in M$.*

A map $T : \mathbb{T}M \oplus \mathbb{T}M \to \mathbb{T}M$ associated to a retraction map R is a vector transport *if it satisfies the following three conditions:*

- *$T(x, v, w) = \bigl(R_x(v), T_{x,v}(w)\bigr)$ for all $x \in M$ and $v, w \in \mathbb{T}_x M$;*
- *$T_{x,0}(w) = w$ for all $x \in M$ and $w \in \mathbb{T}_x M$;*
- *$T_{x,v}(a_1 w_1 + a_2 w_2) = a_1 T_{x,v}(w_1) + a_2 T_{x,v}(w_2)$ for all $a_1, a_2 \in \mathbb{R}$, $x \in M$, and $v, w_1, w_2 \in \mathbb{T}_x M$.*

Here $\mathbb{T}M \oplus \mathbb{T}M$ is a direct sum of vector bundles and each element is parametrized by a point $x \in M$ and two tangent vectors $v, w \in \mathbb{T}_x M$. The condition (3.6.1) says that the vector transport T is compatible with its retraction map R, and also defines the map $T_{x,v} : \mathbb{T}_x M \to \mathbb{T}_x M$. Note that v is the direction to move in while w is the vector to be transported.

For the purpose of optimization, we just need R and T to be well-defined on a neighbourhood of $M \cong \{(x, 0) \in \mathbb{T}M\} \subseteq \mathbb{T}M$ and $M \cong \{(x, 0, 0) \in \mathbb{T}M \oplus \mathbb{T}M\} \subseteq \mathbb{T}M \oplus \mathbb{T}M$ respectively. If R and T are C^1 maps, then various optimization algorithms relying on R and T can be shown to converge [4], possibly under the additional assumption that M has nonnegative [50] or bounded sectional curvature [190]. In particular, these results apply in our case since being a compact symmetric space, $\mathrm{Gr}(k,n)$ has both nonnegative and bounded sectional curvature [36, 212].

Example 3.6.2 (Projection as retraction)**.** *For a manifold M embedded in Euclidean space $\mathbb{R}^n$ or $\mathbb{R}^{m\times n}$, we may regard tangent vectors in $\mathbb{T}_x M$ to be of the form $x+v$. In this case an example of a retraction map is given by the projection of tangent vectors onto M,*

$$R_x(v) = \operatorname{argmin}_{y\in M} \|x+v-y\|,$$

where $\|\cdot\|$ is either the 2- or Frobenius norm. By [5, Lemma 3.1], the map R_x is well-defined for small v and is a retraction.

We will give three retraction maps for $\mathrm{Gr}(k,n)$ that are readily computable in the involution model with EVD, block QR, and Cayley transform respectively. The latter two are inspired by similar maps defined for the projection model in [100] although our motivations are somewhat different.

We begin by showing how one may compute the projection $\operatorname{argmin}\bigl\{\|A-Q\|_{\mathrm{Fro}} : Q \in \mathrm{Gr}(k,n)\bigr\}$ for an arbitrary matrix $A\in\mathbb{R}^{n\times n}$ in the involution model, a result that may be of independent interest.

Lemma 3.6.3. *Let $A\in\mathbb{R}^{n\times n}$ and*

$$\frac{A+A^\top}{2} = VDV^\top \tag{3.43}$$

be an eigendecomposition with $V\in\mathrm{O}(n)$ and $D=\operatorname{diag}(\lambda_1,\dots,\lambda_n)$, $\lambda_1\ge\cdots\ge\lambda_n$. Then $Q=VI_{k,n-k}V^\top$ is a minimizer of

$$\min\bigl\{\|A-Q\|_{\mathrm{Fro}} : Q^\top Q = I,\ Q^\top = Q,\ \operatorname{tr}(Q)=2k-n\bigr\}.$$

Proof. Since Q is symmetric, $\|A-Q\|^2_{\mathrm{Fro}} = \|(A+A^\top)/2-Q\|^2_{\mathrm{Fro}} + \|(A-A^\top)/2\|^2_{\mathrm{Fro}}$, a best approximation to A is also a best approximation to $(A + A^\top)/2$. By (3.43), $\|(A+A^\top)/2 - Q\|_{\mathrm{Fro}} = \|D - V^\top QV\|_{\mathrm{Fro}}$ and so for a best approximation $V^\top QV$ must be a diagonal matrix. Since the eigenvalues $\delta_1, \dots, \delta_n$ of a symmetric orthogonal Q must be ± 1 and $\operatorname{tr}(Q) = 2k - n$, the multiplicities of $+1$ and -1 are k and $n - k$ respectively. By assumption, $\lambda_1 \ge \cdots \ge \lambda_n$, so

$$\min_{\delta_1 + \cdots + \delta_n = 2k-n} (\lambda_1 - \delta_1)^2 + \cdots + (\lambda_n - \delta_n)^2$$

is attained when $\delta_1 = \cdots = \delta_k = +1$ and $\delta_{k+1} = \cdots = \delta_n = -1$. Hence $V^\top QV = \operatorname{diag}(\delta_1, \dots, \delta_n) = I_{k,n-k}$ as required. □

It is clear from the proof, which is a variation of standard arguments [106, Section 8.1], that a minimizer is not unique if and only if $\lambda_k = \lambda_{k+1}$, i.e., the kth and $(k+1)$th eigenvalues of $(A + A^\top)/2$ coincide. Since any $Q \in \operatorname{Gr}(k, n)$ by definition has $\lambda_k = +1 \neq -1 = \lambda_{k+1}$, the projection is always unique in a small enough neighborhood of Q in $\mathbb{R}^{n\times n}$.

In the following, let $\mathcal{E} : \mathbb{R}^{n\times n} \to \mathrm{O}(n)$ be the map that takes any $A \in \mathbb{R}^{n\times n}$ to an orthogonal matrix of eigenvectors of $(A + A^\top)/2$.

Proposition 3.6.4 (Retraction I)**.** *Let $Q \in \operatorname{Gr}(k, n)$ and $X, Y \in \mathbb{T}_Q \operatorname{Gr}(k, n)$ with*

$$Q = VI_{k,n-k}V^\top, \qquad X = V \begin{bmatrix} 0 & B \\ B^\top & 0 \end{bmatrix} V^\top, \qquad Y = V \begin{bmatrix} 0 & C \\ C^\top & 0 \end{bmatrix} V^\top, \tag{3.44}$$

where $V \in \mathrm{O}(n)$ and $B, C \in \mathbb{R}^{k\times(n-k)}$. Then

$$R_Q^{\mathcal{E}}(X) = V\mathcal{E}\Bigl(\begin{bmatrix} I & B \\ B^\top & -I \end{bmatrix}\Bigr) I_{k,n-k} \mathcal{E}\Bigl(\begin{bmatrix} I & B \\ B^\top & -I \end{bmatrix}\Bigr)^\top V^\top$$

defines a retraction and

$$T_Q^{\mathcal{E}}(X,Y) = V\mathcal{E}\Bigl(\begin{bmatrix} I & B \\ B^{\top} & -I \end{bmatrix}\Bigr)\begin{bmatrix} 0 & C \\ C^{\top} & 0 \end{bmatrix}\mathcal{E}\Bigl(\begin{bmatrix} I & B \\ B^{\top} & -I \end{bmatrix}\Bigr)^{\top} V^{\top}$$

defines a vector transport.

Proof. It follows from Lemma 3.6.3 that $R_Q^{\mathcal{E}}$ defines a projection. The properties in Definition 3.6.1 are routine to verify. □

The retraction map above requires an EVD, which is relatively expensive. Furthermore, the eigenvector map $\mathcal{E}$ is generally discontinuous [121], which can present a problem. One alternative would be to approximate the map $\mathcal{E}$ with a QR decomposition — one should think of this as the first step of Francis's QR algorithm for EVD. In fact, we will not even require a full QR decomposition, a 2×2 block QR decomposition suffices. Let $\mathcal{Q} : \mathbb{R}^{n\times n} \to \mathrm{O}(n)$ be a map that takes a matrix A to its orthogonal factor in a 2×2 block QR decomposition, i.e.,

$$A = \mathcal{Q}(A)\begin{bmatrix} R_1 & R_2 \\ 0 & R_3 \end{bmatrix}, \quad R_1 \in \mathbb{R}^{k\times k},\ R_2 \in \mathbb{R}^{k\times(n-k)},\ R_3 \in \mathbb{R}^{(n-k)\times(n-k)}.$$

Note that $\mathcal{Q}(A)$ is an orthogonal matrix but the second factor just needs to be block upper triangular, i.e., R_1 and R_3 are not required to be upper triangular matrices. We could compute $\mathcal{Q}(A)$ with, say, the first k steps of Householder QR applied to A.

Proposition 3.6.5 (Retraction II). *Let $Q \in \mathrm{Gr}(k,n)$ and $X, Y \in \mathbb{T}_Q \mathrm{Gr}(k,n)$ be as in* (3.44)*. If $\mathcal{Q}$ is well-defined and differentiable near $I_{k,n-k}$ and $\mathcal{Q}(I_{k,n-k}) = I$, then*

$$R_Q^{\mathcal{Q}}(X) = V\mathcal{Q}\Bigl(\frac{1}{2}\begin{bmatrix} I & B \\ B^{\top} & -I \end{bmatrix}\Bigr) I_{k,n-k}\mathcal{Q}\Bigl(\frac{1}{2}\begin{bmatrix} I & B \\ B^{\top} & -I \end{bmatrix}\Bigr)^{\top} V^{\top}$$

defines a retraction and

$$T_Q^{\mathcal{Q}}(X,Y) = V\mathcal{Q}\Bigl(\frac{1}{2}\begin{bmatrix} I & B \\ B^\top & -I \end{bmatrix}\Bigr)\begin{bmatrix} 0 & C \\ C^\top & 0 \end{bmatrix}\mathcal{Q}\Bigl(\frac{1}{2}\begin{bmatrix} I & B \\ B^\top & -I \end{bmatrix}\Bigr)^{\top} V^\top$$

defines a vector transport.

Proof. Only property (3.6.1) in Definition 3.6.1 is not immediate and requires checking. Let the following be a block QR decomposition:

$$\frac{1}{2}\begin{bmatrix} I & tB \\ tB^\top & -I \end{bmatrix} = \begin{bmatrix} Q_1(t) & Q_2(t) \\ Q_3(t) & Q_4(t) \end{bmatrix}\begin{bmatrix} R_1(t) & R_2(t) \\ 0 & R_3(t) \end{bmatrix} = Q(t)R(t), \tag{3.45}$$

with $Q(t) \in \mathrm{O}(n)$. Since $Q(t)Q(t)^\top = 1$ and $Q(0) = I$, $Q'(0)$ is skew-symmetric and

$$\frac{d}{dt}Q(t)I_{k,n-k}Q(t)^\top\Big|_{t=0} = \begin{bmatrix} Q_1'(0) + Q_1'(0)^\top & -Q_2'(0) + Q_3'(0)^\top \\ Q_3'(0) - Q_2'(0)^\top & -Q_4'(0) - Q_4'(0)^\top \end{bmatrix} = \begin{bmatrix} 0 & 2Q_3'(0)^\top \\ 2Q_3'(0) & 0 \end{bmatrix}.$$

Comparing the $(1,1)$ and $(2,1)$ entries in (3.45), we get

$$Q_1(t)R_1(t) = I, \qquad Q_3(t)R_1(t) = tB^\top/2.$$

Hence $Q_3(t) = tB^\top Q_1(t)/2$, $Q_3'(0) = B^\top Q_1(0)/2 = B^\top/2$, and we get

$$\frac{d}{dt}Q(t)I_{k,n-k}Q(t)^\top\Big|_{t=0} = \begin{bmatrix} 0 & B \\ B^\top & 0 \end{bmatrix},$$

as required. □

If we use a first-order Padé approximation $\exp(X) \approx (I + X)(I - X)^{-1}$ for the matrix exponential terms in the exponential map (3.29) and parallel transport (3.33), we obtain another retraction map and vector transport. This Padé approximation is the well-known *Cayley transform* $\mathcal{C}$, which takes a skew-symmetric matrix to an orthogonal matrix and vice versa:

$$\mathcal{C} : \mathsf{\Lambda}^2(\mathbb{R}^n) \to \mathrm{O}(n), \quad \Lambda \to (I + \Lambda)(I - \Lambda)^{-1}.$$

Proposition 3.6.6 (Retraction III). *Let $Q \in \mathrm{Gr}(k, n)$ and $X, Y \in \mathbb{T}_Q \mathrm{Gr}(k, n)$ be as in* (3.44)*. Then*

$$R_Q^{\mathcal{C}}(X) = V\mathcal{C}\Big(\frac{1}{4}\begin{bmatrix} 0 & -B \\ B^\top & 0 \end{bmatrix}\Big) I_{k,n-k} \mathcal{C}\Big(\frac{1}{4}\begin{bmatrix} 0 & -B \\ B^\top & 0 \end{bmatrix}\Big)^\top V^\top$$

defines a retraction and

$$T_Q^{\mathcal{C}}(X, Y) = V\mathcal{C}\Big(\frac{1}{4}\begin{bmatrix} 0 & -B \\ B^\top & 0 \end{bmatrix}\Big) \begin{bmatrix} 0 & C \\ C^\top & 0 \end{bmatrix} \mathcal{C}\Big(\frac{1}{4}\begin{bmatrix} 0 & -B \\ B^\top & 0 \end{bmatrix}\Big)^\top V^\top$$

defines a vector transport.

Proof. Again, only property (3.6.1) in Definition 3.6.1 is not immediate and requires checking. But this is routine we omit the details. □

Another alternative that avoids computing an EVD and takes advantage of the skew-symmetry involves the so-called *Strang splitting [186]. Observe that a matrix in the exponent of* (3.49) *may be written as a unique linear combination*

$$\begin{bmatrix} 0 & B \\ -B^\top & 0 \end{bmatrix} = \sum_{i=1}^{k} \sum_{j=1}^{n-k} \alpha_{ij} \begin{bmatrix} 0 & E_{ij} \\ -E_{ij}^\top & 0 \end{bmatrix}$$

where $\alpha_{ij} \in \mathbb{R}$ and E_{ij} is the matrix whose (i, j) entry is one and other entries are zero.

Since

$$\exp\left(\theta\begin{bmatrix}0 & E_{ij}\\ -E_{ij}^{\top} & 0\end{bmatrix}\right)=\begin{bmatrix}I+(\cos\theta-1)E_{ii} & (\sin\theta)E_{ij}\\ -(\sin\theta)E_{ji} & I+(\cos\theta-1)E_{jj}\end{bmatrix}=:G_{i,j+k}(\theta)$$

is a Givens rotation in the ith and $(j+k)$th plane of θ radians [86, p. 240], we may approximate the required matrix exponential as a sequence of Givens rotation

$$\mathcal{S}\left(\begin{bmatrix}0 & B\\ -B^{\top} & 0\end{bmatrix}\right):=G_{1,1+k}\big(\tfrac{1}{2}\alpha_{11}\big)G_{1,2+k}\big(\tfrac{1}{2}\alpha_{12}\big)\cdots G_{k,n-1}\big(\tfrac{1}{2}\alpha_{k,n-k-1}\big)G_{k,n}(\alpha_{k,n-k})$$

$$G_{k,n-1}\big(\tfrac{1}{2}\alpha_{k,n-k-1}\big)\cdots G_{1,2+k}\big(\tfrac{1}{2}\alpha_{12}\big)G_{1,1+k}\big(\tfrac{1}{2}\alpha_{11}\big)\approx\exp\left(\begin{bmatrix}0 & B\\ -B^{\top} & 0\end{bmatrix}\right). \tag{3.46}$$

This is called the Strang splitting [186], which approximates the matrix exponential up to first order, thereby giving us a retraction and a vector transport:

Proposition 3.6.7 (Retraction IV). *Let $Q\in\operatorname{Gr}(k,n)$ and $X,Y\in\mathbb{T}_Q\operatorname{Gr}(k,n)$ be as in* (3.44). *Then*

$$R_Q^{\mathcal{S}}(X)=V\mathcal{S}\left(\frac{1}{2}\begin{bmatrix}0 & -B\\ B^{\top} & 0\end{bmatrix}\right)I_{k,n-k}\mathcal{S}\left(\frac{1}{2}\begin{bmatrix}0 & -B\\ B^{\top} & 0\end{bmatrix}\right)^{\top}V^{\top}$$

defines a retraction and

$$T_Q^{\mathcal{S}}(X,Y)=V\mathcal{S}\left(\frac{1}{2}\begin{bmatrix}0 & -B\\ B^{\top} & 0\end{bmatrix}\right)\begin{bmatrix}0 & C\\ C^{\top} & 0\end{bmatrix}\mathcal{S}\left(\frac{1}{2}\begin{bmatrix}0 & -B\\ B^{\top} & 0\end{bmatrix}\right)^{\top}V^{\top}$$

defines a vector transport.

An efficient way to compute the exponential map is the same as an efficient retraction map. In fact, numerical algorithms for matrix exponential is often contructed by using a local retraction. If there is a retraction map R_I which approximates the matrix exponential locally, then we can compute the matrix exponential up to arbitrary precision by the scaling and squaring method [105]:

$$e^A = \lim_{n\to\infty} \bigl(R_I(A/n)\bigr)^n.$$

The retraction based on Cayley transform in Proposition 3.6.6 is a special case of the Padé approximation method in [32]. Conversely, the Strang splitting in Proposition 3.6.7 can be used to compute the matrix exponential and take time at most $12nk(n-k)$. In fact, computing the product in (3.46) is equivalent to computing a sequence of $2k(n-k)-1$ Givens rotations, which takes time $12nk(n-k)-6n$. For comparison, directly evaluating (3.49) via an SVD of B would have taken time $4k(n-k)^2 + 22k^3 + 2n^3$ (first two summands for SVD [86, p. 493], last summand for two matrix-matrix products).

3.7 Algorithms

We will now discuss optimization algorithms for minimizing a function $f : \mathrm{Gr}(k,n) \to \mathbb{R}$ in the involution model. In principle, this is equivalent to a quadratically constrained optimization problem in n^2 variables $[q_{ij}]_{i,j=1}^n = Q \in \mathbb{R}^{n\times n}$:

$$\begin{aligned} &\text{minimize} && f(Q) \\ &\text{subject to} && Q^\top Q = I,\ Q^\top = Q,\ \operatorname{tr}(Q) = 2k-n. \end{aligned} \tag{3.47}$$

Nevertheless, if one attempts to minimize any of the objective functions f in Section 3.8 by treating (3.47) as a general nonlinear constrained optimization problem using, say, the MATLAB Optimization Toolbox, every available method — interior point, trust region, sequential quadratic programming, active set — will fail without even finding a feasible point, never mind a minimizer. The Riemannian geometric objects and operations of the last few sections are essential to solving (3.47).

We will distinguish between two types of optimization algorithms. The *retraction algorithms*, as its name implies, will be based on various retractions and vector transports discussed in Section 3.6. The *Riemannian algorithms*, on the other hand, are built upon true Riemannian geodesics and parallel transports discussed in Section 3.4. Both types of algorithms will rely on the materials on points in Section 3.2, tangent vectors and metric in Section 3.3, and Riemannian gradients and Hessians in Section 3.5.

For both types of algorithms, the involution model offers one significant advantage over other existing models. In the involution model, explicit parallel transport and exponential map can be avoided. Instead of $\nabla f(Q)$ and $\exp_Q(X)$, it suffices to work with the matrices $G, B \in \mathbb{R}^{k\times(n-k)}$ that we will call *effective gradient* and *effective step* respectively, and doing so leads to extraordinarily simple and straightforward expressions in our algorithms. We will highlight this simplicity at appropriate junctures in Sections 3.7.2 and 3.7.3. Aside from simplicity, a more important consequence is that all key computations in our algorithms are performed at the *intrinsic dimension* of $\operatorname{Gr}(k, n)$. Our steepest descent direction, conjugate direction, Barzilai–Borwein step, Newton step, quasi-Newton step, etc, would all be represented as $k(n-k)$-dimensional objects. This is a feature not found in the algorithms of [3, 65, 100].

3.7.1 Initialization, eigendecomposition, and exponentiation

We begin by addressing three issues that we will frequently encounter in our optimization algorithms.

First observe that it is trivial to generate a point $Q \in \operatorname{Gr}(k, n)$ in the involution model: Take any orthogonal matrix $V \in \operatorname{O}(n)$, generated by say a QR decomposition of a random $n \times n$ matrix. Then we always have $Q := VI_{k,n-k}V^{\top} \in \operatorname{Gr}(k, n)$. We may easily generate as many random feasible initial points for our algorithms as we desire or simply take $I_{k,n-k}$ as our initial point.

The inverse operation of obtaining a $V \in \mathrm{O}(n)$ from a given $Q \in \mathrm{Gr}(k,n)$ so that $Q = VI_{k,n-k}V^{\top}$ seems more expensive as it appears to require an EVD. In fact, by the following observation, the cost is the same — a single QR decomposition.

Lemma 3.7.1. *Let $Q \in \mathbb{R}^{n\times n}$ with $Q^{\top}Q = I$, $Q^{\top} = Q$, $\operatorname{tr}(Q) = 2k - n$. If*

$$\frac{1}{2}(I+Q) = V\begin{bmatrix} R_1 & R_2 \\ 0 & 0 \end{bmatrix}, \qquad V \in \mathrm{O}(n),\ R_1 \in \mathbb{R}^{k\times k},\ R_2 \in \mathbb{R}^{k\times(n-k)}, \tag{3.48}$$

is a QR *decomposition, then $Q = VI_{k,n-k}V^{\top}$.*

Proof. Recall from (3.6) that for such a Q, we may write $V = [Y, Z]$ where $Y \in \mathrm{V}(k,n)$ and $Z \in \mathrm{V}(n-k,n)$ are a $+1$-eigenbasis and a -1-eigenbasis of Q respectively. By Proposition 3.2.3, $\frac{1}{2}(I+Q)$ is the projection matrix onto the $+1$-eigenspace $\operatorname{im}(Y) = \operatorname{im}\bigl(\frac{1}{2}(I+Q)\bigr)$, i.e., Y is an orthonormal column basis for $\frac{1}{2}(I+Q)$ and is therefore given by its condensed QR decomposition. As for Z, note that any orthonormal basis for $\operatorname{im}(Y)^{\perp}$ would serve the role, i.e., Z can be obtained from the full QR decomposition. In summary,

$$\frac{1}{2}(I+Q) = Y\begin{bmatrix} R_1 \\ 0 \end{bmatrix} = \begin{bmatrix} Y & Z \end{bmatrix}\begin{bmatrix} R_1 & R_2 \\ 0 & 0 \end{bmatrix}.$$

As a sanity check, note that

$$\frac{1}{2}(I+Q) = YY^{\top} = \begin{bmatrix} Y & Z \end{bmatrix}\begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} Y \\ Z \end{bmatrix} = V\begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix}V^{\top},$$

and therefore

$$Q = V\begin{bmatrix} I_k & 0 \\ 0 & -I_{n-k} \end{bmatrix}V^{\top} = VI_{k,n-k}V^{\top}.$$

□

Our expressions for tangent vector, exponential map, geodesic, parallel transport, retraction, etc, at a point $Q \in \operatorname{Gr}(k, n)$ all involve its matrix of eigenvectors $V \in \mathrm{O}(n)$. So Lemma 3.7.1 plays an important role in our algorithms. In practice, numerical stability considerations in the presence of rounding errors [55, Section 3.5.2] require that we perform our QR decomposition with *column pivoting* so that (3.48) becomes

$$\frac{1}{2}(I + Q) = V \begin{bmatrix} R_1 & R_2 \\ 0 & 0 \end{bmatrix} \Pi^\top$$

where Π is a permutation matrix. This does not affect our proof above; in particular, note that we have no need for R_1 nor R_2 nor Π in any of our algorithms.

The most expensive step in our Riemannian algorithms is the evaluation

$$B \mapsto \exp\left(\begin{bmatrix} 0 & B \\ -B^\top & 0 \end{bmatrix}\right) \tag{3.49}$$

for $B \in \mathbb{R}^{k \times (n-k)}$. General algorithms for computing matrix exponential [106, 147] do not exploit structures aside from normality. There are specialized algorithms that take advantage of skew-symmetry [32] or both skew-symmetry and sparsity [54] or the fact (3.49) may be regarded as the exponential map of a Lie algebra to a Lie group [34], but all of them require $O(n^3)$ cost. In [65], the exponential is computed via an SVD of B.

3.7.2 Retraction algorithms

In manifold optimization algorithms, an iterate is a point on a manifold and a search direction is a tangent vector at that point. Retraction algorithms rely on the retraction map R_Q for updating iterates and vector transport T_Q for updating search directions. Our interest in retraction algorithms is primarily to use them to initialize the Riemannian algorithms in the next section, and as such we limit ourselves to the least expensive ones.

A retraction-based steepest descent avoids even vector transport and takes the simple form

$$Q_{i+1} = R_{Q_i}\bigl(-\alpha_i \nabla f(Q_i)\bigr),$$

an analogue of the usual $x_{i+1} = x_i - \alpha_i \nabla f(x_i)$ in Euclidean space. As for our choice of retraction map, again computational costs dictate that we exclude the projection $R_Q^{\mathcal{E}}$ in Proposition 3.6.4 since it requires an EVD, and limit ourselves to the QR retraction $R_Q^{\mathcal{Q}}$ or Cayley retraction $R_Q^{\mathcal{C}}$ in Propositions 3.6.5 and 3.6.6 respectively. We present the latter in Algorithm 1 as an example.

We select our step size α_i using the well-known Barzilai–Borwein formula [15] but any line search procedure may be used instead. Recall that over Euclidean space, there are two choices for the Barzilai–Borwein step size:

$$\alpha_i = \frac{s_{i-1}^\top s_{i-1}}{(g_i - g_{i-1})^\top s_{i-1}}, \qquad \alpha_i = \frac{(g_i - g_{i-1})^\top s_{i-1}}{(g_i - g_{i-1})^\top (g_i - g_{i-1})}, \tag{3.50}$$

where $s_{i-1} \coloneqq x_i - x_{i-1}$. On a manifold M, the gradient $g_{i-1} \in \mathbb{T}_{x_{i-1}} M$ would have to be first parallel transported to $\mathbb{T}_{x_i} M$ and the step s_{i-1} would need to be replaced by a tangent vector in $\mathbb{T}_{x_{i-1}} M$ so that the exponential map $\exp_{x_{i-1}}(s_{i-1}) = x_i$. Upon applying this procedure, we obtain

$$\alpha_i = \frac{\operatorname{tr}(S_{i-1}^\top S_{i-1})}{\operatorname{tr}\bigl((G_i - G_{i-1})^\top S_i)\bigr)}, \qquad \alpha_i = \frac{\operatorname{tr}\bigl((G_i - G_{i-1})^\top S_{i-1}\bigr)}{\operatorname{tr}\bigl((G_i - G_{i-1})^\top (G_i - G_{i-1})\bigr)}. \tag{3.51}$$

In other words, it is as if we have naively replaced the g_i and s_i in (3.50) by the effective gradient G_i and the effective step S_i. But (3.51) is indeed the correct Riemannian expressions for Barzilai–Borwein step size in the involution model — the parallel transport and exponential map have already been taken into account when we derive (3.51). This is an example of the extraordinary simplicity of the involution model that we mentioned earlier and will see again in Section 3.7.3.

Algorithm 1 Steepest descent with Cayley retraction

```
Initialize $Q_0 = V_0 I_{k,n-k} V_0^\top \in \mathrm{Gr}(k, n)$.
for $i = 0, 1, \dots$ do
    compute effective gradient $G_i$ at $Q_i$                ▷ entries $*$ not needed
        $$V_i^\top (f_{Q_i} + f_{Q_i}^\top) V_i = \begin{bmatrix} * & 2G_i \\ 2G_i^\top & * \end{bmatrix};$$
    if $i = 0$ then
        initialize $S_0 = -G_0$, $\alpha_0 = 1$;
    else
        compute Barzilai–Borwein step                ▷ or get $\alpha_i$ from line search
            $$\alpha_i = \mathrm{tr}\big((G_i - G_{i-1})^\top S_{i-1}\big) / \mathrm{tr}\big((G_i - G_{i-1})^\top (G_i - G_{i-1})\big);$$
            $$S_i = -\alpha_i G_i;$$
    end if
    perform Cayley transform
        $$C_i = \begin{bmatrix} I & S_i/4 \\ -S_i^\top/4 & I \end{bmatrix} \begin{bmatrix} I & -S_i/4 \\ S_i^\top/4 & I \end{bmatrix}^{-1};$$
    update eigenbasis                ▷ effective vector transport
        $$V_{i+1} = V_i C_i;$$
    update iterate
        $$Q_{i+1} = V_{i+1} I_{k,n-k} V_{i+1}^\top;$$
end for
```

Of the two expressions for α_i in (3.51), we chose the one on the right because our effective gradient G_i, which is computed directly, is expected to be slightly more accurate than our effective step size S_i, which is computed from G_i. Other more sophisticated retraction algorithms [4] can be readily created for the involution model using the explicit expressions derived in Section 3.6.

3.7.3 Riemannian algorithms

Riemannian algorithms, called "geometric algorithms" in [65], are true geometric analogues of those on Euclidean spaces — straight lines are replaced by geodesic curves, displacements by parallel transports, inner products by Riemannian metrics, gradients and Hessians by their Riemannian counterparts. Every operation in a Riemannian algorithm is intrinsic: iterates stay on the manifold, conjugate and search directions stay in tangent spaces, and there are no geometrically meaningless operations like adding a point to a tangent vector or subtracting tangent vectors from two different tangent spaces.

The involution model, like other models in [3, 65, 100], supplies a system of extrinsic coordinates that allow geometric objects and operations to be computed with standard numerical linear algebra but it offers a big advantage, namely, one can work entirely with the effective gradients and effective steps. For example, it looks as if parallel transport is missing from our Algorithms 2–5, but that is only because the expressions in the involution model can be simplified to an extent that gives such an illusion. Our parallel transport is effectively contained in the step where we update the eigenbasis V_i to V_{i+1}.

We begin with steepest descent in Algorithm 2, the simplest of our four Riemannian algorithms. As in the case of Algorithm 1, we will use Barzilai–Borwein step size but any line search procedure may be used to produce α_i. In this case, any conceivable line search procedure would have required us to search over a geodesic curve and thus having to evaluate matrix exponential multiple times, using the Barzilai–Borwein step size circumvents this problem entirely.

Unlike its retraction-based counterpart in Algorithm 1, here the iterates descent along geodesic curves. Algorithm 1 may in fact be viewed as an approximation of Algorithm 2 where the matrix exponential in Step 9 is replaced with its first-order Padé approximation, i.e., a Cayley transform.

Newton method, shown in Algorithm 3, is straightforward with the computation of New-

Algorithm 2 Steepest descent

```
Initialize $Q_0 = V_0 I_{k,n-k} V_0^\top \in \mathrm{Gr}(k,n)$.
for $i = 0, 1, \ldots$ do
    compute effective gradient $G_i$ at $Q_i$                ▷ entries $*$ not needed
        $V_i^\top (f_{Q_i} + f_{Q_i}^\top) V_i = \begin{bmatrix} * & 2G_i \\ 2G_i^\top & * \end{bmatrix};$
    if $i = 0$ then
        initialize $S_0 = -G_0$, $\alpha_0 = 1$;
    else
        compute Barzilai–Borwein step                        ▷ or get $\alpha_i$ from line search
            $\alpha_i = \mathrm{tr}\big((G_i - G_{i-1})^\top S_{i-1}\big) / \mathrm{tr}\big((G_i - G_{i-1})^\top (G_i - G_{i-1})\big);$
            $S_i = -\alpha_i G_i;$
    end if
    update eigenbasis                                        ▷ effective parallel transport
        $V_{i+1} = V_i \exp\left(\begin{bmatrix} 0 & -S_i/2 \\ S_i^\top/2 & 0 \end{bmatrix}\right);$
    update iterate
        $Q_{i+1} = V_{i+1} I_{k,n-k} V_{i+1}^\top;$
end for
```

ton step as in (3.42). In practice, instead of a direct evaluation of $H_Q \in \mathbb{R}^{k(n-k)\times k(n-k)}$ as in (3.41), we determine H_Q in a manner similar to Corollary 3.5.3. When regarded as a linear map $H_Q : \mathbb{T}_Q \operatorname{Gr}(k,n) \to \mathbb{T}_Q \operatorname{Gr}(k,n)$, its value on a basis vector X_{ij} in (3.38) is

$$H_Q(X_{ij}) = \frac{1}{4} V \begin{bmatrix} 0 & B_{ij} + AE_{ij} - E_{ij}C \\ (B_{ij} + AE_{ij} - E_{ij}C)^\top & 0 \end{bmatrix} V^\top, \tag{3.52}$$

where A, C are as in (3.39) and B_{ij} is given by

$$V^\top \big(f_{QQ}(X_{ij}) + f_{QQ}(X_{ij})^\top \big) V = \begin{bmatrix} * & B_{ij} \\ B_{ij}^\top & * \end{bmatrix},$$

for all $i = 1, \dots, k$, $j = 1, \dots, n-k$. Note that these computations can be performed completely in parallel — with $k(n-k)$ cores, entries of H_Q can be evaluated all at once.

Algorithm 3 Newton's method

```
Initialize $Q_0 = V_0 I_{k,n-k} V_0^\top \in \operatorname{Gr}(k,n)$.
for $i = 0, 1, \dots$ do
    compute effiective gradient $G_i$ at $Q_i$
        $V_i^\top (f_{Q_i} + f_{Q_i}^\top) V_i = \begin{bmatrix} A_i & 2G_i \\ 2G_i^\top & C_i \end{bmatrix};$
    generate Hessian matrix $H_Q$ by (3.41) or (3.52);
    solve for effective Newton step $S_i$
        $H_Q \operatorname{vec}(S_i) = -\operatorname{vec}(G_i);$
    update eigenbasis                                    ▷ effective parallel transport
        $V_{i+1} = V_i \exp\left( \begin{bmatrix} 0 & S_i/2 \\ -S_i^\top/2 & 0 \end{bmatrix} \right);$
    update iterate
        $Q_{i+1} = V_{i+1} I_{k,n-k} V_{i+1}^\top;$
end for
```

Our conjugate gradient uses the Polak–Ribière formula [160] for conjugate step size; it is straightforward to replace that with the formulas of Dai–Yuan [51], Fletcher–Reeves [76], or Hestenes–Stiefel [103]. For easy reference:

$$\begin{aligned}
\beta_i^{\mathrm{PR}} &= \operatorname{tr}\bigl(G_{i+1}^{\top}(G_{i+1} - G_i)\bigr) / \operatorname{tr}(G_i^{\top} G_i),\\
\beta_i^{\mathrm{HS}} &= -\operatorname{tr}\bigl(G_{i+1}^{\top}(G_{i+1} - G_i)\bigr) / \operatorname{tr}\bigl(P_i^{\top}(G_{i+1} - G_i)\bigr),\\
\beta_i^{\mathrm{FR}} &= \operatorname{tr}(G_{i+1}^{\top} G_{i+1}) / \operatorname{tr}(G_i^{\top} G_i),\\
\beta_i^{\mathrm{DY}} &= -\operatorname{tr}(G_{i+1}^{\top} G_{i+1}) / \operatorname{tr}\bigl(P_i^{\top}(G_{i+1} - G_i)\bigr).
\end{aligned} \tag{3.53}$$

It may appear from these formulas that we are subtracting tangent vectors from tangent spaces at different points but this is an illusion. The effective gradients G_i and G_{i+1} are defined by the Riemannian gradients $\nabla f(Q_i) \in \mathbb{T}_{Q_i} \operatorname{Gr}(k, n)$ and $\nabla f(Q_{i+1}) \in \mathbb{T}_{Q_{i+1}} \operatorname{Gr}(k, n)$ as in (3.40) but they are not Riemannian gradients themselves. The formulas in (3.53) have in fact already accounted for the requisite parallel transports. This is another instance of the simplicity afforded by the involution model that we saw earlier in our Barzilai–Borwein step size (3.51) — our formulas in (3.53) are no different from the standard formulas for Euclidean space in [51, 76, 103, 160]. Contrast these with the formulas in [65, Equations 2.80 and 2.81], where the parallel transport operator τ makes an explicit appearance and cannot be avoided.

Our quasi-Newton method, given in Algorithm 5, uses L-BFGS updates with two loops recursion [158]. Observe that a minor feature of Algorithms 1, 2, 4, 5 is that they do not require vectorization of matrices; everything can be computed in terms of matrix-matrix products, allowing for Strassen-style fast algorithms. While it is straightforward to replace the L-BFGS updates with full BFGS, DFP, SR1, or Broyden class updates, doing so will require that we vectorize matrices like in Algorithm 3.

Algorithm 4 Conjugate gradient

```
Initialize $Q_0 = V_0 I_{k,n-k} V_0^\top \in \mathrm{Gr}(k,n)$.
Compute effective gradient $G_0$ at $Q_0$                    ▷ entries $*$ not needed
    $V_0^\top (f_{Q_0} + f_{Q_0}^\top) V_0 = \begin{bmatrix} * & 2G_0 \\ 2G_0^\top & * \end{bmatrix};$
initialize $P_0 = S_0 = -G_0$, $\alpha_0 = 1$;
**for** $i = 0, 1, \dots$ **do**
    compute $\alpha_i$ from line search of direction $P_i$ and set
        $S_i = \alpha_i P_i;$
    update eigenbasis                                        ▷ effective parallel transport
        $V_{i+1} = V_i \exp\left(\begin{bmatrix} 0 & -S_i/2 \\ S_i^\top/2 & 0 \end{bmatrix}\right);$
    update iterate
        $Q_{i+1} = V_{i+1} I_{k,n-k} V_{i+1}^\top;$
    compute effective gradient $G_{i+1}$ at $Q_{i+1}$            ▷ entries $*$ not needed
        $V_{i+1}^\top (f_X(Q_{i+1}) + f_X(Q_{i+1})^\top) V_{i+1} = \begin{bmatrix} * & 2G_{i+1} \\ 2G_{i+1}^\top & * \end{bmatrix};$
    compute Polak–Ribière conjugate step size
        $\beta_i = \mathrm{tr}\big((G_{i+1} - G_i)^\top G_{i+1}\big) / \mathrm{tr}(G_i^\top G_i);$
    update conjugate direction
        $P_{i+1} = -G_{i+1} + \beta_i P_i;$
**end for**
```

Algorithm 5 Quasi-Newton with L-BFGS updates

```
Initialize $Q_0 = V_0 I_{k,n-k} V_0^\top \in \mathrm{Gr}(k,n)$.
for $i = 0, 1, \dots$ do
    Compute effective gradient $G_i$ at $Q_i$                ▷ entries $*$ not needed
        $V_i^\top\big(f_X(Q_i) + f_X(Q_i)^\top\big)V_i = \begin{bmatrix} * & 2G_i \\ 2G_i^\top & * \end{bmatrix};$
    if $i = 0$ then
        initialize $S_0 = -G_0$;
    else
        set $Y_{i-1} = G_i - G_{i-1}$ and $P = G_i$;          ▷ $P$ is temporary variable for loop
        for $j = i-1, \dots, \max(0, i-m)$ do
            $\alpha_j = \mathrm{tr}(S_j^\top P)/\,\mathrm{tr}(Y_j^\top S_j)$;
            $P = P - \alpha_j Y_j$;
        end for
        set $Z = \mathrm{tr}(Y_{i-1}^\top S_{i-1})/\,\mathrm{tr}(Y_{i-1}^\top Y_{i-1})P$;   ▷ $Z$ is temporary variable for loop
        for $j = \max(0, i-m), \dots, i-1$ do
            $\beta_j = \mathrm{tr}(Y_j^\top Z)/\,\mathrm{tr}(Y_j^\top S_j)$;
            $Z = Z + (\alpha_j - \beta_j)S_j$;
        end for
        set effective quasi-Newton step $S_i = -Z$;
    end if
    update eigenbasis                                        ▷ effective parallel transport
        $V_{i+1} = V_i \exp\left(\begin{bmatrix} 0 & -S_i/2 \\ S_i^\top/2 & 0 \end{bmatrix}\right);$
    update iterate
        $Q_{i+1} = V_{i+1} I_{k,n-k} V_{i+1}^\top;$
end for
```

3.7.4 Exponential-free algorithms

In our algorithms, an exponential matrix $U := \exp\bigl(\bigl[\begin{smallmatrix} 0 & B \\ -B^{\top} & 0 \end{smallmatrix}\bigr]\bigr)$ is always[3] applied as a conjugation of some *symmetric* matrix $X \in \mathbb{R}^{n \times n}$:

$$X \mapsto UXU^{\top} \quad \text{or} \quad X \mapsto U^{\top}XU. \tag{3.54}$$

In other words, the Givens rotations in (3.46) are applied in the form of *Jacobi rotations* [86, p. 477]. For a symmetric X, a Jacobi rotation $X \mapsto G_{ij}(\theta)XG_{ij}(\theta)^{\top}$ takes the same number (as opposed to twice the number) of floating point operations as a Givens rotation applied on the left, $X \mapsto G_{ij}(\theta)X$, or on the right, $X \mapsto XG_{ij}(\theta)$. Thus with Strang splitting the operations in (3.54) take time $12nk(n-k)$. To keep our algorithms simple, we did not take advantage of this observation.

In principle, one may avoid any actual computation of matrix exponential by simply storing the $k(n-k)$ Givens rotations in (3.46) without actually forming the product, and apply them as Jacobi rotations whenever necessary. The storage of $G_{ij}(\theta)$ requires just a single floating point number θ and two indices but one would need to figure out how to update these $k(n-k)$ Givens rotations from one iteration to the next. We leave this as an open problem for interested readers.

3.7.5 Drawbacks of quotient models

This section, which may again be safely skipped, discusses the pitfalls of modeling a manifold as a homogeneous space of matrices. In our context, this would be the orthogonal, Stiefel, and full-rank models:

$$\operatorname{Gr}(k,n) \cong \operatorname{O}(n)/\bigl(\operatorname{O}(n-k) \times \operatorname{O}(k)\bigr) \cong \operatorname{V}(k,n)/\operatorname{O}(k) \cong \mathbb{R}^{n \times k}_{k}/\operatorname{GL}(k). \tag{3.55}$$

[3] See steps 3, 10 in Algorithm 2; steps 3, 7 in Algorithm 3; steps 7, 8 in Algorithm 4; steps 3, 20 in Algorithm 5.

Such homogeneous space models take the form of a quotient $B = E/G$ where B is the manifold we want to optimize over, E is some other manifold on which we have optimization algorithms, and G is some Lie group. The quotient map $\pi : E \to B$, $x \mapsto [x]$ defines a *principal bundle*. The idea of Riemannian optimization algorithms for such models is to lift every point on B up to the total space E so that optimization algorithms on E can be applied. In Section 3.1, we only mentioned the computational costs that come with lifting a point $[x] \in B$ to $x \in E$ and with checking equality of points $[x_1] = [x_2]$ given x_1, x_2. Here we focus on a more serious mathematical difficulty.

Since our goal is optimization, we cannot simply lift points on B to E in arbitrary ways. Ideally, whatever method of lifting should at least be continuous, i.e., nearby points in B are lifted to nearby points in E. In fact if we need first or second derivatives for the purpose of optimization, then the lifting has to be differentiable to first or second order, as we will see below. In differential geometric lingo, finding a lifting j is called finding a *global section* and it is impossible for any of the models in (3.55). Take the Stiefel model for illustration, the quotient map $\pi : \mathrm{V}(k, n) \to \mathrm{V}(k, n)/\,\mathrm{O}(k)$, $Y \mapsto [Y]$, defines $\mathrm{Gr}(k, n)$ as an $\mathrm{O}(k)$-principal bundle. This is not a trivial bundle, which is equivalent to π not admitting a global section [108]. The consequence is that there is no global 'Stiefel coordinates' for $\mathrm{Gr}(k, n)$, i.e., we cannot represent all points of $\mathrm{Gr}(k, n)$ by points of $\mathrm{V}(k, n)$ in a continuous manner.

We will use the Stiefel model as an illustration of the above discussion. Let $f : \mathrm{V}(k, n)/\,\mathrm{O}(k) \to \mathbb{R}$ be a differentiable objective function and let $\pi : \mathrm{V}(k, n) \to \mathrm{V}(k, n)/\,\mathrm{O}(k)$ be the quotient map, which is always smooth. A lifting is a right inverse $j : \mathrm{V}(k, n)/\,\mathrm{O}(k) \to \mathrm{V}(k, n)$ to π, i.e., $\pi \circ j([Y]) = [Y]$ for any $[Y] \in \mathrm{V}(k, n)/\,\mathrm{O}(k)$. Note that such a map j is not unique. In order to use standard numerical linear algebra, we have to work with actual matrices in $\mathrm{V}(k, n)$ as opposed to equivalence classes of matrices in $\mathrm{V}(k, n)/\,\mathrm{O}(k)$, so we need a way to assign to any equivalence class in $\mathrm{V}(k, n)/\,\mathrm{O}(k)$ a matrix in $\mathrm{V}(k, n)$ that represents that equivalence class — any such assignment gives us a lifting.

Upon selecting a lifting j, we may then transform the problem of optimzing $f : \operatorname{V}(k,n)/\operatorname{O}(k) \to \mathbb{R}$ to optimizing $f \circ \pi : j(\operatorname{V}(k,n)/\operatorname{O}(k)) \to \mathbb{R}$. Note that $j(\operatorname{V}(k,n)/\operatorname{O}(k)) \subseteq \operatorname{V}(k,n)$ is a set of actual matrices. The issue is that in order to have gradients and Hessians we also need $f \circ \pi$ to be a differentiable function — this is only possible if its domain $j(\operatorname{V}(k,n)/\operatorname{O}(k))$ is a differential manifold. But as we saw, we cannot even choose j to be continous, so $j(\operatorname{V}(k,n)/\operatorname{O}(k))$ is not even a topological manifold. In the involution and projection models, we do not face these issues as a point in $\operatorname{Gr}(k,n)$ is already represented by a matrix.

3.8 Numerical experiments

We will describe three sets of numerical experiments, testing Algorithms 1–5 on three different objective functions, the first two are chosen because their true solutions can be independently determined in closed-form, allowing us to ascertain that our algorithms have converged to the global optimizer. All our codes are open source and publicly available at:

https://github.com/laizehua/Simpler-Grassmannians

The goal of these numerical experiments is to compare our algorithms for the involution model in Section 3.7 with the corresponding algorithms for the Stiefel model in [65]. Algorithm 5, although implemented in our codes, is omitted from our comparisons as quasi-Newton methods are not found in [65].

3.8.1 Quadratic function

The standard test function for Grassmannian optimization is the quadratic form in [65, Section 4.4] which, in the Stiefel model, takes the form $\operatorname{tr}(Y^{\top}FY)$ for a symmetric $F \in \mathbb{R}^{n\times n}$ and $Y \in \operatorname{V}(k,n)$. By Proposition 3.2.4, we write $Q = 2YY^{\top} - I$, then $\operatorname{tr}(Y^{\top}FY) = \bigl(\operatorname{tr}(FQ) + \operatorname{tr}(F)\bigr)/2$. Therefore, in the involution model, this optimization problem takes an even simpler form

$$f(Q) = \operatorname{tr}(FQ) \tag{3.56}$$

for $Q \in \operatorname{Gr}(k,n)$. What was originally quadratic in the Stiefel model becomes linear in the involution model. The minimizer of f,

$$Q_* \coloneqq \operatorname{argmin}\bigl\{\operatorname{tr}(FQ) : Q^\top Q = I,\ Q^\top = Q,\ \operatorname{tr}(Q) = 2k - n\bigr\},$$

is given by $Q_* = \Pi V I_{k,n-k} V^\top \Pi^\top$ where

$$\Pi = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix} \qquad \text{and} \qquad \frac{F + F^\top}{2} = VDV^\top$$

is an eigendecomposition with eigenbasis $V \in \mathrm{O}(n)$ and eigenvalues $D \coloneqq \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ in descending order. This follows from essentially the same argument[4] used in the proof of Lemma 3.6.3 and the corresponding minimum is $f(Q_*) = -\lambda_1 - \cdots - \lambda_k + \lambda_{k+1} + \cdots + \lambda_n$.

For the function $f(Q) = \operatorname{tr}(FQ)$, the effective gradient $G_i \in \mathbb{R}^{k \times (n-k)}$ in Algorithms 2, 4, 5 at the point $Q_i = V_i I_{k,n-k} V_i^\top \in \operatorname{Gr}(k,n)$ is given by

$$V_i^\top F V_i = \begin{bmatrix} A & G_i \\ G_i^\top & C \end{bmatrix}.$$

The matrices $A \in \mathbb{R}^{k \times k}$ and $C \in \mathbb{R}^{(n-k)\times(n-k)}$ are not needed for Algorithms 2, 4, 5 but they are required in Algorithm 3. Indeed, the effective Newton step $S_i \in \mathbb{R}^{k \times (n-k)}$ in

[4] Recall also that for any real numbers $a_1 \le \cdots \le a_n$, $b_1 \le \cdots \le b_n$, and any permutation π, one always have that $a_1 b_n + a_2 b_{n-1} + \cdots + a_n b_1 \le a_1 b_{\pi(1)} + a_2 b_{\pi(2)} + \cdots + a_n b_{\pi(n)} \le a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$.

Algorithm 3 is obtained by solving the Sylvester equation

$$AS_i - S_iC = 2G_i.$$

To see this, note that by Proposition 3.5.2, for any $B \in \mathbb{R}^{k\times(n-k)}$,

$$\begin{aligned}
&\nabla^2 f(Q_i)\left(V_i \begin{bmatrix} 0 & B \\ B^\top & 0 \end{bmatrix} V_i^\top, V_i \begin{bmatrix} 0 & S_i \\ S_i^\top & 0 \end{bmatrix} V_i^\top\right) \\
&= -\frac{1}{2}\operatorname{tr}\left(\begin{bmatrix} A & G_i \\ G_i^\top & C \end{bmatrix}\begin{bmatrix} XS_i^\top + S_iB^\top & 0 \\ 0 & -B^\top S_i - S_i^\top B \end{bmatrix}\right) \\
&= -\operatorname{tr}\bigl(B^\top(AS_i - S_iC)\bigr),
\end{aligned}$$

and to obtain the effective Newton step (3.42), we simply set the last term to be equal to $-2\operatorname{tr}(B^\top G_i)$.

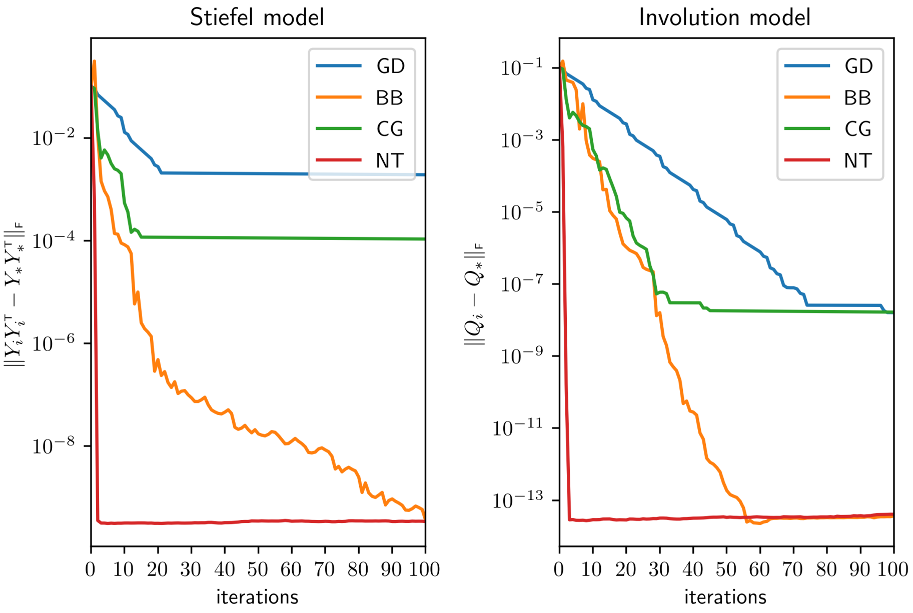


Figure 3.1: Convergence of algorithms in the Stiefel and involution models.

Figure 3.1 compares the convergence behaviors of the algorithms in [65] for the Stiefel model and our Algorithms 2, 3, 4 in the involution model: steepest descent with line search (GD) and with Barzilai–Borwein step size (BB), conjugate gradient (CG), and Newton's method (NT) for $k = 6$, $n = 16$. We denote the ith iterate in the Stiefel and involution models by Y_i and Q_i respectively — note that Y_i is a 16×6 matrix with orthonormal columns whereas Q_i is a 16×16 symmetric orthogonal matrix. All algorithms are fed the same initial point obtained from 20 iterations of Algorithm 1. The matrix F is generated randomly with standard normal entries. We use the Armijo conditions for linesearch in GD and CG; but as in the Euclidean case, BB and NT are used as is without linesearch. Since we have the true global minimizer in closed form, denoted by Y_* and Q_* in the respective model, the error is given by geodesic distance to the true solution. For convenience we compute $\|Y_i Y_i^\top - Y_* Y_*^\top\|_{\mathrm{Fro}}$ and $\|Q_i - Q_*\|_{\mathrm{Fro}}$, which are constant multiples of the chordal distance [203, Table 2] (also called projection F-norm [65, p. 337]) and are equivalent, in the sense of metrics, to the geodesic distance. Since we use a log scale, the vertical axes of the two graphs in Figure 3.1 are effectively both geodesic distance and, in particular, their values may be compared. The conclusion is clear: While Algorithms 2 (BB) and 3 (NT) in the involution model attain a level of accuracy on the order of machine precision, the corresponding algorithms in the Stiefel model do not. The reason is numerical stability, as we will see next.

Figure 3.2 shows the loss of orthogonality for various algorithms in the Stiefel and involution models, measured respectively by $\|Y_i^\top Y_i - I\|_{\mathrm{Fro}}$ and $\|Q_i^2 - I\|_{\mathrm{Fro}}$. In the Stiefel model, the deviation from orthogonality $\|Y_i^\top Y_i - I\|_{\mathrm{Fro}}$ grows exponentially. In the worst case, the GD iterates Y_i, which of course ought to be of rank $k = 6$, actually converged to a rank-one matrix. In the involution model, the deviation from orthogonality $\|Q_i^2 - I\|_{\mathrm{Fro}}$ remains below 10^{-13} for all algorithms — the loss-of-orthogonality is barely noticeable.

A closer inspection of the algorithms for NT [65, p. 325] and CG [65, p. 327] in the Stiefel model reveals why: A point Y_i and the gradient G_i at that point are highly dependent on

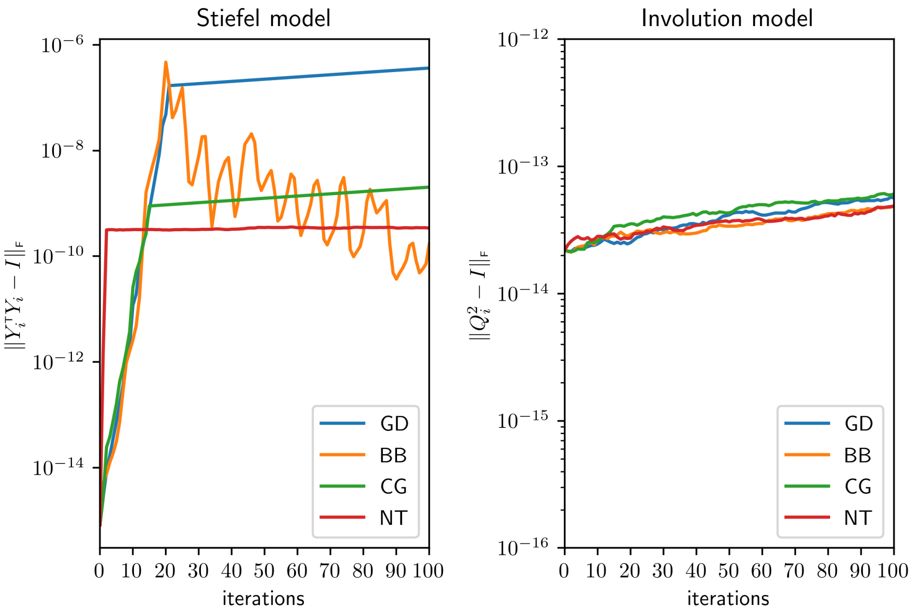


Figure 3.2: Loss of orthogonality in the Stiefel and involution models.

each other — an ε-deviation from orthogonality in Y_i results in an ε-error in G_i that in turn becomes a 2ε-deviation from orthogonality in Y_{i+1}, i.e., one loses orthogonality at an exponential rate. We may of course reorthogonalize Y_i at every iteration in the Stiefel model to artificially enforce the orthonormality of its columns but this incurs additional cost and turns a Riemannian algorithm into a retraction algorithm, as reorthogonalization of Y_i is effectively a QR retraction.

Contrast this with the involution model: In Algorithms 3 (NT) and 4 (CG), the point Q_i and the effective gradient G_i are both computed directly from the eigenbasis V_i, which is updated to V_{i+1} by an orthogonal matrix, or a sequence of Givens rotations if one uses Strang splitting as in (3.46). This introduces a small (constant order) deviation from orthogonality each step. Consequently, the deviation from orthogonality at worst grows linearly.

3.8.2 Grassmann Procrustes problem

Let $k, m, n \in \mathbb{N}$ with $k \le n$. Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{m\times k}$. The minimization problem

$$\min_{Q^\top Q = I} \|AQ - B\|_{\mathrm{Fro}},$$

is called the Stiefel Procrustes problem [65, Section 3.5.2] and the special case $k = n$ is the usual orthogonal Procrustes problem [86, Section 6.4.1]. Respectively, these are

$$\min_{Q\in \mathrm{V}(k,n)} \|AQ - B\|_{\mathrm{Fro}} \qquad \text{and} \qquad \min_{Q\in \mathrm{O}(n)} \|AQ - B\|_{\mathrm{Fro}}.$$

One might perhaps wonder if there is also a *Grassmann Procrustes problem*

$$\min_{Q\in \mathrm{Gr}(k,n)} \|AQ - B\|_{\mathrm{Fro}}. \tag{3.57}$$

In fact, with the involution model for $\mathrm{Gr}(k, n)$, the problem (3.57) makes perfect sense, with the requirement that $k = n$. The same argument in the proof of Lemma 3.6.3 shows that the minimizer Q_* of (3.57) is given by $Q_* = VI_{k,n-k}V^\top$ where

$$\frac{A^\top B + B^\top A}{2} = VDV^\top$$

is an eigendecomposition with eigenbasis $V \in \mathrm{O}(n)$ and eigenvalues $D \coloneqq \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ in descending order. The convergence and loss-of-orthogonality behaviors for this problem are very similar to those in Section 3.8.1 and provides further confirmation for the earlier numerical results. The plots from solving (3.57) for arbitrary A, B using any of Algorithms 2–5 are generated in our codes but as they are nearly identical to Figures 3.1 and 3.2 we omit them here.

3.8.3 Fréchet mean and Karcher mean

Let $Q_1, \dots, Q_m \in \operatorname{Gr}(k, n)$ and consider the sum-of-square-distances minimization problem:

$$\min_{Q \in \operatorname{Gr}(k,n)} \sum_{j=1}^{m} d^2(Q_j, Q), \tag{3.58}$$

where d is the geodesic distance in (3.32). The global minimizer of this problem is called the *Fréchet mean* and a local minimizer is called a *Karcher mean* [118]. For the case $m = 2$, a Fréchet mean is the midpoint, i.e., $t = 1/2$, of the geodesic connecting Q_1 and Q_2 given by the closed-form expression in Proposition 3.4.5. The objective function f in (3.58) is differentiable almost everywhere[5] with its Riemannian gradient [117] given by

$$\nabla f(Q) = 2 \sum_{j=1}^{m} \log_Q(Q_j),$$

where the logarithmic map is as in Corollary 3.4.7. To the best of our knowledge, there is no simple expression for $\nabla^2 f(Q)$ and as such we exclude Newton method from consideration below.

We will set $k = 6$, $n = 16$, and $m = 3$. Unlike the problems in Sections 3.8.1 and 3.8.2, the problem in (3.58) does not have a closed-form solution when $m > 2$. Consequently we quantify convergence behavior in Figure 3.3 by the rate gradient goes to zero. The deviation from orthogonality is quantified as in Section 3.8.1 and shown in Figure 3.4. The instability of the algorithms in the Stiefel model is considerably more pronounced here — both GD and CG failed to converge to a stationary point as we see in Figure 3.3. The cause, as revealed by Figure 3.4, is a severe loss-of-orthogonality that we will elaborate below.

The expression for geodesic distance $d(Y, Y')$ between two points Y, Y' in the Stiefel model (see [3, Section 3.8] or [203, Equation 7]) is predicated on the crucial assumption that

[5] f is nondifferentiable only when Q falls on the cut locus of Q_i for some i but the union of all cut loci of $Q_1, \dots, Q_m$ has codimension ≥ 1.

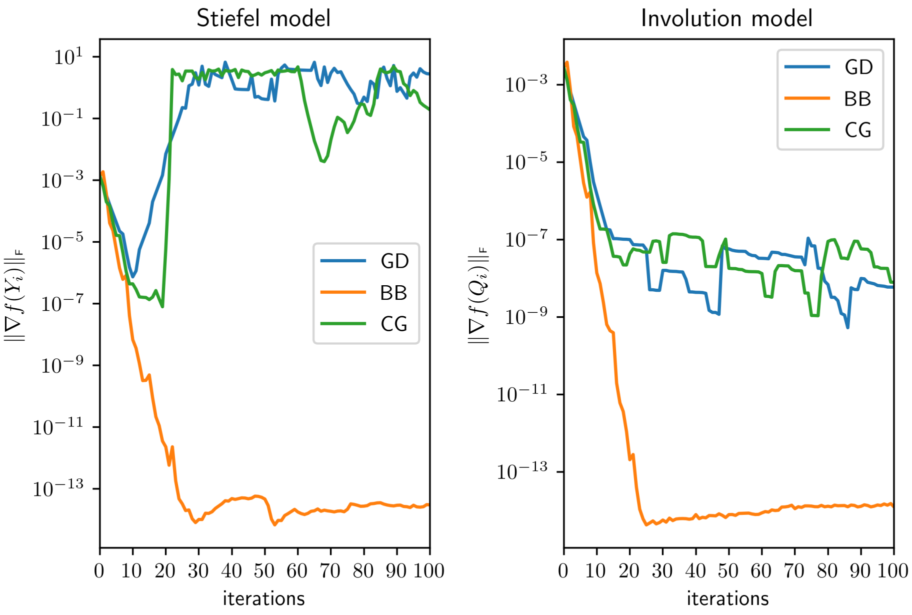


Figure 3.3: Convergence of algorithms in the Stiefel and involution models.

each of these matrices has orthonormal columns. As a result, a moderate deviation from orthonormality in an iterate Y leads to vastly inaccurate values in the objective function value $f(Y)$, which is a sum of m geodesic distances *squared*. This is reflected in the graphs on the left of Figure 3.3 for the GD and CG algorithms, whose step sizes come from line search and depend on these function values. Using the BB step size, which does not depend on objective function values, avoids the issue. But for GD and CG, the reliance on inaccurate function values leads to further loss-of-orthogonality, and when the columns of an iterate Y are far from orthonormal, plugging Y into the expression for gradient simply yields a nonsensical result, at times even giving an *ascent* direction in a minimization problem.

For all three algorithms in the involution model, the deviation from orthogonality in the iterates is kept at a negligible level of under 10^{-13} over the course of 100 iterations.

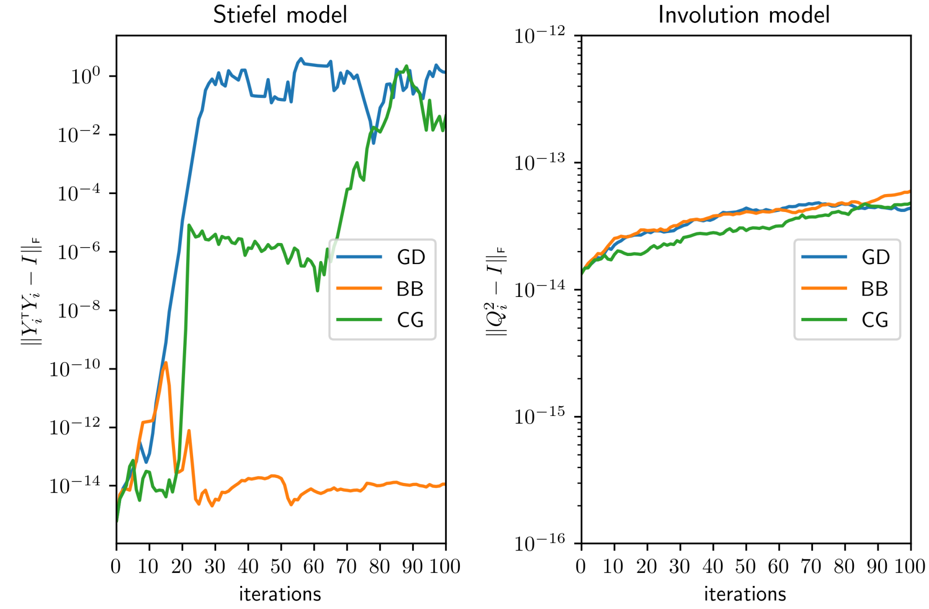


Figure 3.4: Loss of orthogonality in the Stiefel and involution models.

CHAPTER 4

SIMPLER FLAG OPTIMIZATION

This is a joint work with Lek-Heng Lim and Ke Ye.

4.1 Introduction

Let $d \leq n$ be positive integers and let $(n_1, \ldots, n_d)$ be a sequence integers such that $0 < n_1 < \cdots < n_d < n$. We denote by $\operatorname{Flag}(n_1, \ldots, n_d; \mathbb{R}^n)$ the set of all flags in $\mathbb{R}^n$ of type $(n_1, \ldots, n_d)$:

$$\operatorname{Flag}(n_1, \ldots, n_d; n) := \left\{ \{\mathbb{V}_k\}_{k=1}^d : \mathbb{V}_k \subsetneq \mathbb{V}_{k+1} \subsetneq \mathbb{R}^n, \dim \mathbb{V}_k = n_k, k = 1, \ldots, d-1 \right\}.$$

The set $\operatorname{Flag}(n_1, \ldots, n_d; n)$ is in fact a homogeneous space, variety, and manifold [107]. When $d = 1$, it is also called the Grassmannian manifold. If $n_i = i$ for $1 \leq i \leq d$ and $d = n - 1$, it is called the complete flag manifold.

In Riemannian optimization, we are interested in differential geometric objects and operations — tangent vector, metric, normal vector, exponential map, geodesic, parallel transport, gradient, Hessian, etc. And if those objects have closed-form analytic expressions that are computable with standard numerical linear algebra, then it is straightforward to design optimization algorithms on a manifold. For a same manifold, there can be multiple ways to represent it using different frameworks and different models can leads to difference in efficiency or accuracy of optimization algorithms. In [204], a quotient model for flag manifold is propose. In this work we intend to give a different *embedding* model that is simpler and more efficient for certain kinds of problems.

A key difficulty of optimization on flag manifolds is that there is no explicit formula for computing parallel transport or logarithm (or equivalently, the geodesic between two given points) [141, 213, 30, 156]. It implies that the Karcher mean problem on flag manifold is difficult to solve because the standard algorithm for Karcher mean problem requires the computation of logarithm. On the other hand, the extrinsic mean problem requires the computation of projection, i.e., finding the nearest point in a manifold. The projection problem can be easily solved for Stiefel manifolds and Grassmannian manifolds [65, 131], but there is no explicit formular for flag manifolds as far as we know. We provide an efficient algorithm for solving the projection problem, thus also solving the extrinsic mean problem.

4.2 Preliminaries

4.2.1 Some useful functions

We recall the *Peano–Baker series associated to a matrix function* $\Phi : [a, b] \to \mathbb{R}^{n\times n}$. To define the Peano–Baker series, we first recursively define a sequence $\{M_k(t)\}_{k=0}^{\infty}$ of matrix functions

$$\begin{aligned} M_0(t) &= I_n, \\ M_k(t) &= I_n + \int_a^t \Phi(s)M_{k-1}(s)ds, \quad k \in \mathbb{N}. \end{aligned}$$

We have the following:

Theorem 4.2.1. *[29, Section 3, Theorem 1] The sequence $\{M_k(t)\}_{k=0}^{\infty}$ converges to a matrix function $M(t)$ uniformly on $[a, b]$, which solves the differential equation*

$$\frac{d}{dt}X(t) = \Phi(t)X(t), \quad X(a) = I_n.$$

In particular, given any column vector $u \in \mathbb{R}^n$, $M(t)u$ solves the differential equation

$$\frac{d}{dt}x(t) = \Phi(t)x(t), \quad x(a) = u.$$

The limit matrix function $M(t)$ in Theorem 4.2.1 is defined to be the Peano–Baker series associated to $\Phi(t)$.

4.2.2 Vectorization of a matrix

Let m, n be positive integers and let $A[a_1, \dots, a_n]$ be a matrix of size $m \times n$ where $a_1, \dots, a_n \in \mathbb{R}^m$ are column vectors of A. We define the *vectorization* of A to be the column vector

$$\operatorname{vec}(A) \coloneqq \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \in \mathbb{R}^{mn}.$$

We recall that using vectorizations of matrices, we can express the matrix-matrix product in terms of matrix-vector product. Namely, for $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times l}$, we have

$$\operatorname{vec}(AB) = (I_l \otimes A) \operatorname{vec}(B) = (B^\top \otimes I_m) \operatorname{vec}(A). \tag{4.1}$$

Moreover, for any positive integers m, n, there exists a permutation matrix $K^{(m,n)} \in \mathbb{R}^{mn \times mn}$, called *the commutation matrix* such that

$$K^{(m,n)} \operatorname{vec}(A) = \operatorname{vec}(A^\top), \quad A \in \mathbb{R}^{m \times n}. \tag{4.2}$$

4.3 Sub-Riemannian geometry of flag manifolds with classical embeddings

According to [204, Proposition 3.2], $\operatorname{Flag}(n_1,\dots,n_d;n)$ can be naturally embedded into a product of Grassmann manifolds via

$$\begin{aligned} \iota : \operatorname{Flag}(n_1,\dots,n_d;n) &\hookrightarrow \operatorname{Gr}(n_1,n)\times\operatorname{Gr}(n_2-n_1,n)\cdots\times\operatorname{Gr}(n_d-n_{d-1},n)\\ (\{\mathbb{V}_k\}_{k=1}^d) &\mapsto (\mathbb{W}_1,\mathbb{W}_2,\dots,\mathbb{W}_d). \end{aligned} \tag{4.3}$$

Here $\mathbb{W}_1=\mathbb{V}_1$ and $\mathbb{W}_k$ is the orthogonal complement of $\mathbb{V}_{k-1}$ in $\mathbb{V}_k, k=2,\dots,d$. For simplicity, we denote

$$\boxed{m_1 \coloneqq n_1, \quad m_{d+1}\coloneqq n-n_d, \quad m_k \coloneqq n_k-n_{k-1}, \quad k=2,\dots,d} \tag{4.4}$$

so that ι is an embedding of $\operatorname{Flag}(n_1,\dots,n_d;n)$ into $\prod_{k=1}^d \operatorname{Gr}(m_k,n)$.

4.3.1 An embedding of a flag manifold into a matrix manifold

Using the involution model described in the last section, we can embed each $\operatorname{Gr}(m_k,n)$ into $O(n)$ and hence we can write $\mathbb{W}_k$ in (4.3) as $V_k I_{m_k,n-m_k} V_k^\top$ for some $V_k\in O(n)$. We denote by τ the induced embedding of $\prod_{k=1}^d \operatorname{Gr}(m_k,n)$ into $O(n)^d$. In the following, we will explicitly characterize the image $\tau\circ\iota\,(\operatorname{Flag}(n_1,\dots,n_d;n))$ in $O(n)^d$.

Proposition 4.3.1 (embedding)**.** *The image of the embedding*

$$\varepsilon : \operatorname{Flag}(n_1,\dots,n_d;n) \overset{\iota}{\hookrightarrow} \prod_{k=1}^d \operatorname{Gr}(m_k,n) \overset{\tau}{\hookrightarrow} O(n)^d \tag{4.5}$$

is given by

$$\varepsilon\left(\mathrm{Flag}(n_1,\ldots,n_d;n)\right) = \{(Q_1,\ldots,Q_d) \in O(n)^d : \mathrm{tr}(Q_k) = 2m_k - n, Q_k^\top = Q_k$$
$$(I_n + Q_k)(I_n + Q_{k+1}) = 0, k = 1,\ldots,d\}. \quad (4.6)$$

In particular, we have

$$\varepsilon\left(\mathrm{Flag}(n_1,\ldots,n_d;n)\right) = \left\{\left(VJ_1V^\top,\ldots,VJ_dV^\top\right) : V \in O(n)\right\}, \quad (4.7)$$

where $J_k = \mathrm{diag}(-I_{m_1},\cdots,-I_{m_{k-1}},I_{m_k},-I_{m_{k+1}},\cdots,-I_{m_d},-I_{m_{d+1}})$ *is obtained by permuting diagonal blocks of* $I_{m_k,n-m_k}$.

Proof. We must have $\varepsilon(\{\mathbb{R}^n_k\}_{k=1}^d) = (Q_1,\ldots,Q_d) \in (O(n) \cap S_n)$ with $\mathrm{rank}\, Q_k = 2m_k - n$. Moreover, since $\mathbb{W}_k$ is perpendicular to $\mathbb{W}_{k+1}$, we must have $P_{\mathbb{W}_k} \circ P_{\mathbb{W}_{k+1}} = 0$ where $P_{\mathbb{U}}$ is the orthogonal projection from $\mathbb{R}^n$ onto a subspace $\mathbb{U}$. Now by [131, Proposition 2.3], we have $P_{\mathbb{W}_k} = \frac{1}{2}(I_n + Q_k)$ which proves (4.6). To see (4.7), we notice that the relation $(I_n+Q_k)(I_n+Q_{k+1}) = 0$ implies that $Q_kQ_{k+1} = Q_{k+1}Q_k$ and hence there exists $V_0 \in O(n)$ diagonalizing Q_k's simultaneously, i.e., $Q_k = V_0D_kV_0^\top$ where D_k is a diagonal matrix with m_k -1's and $(n - m_k)$ 1's along its diagonal. The restriction $(I_n + Q_k)(I_n + Q_{k+1}) = 0$ forces $D_k = \sigma^\top J_k\sigma$ for some permutation matrix σ and hence $V := \sigma V_0$ gives us the desired expression of $\varepsilon(\{\mathbb{R}^n_k\}_{k=1}^d)$ in (4.7). □

In fact, (4.7) is a special case of the general fact [79, page 384] that G/P is an adjoint orbit of G if P is a parabolic subgroup of a semi-simple Lie group G. In our case, we have $G = O(n)$ and $P = O(m_1)\times\cdots\times O(m_{d+1})$ so that $G/P \simeq \mathrm{Flag}(n_1,\ldots,n_d;n)$ is the adjoint orbit of $(J_1,\ldots,J_d) \in O(n)^d$.

Due to Proposition 4.3.1, in the sequel we abuse the notation by using $\mathrm{Flag}(n_1,\ldots,n_d;n)$ to denote $\varepsilon\left(\mathrm{Flag}(n_1,\ldots,n_d;n)\right)$. Accordingly, an element in $\mathrm{Flag}(n_1,\ldots,n_d;n)$ is written as a d-tuple

$$(VJ_1V^\top, \dots, VJ_dV^\top) = V(J_1, \dots, J_d)V^\top$$

for some $V \in O(n)$, where $m_1 = n_1$ and $m_k = n_k - n_{k-1}$ for $k = 2, \dots, d$.

4.3.2 Tangent space, Riemannian metric and normal space

We first consider the tangent space of $\mathrm{Flag}(n_1, \dots, n_d; n)$ at a point $V(J_1, \dots, J_d)V^\top$. To do this, we take a curve $V(t)$ on $O(n)$ such that $V(0) = V$. It is clear that $\Lambda \coloneqq V(0)^\top \dot{V}(0) \in \mathfrak{so}(n)$ and hence the tangent vector determined by the curve $V(t)(J_1, \dots, J_d)V(t)^\top$ is simply

$$\dot{V}(0)(J_1, \dots, J_d)V(0)^\top + V(0)(J_1, \dots, J_d)\dot{V}(0)^\top$$

which can be further written as

$$V(0)\left(\Lambda(J_1, \dots, J_d) - (J_1, \dots, J_d)\Lambda\right)V(0)^\top.$$

We partition Λ as $\Lambda = (\Lambda(p,q))_{p,q=1}^{d+1}$ where $\Lambda(p,q)$ is a $m_p \times m_q$ matrix such that $\Lambda(q,p) = -\Lambda(p,q)^\top$. This implies that

$$\Lambda J_k - J_k\Lambda = -2\begin{bmatrix} 0 & \cdots & 0 & \Lambda(k,1)^\top & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \Lambda(k,k-1)^\top & 0 & \cdots & 0 \\ \Lambda(k,1) & \cdots & \Lambda(k,k-1) & 0 & \Lambda(k,k+1) & \cdots & \Lambda(k,d+1) \\ 0 & \cdots & 0 & \Lambda(k,k+1)^\top & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \Lambda(k,d+1)^\top & 0 & \cdots & 0 \end{bmatrix}. \tag{4.8}$$

We notice that there is a natural identification $\prod_{1\le j<k\le d+1} \mathbb{R}^{m_j\times m_k} \simeq \mathfrak{so}(n)$ and hence we have an injective map

$$\psi : \prod_{1\le j<k\le d+1} \mathbb{R}^{m_j\times m_k} \simeq \mathfrak{so}(n) \hookrightarrow \prod_{j=1}^{d} S_n, \quad \psi((A_{jk})_{1\le j<k\le d+1}) = \frac{1}{2}(AJ - JA),$$

where $J = (J_1, \ldots, J_d)$ and $A \in \mathfrak{so}(n)$ is the skew-symmetric matrix uniquely determined by $(A_{jk})_{1\le j<k\le d+1}$. The above calculations can be summarized as the following

Proposition 4.3.2. *Given a point $\mathfrak{f} \coloneqq V(J_1,\ldots,J_d)V^\top \in \operatorname{Flag}(n_1,\ldots,n_d;n)$, the tangent space of $\operatorname{Flag}(n_1,\ldots,n_d;n)$ at $\mathfrak{f}$ is*

$$\mathbb{T}_{\mathfrak{f}} \operatorname{Flag}(n_1,\ldots,n_d;n) = V \left\{\psi((A_{jk})_{1\le j<k\le d+1}) : A_{jk} \in \mathbb{R}^{m_j\times m_k}, 1 \le j < k \le d+1\right\} V^\top.$$

In other words, $\mathbb{T}_{\mathfrak{f}} \operatorname{Flag}(n_1,\ldots,n_d;n)$ consists of vectors $V(X_1,\ldots,X_d)V^\top \in \prod_{j=1}^{d} S_n$ satisfying

$$X_k(k,l) = -X_l(k,l), X_k(p,q) = 0, X_k(k,k) = 0, \tag{4.9}$$

for all $1 \le k,l \le d, 1 \le p,q \le d+1$, and $p,q,l \ne k$. Here for each $1 \le s,t \le d+1$, $X_k(s,t) \in \mathbb{R}^{m_s\times m_t}$ denotes the (s,t)-th block of $X_k \in S_n$ when we partition X_k with respect to $n = m_1 + \cdots + m_d + m_{d+1}$.

Due to Proposition 4.3.2, we are able to parametrize a curve on $\operatorname{Flag}(n_1,\ldots,n_d;n)$ easily.

Corollary 4.3.3 (curves)**.** *If $c : (-\varepsilon,\varepsilon) \to \operatorname{Flag}(n_1,\ldots,n_d;n)$ is a differentiable curve such that $c(0) = V(J_1,\ldots,J_d)V^\top$, then there exists a differentiable curve $\Lambda : (-\varepsilon,\varepsilon) \to \mathfrak{so}(n)$ such that $\Lambda(k,k)(t) \equiv 0, k = 1,\ldots,d+1$ and*

$$c(t) = V \exp(\Lambda(t))(J_1,\ldots,J_d)\exp(-\Lambda(t))V^\top,$$

where $\Lambda(t) = (\Lambda(j,k))_{j,k=1}^{d+1,d+1}$ is the partition of $\Lambda(t)$ with respect to $n = m_1 + \cdots + m_{d+1}$.

As a submanifold of $\prod_{k=1}^{d} \operatorname{Gr}(m_k, n)$ (or equivalently, $\prod_{k=1}^{d} O(n)$), $\operatorname{Flag}(n_1, \dots, n_d; n)$ is equipped with an induced Riemannian metric:

$$\langle V(X_1, \dots, X_d)V^\top, V(Y_1, \dots, Y_d)V^\top \rangle_{\mathfrak{f}} \coloneqq \sum_{k=1}^{d} \operatorname{tr}(X_k Y_k), \tag{4.10}$$

where $\mathfrak{f} = V(J_1, \dots, J_d)V^\top$ is a point in $\operatorname{Flag}(n_1, \dots, n_d; n)$, and $V(X_1, \dots, X_d)V^\top$ and $V(Y_1, \dots, Y_d)V^\top$ are tangent vectors of $\operatorname{Flag}(n_1, \dots, n_d; n)$ at $\mathfrak{f}$. More explicitly, we can write

$$\begin{aligned} &\langle V(X_1, \dots, X_d)V^\top, V(Y_1, \dots, Y_d)V^\top \rangle_{\mathfrak{f}} \\ &= 2\sum_{k=1}^{d} \sum_{l<k<m} \operatorname{tr}(X_k(l,k)Y_k(k,l) + X_k(m,k)Y_k(k,m)). \end{aligned} \tag{4.11}$$

We remark that summands in the formula (4.11) are not evenly counted. For example, if $d = 2$, then $\langle V(X_1, X_2)V^\top, V(Y_1, Y_2)V^\top \rangle_{\mathfrak{f}}$ is

$$2(\operatorname{tr}(X_1(2,1)Y_1(1,2)) + \operatorname{tr}(X_1(3,1)Y_1(1,3)) + \operatorname{tr}(X_2(1,2)Y_2(2,1)) + \operatorname{tr}(X_2(3,2)Y_2(2,3))), \tag{4.12}$$

in which there is a repeated term $\operatorname{tr}(X_2(1,2)Y_2(2,1)) = \operatorname{tr}(X_1(2,1)Y_1(1,2))$.

For each $Q \in O(n)$, we have $\mathbb{T}_Q O(n) = Q\mathfrak{so}(n)$ and hence for each $(Q_1, \dots, Q_d) \in \prod_{k=1}^{d} O(n)$, we obtain

$$\mathbb{T}_{(Q_1,\dots,Q_d)} \left(\prod_{k=1}^{d} O(n) \right) = \bigoplus_{k=1}^{d} Q_k \mathfrak{so}(n).$$

To calculate the normal space of $\operatorname{Flag}(n_1, \dots, n_d; n)$ in $\prod_{k=1}^{d} O(n)$ at $\mathfrak{f} = V(J_1, \dots, J_d)V^\top$, we need to determine $Y_1, \dots, Y_d \in \mathfrak{so}(n)$ such that $Y \coloneqq (VJ_1V^\top Y_1, \dots, VJ_dV^\top Y_d)$ is perpendicular to $\mathbb{T}_{\mathfrak{f}} \operatorname{Flag}(n_1, \dots, n_d; n)$, i.e., $\langle X, Y \rangle_{\mathfrak{f}, O(n)} = 0$ for all $X \in \mathbb{T}_{\mathfrak{f}} \operatorname{Flag}(n_1, \dots, n_d; n)$.

Here the inner product $\langle \cdot, \cdot \rangle_{\mathfrak{f}, \prod_{k=1}^{d} O(n)}$ is the canonical Riemannian metric on $\prod_{k=1}^{d} O(n)$ at the point $\mathfrak{f}$, which induces (4.10). We notice that for $X = V(X_1, \ldots, X_d)V^\top$,

$$VX_kV^\top = (VJ_kV^\top)VJ_kX_kV^\top, \quad k = 1, \ldots, d,$$

which implies that

$$\begin{aligned}
\langle X, Y \rangle_{\mathfrak{f}, \prod_{k=1}^{d} O(n)} &= \sum_{k=1}^{d} \operatorname{tr}((VJ_kX_kV^\top)^\top Y_k) \\
&= \sum_{k=1}^{d} \operatorname{tr}((VX_kJ_kV^\top)Y_k) \\
&= \sum_{k=1}^{d} \operatorname{tr}((X_kJ_k)V^\top Y_kV)
\end{aligned}$$

Since $\langle X, Y \rangle_{\mathfrak{f}, \prod_{k=1}^{d} O(n)} = 0$ holds for any $X \in \mathbb{T}_{\mathfrak{f}} \operatorname{Flag}(n_1, \ldots, n_d; n)$, we can equivalently write this condition as

$$\sum_{k=1}^{d} \operatorname{tr}((X_kJ_k)Z_k) = 0, \quad (X_1, \ldots, X_d) \in \mathbb{T}_{\mathfrak{f}_0} \operatorname{Flag}(n_1, \ldots, n_d; n),$$

where $\mathfrak{f}_0 = (J_1, \ldots, J_d) \in \operatorname{Flag}(n_1, \ldots, n_d; n)$ and $Z_k = V^\top Y_kV, k = 1, \ldots, d$. If we fix a pair (k, l) such that $1 \leq k \leq d, 1 \leq l \leq d+1, k \neq l$ and set $X_m(p, q) = 0$ for

$$(m, p, q) \notin \{(k, k, l), (k, l, k), (l, k, l), (l, l, k)\},$$

then since $X_k J_{m_k,n-m_k}$ is skew-symmetric, we have

$$\begin{aligned}
0 &= \langle X, Y\rangle_{\mathfrak{f},\prod_{k=1}^{d} O(n)} \\
&= \operatorname{tr}(X_k(k,l)Z_k(l,k)) - \operatorname{tr}(X_k(k,l)^\top Z_k(k,l)) - \operatorname{tr}(X_l(k,l)^\top Z_l(k,l)) + \operatorname{tr}(X_l(k,l)Z_l(l,k)) \\
&= \operatorname{tr}(X_k(k,l)(Z_k(l,k)) - Z_l(l,k))) - \operatorname{tr}(X_k(k,l)^\top(-Z_k(k,l)) + Z_l(k,l))) \\
&= \operatorname{tr}(X_k(k,l)(Z_k(l,k)) - Z_l(l,k))) + \operatorname{tr}(X_k(k,l)^\top(Z_k(k,l)) - Z_l(k,l)) \\
&= 2\operatorname{tr}(X_k(k,l)(Z_k(l,k) - Z_l(l,k))).
\end{aligned}$$

Therefore, we may derive the following characterization of $\mathbb{N}_{\mathfrak{f}}\operatorname{Flag}(n_1,\dots,n_d;n)$:

Proposition 4.3.4. *At a point $\mathfrak{f} := V(J_1,\dots,J_d)V^\top \in \operatorname{Flag}(n_1,\dots,n_d;\mathbb{R}^n)$, the normal space $\mathbb{N}_{\mathfrak{f}}\operatorname{Flag}(n_1,\dots,n_d;n)$ consists of vectors*

$$(VJ_1Z_1V^\top,\dots,VJ_dZ_dV^\top)$$

where $Z_1,\dots,Z_d \in \mathfrak{so}(n)$ satisfy the relations

- *$Z_k(k,l) - Z_l(k,l) = 0$ for all $1 \le k \ne l \le d$.*
- *$Z_k(k,d+1) = 0, Z_k(d+1,k) = 0$ for all $1 \le k \le d$.*

In particular, we have a decomposition

$$\mathbb{N}_{\mathfrak{f}}\operatorname{Flag}(n_1,\dots,n_d;n) = N_{\mathfrak{f}}\left(\prod_{k=1}^{d}\operatorname{Gr}(m_k,n)\right)\bigoplus N_{\mathfrak{f}}^0 \tag{4.13}$$

where $\mathbb{N}_{\mathfrak{f}}\left(\prod_{k=1}^{d}\operatorname{Gr}(m_k,n)\right) \coloneqq \prod_{k=1}^{d}\mathbb{N}_{VJ_{m_k,n-m_k}V^{\top}}\operatorname{Gr}(m_k,n)$ *and*

$$\begin{aligned}\mathbb{N}_{\mathfrak{f}}^{0} \coloneqq \{&(VJ_{m_1,n-m_1}Z_1V^{\top},\dots,VJ_{m_d,n-m_d}Z_dV^{\top}) : Z_k\in\mathfrak{so}(n), Z_k(k,l)-Z_l(k,l)=0,\\ &Z_k(k,k)=0, Z_k(p,q)=0, Z_k(k,d+1)=0, Z_k(d+1,k)=0,\\ &1\le k,l\le d, 1\le p,q\le d+1, p,q\ne k\}. \end{aligned} \tag{4.14}$$

We recall that $\operatorname{Flag}(n_1,\dots,n_d;n)$ can also be embedded into $\prod_{k=1}^{d}\operatorname{Gr}(n_k,n)$ as a Riemannian submanifold. Hence we may also characterize the normal space with respect to this embedding.

Corollary 4.3.5. *The normal space of* $\operatorname{Flag}(n_1,\dots,n_d;n)$ *in* $\prod_{k=1}^{d}\operatorname{Gr}(m_k,n)$ *at a point* $\mathfrak{f}$ *is* $\mathbb{N}_{\mathfrak{f}}^{0}$.

Proposition 4.3.6. *Projections from* $\mathbb{T}_{\mathfrak{f}}\left(\prod_{k=1}^{d}O(n)\right)$ *onto* $\mathbb{T}_{\mathfrak{f}}\operatorname{Flag}(n_1,\dots,n_d;n)$ *and* $\mathbb{N}_{\mathfrak{f}}\operatorname{Flag}(n_1,\dots,n_d;n)$ *are respectively given by*

$$\begin{aligned}\operatorname{proj}_{\mathfrak{f}}^{\mathbb{T}} : \mathbb{T}_{\mathfrak{f}}\left(\prod_{k=1}^{d}O(n)\right) &\to \mathbb{T}_{\mathfrak{f}}\operatorname{Flag}(n_1,\dots,n_d;n)\\ V(J_1\Lambda_1,\dots,J_d\Lambda_d)V^{\top} &\mapsto V(X_1,\dots,X_d)V^{\top}. \end{aligned} \tag{4.15}$$

and

$$\begin{aligned}\operatorname{proj}_{\mathfrak{f}}^{\mathbb{N}} : \mathbb{T}_{\mathfrak{f}}\left(\prod_{k=1}^{d}O(n)\right) &\to \mathbb{N}_{\mathfrak{f}}\operatorname{Flag}(n_1,\dots,n_d;n)\\ V(J_1\Lambda_1,\dots,J_d\Lambda_d)V^{\top} &\mapsto V(Z_1,\dots,Z_d)V^{\top} \end{aligned} \tag{4.16}$$

where for each $k=1,\dots,d$, $X_k\in S_n$ *(resp.* $Z_k\in\mathbb{R}^{n\times n}$*) is partitioned as* $(X_k(p,q))_{p,q=1}^{d+1}$

(resp. $(Z_k(p,q))_{p,q=1}^{d+1}$*) with respect to* $n = m_1 + \cdots + m_{d+1}$ *and*

$$X_k(p,q) = \begin{cases} \frac{1}{2}(\Lambda_k(k,q) - \Lambda_q(k,q)), & \text{if } p = k \neq q \leq d \\ \Lambda_k(k, d+1), & \text{if } p = k, q = d+1 \\ -\frac{1}{2}(\Lambda_k(p,k) - \Lambda_p(p,k)), & \text{if } q = k \neq p \leq d \\ -\Lambda_k(d+1, q), & \text{if } q = k, p = d+1 \\ 0, & \text{otherwise.} \end{cases}$$

$$Z_k(p,q) = \begin{cases} \frac{1}{2}(\Lambda_k(k,q) + \Lambda_q(k,q)), & \text{if } p = k \neq q \leq d \\ 0, & \text{if } p = k, q = d+1 \\ -\frac{1}{2}(\Lambda_k(p,k) + \Lambda_p(p,k)), & \text{if } q = k \neq p \leq d \\ 0, & \text{if } q = k, p = d+1 \\ \Lambda_k(p,q), & \text{otherwise.} \end{cases}$$

Before we proceed, we work out the case for $d = 2$ to exhibit our calculations above. In this case, our flag manifold is $\operatorname{Flag}(n_1, n_2; n)$ and hence $m_1 = n_1, m_2 = n_2 - n_1, m_3 = n - n_2$. A point $\mathfrak{f}$ in $\operatorname{Flag}(n_1, n_2; n)$ is written as

$$V(J_1, J_2)V^\top = V\left(\begin{bmatrix} I_{m_1} & 0 & 0 \\ 0 & -I_{m_2} & 0 \\ 0 & 0 & -I_{m_3} \end{bmatrix}, \begin{bmatrix} -I_{m_1} & 0 & 0 \\ 0 & I_{m_2} & 0 \\ 0 & 0 & -I_{m_3} \end{bmatrix}\right)V^\top, \quad V \in O(n).$$

A tangent vector of $\operatorname{Flag}(n_1, n_2; n)$ at $\mathfrak{f}$ is of the form

$$V\left(\begin{bmatrix} 0 & A & B \\ A^\top & 0 & 0 \\ B^\top & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -A & 0 \\ -A^\top & 0 & C \\ 0 & C^\top & 0 \end{bmatrix}\right)V^\top, \quad A \in \mathbb{R}^{m_1 \times m_2}, B \in \mathbb{R}^{m_1 \times m_3}, C \in \mathbb{R}^{m_2 \times m_3}.$$

The normal space of $\operatorname{Flag}(n_1, n_2; n)$ as a submanifold of $O(n) \times O(n)$ at $\mathfrak{f}$ consists of vectors

$$V \left(\begin{bmatrix} X & Y & 0 \\ Y^\top & Z & W \\ 0 & -W^\top & U \end{bmatrix}, \begin{bmatrix} R & Y & S \\ Y^\top & T & 0 \\ -S^\top & 0 & K \end{bmatrix} \right) V^\top,$$

where $X, R \in \mathfrak{so}(m_1),\ Z, T \in \mathfrak{so}(m_2),\ U, K \in \mathfrak{so}(m_3),\ Y \in \mathbb{R}^{m_1 \times m_2}, W \in \mathbb{R}^{m_2 \times m_3}, S \in \mathbb{R}^{m_1 \times m_3}$.

A tangent vector ξ of $O(n) \times O(n)$ at $\mathfrak{f}$ can be written as

$$\xi \coloneqq V \left(\begin{bmatrix} A & B & C \\ B^\top & D & E \\ C^\top & -E^\top & F \end{bmatrix}, \begin{bmatrix} X & Y & Z \\ Y^\top & W & S \\ -Z^\top & S^\top & T \end{bmatrix} \right) V^\top,$$

where $A, X \in \mathfrak{so}(m_1), D, W \in \mathfrak{so}(m_2), F, T \in \mathfrak{so}(m_3), B, Y \in \mathbb{R}^{m_1 \times m_2}, C, Z \in \mathbb{R}^{m_1 \times m_3}$, $E, S \in \mathbb{R}^{m_2 \times m_3}$. The projection of ξ onto $T_{\mathfrak{f}} \operatorname{Flag}(n_1, n_2; n)$ is

$$\operatorname{proj}_{\mathfrak{f}}^{\mathbb{T}}(\xi) = V \left(\begin{bmatrix} 0 & \frac{B-Y}{2} & C \\ \frac{B^\top - Y^\top}{2} & 0 & 0 \\ C^\top & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -\frac{B-Y}{2} & 0 \\ -\frac{B^\top - Y^\top}{2} & 0 & S \\ 0 & S^\top & 0 \end{bmatrix} \right) V^\top$$

and its projection onto $N_{\mathfrak{f}} \operatorname{Flag}(n_1, n_2; n)$ is

$$\operatorname{proj}_{\mathfrak{f}}^{\mathbb{N}}(\xi) = V \left(\begin{bmatrix} A & \frac{B+Y}{2} & 0 \\ \frac{B^\top + Y^\top}{2} & D & E \\ 0 & -E^\top & F \end{bmatrix}, \begin{bmatrix} X & \frac{B+Y}{2} & Z \\ \frac{B^\top + Y^\top}{2} & W & 0 \\ -Z^\top & 0 & T \end{bmatrix} \right) V^\top$$

The normal space $\mathbb{N}_f^0$ of $\operatorname{Flag}(n_1, n_2; n)$ as a submanifold of $\operatorname{Gr}(m_1, n) \times \operatorname{Gr}(m_2, n)$ at $\mathfrak{f}$

consists of vectors

$$V\left(\begin{bmatrix}0 & Y & 0\\ Y^\top & 0 & 0\\ 0 & 0 & 0\end{bmatrix}, \begin{bmatrix}0 & Y & 0\\ Y^\top & 0 & 0\\ 0 & 0 & 0\end{bmatrix}\right)V^\top, \quad Y \in \mathbb{R}^{m_1 \times m_2}.$$

We also recall that the tangent space $\mathbb{T}_{\mathfrak{f}}(\mathrm{Gr}(m_1, n) \times \mathrm{Gr}(m_2, n))$ consists of vectors

$$V\left(\begin{bmatrix}0 & A & B\\ A^\top & 0 & 0\\ B^\top & 0 & 0\end{bmatrix}, \begin{bmatrix}0 & D & 0\\ D^\top & 0 & C\\ 0 & C^\top & 0\end{bmatrix}\right)V^\top,$$

where $V \in O(n), A, D \in \mathbb{R}^{m_1\times m_2}, B \in \mathbb{R}^{m_1 \times m_3}, C \in \mathbb{R}^{m_2\times m_3}$. The following identities can be directly verified by the above computations.

$$\begin{aligned}\mathbb{T}_{\mathfrak{f}}\left(O(n) \times O(n)\right) &= \mathbb{T}_{\mathfrak{f}}\,\mathrm{Flag}(n_1, n_2; n) \bigoplus \mathbb{N}_{\mathfrak{f}}\,\mathrm{Flag}(n_1, n_2; n),\\ \mathbb{T}_{\mathfrak{f}}(\mathrm{Gr}(m_1, n) \times \mathrm{Gr}(m_2, n)) &= \mathbb{T}_{\mathfrak{f}}\,\mathrm{Flag}(n_1, n_2; n) \bigoplus \mathbb{N}^0_{\mathfrak{f}}.\end{aligned}$$

4.3.3 Geodesics

Recall that we may parametrize a curve $c(t)$ on $\mathrm{Flag}(n_1, \dots, n_d; n)$ as

$$c(t) = V(t)(J_1, \dots, J_d)V^\top(t),$$

where $V(t)$ is a curve in $O(n)$. By differentiating the equation $V(t)^\top V(t) = I_n$, we obtain

$$\dot{V}(t)^\top V(t) + V(t)^\top \dot{V}(t) = 0,$$

from which we may write $\dot{V}(t)$ as

$$\dot{V}(t) = V(t)\Lambda(t),$$

for some $\Lambda(t) \in \mathfrak{so}(n)$. According to Proposition 4.3.2, we may further partition $\Lambda(t)$ as

$$\Lambda(t) = (\Lambda_{jk})_{j,k=1}^{d+1,d+1}$$

with respect to $n = m_1 + \cdots + m_{d+1}$ and $\Lambda_{kk}(t) \equiv 0, k = 1, \ldots, d+1$. Hence the second derivative of $c(t)$ is

$$\ddot{c}(t) = V(t)\left(\Delta_1(t), \ldots, \Delta_d(t)\right) V(t)^\top$$

where

$$\Delta_k(t) = (\dot{\Lambda}(t)J_k - J_k\dot{\Lambda(t)}) + (\Lambda^2(t)J_k + J_k\Lambda^2(t)) + \left(-2\Lambda(t)J_k\Lambda(t)\right), \quad k = 1, \ldots, d. \tag{4.17}$$

We may rewrite $\ddot{c}(t)$ as

$$\ddot{c}(t) = T_1(t) + T_2(t) - 2T_3(t)$$

where $T_j(t)$ is the j-summand of $V(t)\left(\Delta_1(t), \ldots, \Delta_d(t)\right) V(t)^\top$ with respect to the decomposition of $\Delta_k(t)$ given in (4.17). More precisely,

$$\begin{aligned} T_1(t) &= V(t)(\dot{\Lambda}(t)J_1 - J_1\dot{\Lambda}(t), \ldots, \dot{\Lambda}(t)J_d - J_d\dot{\Lambda}(t))V(t)^\top, & (4.18)\\ T_2(t) &= V(t)(\Lambda^2(t)J_1 + J_1\Lambda^2(t), \ldots, \Lambda^2(t)J_d + J_d\Lambda^2(t))V(t)^\top, & (4.19)\\ T_3(t) &= V(t)(\Lambda(t)J_1\Lambda(t), \ldots, \Lambda(t)J_d\Lambda(t))V(t)^\top. & (4.20) \end{aligned}$$

We recall that the geodesic equation on $\operatorname{Flag}(n_1, \ldots, n_d; n)$ is given by

$$\operatorname{proj}^{\mathbb{T}}_{c(t)}(\ddot{c}(t)) = 0.$$

Therefore, to determine the geodesic equation explicitly, we need to compute the projections of $T_1(t), T_2(t), T_3(t)$ to $T_{c(t)} \operatorname{Flag}(n_1, \ldots, n_d; n)$ respectively. From Proposition 4.3.2, $T_1(t)$ already lies in the tangent space $T_{c(t)} \operatorname{Flag}(n_1, \ldots, n_d; n)$. Hence it is sufficient to determine the projections of $T_2(t)$ and $T_3(t)$.

Lemma 4.3.7. *Let $c(t), T_2(t)$ be as above. The projection of $\operatorname{proj}^{\mathbb{T}}_{c(t)}(T_2(t))$ is zero.*

Proof. We first compute $\Lambda^2(t)J_k + J_k\Lambda^2(t)$ for each $k = 1, \ldots, d$. To do this, we partition $\Lambda^2(t)$ (resp. J_k) as $(\Gamma_{p,q}(t))$ (resp. $(J_k(p,q))$) with respect to the partition $n = m_1 + \cdots + m_{d+1}$ and we recall that

$$J_k(p,q) = \begin{cases} (2\delta_{pk} - 1)I_{m_p}, & \text{if } q = p, \\ 0, \quad \text{otherwise.} \end{cases}$$

Here δ_{pk} is the Kronecker delta function. Since $\Lambda(t)$ is skew-symmetric, $\Lambda^2(t)$ is symmetric. We have $\Gamma_{q,p} = \Gamma_{p,q}^{\top}$. Now the (p,q)-th block of $\Lambda^2(t)J_k$ is

$$\sum_{l=1}^{m+1} \Gamma_{p,l} J_k(l,q) = \Gamma_{p,q} J_k(q,q) = (2\delta_{qk} - 1)\Gamma_{p,q}$$

and the (p,q)-th block of $J_k\Lambda^2(t) = (\Lambda^2(t)J_k)^{\top}$ is $(2\delta_{pk} - 1)\Gamma_{p,q}$. This implies that the (p,q)-th block of $\Lambda^2(t)J_k + J_k\Lambda^2(t)$ is

$$(2\delta_{qk} - 1)\Gamma_{p,q} + (2\delta_{pk} - 1)\Gamma_{p,q} = (-2)(1 - \delta_{pk} - \delta_{qk})\Gamma_{p,q}.$$

In particular, if either $q \neq p = k$ or $p \neq q = k$, we obtain that the (p,q)-th block of $\Lambda^2(t)J_k + J_k\Lambda^2(t)$ is zero and this implies that $\operatorname{proj}^{\mathbb{T}}_{c(t)}(T_2(t)) = 0$. □

Lemma 4.3.8. *Let $c(t), T_3(t)$ be as before. The projection $\operatorname{proj}^{\mathbb{T}}_{c(t)}(T_3(t))$ is*

$$V(t)(X_1, \ldots, X_d)V(t)^{\top}$$

where for each $1 \leq k \leq d$, X_k is a symmetric matrix whose (p,q)-th block vanishes for any (p,q) except $(k, d+1)$ and $(d+1,k)$. Moreover if we partition $\Lambda(t)$ as

$$\Lambda(t) = \begin{bmatrix} \Lambda_0(t) & \Lambda_1(t) \\ -\Lambda_1(t)^\top & 0 \end{bmatrix},$$

where $\Lambda_0(t) \in \mathfrak{so}(n - m_{d+1})$ and $\Lambda_1(t) \in \mathbb{R}^{(n-m_{d+1})\times m_1}$ we have

$$\begin{bmatrix} X_1(1, d+1) \\ \vdots \\ X_d(d, d+1) \end{bmatrix} = -\Lambda_0(t)\Lambda_1(t).$$

Proof. It is sufficient to compute $X_k \coloneqq \Lambda(t) J_k \Lambda(t)$ for each $k = 1, \dots, d$. We again partition $\Lambda(t)$ as $(\Lambda(p,q)(t))_{p,q=1}^d$ with respect to $n = m_1 + \cdots + m_{d+1}$. The (p,q)-th block of $\Lambda(t)J_k\Lambda(t)$ is

$$\begin{aligned}\sum_{l,s=1}^{d+1} \Lambda(p,l)(t) J_k(l,s)\Lambda(s,q)(t) &= \sum_{l=1}^{d+1} \Lambda(p,l)(t) J_k(l,l)\Lambda(l,q)(t) \\ &= \sum_{l=1}^{d+1} (2\delta_{kl} - 1)\Lambda(p,l)(t)\Lambda(l,q)(t). \qquad (4.21)\end{aligned}$$

In particular, for $1 \leq q \neq k \leq d$, the (k,q)-th block of $\Lambda(t)J_k\Lambda(t)$ is

$$\sum_{l=1}^{d+1} (2\delta_{kl} - 1)\Lambda(k,l)(t)\Lambda(l,q)(t),$$

while the (k,q)-th block of $\Lambda(t)J_q\Lambda(t)$ is

$$\sum_{l=1}^{d+1} (2\delta_{ql} - 1)\Lambda(k,l)(t)\Lambda(l,q)(t).$$

Using Proposition 4.3.6, we may conclude that the (k, q)-th block of X_k is zero if $1 \leq k, q \leq d$.

If we take $q = d + 1$ and $p = k$ in (4.21), then the $(k, d + 1)$-th block of X_k is

$$X_k(k, d+1) = -\sum_{1 \leq l \neq k \leq d} \Lambda(k, l)(t)\Lambda(l, d+1)(t).$$

We observe that $X_k(k, d + 1)$ is the k-th block of the product

$$-\begin{bmatrix} 0 & \Lambda(1,2)(t) & \dots & \Lambda(1,d-1)(t) & \Lambda(1,d)(t) \\ \Lambda(2,1)(t) & 0 & \dots & \Lambda(2,d-1)(t) & \Lambda(2,d)(t) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \Lambda(d-1,1)(t) & \Lambda(d-1,2)(t) & \dots & 0 & \Lambda(d-1,d)(t) \\ \Lambda(d,1)(t) & \Lambda(d,2)(t) & \dots & \Lambda(d,d-1)(t) & 0 \end{bmatrix} \begin{bmatrix} \Lambda(1,d+1)(t) \\ \Lambda(2,d+1)(t) \\ \vdots \\ \Lambda(d-1,d+1)(t) \\ \Lambda(d,d+1)(t) \end{bmatrix},$$

which can be written in a compact form $-\Lambda_0(t)\Lambda_1(t)$. □

By assembling Lemmas 4.3.7 and 4.3.8, we can easily derive the geodesic equation on a flag manifold, from which we can even obtain an explicit formula for the geodesic curve. In fact, we have the following:

Proposition 4.3.9 (geodesics). *Let $c(t)$ be a curve on $\operatorname{Flag}(n_1, \dots, n_d; n)$. We parametrize $c(t)$ as*

$$c(t) = V(t)(J_1, \dots, J_d)V(t)^\top,$$

where $V(t)$ is a curve in $O(n)$. We have the following:

1. *There exists a unique $\Lambda(t) \in \mathfrak{so}(n)$ such that $\dot{V}(t) = V(t)\Lambda(t)$.*

2. *If we partition $\Lambda(t)$ as $\Lambda(t) = (\Lambda(p, q)(t))_{p,q=1}^{d+1,d+1} \in \mathfrak{so}(n)$ with respect to $n = m_1 + \cdots + m_{d+1}$, then $\Lambda(p, p)(t) \equiv 0, p = 1, \dots, d + 1$.*

3. $c(t)$ is a geodesic curve if and only if

$$\dot{\Lambda}_0(t) = 0, \quad \dot{\Lambda}_1(t) = \Lambda_0(t)\Lambda_1(t). \tag{4.22}$$

where $\Lambda_0(t) \coloneqq (\Lambda(p,q)(t))_{p,q=1}^{d,d}$ and $\Lambda_1(t) \coloneqq (\Lambda(d+1,q)(t))_{q=1}^{d}$.

4. The solution to (4.22) is

$$\Lambda_0(t) = \Lambda_0(0), \quad \Lambda_1(t) = \exp(t\Lambda_0(0))\Lambda_1(0).$$

Hence a geodesic curve $c(t)$ is

$$c(t) = V(t)(J_1, \ldots, J_d)V^\top(t),$$

where $V(t)$ is a curve in $O(n)$ written as

$$V(t) = V(0)\exp\left(t\begin{bmatrix} 2X_0 & X_1 \\ -X_1^\top & 0 \end{bmatrix}\right)\begin{bmatrix} \exp(-tX_0) & 0 \\ 0 & I_{m_{d+1}} \end{bmatrix} \tag{4.23}$$

for some $X_0 \in \mathfrak{so}(n - m_{d+1})$ satisfying $X_0(k,k) = 0, k = 1, \ldots, d$ and $X_1 \in \mathbb{R}^{(n-m_{d+1})\times m_{d+1}}$.

Proof. (1)–(3) and the first half of (4) are obvious from our earlier discussions, hence it is only left to prove the second part of (4). To that end, we notice that $V(t)$ must satisfy the equation

$$\dot{V}(t) = V(t)\begin{bmatrix} X_0 & \exp(tX_0)X_1 \\ -X_1^\top \exp(-tX_0) & 0 \end{bmatrix} \tag{4.24}$$

and

$$\begin{bmatrix} X_0 & \exp(tX_0)X_1 \\ -X_1^\top \exp(-tX_0) & 0 \end{bmatrix}$$
$$= \begin{bmatrix} \exp(tX_0) & 0 \\ 0 & I_{m_{d+1}} \end{bmatrix} \begin{bmatrix} X_0 & X_1 \\ -X_1^\top & 0 \end{bmatrix} \begin{bmatrix} \exp(-tX_0) & 0 \\ 0 & I_{m_{d+1}} \end{bmatrix}.$$

If we set $W(t) = V(t) \begin{bmatrix} \exp(tX_0) & 0 \\ 0 & I_{m_{d+1}} \end{bmatrix}$, then (4.24) becomes

$$\dot{W}(t) = W(t) \begin{bmatrix} 2X_0 & X_1 \\ -X_1^\top & 0 \end{bmatrix}$$

whose solution is simply

$$W(t) = W(0) \exp\left(t \begin{bmatrix} 2X_0 & X_1 \\ -X_1^\top & 0 \end{bmatrix} \right) = V(0) \exp\left(t \begin{bmatrix} 2X_0 & X_1 \\ -X_1^\top & 0 \end{bmatrix} \right).$$

Hence we obtain that

$$V(t) = V(0) \exp\left(t \begin{bmatrix} 2X_0 & X_1 \\ -X_1^\top & 0 \end{bmatrix} \right) \begin{bmatrix} \exp(-tX_0) & 0 \\ 0 & I_{m_{d+1}} \end{bmatrix}.$$

□

We remark that if $d = 1$, then $X_0 = 0$ in (4.23) and a geodesic curve on $\mathrm{Gr}(n_1, n)$ passing through $VJ_1V^\top$ is

$$c(t) = V \exp\left(t \begin{bmatrix} 0 & X_1 \\ -X_1^\top & 0 \end{bmatrix}\right) I_{n_1, n-n_1} \left(-t \begin{bmatrix} 0 & X_1 \\ -X_1^\top & 0 \end{bmatrix}\right) V^\top,$$

which coincides with the formula derived in [131].

We again work out the case $d = 2$ to illustrate the proof of Proposition 4.3.9. To this end, we write

$$\Lambda(t) = \begin{bmatrix} 0 & A(t) & B(t) \\ -A^\top(t) & 0 & C(t) \\ -B^\top(t) & -C^\top(t) & 0 \end{bmatrix}, \quad A(t) \in \mathbb{R}^{m_1 \times m_2}, B(t) \in \mathbb{R}^{m_1 \times m_3}, C(t) \in \mathbb{R}^{m_2 \times m_3}$$

and suppose that the curve

$$c(t) = V(t)(J_1, J_2)V(t)^\top, \quad \dot{V}(t) = V(t)\Lambda(t), \quad V(t) \in O(n)$$

is a curve passing through (J_1, J_2) with the direction

$$\begin{aligned} &(\Lambda(0)J_1 - J_1\Lambda(0), \Lambda(0)J_2 - J_2\Lambda(0)) \\ &= -2\left(\begin{bmatrix} 0 & A(0) & B(0) \\ A(0)^\top & 0 & 0 \\ B^\top(0) & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -A(0) & 0 \\ -A^\top(0) & 0 & C(0) \\ 0 & C^\top(0) & 0 \end{bmatrix}\right). \end{aligned}$$

We write $\ddot{c}(t) = V(t)\left(\Delta_1(t), \Delta_2(t)\right)V(t)^\top$ where

$$\Delta_k(t) = (\dot{\Lambda}(t)J_k - J_k\dot{\Lambda(t)}) + (\Lambda^2(t)J_k + J_k\Lambda^2(t)) + (-2\Lambda(t)J_k\Lambda(t)).$$

It is sufficient to compute the projection of $\Lambda(t)J_k\Lambda(t)$ onto $T_{c(t)}\operatorname{Flag}(n_1, n_2; n)$, which is

$$\begin{aligned} &\Lambda(t)J_1\Lambda(t) \\ &= \left(\begin{bmatrix} * & B(t)C(t)^\top & -A(t)C(t) \\ C(t)B(t)^\top & * & * \\ -C(t)^\top A(t)^\top & * & * \end{bmatrix}, \begin{bmatrix} * & B(t)C(t)^\top & * \\ C(t)B(t)^\top & * & A(t)^\top B(t) \\ * & B(t)^\top A(t) & * \end{bmatrix} \right), \end{aligned}$$

where $*$ denotes those irrelevant blocks. Eventually, we obtain

$$\begin{aligned} \operatorname{proj}^{\mathbb{T}}_{c(t)}(\dot{c}(t)) = -2(&\begin{bmatrix} 0 & \dot{A}(t) & \dot{B}(t) - A(t)C(t) \\ \dot{A}(t)^\top & 0 & 0 \\ \dot{B}(t)^\top - C(t)^\top A(t)^\top & 0 & 0 \end{bmatrix}, \\ &\begin{bmatrix} 0 & -\dot{A}(t) & 0 \\ -\dot{A}^\top(t) & 0 & \dot{C}(t) + A(t)^\top B(t) \\ 0 & \dot{C}(t)^\top + B(t)^\top A(t) & 0 \end{bmatrix}). \end{aligned}$$

Hence the geodesic equation for $\operatorname{Flag}(n_1, n_2; n)$ is

$$\dot{A}(t) = 0, \quad \dot{B}(t) - A(t)C(t) = 0, \quad \dot{C}(t) + A(t)^\top B(t) = 0,$$

which can be rewritten in a more compact form:

$$\dot{A} = 0, \quad \begin{bmatrix} \dot{B}(t) \\ \dot{C}(t) \end{bmatrix} = \begin{bmatrix} 0 & A(t) \\ -A^\top(t) & 0 \end{bmatrix} \begin{bmatrix} B(t) \\ C(t) \end{bmatrix}. \tag{4.25}$$

The solution to (4.25) is

$$A(t) = A(0), \quad \begin{bmatrix} B(t) \\ C(t) \end{bmatrix} = \exp\left(t \begin{bmatrix} 0 & A(0) \\ -A^\top(0) & 0 \end{bmatrix} \right) \begin{bmatrix} B(0) \\ C(0) \end{bmatrix}.$$

4.4 Sub-Riemannian geometry of flag manifolds with modified embeddings

In this section, we discuss the embedded geometry of flag manifolds with respect to a modified version of the embedding (4.3). Namely, we define

$$\begin{aligned}\tilde{\iota}: \operatorname{Flag}(n_1,\dots,n_d;n) &\hookrightarrow \operatorname{Gr}(n_1,n)\times\operatorname{Gr}(n_2-n_1,n)\times\cdots\\ &\qquad\times\operatorname{Gr}(n_d-n_{d-1},n)\times\operatorname{Gr}(n-n_d,n)\\ (\{\mathbb{V}_k\}_{k=1}^d) &\mapsto (\mathbb{W}_1,\mathbb{W}_2,\dots,\mathbb{W}_d,\mathbb{W}_{d+1}),\end{aligned}\tag{4.26}$$

Here $\mathbb{W}_k$ is the orthogonal complement of $\mathbb{V}_{k-1}$ in $\mathbb{V}_k$ for $2\le k\le d$, $\mathbb{W}_1=\mathbb{R}_1^n$ and $\mathbb{W}_{d+1}$ is the orthogonal complement of $\mathbb{V}_d$ in $\mathbb{R}^n$. We observe that

$$\tilde{\iota}(\{\mathbb{V}_k\}_{k=1}^d)=(\iota(\{\mathbb{V}_k\}_{k=1}^d),\mathbb{W}_{d+1}).$$

In other words, $\tilde{\iota}$ is simply an extension of ι by tautologically adding the orthogonal complement of $\mathbb{V}_d$. Since ι is already an embedding, we may easily conclude that $\tilde{\iota}$ is also an embedding. Adopting the convention (4.4), $\tilde{\iota}$ embeds $\operatorname{Flag}(n_1,\dots,n_d;n)$ into $\prod_{j=1}^{d+1}\operatorname{Gr}(m_j,n)$. Moreover, by Proposition 4.3.1 we have the following:

Proposition 4.4.1 (embedding)**.** *The image of the embedding*

$$\tilde{\varepsilon}: \operatorname{Flag}(n_1,\dots,n_d;n)\overset{\tilde{\iota}}{\hookrightarrow}\prod_{j=1}^{d+1}\operatorname{Gr}(m_j,n)\overset{\tilde{\tau}}{\hookrightarrow}O(n)^{d+1}\tag{4.27}$$

is given by

$$\tilde{\varepsilon}\left(\mathrm{Flag}(n_1,\dots,n_d;n)\right) = \{(Q_1,\dots,Q_{d+1}) \in \prod_{j=1}^{d+1} O(n) : \mathrm{tr}(Q_j) = 2m_j - n, Q_j^\top = Q_j$$

$$(I_n + Q_j)(I_n + Q_{j+1}) = 0, j = 1,\dots,d+1\}. \quad (4.28)$$

In particular, we also have

$$\tilde{\varepsilon}\left(\mathrm{Flag}(n_1,\dots,n_d;n)\right) = \left\{V(J_1,\dots,J_{d+1})V^\top : V \in O(n)\right\}, \quad (4.29)$$

where $J_k = \mathrm{diag}(-I_{m_1},\cdots,-I_{m_{k-1}},I_{m_k},-I_{m_{k+1}},\cdots,-I_{m_{d+1}})$ *is obtained by permuting diagonal blocks of* $I_{m_k,n-m_k}, k = 1,\dots,d+1$ *and*

$$V(J_1,\dots,J_{d+1})V^\top \coloneqq \left(VJ_1V^\top,\dots,VJ_{d+1}V^\top\right).$$

Similarly to Proposition 4.3.2 and Corollary 4.3.3, we also have:

Proposition 4.4.2. *Given a point* $\tilde{\mathfrak{f}} \coloneqq V(J_1,\dots,J_{d+1})V^\top$*, the tangent space of the flag manifold* $\mathbb{T}_{\tilde{\mathfrak{f}}}\mathrm{Flag}(n_1,\dots,n_d;n)$ *consists of vectors* $V(X_1,\dots,X_{d+1})V^\top \in \prod_{j=1}^{d+1} S_n$ *satisfying*

$$X_k(k,l) = -X_l(k,l), X_k(p,q) = 0, X_k(k,k) = 0, \quad 1 \le k,l,p,q \le d+1 \text{ and } p,q,l \ne k. \quad (4.30)$$

Here $X_k(s,t) \in \mathbb{R}^{m_s \times m_t}$ *is the* (s,t)*-th block of* $X_k \in S_n$ *when we partition* X_k *with respect to* $n = \sum_{j=1}^{d+1} m_j$*. Moreover, a curve* $c(t)$ *passing through* $c(0) = V(J_1,\dots,J_{d+1})V^\top$ *on* $\mathrm{Flag}(n_1,\dots,n_d;n)$ *can be locally parametrized as*

$$c(t) = V\exp(\Lambda(t))(J_1,\dots,J_{d+1})\exp(-\Lambda(t))V^\top.$$

For some differentiable curve $\Lambda : (-\varepsilon, \varepsilon) \to \mathfrak{so}(n)$ *such that* $\Lambda(k,k)(t) \equiv 0$.

If $d = 2$, then a tangent vector of $\operatorname{Flag}(n_1, n_2; n)$ at $\tilde{f}$ can be written as

$$V\left(\begin{bmatrix} 0 & A & B \\ A^\top & 0 & 0 \\ B^\top & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -A & 0 \\ -A^\top & 0 & C \\ 0 & C^\top & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & -B \\ 0 & 0 & -C \\ -B^\top & -C^\top & 0 \end{bmatrix}\right) V^\top,$$

where $A \in \mathbb{R}^{m_1 \times m_2}, B \in \mathbb{R}^{m_1 \times m_3}, C \in \mathbb{R}^{m_2 \times m_3}$.

4.4.1 Induced Riemannian metric, normal space and projections

As a submanifold of $\prod_{j=1}^{d+1} O(n)$, $\operatorname{Flag}(n_1, \dots, n_d; n)$ is equipped with a naturally induced Riemannian metric:

$$\begin{aligned} &\langle V(X_1, \dots, X_{d+1})V^\top, V(Y_1, \dots, Y_{d+1})V^\top \rangle_{\tilde{\mathfrak{f}}} \\ &:= \sum_{j=1}^{d+1} \operatorname{tr}(X_j Y_j) \qquad (4.31) \\ &= 2 \sum_{k=1}^{d+1} \sum_{l<k<m} \operatorname{tr}(X_k(l,k)Y_k(k,l) + X_k(m,k)Y_k(k,m)). \end{aligned}$$

Unlike (4.10) in which some summands are weighted differently, all summands in the new metric (4.31) are evenly weighted. For instance, if we take $d = 2$ then $\langle V(X_1, X_2, X_3)V^\top$, $V(Y_1, Y_2, Y_3)V^\top \rangle_{\tilde{\mathfrak{f}}}$ is simply

$$4\left(\operatorname{tr}(X_1(2,1)Y_1(1,2)) + \operatorname{tr}(X_1(3,1)Y_1(1,3)) + \operatorname{tr}(X_2(3,2)Y_2(2,3))\right). \qquad (4.32)$$

The distinction between (4.31) and (4.10) can be easily observed by comparing (4.32) with (4.12).

We notice that the tangent space of $\prod_{j=1}^{d+1} O(n)$ at $\tilde{\mathfrak{f}} = V(J_1, \dots, J_{d+1})V^\top$ is

$$\mathbb{T}_{\tilde{\mathfrak{f}}} \left(\prod_{j=1}^{d+1} O(n) \right) = \bigoplus_{j=1}^{d+1} \left(V J_j V^\top \mathfrak{so}(n) \right) = \bigoplus_{j=1}^{d+1} \left(V J_j \mathfrak{so}(n) V^\top \right).$$

Proposition 4.4.3. *At a point* $\tilde{\mathfrak{f}} \coloneqq V(J_1, \dots, J_{d+1})V^\top \in \operatorname{Flag}(n_1, \dots, n_d; \mathbb{R}^n)$*, the normal space* $\mathbb{N}_{\tilde{\mathfrak{f}}} \operatorname{Flag}(n_1, \dots, n_d; n)$ *consists of vectors*

$$V(J_1 Z_1, \dots, J_{d+1} Z_{d+1})V^\top$$

where $Z_1, \dots, Z_d \in \mathfrak{so}(n)$ *satisfy the relation*

$$Z_k(k,l) - Z_l(k,l) = 0, \quad \text{for all } 1 \le k \ne l \le d+1.$$

In particular, we have a decomposition

$$\mathbb{N}_{\tilde{\mathfrak{f}}} \operatorname{Flag}(n_1, \dots, n_d; n) = N_{\tilde{\mathfrak{f}}} \left(\prod_{k=1}^{d+1} \operatorname{Gr}(m_k, n) \right) \bigoplus N_{\tilde{\mathfrak{f}}}^0, \tag{4.33}$$

where $\mathbb{N}_{\tilde{\mathfrak{f}}} \left(\prod_{k=1}^{d+1} \operatorname{Gr}(m_k, n) \right) = \bigoplus_{k=1}^{d+1} \mathbb{N}_{V J_{m_k, n-m_k} V^\top} \operatorname{Gr}(m_k, n)$ *and*

$$\begin{aligned} \mathbb{N}_{\tilde{\mathfrak{f}}}^0 = \{ V(J_{m_1, n-m_1} Z_1, \dots, J_{m_{d+1}, n-m_{d+1}} Z_{d+1}) V^\top : Z_k \in \mathfrak{so}(n), Z_k(k,l) - Z_l(k,l) = 0, \\ Z_k(k,k) = 0, Z_k(p,q) = 0, 1 \le k, l, p, q \le d+1, p, q \ne k \}. \end{aligned} \tag{4.34}$$

Proposition 4.4.4. *] Projections from* $\mathbb{T}_{\tilde{\mathfrak{f}}} \left(\prod_{k=1}^{d+1} O(n) \right)$ *onto* $\mathbb{T}_{\tilde{\mathfrak{f}}} \operatorname{Flag}(n_1, \dots, n_d; n)$ *and*

$\mathbb{N}_{\tilde{\mathfrak{f}}} \operatorname{Flag}(n_1, \ldots, n_d; n)$ *are respectively given by*

$$\begin{aligned} \operatorname{proj}_{\tilde{\mathfrak{f}}}^{\mathbb{T}} : \mathbb{T}_{\tilde{\mathfrak{f}}} \left(\prod_{k=1}^{d+1} O(n) \right) &\to \mathbb{T}_{\tilde{\mathfrak{f}}} \operatorname{Flag}(n_1, \ldots, n_d; n) \\ V(J_1\Lambda_1, \ldots, J_{d+1}\Lambda_{d+1})V^\top &\mapsto V(X_1, \ldots, X_{d+1})V^\top, \end{aligned} \tag{4.35}$$

and

$$\begin{aligned} \operatorname{proj}_{\tilde{\mathfrak{f}}}^{\mathbb{N}} : \mathbb{T}_{\tilde{\mathfrak{f}}} \left(\prod_{k=1}^{d+1} O(n) \right) &\to \mathbb{N}_{\tilde{\mathfrak{f}}} \operatorname{Flag}(n_1, \ldots, n_d; n) \\ V(J_1\Lambda_1, \ldots, J_{d+1}\Lambda_{d+1})V^\top &\mapsto V(Z_1, \ldots, Z_{d+1})V^\top, \end{aligned} \tag{4.36}$$

where for each $k = 1, \ldots, d$, $X_k \in S_n$ *(resp.* $Z_k \in \mathbb{R}^{n \times n}$*) is partitioned as* $(X_k(p, q))_{p,q=1}^{d+1}$ *(resp.* $(Z_k(p, q))_{p,q=1}^{d+1}$*) with respect to* $n = m_1 + \cdots + m_{d+1}$ *and*

$$X_k(p, q) = \begin{cases} \frac{1}{2}(\Lambda_k(k, q) - \Lambda_q(k, q)), \ \textit{if } p = k \neq q \\ -\frac{1}{2}(\Lambda_k(p, k) - \Lambda_p(p, k)), \ \textit{if } q = k \neq p \\ 0, \ \textit{otherwise.} \end{cases}$$

$$Z_k(p, q) = \begin{cases} \frac{1}{2}(\Lambda_k(k, q) + \Lambda_q(k, q)), \ \textit{if } p = k \neq q \\ -\frac{1}{2}(\Lambda_k(p, k) + \Lambda_p(p, k)), \ \textit{if } q = k \neq p \\ \Lambda_k(p, q), \ \textit{otherwise.} \end{cases}$$

As an illustrative example, we take a tangent vector ξ of $O(n) \times O(n) \times O(n)$ at some point $\mathfrak{f} = V(J_1, J_2, J_3)V^\top$, which can be written as

$$\xi \coloneqq V\left(\begin{bmatrix} A & B & C \\ B^\top & D & E \\ C^\top & -E^\top & F \end{bmatrix}, \begin{bmatrix} X & Y & Z \\ Y^\top & W & S \\ -Z^\top & S^\top & T \end{bmatrix}, \begin{bmatrix} L & M & N \\ -M^\top & P & Q \\ N^\top & Q^\top & R \end{bmatrix}\right)V^\top,$$

where $A, X, L \in \mathfrak{so}(m_1)$, $D, W, P \in \mathfrak{so}(m_2)$, $F, T, R \in \mathfrak{so}(m_3)$, $B, Y, M \in \mathbb{R}^{m_1 \times m_2}$, $C, Z, N \in \mathbb{R}^{m_1 \times m_3}$, $E, S, Q \in \mathbb{R}^{m_2 \times m_3}$. The projection of ξ to $T_{\tilde{\mathfrak{f}}} \operatorname{Flag}(n_1, n_2; n)$ is

$$\operatorname{proj}^{\mathbb{T}}_{\tilde{\mathfrak{f}}}(\xi) = V(\begin{bmatrix} 0 & \frac{B-Y}{2} & \frac{C-N}{2} \\ \frac{B^\top - Y^\top}{2} & 0 & 0 \\ \frac{C^\top - N^\top}{2} & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -\frac{B-Y}{2} & 0 \\ -\frac{B^\top - Y^\top}{2} & 0 & \frac{S-Q}{2} \\ 0 & \frac{S^\top - Q^\top}{2} & 0 \end{bmatrix},$$
$$\begin{bmatrix} 0 & 0 & -\frac{C-N}{2} \\ 0 & 0 & -\frac{S-Q}{2} \\ -\frac{C^\top - N^\top}{2} & -\frac{S^\top - Q^\top}{2} & 0 \end{bmatrix})V^\top$$

and its projection to $N_{\mathfrak{f}} \operatorname{Flag}(n_1, n_2; n)$ is

$$\operatorname{proj}^{\mathbb{N}}_{\mathfrak{f}}(\xi) = V(\begin{bmatrix} A & \frac{B+Y}{2} & \frac{C+N}{2} \\ \frac{B^\top + Y^\top}{2} & D & E \\ \frac{C^\top + N^\top}{2} & -E^\top & F \end{bmatrix}, \begin{bmatrix} X & \frac{B+Y}{2} & Z \\ \frac{B^\top + Y^\top}{2} & W & \frac{S+Q}{2} \\ -Z^\top & \frac{S^\top + Q^\top}{2} & T \end{bmatrix},$$
$$\begin{bmatrix} L & M & \frac{C+N}{2} \\ -M^\top & P & \frac{S+Q}{2} \\ \frac{C^\top + N^\top}{2} & \frac{S^\top + Q^\top}{2} & T \end{bmatrix})V^\top$$

4.4.2 Geodesics

Assume that $c(t)$ is a curve in $\operatorname{Flag}(n_1, \ldots, n_d; n)$, then according to Proposition 4.4.2 we may parametrize $c(t)$ as

$$c(t) = V(t)(J_1, \ldots, J_{d+1})V(t)^\top \tag{4.37}$$

for some differentiable curve $V(t)$ in $O(n)$. Moreover, we have $\dot{V}(t) = V(t)\Lambda(t)$ where $\Lambda(t)$ is a curve in $\mathfrak{so}(n)$ partitioned as $\Lambda(t) = (\Lambda(p,q))_{p,q=1}^{d+1}$ with respect to $m_1 + \cdots + m_{d+1} = n$ and and $\Lambda(k,k)(t) \equiv 0, k = 1, \ldots, d+1$. This implies that we have

$$\ddot{c}(t) = T_1(t) + T_2(t) - 2T_3(t),$$

where $T_j(t)$'s are respectively given by

$$T_1(t) = V(t)(\dot{\Lambda}(t)J_1 - J_1\dot{\Lambda(t)}, \ldots, \dot{\Lambda}(t)J_{d+1} - J_{d+1}\dot{\Lambda(t)})V^\top(t), \tag{4.38}$$

$$T_2(t) = V(t)(\Lambda^2(t)J_1 + J_1\Lambda^2(t), \ldots, \Lambda^2(t)J_{d+1} + J_{d+1}\Lambda^2(t))V^\top(t), \tag{4.39}$$

$$T_3(t) = V(t)(\Lambda(t)J_1\Lambda(t), \ldots, \Lambda(t)J_{d+1}\Lambda(t))V^\top(t). \tag{4.40}$$

By similar calculations in proofs of Lemmas 4.3.7 and 4.3.8, we may easily obtain the following characterizations of $\operatorname{proj}_{c(t)}^{\mathbb{T}}(T_j(t)), j = 1, 2, 3$.

Lemma 4.4.5. *Let $c(t), \Lambda(t), T_1(t), T_2(t), T_3(t)$ be as above. We have*

1. $T_1(t) \in \mathbb{T}_{c(t)} \operatorname{Flag}(n_1, \ldots, n_d; n)$.

2. $\operatorname{proj}_{c(t)}^{\mathbb{T}}(T_2(t)) = 0$.

3. $\operatorname{proj}_{c(t)}^{\mathbb{T}}(T_3(t)) = 0$.

Proposition 4.4.6. *Let $c(t)$ be a curve on $\operatorname{Flag}(n_1, \ldots, n_d; n)$ parametrized as*

$$c(t) = V(t)(J_1, \ldots, J_{d+1})V(t)^\top$$

for some differentiable curve $V(t)$ in $O(n)$. Let $\Lambda(t)$ be the curve in $\mathfrak{so}(n)$ such that $\dot{V}(t) = V(t)\Lambda(t)$, where $\Lambda(t)$ is a curve in $\mathfrak{so}(n)$ partitioned as $\Lambda(t) = (\Lambda(p,q))_{p,q=1}^{d+1}$ with respect to $m_1 + \cdots + m_{d+1} = n$ and and $\Lambda(k,k)(t) \equiv 0, k = 1, \dots, d+1$. Then $c(t)$ is a geodesic curve if and only if $V(t) = V(0)\exp(t\Lambda(0))$.

Proof. Since $c(t)$ is a geodesic if and only if $\mathrm{proj}_{c(t)}(\ddot{c}(t)) \equiv 0$, Lemma 4.4.5 implies that $c(t)$ is a geodesic curve if and only if

$$\dot{\Lambda}(t)J_k - J_k\dot{\Lambda(t)} = 0, \quad k = 1, \dots, d+1.$$

By (4.8), we may conclude that $c(t)$ is a geodesic if and only if $\dot{\Lambda}(t) \equiv 0$, i.e., $\Lambda(t) = \Lambda(0)$. This implies that $V(t)$ is determined by the equation $\dot{V}(t) = V(t)\Lambda(0)$, from which we may conclude that $V(t) = V(0)\exp(t\Lambda(0))$. □

4.5 The comparison of Riemannian metrics on flag manifolds

The goal of this section is to discuss relations among three Riemannian metrics on a flag manifold $\mathrm{Flag}(n_1, \dots, n_d; n)$. We recall that the two metrics discussed in this paper are respectively induced by the embedding $\varepsilon : \mathrm{Flag}(n_1, \dots, n_d; n) \hookrightarrow \prod_{k=1}^{d} O(n)$ given in (4.5) and $\tilde{\varepsilon} : \mathrm{Flag}(n_1, \dots, n_d; n) \hookrightarrow \prod_{k=1}^{d+1} O(n)$ given in (4.27). For notational simplicity, we denote the two induced metrics by g^e and $\tilde{g}^e$, respectively. Yet there is another metric induced from the homogeneous space structure of $\mathrm{Flag}(n_1, \dots, n_d; n)$, which is discussed thoroughly in [204]. We denote this quotient metric by g^q.

Proposition 4.5.1. *The Riemaannian metrics $\tilde{g}^e$ and g^q coincide. Moreover, $\tilde{g}^e$ and g^e coincide with g^q when $d = 1$, in which case* $\mathrm{Flag}(n_1, \dots, n_d; n)$ *is simply the Grassmann manifold* $\mathrm{Gr}(n_1; n)$.

We will see in Proposition 4.5.2 that both g^e and $\tilde{g}^e = g^q$ can be constructed by a uniform method. To begin with, we notice that in general, any smooth map

$$\varphi : \left(\mathbb{R}^{n\times n}\right)^d \to \mathbb{R}^{n\times n}$$

induces an embedding $\kappa_\varphi : \left(\mathbb{R}^{n\times n}\right)^d \to \left(\mathbb{R}^{n\times n}\right)^{d+1}$ defined by

$$\kappa_\varphi(A_1,\dots,A_d) = (A_1,\dots,A_d,\varphi(A_1,\dots,A_d)), \quad A_j \in \mathbb{R}^{n\times n}, j = 1,\dots,d.$$

Hence we have another embedding $\kappa_\varphi \circ \varepsilon$ of $\operatorname{Flag}(n_1,\dots,n_d;n)$ into $O(n)^{d+1} \subseteq \left(\mathbb{R}^{n\times n}\right)^{d+1}$, which induces a metric g^φ on $\operatorname{Flag}(n_1,\dots,n_d;n)$ from the Euclidean metric on $\left(\mathbb{R}^{n\times n}\right)^{d+1}$.

Proposition 4.5.2. *We have the following:*

- *$g^\varphi = g^e$ if and only if φ is a constant map on $\varepsilon(\operatorname{Flag}(n_1,\dots,n_d;n))$. In particular, $g^\varphi = g^e$ if φ is a constant map.*
- *There exists φ such that $g^\varphi = \widetilde{g}^e$.*

Proof. The "if" part of the first statement can be verified by a straightforward calculation. For the "only if" part, we notice that $g^\varphi = g^e$ implies that the differential map $d_{(Q_1,\dots,Q_d)}\varphi$ must be zero on $T_{(Q_1,\dots,Q_d)}\varepsilon(\operatorname{Flag}(n_1,\dots,n_d;n))$ at any $(Q_1,\dots,Q_d)$. Since $\varepsilon(\operatorname{Flag}(n_1,\dots,n_d;n))$ is connected and φ is continuous, we may conclude that φ is a constant map on $\varepsilon(\operatorname{Flag}(n_1,\dots,n_d;n))$.

For the second statement, we notice that $C \coloneqq \varepsilon(\operatorname{Flag}(n_1,\dots,n_d;n))$ is a compact subset of $X \coloneqq \left(\mathbb{R}^{n\times n}\right)^d$ and we can define

$$\psi : C \to O(n) \subseteq \mathbb{R}^{n\times n}, \quad \psi(Q_1,\dots,Q_d) = Q_{d+1},$$

where $(Q_{d+1} + I_n)/2$ is the projection matrix of $\left(\bigoplus_{j=1}^d \operatorname{im}(Q_j + I_n)\right)^\perp$. We denote by p_{ij} the projection map from $\mathbb{R}^{n\times n}$ onto its (i,j)-th entry, $1 \le i,j \le n$. It is clear that

$p_{ij} \circ \psi : C \to \mathbb{R}$ is a smooth function. The compactness of C in X implies that $p_{ij} \circ \psi$ has a smooth extension $\varphi_{ij} : X \to \mathbb{R}$. Indeed, we can first extend the function $p_{ij} \circ \psi$ smoothly to an open neighbourhood of C and then further extend it smoothly to the whole X by a smooth partition of unity. Now we have a smooth map

$$\varphi \coloneqq (\varphi_{ij}) : \left(\mathbb{R}^{n\times n}\right)^d \to \mathbb{R}^{n\times n}$$

which extends ψ and hence we have $g^{\varphi} = \widetilde{g}^e$. □

4.6 A coordinate minimization method for optimization on flag manifolds

Given a strictly increasing sequence $n_1 < \cdots < n_d$, we define

$$m_1 \coloneqq n_1, \quad m_{d+1} \coloneqq n - n_d, \quad m_j \coloneqq n_j - n_{j-1}, \quad j = 2, \ldots, d+1.$$

We recall from (4.26) that a flag $\{\mathbb{V}_k\}_{k=1}^d \in \operatorname{Flag}(n_1, \ldots, n_d; n)$ can be regarded as $\{\mathbb{W}_j\}_{j=1}^{d+1}$ via the modified embedding $\widetilde{\iota} : \operatorname{Flag}(n_1, \ldots, n_d; n) \hookrightarrow \prod_{j=1}^{d+1} \operatorname{Gr}(m_j, n)$, where $\mathbb{W}_j$ is the orthogonal complement of $\mathbb{V}_{j-1}$ in $\mathbb{V}_j, 2 \le j \le d+1$, $\mathbb{W}_1 = \mathbb{V}_1$ and $\mathbb{V}_{d+1} = \mathbb{R}^n$. Therefore, an optimization problem on $\operatorname{Flag}(n_1, \ldots, n_d; n)$ has the following form:

$$\begin{aligned} \min \quad & f(\mathbb{W}_1, \ldots, \mathbb{W}_{d+1}) \\ \text{s.t.} \quad & \mathbb{W}_j \in \operatorname{Gr}(m_j, n), 1 \le j \le d+1 \\ & \mathbb{W}_j \perp \mathbb{W}_l, 1 \le j < l \le d+1 \end{aligned} \tag{4.41}$$

Here f is a function on $\operatorname{Flag}(n_1, \ldots, n_d; n)$. We propose Algorithm 6, an alternating type algorithm to solve the problem (4.41).

Algorithm 6 Coordinate minimization method for optimization on flag manifolds

```
Input  A differentiable function $f$ on $\operatorname{Flag}(n_1, \dots, n_d; n)$
Output  A critical point of $f$
Initialization  Choose an initial point $(\mathbb{W}_1, \dots, \mathbb{W}_{d+1}) \in \prod_{j=1}^{d+1} \operatorname{Gr}(m_j, n)$
while not converge do
    set $(s, t) = (1, 2)$
    for $1 \le s < t \le d + 1$ do
        Solve the following sub-problem for $(\mathbb{X}_s, \mathbb{X}_t) \in \operatorname{Gr}(m_s, n) \times \operatorname{Gr}(m_t, n)$:
            $$\begin{aligned} \min \quad & f(\mathbb{W}_1, \dots, \mathbb{W}_{s-1}, \mathbb{X}_s, \mathbb{W}_{s+1}, \dots, \mathbb{W}_{t-1}, \mathbb{X}_t, \mathbb{W}_{t+1}, \dots \mathbb{W}_{d+1}) \\ \text{s.t.} \quad & \mathbb{X}_s \perp \mathbb{X}_t \\ & \mathbb{X}_s \perp \mathbb{W}_j, 1 \le j \ne s \le d+1 \\ & \mathbb{X}_t \perp \mathbb{W}_j, 1 \le j \ne t \le d+1 \end{aligned} \tag{4.42}$$
        Update $(\mathbb{W}_s, \mathbb{W}_t)$ by the solution $(\overline{\mathbb{X}}_s, \overline{\mathbb{X}}_t)$ to (4.42).
        Update $(s, t)$ by $(s + 1, t)$ if $s + 1 < t$ and by $(s, t + 1)$ otherwise
    end for
end while
```

We remark that the sub-problem (4.42) in Algorithm 6 is an optimization problem on a Grassmann manifold. Indeed, we notice that $\mathbb{W}_j$ in (4.42) is fixed whenever $j \ne s, t$. This implies

$$\mathbb{X}_s \oplus \mathbb{X}_t = \left(\bigoplus_{j \ne s,t} \mathbb{W}_j \right)^{\perp}$$

is a fixed $(m_s + m_t)$-dimensional subspace of $\mathbb{R}^n$. So the submanifold given by fixed $\mathbb{W}_j, j \ne s, t$ and $\mathbb{X}_s \oplus \mathbb{X}_t$ is isomorphic to $\operatorname{Gr}(m_s, m_s + m_t)$. This submanifold is actually a totally-geodesic manifold, which is clear from the geodesic formulas of flag and Grassmann manifolds. Thus the objective function

$$f(\mathbb{W}_1, \dots, \mathbb{W}_{s-1}, \mathbb{X}_s, \mathbb{W}_{s+1}, \dots, \mathbb{W}_{t-1}, \mathbb{X}_t, \mathbb{W}_{t+1}, \dots \mathbb{W}_{d+1})$$

can be recognized as a function on the submanifold $\operatorname{Gr}(m_s, m_s + m_t)$. Furthermore, at a given point, there are $d(d+1)/2$ such submanifolds indexed by $1 \le s < t \le d + 1$. The tangent spaces of those submanifolds are orthogonal to each other and span the whole tangent space.

Algorithm 6 is a generalization of coordinate minimization algorithm in Euclidean space.

4.6.1 Projection to flag manifolds

Given $d+1$ subspaces $\mathbb{U}_1, \dots, \mathbb{U}_{d+1}$ of some ambient space $\mathbb{R}^N$. The separation problem can be mathematically formulated as the following optimization problem on a flag manifold:

$$\begin{aligned} \min \quad & F(\mathbb{W}) := \sum_{j=1}^{d+1} \|\tau_j(\mathbb{U}_j) - \tau_j(\mathbb{W}_j)\|_F^2 \\ \text{s.t.} \quad & \mathbb{W}_j \in \mathrm{Gr}(m_j, n), 1 \le j \le d+1 \\ & \mathbb{W}_j \perp \mathbb{W}_l, 1 \le j < l \le d+1 \end{aligned} \tag{4.43}$$

Here $m_j = \dim \mathbb{U}_j, 1 \le j \le d+1$, $n = \sum_{j=1}^{d+1} m_j$ and τ_j is the embedding of $\mathrm{Gr}(m_j, n)$ into $O(n) \cap S_n$ defined by

$$\tau_j(\mathbb{W}) = V \begin{bmatrix} -I_p & 0 & 0 \\ 0 & I_{m_j} & 0 \\ 0 & 0 & -I_q \end{bmatrix} V^\top$$

where $p = \sum_{l=1}^{j-1} m_l$, $q = \sum_{l=j+1}^{d+1} m_l$ and $V = [v_1, \dots, v_n] = [V_1, \dots, V_{d+1}] \in O(n)$ such that $[v_{p+1}, \dots, v_{q-1}] = V_j, \mathrm{span}\{v_{p+1}, \dots, v_{q-1}\} = \mathbb{W}_j$.

Equivalently, this is the projection to flag manifold:

$$\begin{aligned} \min \quad & F(\{\mathbb{V}_k\}_{k=1}^d) := \|\tilde{\varepsilon}(\{\mathbb{V}_k\}_{k=1}^d) - A\|_F \\ \text{s.t.} \quad & \{\mathbb{V}_k\}_{k=1}^d \in \mathrm{Flag}(n_1, \dots, n_d; n) \end{aligned} \tag{4.44}$$

where $\tilde{\varepsilon}$ is the embedding given in Prop. 4.4.1 and $A \in O(n)^{d+1}$, or more generally, $A \in M(n)^{d+1}$.

Although Problem 4.43 is a natural problem in Riemannian optimization, there is no explicit solution as far as we know. However, the coordinate minimization subproblem of it is essentially the projection to Grassmannian manifold, which has explicit solution. As a result, Algorithm 6 can be used to compute Problem 4.43 efficiently. The projection problem is very important in Riemannian optimization. For example, the extrinsic sample mean problem [22] for flag manifolds under modified embedding can be solved efficiently.

Lemma 4.6.1. *Consider the maximization of linear function $f(Q) = \langle A, Q\rangle$ on the Grassmann manifold,*

$$\begin{aligned} \max \quad & \langle A, Q\rangle \\ s.t. \quad & Q \in \mathrm{Gr}(k, n) \end{aligned}$$

The gradient of $f(Q)$ is given by

$$\nabla f(Q) = \frac{1}{4}(A + A^\top - QAQ - QA^\top Q).$$

Let $(A + A^\top)/2 = U\Lambda U^\top$ be an eigendecomposition of $(A + A^\top)/2$ such that $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$, $\lambda_1 \geq \cdots \geq \lambda_n$. Then $Q^* = UI_{k,n-k}U^\top$ is a maximizer of $f(Q)$. Furthermore,

$$2\|\Lambda\|(f(Q^*) - f(Q)) \geq \|\nabla f(Q)\|^2.$$

Proof. The formula for gradient is given in [131, Proposition 5.1]. The original problem is equivalent to

$$\begin{aligned} \max \quad & \langle \Lambda, Q\rangle \\ \text{s.t.} \quad & Q \in \mathrm{Gr}(k, n) \end{aligned}$$

and we need to prove $Q = I_{k,n-k}$ is a maximizer. Using gradient formula, we can simplify the first order condition $\nabla f(Q^*) = 0$ to

$$Q^*\Lambda = \Lambda Q^*.$$

So Q^*, A can be simultaneously diagonalized, and we can assume Q^* is diagonalized. The original problem is equivalent to

$$\min_{\delta_1+\cdots+\delta_n=2k-n, \delta_i=\pm 1} \lambda_1\delta_1 + \cdots + \lambda_n\delta_n.$$

It is clear that $\delta_1 = \cdots = \delta_k = 1, \delta_{k+1} = \cdots = \delta_n = -1$ is a maximizer. So $Q^* = I_{k,n-k}$ is a maximizer. Now consider the last inequality. The term $\|\nabla f(Q)\|^2, f(Q^*) - f(Q)$ can be simplified

$$\begin{aligned} \|\nabla f(Q)\|^2 &= \frac{1}{4}\langle \Lambda - Q\Lambda Q, \Lambda - Q\Lambda Q\rangle \\ &= \frac{1}{2}\sum_{i=1}^n \lambda_i^2 - \frac{1}{2}\operatorname{tr}(\Lambda Q\Lambda Q) \\ &= \frac{1}{2}\operatorname{tr}(\Lambda Q^*\Lambda Q^*) - \frac{1}{2}\operatorname{tr}(\Lambda Q\Lambda Q), \end{aligned}$$

$$f(Q^*) - f(Q) = \langle \Lambda, Q^*\rangle - \langle \Lambda, Q\rangle.$$

For any $c > 2\|\Lambda\|$, we have

$$\begin{aligned} c(f(Q^*) - f(Q)) - \|\nabla f(Q)\|^2 &= c\langle \Lambda, Q^*\rangle - c\langle \Lambda, Q\rangle - \operatorname{tr}(\Lambda Q^*\Lambda Q^*)/2 + \operatorname{tr}(\Lambda Q\Lambda Q)/2 \\ &= g(Q^*) - g(Q), \end{aligned}$$

where $g(Q) = c\langle \Lambda, Q\rangle - \operatorname{tr}(\Lambda Q\Lambda Q)/2$. Assume Q^{**} is a maximizer of $g(Q)$. The first order condition of $g(Q)$ is

$$Q^{**}\Lambda Q^{**}\Lambda Q^{**} - cQ^{**}\Lambda Q^{**} - \Lambda Q^{**}\Lambda + c\Lambda = 0,$$

which is equivalent to

$$(Q^{**}\Lambda - \Lambda Q^{**})(Q^{**}\Lambda + \Lambda Q^{**} - cI) = 0.$$

By definition $c > 2\|\Lambda\| \geq \|Q^{**}\Lambda + \Lambda Q^{**}\|$, so $Q^{**}\Lambda + \Lambda Q^{**} - cI$ is invertible and $Q^{**}\Lambda = \Lambda Q^{**}$. So Q^{**}, Λ can be simultaneously diagonalized. We can assume Q^{**} is diagonalized. So

$$g(Q^{**}) = \sum_{i=1}^{n}(2\|\Lambda\|\lambda_i\delta_i - \frac{1}{2}\lambda_i^2),$$

where δ_i is the diagonal of Q^{**}. Again, Q^* is a maximizer of $g(Q)$. So we have proved that $g(Q^*) \geq g(Q)$, i.e.,

$$c(f(Q^*) - f(Q)) - \|\nabla f(Q)\|^2 \geq 0.$$

Because c is any number larger than $2\|\Lambda\|$, it also holds for $c = 2\|\Lambda\|$ and the proof is finished. □

Proposition 4.6.2. *If we apply Algorithm 6 to solve the problem* (4.43)*, then for each* $1 \leq s < t \leq d+1$*, the sub-problem has the form*

$$\begin{aligned} \min \quad & \|A_1 - WI_{m_s,m_t}W^\top\|_F^2 + \|A_2 + WI_{m_s,m_t}W^\top\|_F^2 \\ s.t. \quad & W \in O(m_s + m_t) \end{aligned} \tag{4.45}$$

where $A_1, A_2 \in O(m_s + m_t) \cap S_{m_s+m_t}$ *are some fixed matrices. Moreover, the sub-problem has an explicit solution* W_* *which is given by the SVD of* $A_1 - A_2 = W_*\Sigma W_*^\top$.

We denote the change of the value of F *at this step by* $\Delta_{s,t}$*. By previous discussion, the* *full gradient ∇F can be partition into $d(d+1)/2$ block components, such that one of the block $\nabla_{s,t}F$ corresponds to the subproblem. Then*

$$\|\tau_s(\mathbb{U}_s) - \tau_t(\mathbb{U}_t)\||\Delta_{s,t}| \geq \|\nabla_{s,t}F\|^2.$$

Proof. Given $1 \leq s < t \leq d+1$, the sub-step in (4.43) is

$$\begin{aligned}
\min \quad & \|\tau_s(\mathbb{U}_s) - \tau_s(\mathbb{W}_s)\|_F^2 + \|\tau_t(\mathbb{U}_t) - \tau_t(\mathbb{W}_t)\|_F^2 \\
\text{s.t.} \quad & (\mathbb{W}_s, \mathbb{W}_t) \in \mathrm{Gr}(m_s, m_s+m_t) \times \mathrm{Gr}(m_t, m_s+m_t) \\
& \mathbb{W}_s \perp \mathbb{W}_t
\end{aligned}$$

In particular, $\mathbb{W}_s \oplus \mathbb{W}_t = \left(\bigoplus_{j\neq s,t} \mathbb{W}_j\right)^\perp$ is a fixed (m_s+m_t)-dimensional vector space represented by $V_{s,t} := [V_s, V_t]$. We construct the matrix $V^\perp$ whose columns form an orthonormal basis of $\left(\bigoplus_{j\neq s,t} \mathbb{W}_j\right)^\perp$. The choice of $\mathbb{W}_s, \mathbb{W}_t$ can be further specified by an orthogonal matrix $W \in O(m_s+m_t)$ so that $V_{s,t}W = [W_s, W_t]$ where W_s, W_t span $\mathbb{W}_s, \mathbb{W}_t$ respectively. As a result, the images of $\mathbb{W}_s, \mathbb{W}_t$ can be written as

$$\tau_s(\mathbb{W}_s) = V_{s,t}WI_{m_s,m_t}W^\top V_{s,t}^\top + V^\perp(V^\perp)^\top, \quad \tau_t(\mathbb{W}_t) = -V_{s,t}WI_{m_s,m_t}W^\top V_{s,t}^\top + V^\perp(V^\perp)^\top.$$

(4.45) follows easily by taking $A_1 = V_{s,t}^\top \tau_s(\mathbb{U}_s)V_{s,t}$, $A_2 = V_{s,t}^\top \tau_t(\mathbb{U}_t)V_{s,t}$.

Next we observe that the objective function in (4.45) can further be re-written as

$$\begin{aligned}
&\|A_1\|_F^2 + \|A_2\|_F^2 + 2(m_s+m_t) - 2\langle A_1, WI_{m_s,m_t}W^\top\rangle + 2\langle A_2, WI_{m_s,m_t}W^\top\rangle \\
=&\|A_1\|_F^2 + \|A_2\|_F^2 + 2(m_s+m_t) + 2\langle A_2 - A_1, WI_{m_s,m_t}W^\top\rangle.
\end{aligned}$$

Therefore, the problem (4.45) is equivalent to

$$\begin{aligned} \min \quad & \langle A_2 - A_1, W I_{m_s,m_t} W^\top \rangle \\ \text{s.t.} \quad & W \in O(m_s + m_t) \end{aligned} \tag{4.46}$$

By Lemma 4.6.1, we may conclude that a solution to (4.46) is W_*, which can be obtained by the SVD of $A_1 - A_2$. Furthermore, we have

$$\|\nabla_{s,t} F\|^2 \leq \|A_2 - A_1\| |\Delta_{s,t}| \leq \|\tau_s(\mathbb{U}_s) - \tau_t(\mathbb{U}_t)\| |\Delta_{s,t}| \square$$

Theorem 4.6.3. *Consider a randomized version of Algorithm 6 for problem* (4.43)*. At each step, choose* (s_i, t_i) *uniformly from all possible* (s, t)*. Let* $\mathbb{W}_i$ *be the point at step* i*. Then every cluster point of* $\mathbb{W}_i$ *is a stationary point almost surely. (Because flag manifolds are compact, cluster point exists.)*

Proof. If $\|\tau_s(\mathbb{U}_s) - \tau_t(\mathbb{U}_t)\| = 0$ for all s, t, then the function is trivial and there is nothing to prove. Otherwise, there is a set $A \subseteq \{(s,t) \mid 1 \leq s < t \leq d+1\}$ such that $\|\tau_s(\mathbb{U}_s) - \tau_t(\mathbb{U}_t)\| \neq 0$ if and only if $(s,t) \in A$. At each step i, assume $\operatorname{argmax}_{(s,t)\in A} \|\nabla_{s,t} F(\mathbb{W}_i)\|$ is achieved for (s^*, t^*). If $(s_i, t_i) = (s^*, t^*)$, then

$$\begin{aligned} F(\mathbb{W}_i) - F(\mathbb{W}_{i+1}) &\geq \frac{\|\nabla_{s^*,t^*} F(\mathbb{W}_i)\|^2}{\|\tau_{s^*}(\mathbb{U}_{s^*}) - \tau_{t^*}(\mathbb{U}_{t^*})\|} \\ &\geq \frac{\max \|\nabla_{s,t} F(\mathbb{W}_i)\|^2}{\max \|\tau_s(\mathbb{U}_s) - \tau_t(\mathbb{U}_t)\|} \\ &\geq C \|\nabla F(\mathbb{W}_i)\|^2, \end{aligned}$$

where C is a constant independent of $\mathbb{W}_i$. If $(s_i, t_i) \neq (s^*, t^*)$, at least we have $F(\mathbb{W}_i) - F(\mathbb{W}_{i+1}) \geq 0$. So

$$\mathbb{E} F(\mathbb{W}_i) - \mathbb{E} F(\mathbb{W}_{i+1}) \geq \frac{2C}{n(n-1)} \|\nabla F(\mathbb{W}_i)\|^2.$$

Summing from $i = 0$ to ∞, and take expectation, we have

$$\mathbb{E}[F(\mathbb{W}_0) - \lim_{i\to\infty} F(\mathbb{W}_i)] \geq C'\mathbb{E}\sum_{i=0}^{\infty} \|\nabla F(\mathbb{W}_i)\|^2.$$

So with probability 1, $\sum_{i=0}^{\infty} \|\nabla F(\mathbb{W}_i)\|^2$ exists and $\|\nabla F(\mathbb{W}_i)\|$ converges to 0. Any cluster point must be a stationary point. □

The coordiante minimization method works best for this special choice of $f(V)$ because the optimization sub-problem has explicit solution and can be solved sufficiently. For more general problems, it might not be the case.

4.7 Numerical experiments

In this section, we consider the function

$$f(V) = \sum_{k=1}^{d} \operatorname{tr}(V_k^\top A_k V_k),$$

where A_i is randomly generated symmetric matrix, V_i is the submatrix of V with index $1 \leq i \leq n, n_{k-1} < j \leq n_k$, i.e., the basis of $\mathbb{W}_k$. This function is clearly a function on the flag manifolds $\operatorname{Flag}(n_1, \ldots, n_d; n)$.

We choose $\operatorname{Flag}(5, 5; 200)$ and test five methods: (i) gradient descent method under classical embedding metric; (ii) gradient descent method under modified embedding metric; (iii) gradient descent method using the quotient model proposed in Algorithm 1 in [204]; (iv) coordinate minimization method under modified metric (Algorithm 1). We define $\kappa = \|A_1\| + \|A_2\| + \|A_3\|$ and record the convergence of $\|\nabla f(V)\|/\kappa$. Figure 4.1 shows the convergence rate averaged over 10 simulations. We also record the running time to hit $\|\nabla f(V)\|/\kappa \leq 10^{-5}$, averaged over 10 simulations, as shown in Table 4.1.

Method (ii) is equivalent to (iii), and their convergence rate and running time are similar.

(i) Classic Descent	23.915s
(ii) Modified Descent	20.581s
(iii) Quotient Descent	27.216s
(iv) Coordinate Minimization	0.911s

Table 4.1: Running time to hit $\|\nabla f(V)\|/\kappa \leq 10^{-5}$ of different methods.

In Figure 4.1, their convergence trajectories almost coincide. All three descent method has comparable performance, while the coordinate minimization outperforms them significantly.

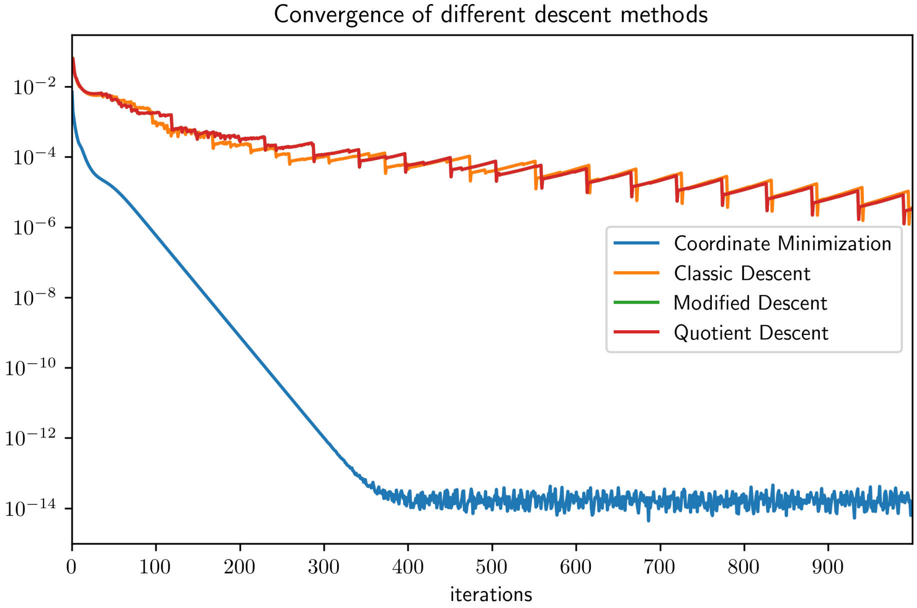


Figure 4.1: Convergence behavior of different descent methods.

CHAPTER 5

ONLINE STATISTICAL INFERENCE FOR STOCHASTIC OPTIMIZATION VIA KIEFER-WOLFOWITZ METHODS

This is a joint work with He Li, Yichen Zhang, and Xi Chen.

5.1 Introduction

Stochastic optimization algorithms, introduced by [168, 123], have been widely used in statistical estimation, especially for large-scale datasets and online learning where the sample arrives sequentially (e.g., web search queries, transactional data). The Robbins-Monro algorithm [168], often known as the stochastic gradient descent, is perhaps the most popular algorithm in stochastic optimization and has found a wide range of applications in statistics and machine learning. Nevertheless, in many modern applications, the gradient information is not available. For example, the objective function may be embedded in a black box and the user can only access the noisy objective value for a given input. In such cases, the Kiefer-Wolfowitz algorithm [123] becomes a natural choice as it is completely free of gradient computation. Despite being equipped with an evident computational advantage to avoid gradient measurements, the Kiefer-Wolfowitz algorithm has been historically out of practice as compared to the Robbins-Monro counterpart. Nonetheless, heralded by the big data era, there has been a restoration of the interest of gradient-free optimization in a wide range of applications in recent years [48, 155]. We briefly highlight a few of them to motivate our paper.

- In some bandit problems, one may only have black-box access to individual objective values but not to their gradients [75, 175]. Other examples include graphical models and variational inference problems, where the objective is defined variationally [195], and the explicit differentiation can be difficult.

- In some scenarios, the computation of gradient information is possible but very expensive. For example, in the online sensor selection problem [114], evaluating the stochastic gradient requires the inverse of matrices, which generates $\mathrm{O}(d^3)$ computation cost per iteration, where d is the number of sensors in the network. In addition, the storage for gradient calculation also requires an $O(d^3)$ memory, which could be practically infeasible.

- In some statistical problems such as quantile regression and its variants [127], the objective function is not differentiable. Extending the gradient definition to nonsmooth functions is generally nontrivial, and techniques of defining sets of local differential characteristics suffer from the incompleteness of chain rule in complex problems [154].

This paper aims to study the asymptotic properties of the Kiefer-Wolfowitz stochastic optimization and conduct online statistical inference. In particular, we consider the problem,

$$\theta^\star = \text{argmin } F(\theta), \quad \text{where } F(\theta) := \mathbb{E}_{\mathcal{P}_\zeta}\left[f(\theta;\zeta)\right] = \int f(\theta;\zeta)\mathrm{d}\mathcal{P}_\zeta, \tag{5.1}$$

where $f(\theta;\zeta)$ is a convex *individual loss function* for a data point ζ, $F(\theta)$ is the *population loss* function, and $\theta^\star$ is the true underlying parameter of a fixed dimension d. Let θ_0 denote any given initial point. Given a sequentially arriving online sample $\{\zeta_n\}$, the [168] algorithm `(RM)`, also known as the stochastic gradient descent (SGD), iteratively updates,

$$\texttt{(RM)} \qquad \theta_n^{\texttt{(RM)}} = \theta_{n-1}^{\texttt{(RM)}} - \eta_n g(\theta_{n-1};\zeta_n), \tag{5.2}$$

where $\{\eta_n\}$ is a positive non-increasing step-size sequence, and $g(\theta;\zeta)$ denotes the stochastic gradient, i.e., $g(\theta;\zeta) = \nabla f(\theta;\zeta)$. In the scenarios that direct gradient measurements are inaccessible to practitioners, the [123] algorithm `(KW)` becomes the natural choice, as

$$\texttt{(KW)} \qquad \theta_n^{\texttt{(KW)}} = \theta_{n-1}^{\texttt{(KW)}} - \eta_n \widehat{g}(\theta_{n-1};\zeta_n), \tag{5.3}$$

where $\widehat{g}(\theta_{n-1};\zeta_n)$ is an estimator of $g(\theta_{n-1};\zeta_n)$. Under the univariate framework ($d = 1$), [123] considered the finite-difference approximation

$$\widehat{g}(\theta_{n-1};\zeta_n) = \frac{f(\theta_{n-1}+h_n;\zeta_n) - f(\theta_{n-1};\zeta_n)}{h_n}, \tag{5.4}$$

where h_n is be a positive deterministic sequence that goes to zero. [23] later extended the algorithm to the multivariate case and proved its almost sure convergence. This pioneering work extended in various directions of statistics and control theory (see, e.g., [67, 69, 91, 170, 42, 161, 182, 38, 183, 92, 57, 146, 28]). In the optimization literature, the Kiefer-Wolfowitz (KW) algorithm is often referred to as the gradient-free stochastic optimization, or zeroth-order SGD [7, 8, 110, 83, 61, 175, 155, 196, among others].

For the (RM) algorithm in (5.2), [172] and [162] characterize the limiting distribution and statistical efficiency of the *averaged iterate* $\overline{\theta}_n^{(\mathtt{RM})} = \frac{1}{n}\sum_{i=1}^n \theta_i^{(\mathtt{RM})}$ by

$$\sqrt{n}\left(\overline{\theta}_n^{(\mathtt{RM})} - \theta^\star\right) \Longrightarrow \mathcal{N}\left(0, H^{-1}SH^{-1}\right), \tag{5.5}$$

where $H = \nabla^2 F(\theta^\star)$ is the Hessian matrix at $\theta = \theta^\star$, and $S = \mathbb{E}[\nabla f(\theta^\star;\zeta)\nabla f(\theta^\star;\zeta)^\top]$ is the covariance matrix of the stochastic gradient at $\theta = \theta^\star$. Under a well-specified model, this asymptotic covariance matrix matches the inverse Fisher information and the averaged (RM) estimator is asymptotically efficient. Based on the limiting distribution result (5.5), there are many recent research efforts devoted to statistical inference for (RM). A brief survey is conducted at the end of the introduction.

For the (KW) scheme, we can similarly construct the averaged Kiefer-Wolfowitz (AKW) estimator

$$\text{(AKW)} \qquad \overline{\theta}_n^{(\mathtt{KW})} = \frac{1}{n}\sum_{i=1}^{n} \theta_i^{(\mathtt{KW})}. \tag{5.6}$$

As compared to well-established asymptotic properties of (RM), study of the asymptotics of (AKW) is limited, particularly with a random sampling direction in multivariate (KW). In this paper, we study the `(KW)` algorithm (5.3) with random search directions $\{v_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{P}_v$, i.e., at each iteration $i = 1, 2, \ldots, n$, a random direction v_i is sampled independently from $\mathcal{P}_v$, and the `(KW)` gradient

$$\widehat{g}_{h_n,v_n}(\theta_{n-1};\zeta_n) = \frac{f(\theta_{n-1} + h_n v_n;\zeta_n) - f(\theta_{n-1};\zeta_n)}{h_n} v_n. \tag{5.7}$$

Compared to the `(RM)` scheme, `(KW)` introduces additional randomness into the stochastic gradient estimator through $\{v_n\}$. Indeed, as one can see from our main result in Theorem 5.3.10, `(AKW)` is no longer statistically efficient and its asymptotic covariance structure depends on the distribution $\mathcal{P}_v$. It opens the room for the investigation on the impact of $\mathcal{P}_v$ (see Section 5.3.1 for details). We further extend the estimator to utilize multiple function-value queries per step and establish an online statistical inference framework. We summarize our main results and contributions as follows,

- First, we quantify the asymptotic covariance structure of `(AKW)` in Theorem 5.3.10. Since the asymptotic distribution depends on the choice of the direction variable v, we provide an introductory analysis on the asymptotic performance for different choices of random directions for constructing `(AKW)` estimators (see Section 5.3.1).

- The efficiency loss of `(AKW)` is due to the information constraint as one evaluates only *two* function values at each iteration. We analyze the `(AKW)` estimators in which multiple function queries can be assessed at each iteration, and show that the asymptotic covariance matrix decreases as the number of function queries $m + 1$ increases (see Section 5.3.3). Moreover, `(AKW)` achieves asymptotic statistical efficiency as $m \to \infty$. We further show that when v is sampled without replacement from $\mathcal{P}_v$ with a discrete uniform distribution of any orthonormal basis, `(AKW)` achieves asymptotic statistical efficiency with $d + 1$ function queries per iteration.

- Based on the asymptotic distribution, we propose two online statistical inference procedures. The first one is using a plug-in estimator of the asymptotic covariance matrix, which separately estimates the Hessian matrix and Gram matrix of the (KW) gradients (with additional function-value queries, see Theorem 5.4.4). The second procedure is to characterize the distribution of intermediate (KW) iterates as a stochastic process and construct an asymptotically pivotal statistic by normalizing the (AKW) estimator, without directly estimating the covariance matrix. This inference procedure is inspired by the "random scaling" method proposed in [136] that considers the online inference for the (RM) scheme. These two procedures have their advantages and disadvantages: the plug-in approach leads to better empirical performance but requires additional function-value queries to estimate the Hessian matrix, while the other one is more efficient in both computation and storage, though its finite-sample performance is inferior in practice when the dimension is large. A practitioner may choose the approach suitable to her computational resources and requirement of the inference accuracy.

Lastly, we provide a brief literature survey on the recent works for statistical inference for the (RM)-type SGD algorithms. [43] developed a batch-means estimator of the limiting covariance matrix $H^{-1}SH^{-1}$ in (5.5), which only uses the stochastic gradient information (i.e., without estimating any Hessian matrices). [210] further extended the batch-means method in [43] to a fully online covariance estimator. [136] extended the results in [162] to a functional central limit theorem and utilize it to propose a novel online inference procedure that allows for efficient implementation. [71] presented a perturbation-based resampling procedure for inference. [189] proposed a tree-structured inference scheme, which splits the SGD into several threads to construct confidence intervals. [139] introduced a moment-adjusted method and its corresponding inference procedure. [191] considered the implicit SGD, and investigate the statistical inference problem under the variant. [62] studied the stochastic optimization problem with constraints and investigate its optimality properties.

[35] proposed a class of generalized regularized dual averaging (RDA) algorithms and make uncertainty quantification possible for online ℓ_1-penalized problems. [177] developed an online estimation procedure for high-dimensional statistical inference. [40] studied statistical inference of online decision-making problems via SGD in a contextual bandit setting.

5.1.1 Notation

We write vectors in boldface letters (e.g., θ and v) and scalers in lightface letters (e.g., η). For any positive integer n, we use $[n]$ as a shorthand for the discrete set $\{1, 2, \cdots, n\}$. Let $\{e_k\}_{k=1}^d$ be the standard basis in $\mathbb{R}^d$ with the k-th coordinate as 1 and the other coordinates as 0. Denote I_d as the identity matrix in $\mathbb{R}^{d\times d}$. Let $\|\cdot\|$ denote the standard Euclidean norm for vectors and the spectral norm for matrices. We use $A_{k\ell}$ and $A_{n,k\ell}$ to denote the (k, ℓ)-th element of matrices $A, A_n \in \mathbb{R}^{d\times d}$, respectively, for all $k, \ell \in [d]$. Furthermore, we denote by $\operatorname{diag}(v)$ a matrix in $\mathbb{R}^{d\times d}$ whose main diagonal is the same as the vector v and off-diagonal elements are zero, for some vector $v \in \mathbb{R}^d$. With a slight abuse of notation, for a matrix $M \in \mathbb{R}^{d\times d}$, we also let $\operatorname{diag}(M)$ denote a $\mathbb{R}^{d\times d}$ diagonal matrix with same diagonal elements as matrix M. We use the standard Loewner order notation $A \succeq 0$ if a matrix A is positive semi-definite. We use $\theta^{(\mathtt{RM})}$ and $\theta^{(\mathtt{KW})}$ to denote the iterates generated by the (RM) scheme and the (KW) scheme, respectively. We use $\widehat{\theta}^{(\mathtt{ERM})}$ for the offline empirical risk minimizer, i.e., $\widehat{\theta}^{(\mathtt{ERM})} = \operatorname{argmin}_\theta \frac{1}{n}\sum_{i=1}^n f(\theta; \zeta_i)$. As we focus on the (KW) scheme in this paper, we sometimes omit the superscript (KW) in the estimator to make room for the other notations. In derivations of the (KW) estimator, we denote the finite difference of $f(\cdot)$ as,

$$\Delta_{h,v} f(\theta; \zeta) = f(\theta + hv; \zeta) - f(\theta; \zeta), \tag{5.8}$$

for some spacing parameter $h \in \mathbb{R}_+$ and search vector $v \in \mathbb{R}^d$. We use $\mathbb{E}_n$ to denote the conditional expectation with respect to the natural filtration, i.e.,

$$\mathbb{E}_n[\theta_{n+1}] := \mathbb{E}[\theta_{n+1}|\mathcal{F}_n], \;\; \mathcal{F}_n := \sigma\{\theta_k, \zeta_k | k \leq n\}.$$

We use the $\mathrm{O}(\cdot)$ notation to hide universal constants independent of the sample size n.

The remainder of the paper is organized as follows. In Section 5.2, we describe the Kiefer-Wolfowitz algorithm with random search directions along with three illustrative examples of the classical regression problems. We also provide a technical lemma to characterize the limiting behavior of the (KW) gradient, which leads to the distributional constraint of the random direction vector. In Section 5.3, we first introduce the technical assumptions before we present the finite-sample rate of convergence of the (KW) estimator. We further provide the asymptotic distribution of the (AKW) estimator, accompanied by discussions on the statistical (in)efficiency. We highlight a comparison of the choices of the direction distributions in Section 5.3.1, and further extend the theoretical analysis to multi-query settings of the (KW) algorithm in Section 5.3.3. Based on the established asymptotic distribution results, we propose two types of online statistical inference procedures in Section 5.4. A functional extension of the distributional analysis of (KW) as a stochastic process is also provided. Numerical experiments in Section 5.5 lend empirical support to our theory. Further discussions are provided in Section 5.6.

5.2 Kiefer-Wolfowitz algorithm

In this section, we introduce the general form of the Kiefer-Wolfowitz (KW) gradient estimator and the corresponding iterative algorithm $\theta_n = \theta_{n-1} - \eta_n \widehat{g}(\theta_{n-1}; \zeta_n)$. In the seminal work by [23], the (KW) gradient estimator $\widehat{g}(\theta_{n-1}; \zeta_n)$ is constructed by approximating the stochastic gradient $g(\theta_{n-1}; \zeta_n)$ using the canonical basis of $\mathbb{R}^d$, $\{e_1, e_2, \ldots, e_d\}$, as search directions. In particular, given any $\theta \in \mathbb{R}^d$ and $\zeta \sim \mathcal{P}_\zeta$, the k-th coordinate of the (KW) gradient estimator

$$\big(\widehat{g}_{h,e}(\theta;\zeta)\big)_k = \frac{f(\theta + he_k;\zeta) - f(\theta;\zeta)}{h}, \qquad \text{for} \quad k = 1, 2, \ldots, d, \tag{5.9}$$

where h is a spacing parameter for approximation. At each iteration, (5.9) queries $d+1$ function values from d fixed directions $\{e_k\}_{k=1}^d$. To reduce the query complexity, a random difference becomes a natural choice. [128] introduced a random version of the (KW) algorithm using a sequence of random unit vectors that are independent and uniformly distributed on the unit sphere or unit cube. [182] also provided a random direction version of the (KW) algorithm, named as the simultaneous perturbation stochastic approximation (SPSA) algorithm and later extended to several variants [38, 183, 98]. These random direction methods can reduce the bias in gradient estimates as compared to their non-random counterparts. In the following, we write the (KW) algorithm with general random search directions, as in (5.7),

$$\begin{aligned} &\theta_n = \theta_{n-1} - \eta_n \widehat{g}_{h_n,v_n}(\theta_{n-1};\zeta_n), \\ &\qquad \text{where } \widehat{g}_{h,v}(\theta;\zeta) := \frac{1}{h}\Delta_{h,v} f(\theta;\zeta) v = \frac{f(\theta + hv;\zeta) - f(\theta;\zeta)}{h} v. \end{aligned} \tag{5.10}$$

Here $\{v_n\}$ is sampled from an underlying distribution $\mathcal{P}_v$ satisfying certain conditions (see Assumption 5.3.5 in Section 5.3). At each iteration n, the algorithm samples a direction vector v_n independently from P_v, and makes two solitary function-value queries, $f(\theta_{n-1};\zeta_n)$ and $f(\theta_{n-1} + h_n v_n;\zeta_n)$. We refer to the (KW) gradient estimator $\widehat{g}_{h_n,v_n}(\theta_{n-1},\zeta_n)$ in (5.10) as a *two-query* finite-difference approximation of the stochastic gradient. If one is allowed to make additional function-value queries, an averaging of the function values from multiple directions generates a *multi-query* stochastic gradient estimator with reduced variance. In particular, at each iteration n, the practitioner makes $m+1$ queries $\{f(\theta_{n-1};\zeta_n), f(\theta_{n-1} + h_n v_n^{(j)};\zeta_n)\}_{1\le j\le m}$ via m random directions $\{v_n^{(j)}\}$ sampled from $\mathcal{P}_v$. If $\mathcal{P}_v$ is a finite distribution, practitioners may choose to sample *with* or *without replacement.* In summary, an $(m+1)$-*query* (KW) algorithm constructs a stochastic gradient estimator

$$\overline{g}_n^{(m)}(\theta_{n-1};\zeta_n) = \frac{1}{m}\sum_{j=1}^{m} \widehat{g}_{h_n,v_n^{(j)}}(\theta_{n-1};\zeta_n) = \frac{1}{mh_n}\sum_{j=1}^{m} \Delta_{h_n,v_n^{(j)}} f(\theta_{n-1};\zeta_n) v_n^{(j)}, \tag{5.11}$$

at each iteration n, and updates $\theta_n = \theta_{n-1} - \eta_n \overline{g}_n^{(m)}(\theta_{n-1};\zeta_n)$. Here we restrict the procedure to sampling from the same distribution $\mathcal{P}_v$ independently across different iterations. We use $\theta_n^{(m)}$ to denote the final (KW) estimator using the above $(m+1)$-query finite-difference approximation.

We now provide some illustrative examples of the two-query (KW) estimator $\widehat{g}_{h_n,v_n}$ in (5.10) used in popular statistical models, and we will refer to these examples throughout the paper. A multi-query extension of the examples can be constructed accordingly.

Example 5.2.1 (Linear Regression). *Consider a linear regression model* $y_i = x_i^\top \theta^\star + \epsilon_i$ *where* $\{\zeta_i = (x_i, y_i),\ i = 1, 2, \ldots, n\}$ *is an i.i.d. sample of* $\zeta = (x, y)$ *and the noise* $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. *We use a quadratic loss function* $f(\theta;\zeta) = (y - x^\top\theta)^2$. *Therefore, the stochastic gradient* $\nabla f(\theta;\zeta) = \left(x^\top\theta - y\right) x$, *and the* (KW) *gradient estimator* $\widehat{g}_{h,v}(\theta;\{x,y\})$ *in* (5.10) *becomes*

$$\widehat{g}_{h,v}(\theta;\{x,y\}) = \frac{1}{h}\left[\big(y - x^\top(\theta + hv)\big)^2 - \big(y - x^\top\theta\big)^2\right] v = vv^\top\big(x^\top\theta - y\big)x + h(x^\top v)^2 v.$$

Example 5.2.2 (Logistic Regression). *Consider a logistic regression model with a binary response* $y_i \in \{-1, 1\}$ *generated by* $\Pr(y_i|x_i) = \left(1 + \exp\left(-y_i x_i^\top \theta^\star\right)\right)^{-1}$. *The individual loss function* $f(\theta;\zeta) = \log\left(1 + \exp(-yx^\top\theta)\right)$. *The stochastic gradient* $\nabla f(\theta;\zeta) = -yx\left(1 + \exp(yx^\top\theta)\right)^{-1}$, *and the* (KW) *gradient estimator* $\widehat{g}_{h,v}(\theta;\{x,y\})$ *in* (5.10) *becomes*

$$\begin{aligned}\widehat{g}_{h,v}(\theta;\{x,y\}) &= \frac{v}{h}\left[\log\left(1 + \exp(-yx^\top(\theta + hv))\right) - \log\left(1 + \exp(-yx^\top\theta)\right)\right] \\ &= \frac{-yvv^\top x}{1 + \exp(yx^\top\theta)} + \frac{y^2(x^\top v)^2 \exp(yx^\top\theta) hv}{2(1 + \exp(yx^\top\theta))^2} + \mathrm{O}(h^2), \quad \text{as } h \to 0_+,\end{aligned}$$

under some regularity conditions on θ and the distribution of x.

Example 5.2.3 (Quantile Regression)**.** *Consider a quantile regression model $y_i = x_i^\top \theta^\star + \epsilon_i$ where $\{\zeta_i = (x_i, y_i),\ i = 1, 2, \ldots, n\}$ is an* i.i.d. *sample of $\zeta = (x, y)$ and the noise satisfies* $\Pr(\epsilon_i \leq 0 | x_i) = \tau$*. The individual loss $f(\theta; \zeta) = \rho_\tau(y - x^\top \theta)$, where $\rho_\tau(z) = z(\tau - 1_{\{z<0\}})$. Although ρ_τ is non-differentiable, the* `(KW)` *gradient estimator $\widehat{g}_{h,v}$ is well-defined and takes the following form,*

$$
\begin{aligned}
\widehat{g}_{h,v}(\theta; \{x, y\}) =& \frac{v}{h} \Big[\rho_\tau\big(y - x^\top(\theta + hv)\big) - \rho_\tau\big(y - x^\top\theta\big) \Big] \\
=& vv^\top x\big(\tau - 1_{\{y - x^\top\theta < 0\}}\big), \quad \textit{for } 0 < h < \left| \frac{y - x^\top \theta}{x^\top v} \right|.
\end{aligned}
$$

We note that for the `(RM)` scheme with differentiable loss functions, the stochastic gradient is an unbiased estimator of the population gradient under very mild assumption, i.e., $\mathbb{E}_\zeta g(\theta; \zeta) = \nabla F(\theta)$. In contrast, the `(KW)` gradient estimator is no longer an unbiased estimator of $\nabla F(\theta)$. In the following lemma, we precisely quantifies the bias incurred by the `(KW)` gradient estimator.

Lemma 5.2.4. *We assume that the population loss function $F(\cdot)$ is twice continuously differentiable and L_f-smooth, i.e., $\nabla^2 F(\theta) \preceq L_f I_d$ for any $\theta \in \mathbb{R}^d$. Given any fixed parameter $\theta \in \mathbb{R}^d$, suppose the random direction vector v is independent from ζ, we have*

$$
\big\| \mathbb{E}\, \widehat{g}_{h,v}(\theta; \zeta) - \nabla F(\theta) \big\| \leq \Big\| \mathbb{E}\big(vv^\top - I_d\big) \nabla F(\theta) \Big\| + \frac{h}{2} L_f \mathbb{E} \|v\|^3,
$$

where the expectation in $\mathbb{E}\, \widehat{g}_{h,v}(\theta; \zeta)$ takes over both the randomness in v and ζ.

Proof. In all the proofs, we will assume, without loss of generality, $F(\cdot)$ achieves its minimum at $\theta^\star = 0$ and $F(0) = 0$. We now introduce some notations as follows,

$$
\begin{aligned}
&\xi_n = \nabla F(\theta_{n-1}) - \mathbb{E}_{n-1}(\frac{1}{h_n}[F(\theta_{n-1} + h_n v_n) - F(\theta_{n-1})]v_n),\\
&\gamma_n = \mathbb{E}_{n-1}\frac{1}{h_n}[F(\theta_{n-1} + h_n v_n) - F(\theta_{n-1})]v_n - \frac{1}{h_n}[F(\theta_{n-1} + h_n v_n) - F(\theta_{n-1})]v_n,\\
&\varepsilon_n = \frac{1}{h_n}[F(\theta_{n-1} + h_n v_n) - F(\theta_{n-1})]v_n - \frac{1}{h_n}[f(\theta_{n-1} + h_n v_n; \zeta_n) - f(\theta_{n-1}; \zeta_n)]v_n.
\end{aligned}
$$

By definition, $\mathbb{E}_\zeta\ \widehat{g}_{h,v}(\theta;\zeta) = \frac{1}{h}\Delta_{h,v}F(\theta)v = \frac{1}{h}\left[F(\theta + hv) - F(\theta)\right]v$. For the first inequality, we have

$$
\begin{aligned}
\left\|\mathbb{E}\ \widehat{g}_{h,v}(\theta;\zeta) - \nabla F(\theta)\right\| &= \left\|\mathbb{E}\ \frac{1}{h}\left[F(\theta + hv) - F(\theta)\right]v - \nabla F(\theta)\right\|\\
&= \left\|\mathbb{E}\ vv^\top \nabla F(\theta) + \frac{1}{2}h\mathbb{E}\ vv^\top \nabla^2 F(\theta_{h,v})v - \nabla F(\theta)\right\|\\
&= \frac{1}{2}h\left\|\mathbb{E}\ vv^\top \nabla^2 F(\theta_{h,v})v\right\|\\
&\le \frac{1}{2}hL_f\mathbb{E}\|v\|^3,
\end{aligned}
\tag{5.12}
$$

where in the third equality we use the Taylor expansion of $F(\theta)$, and $\theta_{h,v}$ comes from the remainder term of the Taylor expansion. □

To reduce the bias in the (KW) gradient, Lemma 5.2.4 indicates that one should choose the random direction v_n that satisfies the distributional constraint $\mathbb{E}[v_n v_n^\top] = I_d$ (see Assumption 5.3.5 in Section 5.3). We will further conduct a comprehensive analysis in Section 5.3.1 on different choices of distributions $\mathcal{P}_v$ satisfying the condition $\mathbb{E}[v_n v_n^\top] = I_d$. Despite the existence of the bias, as the spacing parameter $h_n \to 0$, the bias convergences to zero asymptotically.

5.3 Theoretical results

We first introduce some regularity assumptions on the population loss $F(\theta)$ and the individual loss $f(\theta;\zeta)$.

Assumption 5.3.1. *The population loss function $F(\theta)$ is twice continuously differentiable. Moreover, there exists $L_f > \lambda > 0$, such that, $\lambda I_d \preceq \nabla^2 F(\theta) \preceq L_f I_d$ for any $\theta \in \mathbb{R}^d$.*

Assumption 5.3.2. *The population loss function $F(\theta)$ is twice continuously differentiable, convex and L_f-smooth. In addition, there exists $\delta_1 > 0$ such that for all θ in the δ_1-ball centered at $\theta^\star$, the Hessian matrix $\nabla^2 F(\theta)$ is positive-definite.*

Assumption 5.3.3. *Assume $\mathbb{E}\left[\nabla f(\theta;\zeta_n)\right] = \nabla F(\theta)$ for any $\theta \in \mathbb{R}^d$. Moreover, for some $0 < \delta \leq 2$, there exists $M > 0$ such that*

$$\mathbb{E}\|\nabla f(\theta;\zeta_n) - \nabla F(\theta)\|^{2+\delta} \leq M\big(\|\theta - \theta^\star\|^{2+\delta} + 1\big).$$

Assumption 5.3.4. *There are constants $L_h, L_p > 0$ such that for any $\theta, y \in \mathbb{R}^d$,*

$$\mathbb{E}\left\|\nabla^2 f(\theta;\zeta_n) - \nabla^2 f(y;\zeta_n)\right\|^2 \leq L_h\|\theta - y\|^2, \qquad \mathbb{E}\left\|[\nabla^2 f(\theta^\star;\zeta_n)]^2 - H^2\right\| \leq L_p,$$

where H is the Hessian matrix of the population loss function $F(\cdot)$, i.e., $H = \nabla^2 F(\theta^\star)$.

Assumption 5.3.5. *We adopt i.i.d. random direction vectors $\{v_n\}$ from some common distribution $v \sim \mathcal{P}_v$ such that $\mathbb{E}[vv^\top] = I_d$. Moreover, assume that the $(6+3\delta)$-th moment of v is bounded.*

We discuss the above assumptions and compare them with the standard conditions in the literature of (RM)-type SGD inference. Assumption 5.3.1 requires the population loss function $F(\cdot)$ to be λ-strongly convex and L_f-smooth, which is a convenient assumption widely used in the existing literature of statistical inference on stochastic optimization [162, 43]. It is possible to replace this assumption with Assumption 5.3.2 that only assumes local strong convexity in the neighborhood of the true parameter $\theta^\star$ [189, 62]. This weaker condition is satisfied in the setting of logistic regression (Example 5.2.2). Assumption 5.3.3 introduces the unbiasedness condition on the stochastic gradient $\nabla f(\theta;\zeta)$ when the individual loss function $f(\theta;\zeta)$ is smooth. The $(2+\delta)$-th moment condition is the classical Lyapunov condition used in the derivation of asymptotic normality. Relaxation to this assumption can be made to handle nonsmooth loss functions $f(\theta;\zeta)$, such as the quantile regression as described in Example 5.2.3. The statements in Assumption 5.3.4 introduce the Lipschitz continuity condition and the concentration condition on the Hessian matrix. Assumption 5.3.5 guarantees that the (KW) gradient $\widehat{g}_{h,v}(\theta;\zeta)$ is an asymptotically unbiased estimator of $\nabla F(\theta)$ when the spacing parameter h_n decreases to 0, as suggested by Lemma 5.2.4. The moment condition of v in Assumption 5.3.5 is imposed for technical simplicity and could be possibly weakened. We provide several examples of $\mathcal{P}_v$ in Section 5.3.1.

Before we derive the asymptotic distribution for (AKW), we first provide a finite sample error bound for the final (KW) iterate θ_n:

Proposition 5.3.6. *consist Assume Assumptions 5.3.2, 5.3.3, and 5.3.5 hold. Set the step size as $\eta_n = \eta_0 n^{-\alpha}$ for some constant $\eta_0 > 0$ and $\alpha \in \left(\frac{1}{2}, 1\right)$ and the spacing parameter as $h_n = h_0 n^{-\gamma}$ for some constant $h_0 > 0$, and $\gamma \in \left(\frac{1}{2}, 1\right)$. The (KW) iterate θ_n converges to $\theta^\star$ almost surely.*

Furthermore, assume Assumptions 5.3.1 holds. For sufficiently large n, we have for $0 < \delta \le 2$,

$$\mathbb{E}\|\theta_n - \theta^\star\|^{2+\delta} \le C n^{-\alpha(2+\delta)/2}, \tag{5.13}$$

where the constant C depends on $d, \lambda, L_f, \alpha, \gamma, \eta_0, h_0$.

Remark 5.3.7. *The parameter dependency in Proposition 5.3.6 could be given explicitly as* *follows,*

$$
\begin{aligned}
\mathbb{E}\|\theta_n - \theta^*\|^2 \leq \exp &\Big(CM_1\eta_0/(2\alpha-1) + CM_2/(2\beta-1) - C\lambda\eta_0 n^{1-\alpha}/(1-\alpha)\Big)\|\theta_0\|^2 \\
&+ M_3\left(\exp\Big(-C\lambda\eta_0 n^{1-\alpha}/(1-\alpha)\Big) + \frac{\eta_0 n^{-\alpha}}{\lambda}\right) \\
&+ \frac{M_3}{M_1}\exp\Big(CM_1\eta_0/(2\alpha-1) + CM_2/(2\beta-1) - C\lambda\eta_0 n^{1-\alpha}/(1-\alpha)\Big),
\end{aligned}
$$

where the constant C above is a universal constant that does not depend on any constant or parameter in the assumptions. The other terms M_1, M_2, M_3 above are given below,

$$
\begin{aligned}
M_1 &= C\left(L_f^2\mathbb{E}\|v\|^4 + M^{\frac{2}{2+\delta}}\mathbb{E}\|v\|^4 + L_f^2\right), \\
M_2 &= CL_f^2\mathbb{E}\|v\|^3, \\
M_3 &= C\left(\mathbb{E}\|v\|^3 + \left(h_n^2L_f^2\mathbb{E}\|v\|^6 + M^{\frac{2}{2+\delta}}\mathbb{E}\|v\|^4(h_n^2\|v\|^2+1)\right)\right).
\end{aligned}
$$

Before we prove Proposition 5.3.6. It is helpful to give the following lemma, due to [10]. We include the proof here.

Lemma 5.3.8 ([10])**.** *Let $\{X_n\}$ be a martingale difference sequence in $\mathbb{R}^d$, in other words, $\mathbb{E}[X_n|X_{n-1}] = 0$. For any $1 \leq p \leq 2$ and any norm $\|\cdot\|$ on $\mathbb{R}^d$, there exists a constant C such that*

$$
\mathbb{E}\left\|\sum_{i=1}^n X_i\right\|^p \leq C\sum_{i=1}^n \mathbb{E}\left[\|X_i\|^p|X_{i-1}\right].
$$

Proof. We would like to show that there exists a constant C (which depends on d and p) such that for any $a, b \in \mathbb{R}^d$,

$$
\frac{1}{2}\left(\|a+b\|_2^p + \|a-b\|_2^p\right) \leq \|a\|_2^p + C\|b\|_2^p,
$$

where $\|\cdot\|_2$ is the 2-norm. To see this, in the one dimensional case, this is equivalent to

$$\frac{1}{2}\left(|1+x|^p + |1-x|^p\right) \leq 1 + C|x|^p.$$

At $x = 1$, the left hand side is differentiable and its first derivative is 0, so there exists a constant C such that the inequality holds in a neighborhood of $x = 1$. At $x \to \pm\infty$, the inequality also holds with some constant C. So it is easy to find a constant C such that the inequality holds for all x. The proof for the d-dimensional case is the same.

Using the above inequality, we have

$$\begin{aligned}\mathbb{E}_{n-1}\left\|\sum_{i=1}^{n} X_i\right\|_2^p &= \mathbb{E}_{n-1}\left\|\sum_{i=1}^{n-1} X_i + X_n\right\|_2^p \\ &\leq 2\left\|\sum_{i=1}^{n-1} X_i\right\|_2^p + 2C\mathbb{E}_{n-1}\|X_n\|_2^p - \mathbb{E}_{n-1}\left\|\sum_{i=1}^{n-1} X_i - X_n\right\|_2^p.\end{aligned}$$

On the other hand,

$$\mathbb{E}_{n-1}\left\|\sum_{i=1}^{n-1} X_i - X_n\right\|_2^p \geq \left\|\sum_{i=1}^{n-1} X_i - \mathbb{E}_{n-1}X_n\right\|_2^p = \left\|\sum_{i=1}^{n-1} X_i\right\|_2^p.$$

So

$$\mathbb{E}_{n-1}\left\|\sum_{i=1}^{n} X_i\right\|_2^p \leq \left\|\sum_{i=1}^{n-1} X_i\right\|_2^p + 2C\mathbb{E}_{n-1}\|X_n\|_2^p.$$

By induction, we then have

$$\mathbb{E}\left\|\sum_{i=1}^{n} X_i\right\|_2^p \leq 2C\sum_{i=1}^{n}\mathbb{E}\left[\|X_i\|_2^p | X_{i-1}\right].$$

For any general norm, there exists a constant C such that

$$\frac{1}{C}\|X\| \leq \|X\|_2 \leq C\|X\|.$$

So the same result holds for any norm. □

Now we return to the proof of both Proposition 5.3.6 and Remark 5.3.7.

Proof. We first assume Assumptions 5.3.2, 5.3.3, and 5.3.5 and give some bounds on ξ_n, γ_n, ε_n. By definition, $\mathbb{E}_{n-1}\gamma_n = \mathbb{E}_{n-1}\varepsilon_n = 0$. From (5.12),

$$\|\xi_n\| \leq \frac{1}{2} h_n L_f \mathbb{E}\|v\|^3. \tag{5.14}$$

We can bound γ_n by the following

$$\begin{aligned}\mathbb{E}\|\gamma_n\|^2 &\leq \mathbb{E}\left\|\frac{1}{h_n}[F(\theta_{n-1} + h_n v_n) - F(\theta_{n-1})]v_n\right\|^2 \\ &\leq \mathbb{E}\|\langle \nabla F(\theta_{n-1}), v_n\rangle v_n\|^2 + \frac{1}{4} h_n^2 L_f^2 \mathbb{E}\|v\|^6 \\ &\leq L_f^2 \mathbb{E}\|v\|^4 \mathbb{E}\|\theta_{n-1}\|^2 + \frac{1}{4} h_n^2 L_f^2 \mathbb{E}\|v\|^6. \end{aligned} \tag{5.15}$$

We also have the following fact for ε.

$$\begin{aligned} &\mathbb{E}_{n-1}\left[\|\varepsilon_n\|^2 | v_n\right] \\ = \;& \mathbb{E}_{n-1}\left[\left\|\frac{1}{h_n}\int_0^{h_n} \langle \nabla F(\theta_{n-1} + s v_n) - \nabla f(\theta_{n-1} + s v_n; \zeta_n), v_n\rangle v_n ds\right\|^2 \middle| v_n\right] \\ \leq \;& \|v_n\|^4 \mathbb{E}_{n-1}\left[\frac{1}{h_n}\int_0^{h_n} \|\nabla F(\theta_{n-1} + s v_n) - \nabla f(\theta_{n-1} + s v_n; \zeta_n)\|^2 ds \middle| v_n\right] \\ \leq \;& M^{\frac{2}{2+\delta}} \|v_n\|^4 \frac{1}{h_n}\int_0^{h_n} (\|\theta_{n-1} + s v_n\|^2 + 1) ds \\ \leq \;& M^{\frac{2}{2+\delta}} \|v_n\|^4 (\|\theta_{n-1}\|^2 + h_n^2 \|v_n\|^2 + 1), \end{aligned} \tag{5.16}$$

where in the second inequality, we use Assumption 5.3.3.

Now decompose the update step as follows,

$$\begin{aligned}\theta_n &= \theta_{n-1} - \eta_n \frac{1}{h_n}[f(\theta_{n-1} + h_n v_n; \zeta_n) - f(\theta_{n-1}; \zeta_n)] \\ &= \theta_{n-1} - \eta_n \nabla F(\theta_{n-1}) + \eta_n(\xi_n + \gamma_n + \varepsilon_n).\end{aligned}$$

We can derive that,

$$\begin{aligned}\|\theta_n\|^2 \leq &\|\theta_{n-1}\|^2 - 2\eta_n \langle \nabla F(\theta_{n-1}), \theta_{n-1}\rangle + 2\eta_n \langle \xi_n + \gamma_n + \varepsilon_n, \theta_{n-1}\rangle \\ &+ \eta_n^2 \|\xi_n + \gamma_n + \varepsilon_n - \nabla F(\theta_{n-1})\|^2. \end{aligned} \tag{5.17}$$

For the first part in the RHS of (5.17), using Lemma B.1 in [189], we have

$$\langle \theta, \nabla F(\theta)\rangle \geq \rho \|\theta\| \min\left\{\|\theta\|, \delta_1\right\}. \tag{5.18}$$

for some $\rho > 0$. For other parts in (5.17), we can bound them as

$$\begin{aligned}&|\eta_n \mathbb{E}_{n-1}\langle \xi_n + \gamma_n + \varepsilon_n, \theta_{n-1}\rangle| \\ &= \eta_n \left|\mathbb{E}_{n-1}\langle \xi_n, \theta_{n-1}\rangle\right| \\ &\leq \frac{1}{2}\eta_n h_n L_f \|\theta_{n-1}\| \mathbb{E}\|v\|^3 \\ &\leq C L_f^2 \mathbb{E}\|v\|^3 h_n^2 \|\theta_{n-1}\|^2 + C\mathbb{E}\|v\|^3 \eta_n^2, \end{aligned} \tag{5.19}$$

$$\begin{aligned}&\mathbb{E}_{n-1}\|\xi_n + \gamma_n + \varepsilon_n - \nabla F(\theta_{n-1})\|^2 \\ &\leq 4\|\xi_n\|^2 + 4\|\gamma_n\|^2 + 4\|\varepsilon_n\|^2 + 4\|\nabla F(\theta_{n-1})\|^2 \\ &\leq h_n^2 L_f^2 \mathbb{E}(\|v\|^3)^2 + 4L_f^2 \mathbb{E}\|v\|^4 \|\theta_{n-1}\|^2 + h_n^2 L_f^2 \mathbb{E}\|v\|^6 \\ &\quad + 4M^{\frac{2}{2+\delta}} \mathbb{E}\|v_n\|^4 (\|\theta_{n-1}\|^2 + h_n^2 \|v_n\|^2 + 1) + 4L_f^2 \|\theta_{n-1}\|^2 \\ &:= M_1 \|\theta_{n-1}\|^2 + M_2 \end{aligned} \tag{5.20}$$

where we use Cauchy-Schwarz inequality in (5.19), (5.20) and

$$M_1 = C\big(L_f^2 \mathbb{E}\|v\|^4 + M^{\frac{2}{2+\delta}} \mathbb{E}\|v\|^4 + L_f^2\big),$$

$$M_2 = C\big(h_n^2 L_f^2 \mathbb{E}\|v\|^6 + M^{\frac{2}{2+\delta}} \mathbb{E}\|v\|^4 (h_n^2 \|v\|^2 + 1)\big).$$

Combining all estimates, we have

$$\mathbb{E}_{n-1}\|\theta_n\|^2 \leq \Big(1 + C\eta_n^2 + Ch_n^2\Big) \|\theta_{n-1}\|^2 - 2\eta_n \rho \|\theta_{n-1}\| \min\{\|\theta_{n-1}\|, \delta_1\} + C\eta_n^2.$$

This is exactly the recursion considered in Part 1 and 2 of the proof of Theorem 2 in [162], and we can yield that θ_n converges almost surely to 0.

Now assume Assumption 5.3.1, we have a stronger estimate,

$$\langle \nabla F(\theta_{n-1}), \theta_{n-1} \rangle \geq F(\theta_{n-1}) + \frac{\lambda}{2} \|\theta_{n-1}\|^2 \geq \lambda \|\theta_{n-1}\|^2.$$

So combining all inequalities, we have

$$\mathbb{E}_{n-1}\|\theta_n\|^2 \leq \Big[1 - 2\lambda\eta_n + M_1\eta_n^2 + M_3 h_n^2\Big] \|\theta_{n-1}\|^2 + M_4 \eta_n^2, \tag{5.21}$$

where M_3, M_4 is defined by $M_3 = CL_f^2 \mathbb{E}\|v\|^3$, $M_4 = C(\mathbb{E}\|v\|^3 + M_2)$. Following the proof of Theorem 1 of [148], we can apply the recursion and get

$$\begin{aligned} \mathbb{E}\|\theta_n\|^2 \leq & \prod_{k=1}^{n} \Big[1 - 2\lambda\eta_k + M_1\eta_k^2 + CM_3 h_k^2\Big] \|\theta_0\|^2 \\ & + M_4 \sum_{k=1}^{n} \prod_{i=k+1}^{n} \Big[1 - 2\lambda\eta_i + M_1\eta_k^2 + M_3 h_k^2\Big] \eta_k^2. \end{aligned}$$

We can then bound the first term on the RHS,

$$\prod_{k=1}^{n}\left[1-2\lambda\eta_k+M_1\eta_k^2+M_3h_k^2\right]\leq\exp\left(-2\lambda\sum_{k=1}^{n}\eta_k\right)\exp\left(M_1\sum_{k=1}^{n}\eta_k^2\right)\exp\left(M_3\sum_{k=1}^{n}h_k^2\right),$$

as well as the second term on the RHS

$$\begin{aligned}&\sum_{k=1}^{n}\prod_{i=k+1}^{n}\left[1-2\lambda\eta_i+M_1\eta_k^2+M_3h_k^2\right]\eta_k^2\\ \leq\ &\exp\left(-\lambda\sum_{k=m+1}^{n}\eta_k\right)\sum_{k=1}^{n}\eta_k^2+\frac{\eta_m}{\lambda}\\ &+\frac{1}{M_1}\exp\left(M_1\sum_{k=1}^{n_0}\eta_k^2\right)\exp\left(M_3\sum_{k=1}^{n_0}h_k^2\right)\exp\left(-\lambda\sum_{k=1}^{n}\eta_k\right),\end{aligned}$$

where we denote by $n_0=\inf\{k\in\mathbb{N},1-2\lambda\eta_k+M_1\eta_k^2+M_3h_k^2\leq 1-\lambda\eta_k\}$ and m is any integer in $\{1,\ldots,n\}$. Choose $m=n/2$ and bound n_0 by n. Notice that $\sum_{k=1}^{n}\eta_k^2$ converge. So we can get

$$\begin{aligned}\mathbb{E}\|\theta_n\|^2\leq&\exp\left(CM_1\eta_0/(2\alpha-1)+CM_3/(2\beta-1)-C\lambda\eta_0n^{1-\alpha}/(1-\alpha)\right)\|\theta_0\|^2\\ &+M_4\left(\exp\left(-C\lambda\eta_0n^{1-\alpha}/(1-\alpha)\right)+\frac{\eta_0n^{-\alpha}}{\lambda}\right)\\ &+\frac{M_4}{M_1}\exp\left(CM_1\eta_0/(2\alpha-1)+CM_3/(2\beta-1)-C\lambda\eta_0n^{1-\alpha}/(1-\alpha)\right).\end{aligned}$$

Only the term $M_4\eta_0n^{-\alpha}/\lambda$ decreases at the order of $O(n^{-\alpha})$ while all the other terms decrease much faster.

Notice that all C's in the above inequality are universal constants which do not depend on any parameters in the assumptions. This proves Remark 5.3.7.

From now on, we will absorb all parameters (other than n) into C to make the asymptotic analysis more clear. By martingale convergence theorem, $\|\theta_n\|$ converges almost surely. Because its second moment converges to 0, it must converge to 0 almost surely.

We now show that,

$$\mathbb{E}\|\theta_n - \theta^\star\|^{2+\delta} \leq Cn^{-\alpha(2+\delta)/2}.$$

By same arguments as in (5.14), (5.15), (5.16), we can get

$$\begin{aligned}
\|\xi_n\|^{2+\delta} &\leq Ch_n^{2+\delta}, \\
\mathbb{E}_{n-1}\|\gamma_n\|^{2+\delta} &\leq \|\theta_{n-1}\|^{2+\delta} + Ch_n^{2+\delta}, \\
\mathbb{E}_{n-1}\left[\|\varepsilon_n\|^{2+\delta}\right] &\leq C(\|\theta_{n-1}\|^{2+\delta} + 1).
\end{aligned}$$

By similar arguments as in Lemma 5.3.8, there exists constants C such that for any a, b,

$$\|a+b\|^{2+\delta} \leq \|a\|^{2+\delta} + (2+\delta)\langle a, b\rangle\|a\|^{\delta} + C\|a\|^{\delta}\|b\|^2 + C\|b\|^{2+\delta}.$$

So we have the bound

$$\begin{aligned}
\mathbb{E}_{n-1}\|\theta_n\|^{2+\delta} \leq{}& \|\theta_{n-1}\|^{2+\delta} + \eta_n(2+\delta)\mathbb{E}_{n-1}\langle\theta_{n-1}, -\nabla F(\theta_{n-1}) + \xi_n + \gamma_n + \varepsilon_n\rangle\|\theta_{n-1}\|^{\delta} \\
&+ C\eta_n^2\|\theta_{n-1}\|^{\delta}\mathbb{E}_{n-1}\| - \nabla F(\theta_{n-1}) + \xi_n + \gamma_n + \varepsilon_n\|^2 \\
&+ C\eta_n^{2+\delta}\mathbb{E}_{n-1}\| - \nabla F(\theta_{n-1}) + \xi_n + \gamma_n + \varepsilon_n\|^{2+\delta} \\
\leq{}& (1 - (2+\delta)\lambda\eta_n)\|\theta_{n-1}\|^{2+\delta} + C\eta_n h_n\|\theta_{n-1}\|^{1+\delta} \\
&+ C\eta_n^2(\|\theta_{n-1}\|^2 + 1)\|\theta_{n-1}\|^{\delta} + C\eta_n^{2+\delta}(\|\theta_{n-1}\|^{2+\delta} + 1).
\end{aligned}$$

If $0 < \delta \leq 1$, by previous bound $\mathbb{E}\|\theta_n\|^2 \leq Cn^{-\alpha}$, we can get $\mathbb{E}\|\theta_n\|^{1+\delta} \leq Cn^{-\alpha(1+\delta)/2}$ and $\mathbb{E}\|\theta_n\|^{\delta} \leq Cn^{-\alpha\delta/2}$ by Hölder's inequality. So we can further get

$$\mathbb{E}\|\theta_n\|^{2+\delta} \leq (1 - Cn^{-\alpha} + Cn^{-2\alpha})\mathbb{E}\|\theta_{n-1}\|^{2+\delta} + Cn^{-(2+\delta)\alpha/2},$$

which implies $\mathbb{E}\|\theta_n\|^{2+\delta} \le Cn^{-(2+\delta)\alpha/2}$ as in the above proof after (5.21).

Now the case for $0 < \delta \le 1$ is proved. We can then use induction. If $\mathbb{E}\|\theta_n\|^{2+\delta} \le Cn^{-(2+\delta)\alpha/2}$ for all $\delta \le n$, then we can use the same method to prove the same inequality holds for $\delta \in (n, n+1]$. Thus the inequality holds for all δ. $\square$

A similar error bound is given by [61] in terms of the function values for $\delta = 0$. We generalize the result to the $(2+\delta)$-moment error bound on the parameter θ, where $\delta \in (0, 2]$ is assumed in Assumption 5.3.3 for the purpose of derivation of asymptotic normality. Proposition 5.3.6 suggests that the asymptotic rate of the (KW) estimator matches the best convergence rate of the (RM) estimator [148] when the spacing parameter $h_n = h_0 n^{-\gamma}$ is a decreasing sequence with $\gamma \in (\frac{1}{2}, 1)$.

Recall that to characterize the asymptotic behavior of (RM) iterates, we denote by S, the Gram matrix of $\nabla f(\theta;\zeta)$ at the true parameter $\theta^\star$, i.e., $S := \mathbb{E}\left[\nabla f(\theta^\star;\zeta)\nabla f(\theta^\star;\zeta)^\top\right]$. Analogously, we define the limiting Gram matrix of the (KW) gradient estimator $\widehat{g}_{h,v}$ at $\theta^\star$ as $h \to 0$ to be Q. The following lemma proves that the limiting Gram matrix takes the form of $Q = \mathbb{E}\left[vv^\top S vv^\top\right]$, and it quantifies the distance between $\widehat{g}_{h,v}(\theta^\star;\zeta)\widehat{g}_{h,v}(\theta^\star;\zeta)^\top$ and Q, as the spacing parameter $h \to 0$.

Lemma 5.3.9. *Under Assumptions 5.3.2, 5.3.3, 5.3.4, and 5.3.5, we have*

$$\left\|\mathbb{E}\big[\widehat{g}_{h,v}(\theta^\star;\zeta)\widehat{g}_{h,v}(\theta^\star;\zeta)^\top\big] - Q\right\| \le Ch(1+h^2), \quad Q = \mathbb{E}\big[vv^\top S vv^\top\big].$$

where $S = \mathbb{E}\left[\nabla f(\theta^\star;\zeta)\nabla f(\theta^\star;\zeta)^\top\right]$ *is defined in Assumption 5.3.3.*

Proof. By Assumption 5.3.3, we know that

$$\mathbb{E}\|\nabla f(\theta;\zeta) - \nabla F(\theta)\|^{2+\delta} \le M(\|\theta\|^{2+\delta} + d^{2+\delta}).$$

Therefore, the following holds for some constant $C > 0$,

$$\mathbb{E}\|\nabla f(\theta;\zeta) - \nabla F(\theta)\|^2 \le C(\|\theta\|^2 + d^2). \tag{5.22}$$

In particular,

$$\mathbb{E}\|\nabla f(0;\zeta) - \nabla F(0)\|^2 \le C. \tag{5.23}$$

From Assumption 5.3.4, we can get the following estimate for the Hessian matrix $\nabla^2 f(\theta;\zeta)$,

$$\begin{aligned}\mathbb{E}\|\nabla^2 f(\theta;\zeta)\|^2 &\le 2\mathbb{E}\|\nabla^2 f(0;\zeta)\|^2 + 2\mathbb{E}\left\|\nabla^2 f(\theta;\zeta) - \nabla^2 f(0;\zeta)\right\|^2 \\ &\le C(1 + \|\theta\|^2).\end{aligned}$$

Using the above observation, we find that

$$\begin{aligned}&\mathbb{E}\|\nabla f(\theta;\zeta) - \nabla F(\theta) - \nabla f(0;\zeta) + \nabla F(0)\|^2 \\ &\le C\|\theta\|^2 + 2\mathbb{E}\|\nabla f(\theta;\zeta) - \nabla f(0;\zeta))\|^2 \\ &= C\|\theta\|^2 + 2\mathbb{E}\left\|\int_0^1 \nabla^2 f(s\theta;\zeta)\theta ds\right\|^2 \\ &\le C\|\theta\|^2 + 2\mathbb{E}\int_0^1 \|\nabla^2 f(s\theta;\zeta)\theta\|^2 ds \\ &\le C\|\theta\|^2(1 + \int_0^1 \mathbb{E}\|\nabla^2 f(s\theta;\zeta)\|^2 ds) \\ &\le C\|\theta\|^2(1 + \|\theta\|^2). \end{aligned} \tag{5.24}$$

Define the function $\Sigma(\theta_1, \theta_2)$ by

$$\Sigma(\theta_1, \theta_2) := \mathbb{E}(\nabla f(\theta_1;\zeta) - \nabla F(\theta_1))(\nabla f(\theta_2;\zeta) - \nabla F(\theta_2))^\top.$$

Then combining inequalities (5.22), (5.23), (5.24), we have

$$
\begin{aligned}
\|\Sigma(\theta_1, \theta_2) - S\| &\leq \mathbb{E}\|(\nabla f(\theta_1; \zeta) - \nabla F(\theta_1))(\nabla f(\theta_2; \zeta) - \nabla F(\theta_2))^\top \\
&\quad - (\nabla f(0; \zeta) - \nabla F(0))(\nabla f(0; \zeta) - \nabla F(0))^\top\| \\
&\leq \mathbb{E}\|\nabla f(\theta_1; \zeta) - \nabla F(\theta_1)\|\|\nabla f(\theta_2; \zeta) - \nabla F(\theta_2) - \nabla f(0; \zeta) + \nabla F(0)\| \\
&\quad + \mathbb{E}\|\nabla f(\theta_1; \zeta) - \nabla F(\theta_2) - \nabla f(0; \zeta) + \nabla F(0)\|\|\nabla f(0; \zeta) - \nabla F(0)\| \\
&\leq C(d + \|\theta_1\|)\|\theta_2\|(1 + \|\theta_2\|) + C\|\theta_1\|(1 + \|\theta_1\|). \qquad (5.25)
\end{aligned}
$$

Notice that

$$
\begin{aligned}
&\mathbb{E}_\zeta \widehat{g}_{h,v}(\theta; \zeta)\widehat{g}_{h,v}(\theta; \zeta)^\top - (\frac{1}{h}\Delta_{h,v}F(\theta)v)(\frac{1}{h}\Delta_{h,v}F(\theta)v)^\top \\
= {} & \mathbb{E}_\zeta(\widehat{g}_{h,v}(\theta; \zeta) - \frac{1}{h}\Delta_{h,v}F(\theta)v)(\widehat{g}_{h,v}(\theta; \zeta) - \frac{1}{h}\Delta_{h,v}F(\theta)v)^\top \\
= {} & \frac{1}{h^2}\mathbb{E}_\zeta v(f(\theta + hv; \zeta) - f(\theta; \zeta) - F(\theta + hv) + F(\theta))^2 v^\top \\
= {} & \frac{1}{h^2}\mathbb{E}_\zeta vv^\top \Bigg[\int_0^h \int_0^h (\nabla F(\theta + s_1 v) - \nabla f(\theta + s_1 v; \zeta)) \\
&(\nabla F(\theta + s_2 v) - \nabla f(\theta + s_2 v; \zeta))^\top ds_1 ds_2 \Bigg| \Bigg] vv^\top \\
= {} & \frac{1}{h^2}\mathbb{E}_\zeta vv^\top \int_0^h \int_0^h \Sigma(\theta + s_1 v, \theta + s_2 v) ds_1 ds_2 vv^\top.
\end{aligned}
$$

We can use (5.25) and derive that

$$
\begin{aligned}
&\|\mathbb{E}_\zeta \widehat{g}_{h,v}(\theta; \zeta)\widehat{g}_{h,v}(\theta; \zeta)^\top - (\frac{1}{h}\Delta_{h,v}F(\theta)v)(\frac{1}{h}\Delta_{h,v}F(\theta)v)^\top - vv^\top S vv^\top\| \\
\leq {} & C\|v\|^4(\|\theta\| + h\|v\|)(1 + \|\theta\| + h\|v\|)(d + \|\theta\| + h\|v\|).
\end{aligned}
$$

Now we have

$$\begin{aligned}
&\|\mathbb{E}\widehat{g}_{h,v}(\theta;\zeta)\widehat{g}_{h,v}(\theta;\zeta)^\top - \mathbb{E}(\frac{1}{h}\Delta_{h,v}F(\theta)v)(\frac{1}{h}\Delta_{h,v}F(\theta)v)^\top - \mathbb{E}vv^\top Svv^\top\| \\
&\leq C\mathbb{E}\|v\|^4(\|\theta\| + h\|v\|)(1 + \|\theta\| + h\|v\|)(d + \|\theta\| + h\|v\|). \qquad (5.26)
\end{aligned}$$

By the same argument,

$$\begin{aligned}
&\|\mathbb{E}(\frac{1}{h}\Delta_{h,v}F(\theta)v)(\frac{1}{h}\Delta_{h,v}F(\theta)v)^\top\| \\
&\leq \frac{1}{h^2}\mathbb{E}\left\|vv^\top\left[\int_0^h\int_0^h (\nabla F(\theta + s_1v))(\nabla F(\theta + s_2v))^\top ds_1 ds_2\right]vv^\top\right\| \\
&\leq C\mathbb{E}\|v\|^4(\|\theta\|^2 + h^2\|v\|^2).
\end{aligned}$$

So we finally get

$$\begin{aligned}
&\|\mathbb{E}\widehat{g}_{h,v}(\theta;\zeta)\widehat{g}_{h,v}(\theta;\zeta)^\top - \mathbb{E}vv^\top Svv^\top\| \\
&\leq C\mathbb{E}\|v\|^4(\|\theta\| + h\|v\|)(1 + \|\theta\| + h\|v\|)(d + \|\theta\| + h\|v\|).
\end{aligned}$$

for some constant $C > 0$. □

With Lemma 5.3.9 in place, we state our first main result that characterizes the limiting distribution of the averaged `(AKW)` iterates defined in (5.1).

Theorem 5.3.10. *Let Assumptions 5.3.2, 5.3.3, 5.3.4, and 5.3.5 hold. Set the step size as $\eta_n = \eta_0 n^{-\alpha}$ for some constant $\eta_0 > 0$ and $\alpha \in \left(\frac{1}{2}, 1\right)$, and the spacing parameter as $h_n = h_0 n^{-\gamma}$ for some constant $h_0 > 0$, and $\gamma \in \left(\frac{1}{2}, 1\right)$. The averaged `(KW)` estimator $\overline{\theta}_n$ satisfies,*

$$\sqrt{n}\ \left(\overline{\theta}_n - \theta^\star\right) \Longrightarrow \mathcal{N}\left(0, H^{-1}QH^{-1}\right), \qquad \text{as} \quad n \to \infty, \qquad (5.27)$$

where $H = \nabla^2 F(\theta^\star)$ is the population Hessian matrix and $Q = \mathbb{E}\left[vv^\top S vv^\top\right]$ is defined in Lemma 5.3.9. Here $\Longrightarrow$ represents the convergence in distribution.

Proof. We follow the proof in [162]. The update step is

$$\begin{aligned}\theta_n &= \theta_{n-1} - \eta_n \nabla F(\theta_{n-1}) + \eta_n(\xi_n + \gamma_n + \varepsilon_n)\\ &= (I_d - \eta_n H)\theta_{n-1} + \eta_n(H\theta_{n-1} - \nabla F(\theta_{n-1}) + \xi_n + \gamma_n + \varepsilon_n).\end{aligned}$$

We only need to prove the following three conditions. First,

$$\sum_{i=1}^{\infty} \frac{1}{\sqrt{i}} \mathbb{E}\|H\theta_{i-1} - \nabla F(\theta_{i-1}) + \xi_i\|, \tag{5.28}$$

is bounded almost surely. Furthermore, we have

$$\mathbb{E}\|\gamma_i + \varepsilon_i\|^2, \tag{5.29}$$

is bounded almost surely, and when $t \to \infty$, the following convergence in probability,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\gamma_i + \varepsilon_i) \Longrightarrow \mathcal{N}(0, Q). \tag{5.30}$$

Condition (5.28) is used in Part 4 of the proof of Theorem 2 in [162]. This implies the error term introduced by the KW estimator is negligible. Conditions (5.29) and (5.30) are used in Part 1 of the proof of Theorem 1 in [162]. They are used to establish the central limit theorem under linear approximation. It is easy to check that as long as those three conditions are proved, the rest of the proof works without change.

Assumption 5.3.4 implies that

$$\|\nabla^2 F(\theta) - \nabla^2 F(y)\|^2 \le L_g \|\theta - y\|^2.$$

By Taylor expansion, we can further derive the bound

$$\|H\theta_{i-1} - \nabla F(\theta_{i-1})\| \leq C\|\theta_{i-1}\|^2.$$

Combining the previous inequality with inequality (5.14), we know that

$$\mathbb{E}\|H\theta_{i-1} - \nabla F(\theta_{i-1}) + \xi_i\| \leq C(\|\theta_{i-1}\|^2 + h_i^2),$$

which indicates that

$$\begin{aligned}\sum_{i=1}^{\infty} \frac{1}{\sqrt{i}} \mathbb{E}\|H\theta_{i-1} - \nabla F(\theta_{i-1}) + \xi_i\| &\leq C\sum_{i=1}^{\infty} \frac{1}{\sqrt{i}}(\|\theta_{i-1}\|^2 + h_i^2) \\ &\leq C + C\sum_{i=1}^{\infty} \frac{1}{\sqrt{i}}\|\theta_{i-1}\|^2 < \infty.\end{aligned}$$

The last step comes from Part 4 of the proof of Theorem 2 in [162].

Because γ_i converges to 0 almost surely and ε_i has bounded variance. So condition (5.29) holds. To prove condition (5.30), it suffices to verify that,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \Longrightarrow \mathcal{N}(0, Q).$$

By martingale central limit theorem [64, Theorem 8.2.8], we only need to verify two conditions,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{i-1}[\varepsilon_i \varepsilon_i^\top] \to Q, \tag{5.31}$$

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\|\varepsilon_i\|^2 1_{\|\varepsilon_i\| > a\sqrt{n}}\right] \to 0, \tag{5.32}$$

in probability for all $a > 0$.

Notice that (5.26) is equivalent to the following inequality,

$$\|\mathbb{E}_{n-1}\varepsilon_n\varepsilon_n^\top - \mathbb{E}vv^\top Svv^\top\| \leq C(\|\theta_{n-1}\| + h_n)(1 + \|\theta_{n-1}\|^3 + h_n^3). \tag{5.33}$$

Thus $\mathbb{E}_{n-1}[\varepsilon_n\varepsilon_n^\top]$ converges almost surely to Q and condition (5.31) holds.

Now consider the quantity in (5.32), by Proposition 5.3.6,

$$\mathbb{E}_{i-1}\left[\|\varepsilon_i\|^2 1_{\|\varepsilon_i\|>a\sqrt{n}}\right] \leq \left[\mathbb{E}_{i-1}\left[\|\varepsilon_i\|^{2+\delta}\right]\right]^{\frac{2}{2+\delta}} \left[\mathbb{E}_{i-1}\left[1_{\|\varepsilon_i\|>a\sqrt{n}}\right]\right]^{\frac{\delta}{2+\delta}}.$$

Note that

$$\mathbb{E}_{i-1}\left[1_{\|\varepsilon_i\|>a\sqrt{n}}\right] = \mathbb{P}_{i-1}\left(\|\varepsilon_i\| > a\sqrt{n}|\theta_{i-1}\right) \leq \frac{1}{a\sqrt{n}}\mathbb{E}_{i-1}\|\varepsilon_i\|.$$

Therefore, it can be bounded by

$$\mathbb{E}_{i-1}\left[\|\varepsilon_i\|^2 1_{\|\varepsilon_i\|>a\sqrt{n}}\right] \leq C\left(\frac{1}{a\sqrt{n}}\right)^{\frac{\delta}{2+\delta}} \left(1 + \|\theta_{i-1}\|^{2+\delta}\right)^{\frac{2}{2+\delta}} \left(1 + \|\theta_{i-1}\|\right)^{\frac{\delta}{2+\delta}},$$

from which we can obtain that

$$\mathbb{E}[\|\varepsilon_i\|^2 1_{\|\varepsilon_i\|>a\sqrt{n}}] \leq C\left(\frac{1}{a\sqrt{n}}\right)^{\frac{\delta}{2+\delta}}. \tag{5.34}$$

We find that condition (5.32) holds when n goes to infinity:

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\varepsilon_i\|^2 1_{\|\varepsilon_i\|>a\sqrt{n}}\right] \leq C\left(\frac{1}{a\sqrt{n}}\right)^{\frac{\delta}{2+\delta}} \to 0.$$

Therefore, we conclude the result. □

We now compare the asymptotic covariance matrix of $\overline{\theta}_n$ with that of the `(RM)` counterpart in (5.5) [1]. As one can see, the asymptotic covariance matrix of (AKW) estimator $\overline{\theta}_n$ exhibits a similar sandwich form as the covariance matrix of (RM), but strictly dominates the latter, regardless of the choice of random direction vectors $\{v_1, v_2, \ldots, v_n\}$. In fact, it is easy to check that

$$H^{-1}QH^{-1} - H^{-1}SH^{-1} = H^{-1}\mathbb{E}_v\left[(vv^\top - I_d)S(vv^\top - I_d)\right]H^{-1} \succ 0, \tag{5.35}$$

which suggests the (AKW) estimator suffers an inevitable loss of efficiency compared to the $\widehat{\theta}^{(\mathtt{RM})}$. In Section 5.3.3, we analyze (AKW) with multiple function-value queries at each iteration. With the price of additional per-iteration computational complexity, one is able to improve the statistical efficiency of (AKW) and achieve the optimal asymptotic variance $H^{-1}SH^{-1}$.

Remark 5.3.11. *To complete the distributional analysis on* (KW) *iterates, we also provide the asymptotic distribution of the n-th iterate* $\theta_n^{(\mathtt{KW})}$ *of* (5.3) *without averaging. Assume the Hessian matrix has decomposition* $H = P\Lambda P^\top$*, where* P *is an orthogonal matrix and* Λ *is a diagonal matrix. Using the proof in [68], we establish the following asymptotic distribution for* $\theta_n^{(\mathtt{KW})}$*,*

$$n^{\alpha/2}(\theta_n^{(\mathtt{KW})} - \theta^\star) \Longrightarrow \mathcal{N}(0, \Sigma), \tag{5.36}$$

where each (k,ℓ)*-th entry of the covariance matrix* Σ *is,*

$$\Sigma_{k\ell} = \eta_0\big(P^\top QP\big)_{kl}\big(\Lambda_{kk} + \Lambda_{\ell\ell}\big)^{-1}, \quad 1 \le k, \ell \le d.$$

Here $\eta_0 > 0$ *and* $\alpha \in (\frac{1}{2}, 1)$ *are specified in the step size* $\eta_n = \eta_0 n^{-\alpha}$*. As* $\alpha < 1$*, the*

[1] Note that the asymptotic covariance $H^{-1}SH^{-1}$ in (5.5) is “optimal” in the sense that it matches the asymptotic covariance for the empirical risk minimizer $\widehat{\theta}^{(\mathtt{ERM})}$ without online computation and gradient information constraint.

n-th iterate $\theta_n^{(\mathtt{KW})}$ without averaging converges at a slower rate $n^{-\alpha/2}$ than that of (AKW) *in Theorem 5.3.10.*

5.3.1 Examples: choices of direction distribution

By Theorem 5.3.10, the asymptotic covariance matrix of (AKW) estimator, $H^{-1}QH^{-1}$, depends on the distribution of search direction $\mathcal{P}_v$ via $Q = \mathbb{E}[vv^\top Svv^\top]$. In this section, we compare the asymptotic covariance matrices of the (AKW) estimator when the random directions $\{v_i\}_{i=1}^n$ are sampled from different $\mathcal{P}_v$'s. Several popular choices of $\mathcal{P}_v$ are listed as follows,

- (G) Gaussian: $v \sim \mathcal{N}(0, I)$.
- (S) Spherical: v is sampled from the uniform distribution on the sphere $\|v\|^2 = d$.
- (I) Uniform in the canonical basis: v is sampled from $\left\{\sqrt{d}e_1, \sqrt{d}e_2, \ldots, \sqrt{d}e_d\right\}$ with equal probability, where $\{e_1, e_2, \ldots, e_d\}$ is the canonical basis of $\mathbb{R}^d$.

It is easy to verify that the above three classical choices of $\mathcal{P}_v$ satisfy Assumption 5.3.5, among which (G) and (S) are continuous distributions, while (I) is a discrete distribution. In particular, (I) is a discrete uniform distribution with equal probability among the d vectors of the standard basis of Euclidean space $\mathbb{R}^n$, which can be generalized in the following two forms.

- (U) Uniform in an arbitrary orthonormal basis U: v_i is sampled uniformly from $\{\sqrt{d}u_1, \sqrt{d}u_2, \ldots, \sqrt{d}u_d\}$, where $\{u_1, u_2, \ldots, u_d\}$ is an arbitrary *orthonormal basis* of $\mathbb{R}^d$, i.e., the matrix $U = (u_1, u_2, \ldots, u_d)$ is a $d \times d$ orthonormal matrix such that $UU^\top = U^\top U = I$.
- (P) Non-uniform in the canonical basis with probability $(p_1, p_2, \ldots, p_d)$: $v = \sqrt{1/p_k}\, e_k$ with probability $p_k > 0$, for $k \in [d]$ and $\sum_{k=1}^d p_k = 1$.

The following proposition provides expressions of the matrix Q for the above five choices of $\mathcal{P}_v$.

Proposition 5.3.12. *Under the assumptions in Theorem 5.3.10, for above examples of $\mathcal{P}_v$, we have*

(G) *Gaussian:* $Q^{(\mathtt{G})} = (2S + \mathrm{tr}(S) I_d)$.

(S) *Spherical:* $Q^{(\mathtt{S})} = \frac{d}{d+2}\left(2S + \mathrm{tr}(S) I_d\right)$.

(I) *Uniform in the canonical basis:* $Q^{(\mathtt{I})} = d \, \mathrm{diag}(S)$.

(U) *Uniform in an arbitrary orthonormal basis U:* $Q^{(\mathtt{U})} = d \, U \, \mathrm{diag}(U^\top S U) U^\top$.

(P) *Non-uniform in a natural coordinate basis:* $Q^{(\mathtt{P})} = \mathrm{diag}(S_{11}/p_1, S_{22}/p_2, \ldots, S_{dd}/p_d)$.

Proof. For $Q^{(\mathtt{G})}$, let $z \sim \mathcal{N}(0, I_d)$, and we now calculate $\mathbb{E} z z^\top S z z^\top$. The (i,i)-th entry is

$$\mathbb{E} \sum_{j,k} z_i z_j S_{jk} z_k z_i = \sum_{j \neq i} S_{jj} + 3 S_{ii} = 2 S_{ii} + \mathrm{tr}(S).$$

For $i \neq j$, the (i,j)th entry is

$$\mathbb{E} \sum_{k,l} z_i z_k S_{kl} z_l z_j = 2 S_{ij}.$$

So $\mathbb{E} z z^\top S z z^\top = 2S + \mathrm{tr}(S) I_d$.

For $Q^{(\mathtt{S})}$, let v be sampled from the uniform distribution on the sphere $\|v\| = d$. The Gaussian vector z can be decomposed into independent radius part and spherical part,

$$\mathbb{E}[z z^\top] = \mathbb{E}\left[\|z\|^2 \frac{z}{\|z\|} \frac{z^\top}{\|z\|} \right] = \mathbb{E} v v^\top,$$

$$\mathbb{E}[z z^\top S z z^\top] = \mathbb{E}\left[\|z\|^4 \frac{z}{\|z\|} \frac{z^\top}{\|z\|} S \frac{z}{\|z\|} \frac{z^\top}{\|z\|} \right] = \frac{d+2}{d} \mathbb{E} v v^\top S v v^\top.$$

Now we have

$$\mathbb{E}vv^\top = I_d, \;\; \mathbb{E}vv^\top Svv^\top = \frac{d}{d+2}(2S + \mathrm{tr}(S)I_d).$$

For $Q^{(\mathtt{U})}$, let u obey the uniform distribution on $\{\sqrt{d}e_1, \dots, \sqrt{d}e_d\}$. By direct calculation, we have

$$\mathbb{E}uu^\top Suu^\top = \sum_{j=1}^{d} \frac{1}{d} \cdot d^2 S_{jj} = d \;\, \mathrm{diag}(S).$$

The final two cases for $Q^{(\mathtt{U})}, Q^{(\mathtt{P})}$ can also be verified by direct calculation. □

From Proposition 5.3.12, one can see that any of the above choices of $\mathcal{P}_v$ leads to a $Q^{(\cdot)}$ that strictly dominates S. Take $S = I_d$ as an example, we have $Q^{(\mathtt{G})} = (d+2)I_d$ and $Q^{(\mathtt{S})} = Q^{(\mathtt{I})} = Q^{(\mathtt{U})} = dI_d$ and $Q^{(\mathtt{P})} = \mathrm{diag}(p_1^{-1}, p_2^{-1}, \dots, p_d^{-1}) \succ I_d$ where $p_1 + p_2 + \cdots + p_d = 1$. Note that $Q^{(\mathtt{G})} \succ Q^{(\mathtt{S})}$ regardless of the dimension d and Gram matrix S. Intuitively, when the direction v is generated by Gaussian `(G)`, it can be decomposed into two independent random variables: the radical part $\|v\|$ and the spherical part $v/\|v\|$. The spherical part $v/\|v\|$ follows the same distribution as the uniform distribution on the sphere with radius d (which is identical to `(S)`). The extra randomness in the radical part $\|v\|^2 \sim \chi^2(d)$ leads to a larger magnitude of Q compared to that of `(S)`. Therefore the `(AKW)` estimator with Gaussian directions `(G)` is always inferior to that with spherical directions `(S)`, asymptotically. However, for the other candidates, they are not directly comparable, and the optimal choice of $\mathcal{P}_v$ depends on the optimality criterion, and Gram matrix S.

As a simple illustration, we consider $S = \mathrm{diag}(1, r_0)$ for some $r_0 > 0$. We have

`(S)` Spherical: $Q^{(\mathtt{S})} = \mathrm{diag}\left(\frac{r_0+3}{2}, \frac{3r_0+1}{2}\right)$.

`(I)` Uniform in a natural coordinate basis: $Q^{(\mathtt{I})} = \mathrm{diag}(2, 2r_0)$.

(U) Uniform in an arbitrary orthonormal basis U: when $U = (\cos\omega, \sin\omega; -\sin\omega, \cos\omega)$ and $\omega = 0$, we have $Q^{(\mathtt{U})} = Q^{(\mathtt{I})} = \mathrm{diag}(2, 2r_0)$; when $\omega = \pi/4$, we have $Q^{(\mathtt{U})} = \mathrm{diag}(1 + r_0, 1 + r_0)$.

(P) Non-uniform in a natural coordinate basis: $\mathrm{diag}\left(\frac{1}{p_1}, \frac{r_0}{1-p_1}\right), p_1 \in (0, 1)$.

From the above we can see that, the choices of the distribution of direction vectors $\mathcal{P}_v$ depends on the optimality-criteria on comparing the covariance matrices. Specifically in the above example, if one seeks to minimize

- the trace of covariance matrix, we have

$$\mathrm{tr}(Q^{(\mathtt{S})}) = \mathrm{tr}(Q^{(\mathtt{I})}) = \mathrm{tr}(Q^{(\mathtt{U})}) = 2 + 2r_0, \;\; \mathrm{tr}(Q^{(\mathtt{P})}) = \frac{1}{p_1} + \frac{r_0}{1 - p_1},$$

  and the optimal distribution that minimizes the trace depends on the value of p_1.

- the determinant of covariance matrix, we have

$$\begin{aligned} \det(Q^{(\mathtt{S})}) &= \frac{3r_0^2 + 10r_0 + 3}{4}, & \det(Q^{(\mathtt{I})}) &= 4r_0, \\ \det(Q^{(\mathtt{U})}) &= \frac{-\cos(4\omega)(r_0 - 1)^2 + r_0^2 + 6r_0 + 1}{2}, & \det(Q^{(\mathtt{P})}) &= \frac{r_0}{p_1(1 - p_1)}. \end{aligned}$$

  By a simple derivation, we have $\det(Q^{(\mathtt{S})}) \geq \det(Q^{(\mathtt{U})}) \geq \det(Q^{(\mathtt{I})})$ and $\det(Q^{(\mathtt{P})}) \geq \det(Q^{(\mathtt{I})})$.

- the operator norm of covariance matrix, i.e., the largest eigenvalue, we have

$$\begin{aligned} \lambda_{\max}(Q^{(\mathtt{S})}) &= \frac{r_0 + 3}{2}, & \lambda_{\max}(Q^{(\mathtt{I})}) &= 2, \\ \lambda_{\max}(Q^{(\mathtt{P})}) &= \max\left\{\frac{1}{p_1}, \frac{r_0}{1 - p_1}\right\}, & \lambda_{\max}(Q^{(\mathtt{U})}) &= r_0 + 1 + (1 - r_0)\left|\cos(2\omega)\right|. \end{aligned}$$

  The smallest operator norm for $Q^{(\mathtt{P})}$ is given by $p_1 = \frac{1}{1+r_0}$. When $r_0 \leq 1$, and $0 \leq \omega \leq$

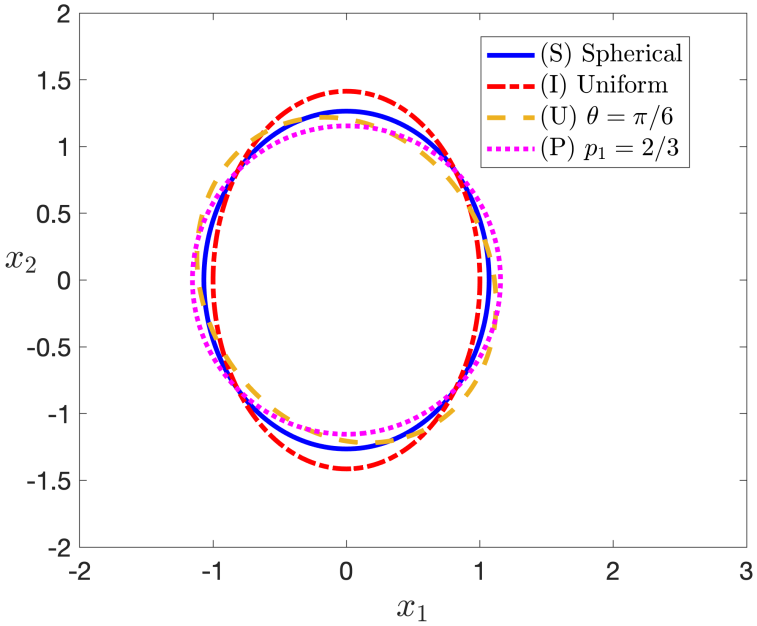


Figure 5.1: Comparison of Q matrices under different direction distributions $\mathcal{P}_v$ when $S = \mathrm{diag}(1, 1/2)$.

$\pi/6$, we have $\lambda_{\max}(Q^{(\mathtt{I})}) \geq \lambda_{\max}(Q^{(\mathtt{U})}) \geq \lambda_{\max}(Q^{(\mathtt{S})}) \geq \lambda_{\max}(Q^{(\mathtt{P})})$. When $r_0 \geq 1$, and $0 \leq \omega \leq \pi/6$, we have $\lambda_{\max}(Q^{(\mathtt{P})}) \geq \lambda_{\max}(Q^{(\mathtt{S})}) \geq \lambda_{\max}(Q^{(\mathtt{U})}) \geq \lambda_{\max}(Q^{(\mathtt{I})})$. For other choices of ω, we can obtain a comparison analogously.

In general, it is natural to use Loewner order to compare two positive semi-definite matrix $A, B \in \mathbb{R}^{d \times d}$, i.e., $A \succeq B$ if $x^\top A x \geq x^\top B x$ for any $x \in \mathbb{R}^d$. It is equivalent to say, for any positive constant $c > 0$, the ellipsoid $\{x \in \mathbb{R}^d : x^\top A x \leq c\}$ contains the ellipsoid $\{x \in \mathbb{R}^d : x^\top B x \leq c\}$. To better illustrate the result, we consider the 2-dimensional case where $S = \mathrm{diag}(1, 1/2)$ and plot the ellipse $\{x \in \mathbb{R}^2 : x^\top Q^{(\cdot)} x = 2\}$. In Figure 5.1, we compare $Q^{(\mathtt{S})}$, $Q^{(\mathtt{I})}$ (as a special case of $Q^{(\mathtt{U})}$ with $\theta = 0$), $Q^{(\mathtt{U})}$ with $\theta = \frac{\pi}{6}$, and $Q^{(\mathtt{P})}$ with $p_1 = \frac{1}{1+r_0} = \frac{2}{3}$. As can be inferred from the plot, none of the ellipsoids contain any other ellipsoids.

As shown in this illustrative example, there is no unique optimal direction distribution, and a practitioner might choose a search direction based on her favorable optimality criterion.

Lastly, in the following Remark 5.3.13, we show that, if the optimality criterion degenerates to one dimension, one may utilize the non-uniform distribution (P) to obtain a smaller limiting variance. In particular, consider the application where we are only interested in the first coordinate of $\theta^\star$, in which cases the optimality criterion of the limiting variance is on $\theta^\star_1$. We will show that the (AKW) estimator with the non-uniform distribution (P) achieves the Cramér-Rao lower bound.

Remark 5.3.13. *Assume the population loss function $F(\cdot)$ has Hessian $H = I_d$. Considering a non-uniform sampling* (P) *from $\{e_k\}_{k=1}^d$ for the direction distribution $\mathcal{P}_v$. We choose $v = e_k$ with probability p_k for $k = 1, 2, \ldots, d$, where $p_1 = 1 - p$ for some constant $p \in (0, 1]$ and $p_k = p/(d-1)$ for $k \neq 1$. Define i.i.d. random variables k_n where $k_n = 1$ with probability $1 - p$ and $k_n = 2, \ldots, d$ uniformly with probability $p/(d-1)$. The gradient estimator is defined by,*

$$\widehat{g}(\theta_{n-1}; \zeta_n) = \frac{f(\theta_{n-1} + h_n e_{k_n}; \zeta_n) - f(\theta_{n-1}; \zeta_n)}{h_n p_n} e_{k_n},$$

where $p_n = 1 - p$ if $k_n = 1$, $p_n = p/(d-1)$ for $k_n > 1$. By the same argument as Proposition 5.3.12, the variance for $\overline{\theta}_n$ in the direction e_1 is,

$$n\mathrm{Var}\left(e_1^\top(\overline{\theta}_n - \theta^\star)\right) = \frac{S_{11}}{1-p}.$$

As $p \to 0$, we approximately obtain the optimal variance given by Cramér-Rao lower bound in the direction e_1. However, in order to approach the optimal variance in the direction e_1, we increase the magnitude of variance in all other directions, where the variance in other directions is given by $n\mathrm{Var}\left(e_k^\top(\overline{\theta}_n - \theta^\star)\right) = (d-1)S_{kk}/p$ for $k = 2, \ldots, d$.

5.3.2 Asymptotic behavior of (AKW) *estimator for nonsmooth loss functions*

The theoretical analysis of the asymptotic distribution of the (AKW) estimator remains valid with a weaken assumption, which is a natural fit to some nonsmooth loss functions $F(\theta)$

including the quantile regression in Example 5.2.3.

Assumption 5.3.14. *Assume there exists $C > 0$ such that $\mathbb{E}\left[\widehat{g}_{h,v}(\theta^\star;\zeta)\widehat{g}_{h,v}(\theta^\star;\zeta)^\top\right] = Q + \Delta_h$ for some matrix $Q \in \mathbb{R}^{d\times d}$ and $\|\Delta_h\| \le Ch$.*

Theorem 5.3.15. *Let Assumption 5.3.2, 5.3.3, 5.3.5, and 5.3.14 hold. Under the step size and spacing parameter conditions specified in Theorem 5.3.10, the averaged estimator $\overline{\theta}_n$ satisfies,*

$$\sqrt{n}\ \left(\overline{\theta}_n - \theta^\star\right) \Longrightarrow \mathcal{N}\left(0, H^{-1}QH^{-1}\right), \qquad as \quad n \to \infty. \tag{5.37}$$

Proof. Under Assumption 5.3.2 and 5.3.14, the conclusions in Lemma 5.2.4 and Lemma 5.3.9 naturally hold. The rest of the proof follows from the proof in Proposition 5.3.6 and Theorem 5.3.10. □

To further illustrate the result, consider the setting of regression with non-smooth loss function. Suppose that the data consists of $\zeta_n = (x_n, y), n = 1, 2, \dots,$ and the loss function is

$$f(\theta;\zeta_n) = \rho(y_n - x_n^\top\theta).$$

We require the following regularity conditions.

Assumption 5.3.16.

- *Assume that $\rho(u)$ is a convex function with a subgradient $\psi(u)$. There exists constant C such that $|\psi(u)| \le C(|u| + 1)$.*
- *Let $\varepsilon_n = y_n - x_n^\top\theta^\star$. Assume $\{(x_n, \varepsilon_n), n = 1, 2, \dots\}$ are i.i.d., x have finite second moments and nondegenerate covariance matrices. Furthermore, ε and x are independent, $S = \mathbb{E}[\psi^2(\varepsilon)xx^\top]$ is positive definite. Define $Q = \mathbb{E}[vv^\top Svv^\top]$.*
- *Assume the probability density function $p(x)$ of ε is in C^3 and its derivatives up to third order are all integrable.*

- *Define $\phi(u) = \mathbb{E}[\psi(u+\varepsilon)]$. Assume $\phi(0) = 0, u\phi(u) > 0$ for any $u \neq 0$. Assume $\phi(u)$ is differentiable and $\phi'(0) > \sigma > 0$ uniformly for all x. Let $H = \mathbb{E}[\phi'(0)xx^\top]$.*
- *Assume that $\phi'(u)$ is Lipschitz at $u = 0$. That is, there exist constants $C > 0$ and $\delta > 0$ such that $|\phi'(u) - \phi'(0)| \leq C|u|$ for $|u| \leq \delta$.*

Those conditions are similar to those in [72], except that we make stronger assumption on the noise ε to ensure the (`KW`) method is stable. A direct computation can show that, under Assumption 5.3.2 and 5.3.16, all assumptions in Theorem 5.3.15 holds. So we have

$$\sqrt{n}\,\left(\overline{\theta}_n^{(\texttt{RM})} - \theta^\star\right) \Longrightarrow \mathcal{N}\left(0, H^{-1}SH^{-1}\right), \qquad \text{as} \quad n \to \infty,$$

and

$$\sqrt{n}\,\left(\overline{\theta}_n^{(\texttt{KW})} - \theta^\star\right) \Longrightarrow \mathcal{N}\left(0, H^{-1}QH^{-1}\right), \qquad \text{as} \quad n \to \infty,$$

in probability. We only verify Assumption 5.3.14 to indicate the difference between (`RM`) and (`KW`) estimators. The verification of the rest are the same for the two estimators. Notice that

$$\begin{aligned}\mathbb{E}\left[\widehat{g}_{h,v}(\theta^\star;\zeta)\widehat{g}_{h,v}(\theta^\star;\zeta)^\top\right] &= \mathbb{E}\frac{vv^\top}{h^2}\left[\rho\big(y - x^\top(\theta^\star + hv)\big) - \rho\big(y - x^\top\theta^\star\big)\right]^2 \\ &= \mathbb{E}\frac{vv^\top}{h^2}\left[\rho\big(\varepsilon - hx^\top v\big) - \rho(\varepsilon)\right]^2.\end{aligned}$$

Define the function $D(h) := \mathbb{E}vv^\top\left[\rho\big(\varepsilon - hx^\top v\big) - \rho(\varepsilon)\right]^2$. By the assumptions on the probability density function of ε, it is easy to see that $D(h)$ is also in C^3 (using integration by parts). Furthermore, direct computation shows that

$$D(0) = 0, D'(0) = 0, D''(0) = \mathbb{E}[\psi^2(\varepsilon)vv^\top xx^\top vv^\top] = 2Q.$$

Hence $\mathbb{E}\left[\widehat{g}_{h,v}(\theta^\star;\zeta)\widehat{g}_{h,v}(\theta^\star;\zeta)^\top\right] = D(h)/h^2 = Q + O(h)$.

5.3.3 Multi-query extension and statistical efficiency

We now consider the `(AKW)` estimator using $(m+1)$ function queries $\overline{\theta}_n^{(m)}$ in (5.11),

$$\overline{\theta}_n^{(m)} = \frac{1}{n}\sum_{i=1}^{n}\theta_i^{(m)}, \qquad \text{where } \theta_i^{(m)} = \theta_{i-1}^{(m)} - \eta_i \overline{g}_n^{(m)}(\theta_{i-1};\zeta_i) = \theta_{i-1}^{(m)} - \frac{\eta_i}{m}\sum_{j=1}^{m}\widehat{g}_{h_i, v_i^{(j)}}(\theta_{i-1};\zeta_i).$$

Here we first consider using the same sampling distribution across m queries and n iterations. In other words, $v_i^{(j)}$ is sampled *i.i.d.* from $\mathcal{P}_v$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$. Analogous to Theorem 5.3.10, we present the asymptotic distribution of the multi-query `(AKW)`,

Theorem 5.3.17. *Under the assumptions in Theorem 5.3.10, the $(m+1)$-query* `(AKW)` *estimator has the following asymptotic distribution, as $n \to \infty$,*

$$\sqrt{n}\left(\overline{\theta}_n^{(m)} - \theta^\star\right) \Longrightarrow \mathcal{N}\left(0, H^{-1}Q_m H^{-1}\right), \quad \textit{where } Q_m = \frac{1}{m}Q + \frac{m-1}{m}S.$$

Proof. The convergence result can be obtained as in the two function evaluation case. The only difference is the following calculation:

$$\mathbb{E}\left(\frac{1}{m}\sum_{i=1}^{m} v_i v_i^\top\right) S \left(\frac{1}{m}\sum_{i=1}^{m} v_i v_i^\top\right) = \frac{1}{m}\mathbb{E}vv^\top S vv^\top + \frac{m-1}{m}S,$$

which implies the desired result. □

Theorem 5.3.17 illustrates a trade-off effect between the statistical efficiency and computational efficiency. When $m = 1$ and only two queries of function evaluations are available, Theorem 5.3.17 reduces to Theorem 5.3.10, and $Q_m = Q$. Conversely, as $m \to \infty$, we have $Q_m \to S$. Therefore, the asymptotic covariance of $(m+1)$-query `(AKW)` estimator $\overline{\theta}_n^{(m)}$ approaches the optimal covariance $H^{-1}SH^{-1}$ as m approaches infinite. Nevertheless, the algorithm requires m function-value queries at each iteration, which significantly increases the computation complexity.

For a finite m, a slight revision of the sampling scheme of the direction vectors provides a remedy to achieve a smaller and indeed optimal asymptotic covariance matrix. Particularly at the i-th iteration, one may sample m direction vectors $\{v_i^{(j)}\}_{j=1,2,\ldots,m}$ from a discrete distribution (such as `(I)` and `(U)`) *without replacement.* In such settings, the direction vectors $\{v_i^{(1)}, v_i^{(2)}, \ldots, v_i^{(m)}\}$ are no longer independent but they have the same marginal distribution. The asymptotic distribution of the multi-query `(KW)` algorithm sampling without replacement is provided in the following theorem of its asymptotic distribution.

Theorem 5.3.18. *Under the assumptions in Theorem 5.3.10, and the direction vectors in all iterations $\{\widetilde{V}_i\}_{i=1}^n$ are* i.i.d. *from $\mathcal{P}_v$ such that $\widetilde{V}_i = (v_i^{(1)}, v_i^{(2)}, \ldots, v_i^{(m)})$ follows discrete sampling scheme in* `(I)` *and* `(U)` *WithOut Replacement* `(WOR)`*, the $(m+1)$-query* `(AKW)` *estimator, referred to as $\overline{\theta}_n^{(m,\texttt{WOR})}$, has the following asymptotic distribution, as $n \to \infty$,*

$$\sqrt{n}\left(\overline{\theta}_n^{(m,\texttt{WOR})} - \theta^\star\right) \Longrightarrow \mathcal{N}\left(0, H^{-1} Q_m^{(\texttt{WOR})} H^{-1}\right),$$

where $Q_m^{(\texttt{WOR})} = \frac{(d-m)}{m(d-1)} Q + \frac{d(m-1)}{m(d-1)} S$.

Proof. It is clear that $Q_m = S$ for $m = d$. We need to compute the quantity

$$Q_m = \frac{d^2}{m^2}\mathbb{E}\Big(\sum_{i=1}^m v_i v_i^\top\Big) S \Big(\sum_{i=1}^m v_i v_i^\top\Big),$$

which can be simplifies to

$$Q_m = \frac{d^2}{m^2}\mathbb{E}\Big(\sum_{i=1}^m v_i v_i^\top S v_i v_i^\top\Big) + \frac{d^2}{m^2}\mathbb{E}\Big(\sum_{i \neq j} v_i v_i^\top S v_j v_j^\top\Big).$$

By symmetry, it equals to

$$Q_m = \frac{d^2}{m}\mathbb{E} v_1 v_1^\top S v_1 v_1^\top + \frac{d^2(m-1)}{m}\mathbb{E} v_1 v_1^\top S v_2 v_2^\top.$$

We know $\mathbb{E}v_1v_1^\top Sv_1v_1^\top = \frac{1}{d^2}Q$ and $Q_d = S$. So we can solve for $\mathbb{E}v_1v_1^\top Sv_2v_2^\top$ and get

$$\mathbb{E}v_1v_1^\top Sv_2v_2^\top = \frac{1}{d(d-1)}(\frac{1}{d}Q - \operatorname{diag}(S)).$$

Therefore,

$$\begin{aligned} Q_m &= \frac{1}{m}Q + \frac{d(m-1)}{m(d-1)}(\frac{1}{d}Q - \operatorname{diag} S) \\ &= \frac{d-m}{m(d-1)}Q + \frac{d(m-1)}{m(d-1)}S. \end{aligned}$$ □

By comparing the asymptotic covariance matrices in Theorems 5.3.17 and 5.3.18, $Q_m^{(\texttt{WOR})}$ for sampling without replacement case is strictly smaller than Q_m in Theorems 5.3.17 when we consider multi-query evaluation ($m \geq 2$). Moreover, when $m = d$, it is easy to see that $Q_m^{(\texttt{WOR})} = S$. Therefore, the $(d+1)$-query (AKW) estimator $\overline{\theta}_n^{(m,\texttt{WOR})}$ achieves the same limiting covariance as that of the averaged (RM) estimator. Furthermore, when the model is well-specified, the limiting covariance matrix $H^{-1}SH^{-1} = H^{-1}$ achieves the Cramér-Rao lower bound. This result indicates that the $(d+1)$-query (AKW) estimator $\overline{\theta}_n$ is asymptotically efficient [193].

Remark 5.3.19. *Though the above* (WOR) *construction is reserved for* (I) *and* (U)*, there is a corresponding* (WOR) *scheme for the spherical direction* (S)*. One may use $\sqrt{d}v_1, \ldots, \sqrt{d}v_m$ to construct the $(m+1)$-query estimator, and $V = (v_1, \ldots, v_m)$ is sampled from the Stiefel manifold using the unique invariant measure, where the Stiefel manifold $V_{m,d}$ consists of all m-frames in d-dimensional vector space, i.e.,*

$$V_{m,d} := \{V \in \mathbb{R}^{d\times m} | V^\top V = I\}.$$

The proof of Theorem 5.3.18 can be used without modification to prove the asymptotic distribution for the (AKW) *estimator under the* (S+WOR) *scheme.*

5.4 Online statistical inference

In the previous section, we provide the asymptotic distribution for the `(AKW)` estimator. For the purpose of conducting statistical inference of $\theta^\star$, we need a consistent estimator of the limiting covariance $H^{-1}QH^{-1}$ in (5.27). A direct way is to construct a pair of consistent estimators $\hat{H}$ and $\hat{Q}$ of H and Q, respectively, and estimate the asymptotic covariance by the *plug-in* estimator $\hat{H}^{-1}\hat{Q}\hat{H}^{-1}$. Offline construction of those estimators is generally straightforward. However, as the `(KW)` scheme typically applies to sequential data, it is ideal to estimate the asymptotic covariance in an online fashion without storing the data. Therefore, one cannot simply replace the true parameter $\theta^\star$ by its estimate $\overline{\theta}_n$ in Q and H in an online setting, since we can no longer access the data stream $\{\zeta_i\}_{i=1}^n$ after the estimator $\overline{\theta}_n$ is obtained. To address this challenge, we first propose the following finite-difference Hessian estimator at each iteration n:

$$\begin{aligned}\widetilde{G}_n &= \sum_{k=1}^{d}\sum_{\ell=1}^{d} \widetilde{G}_{n,kl} e_k e_\ell^\top \\ &= \frac{1}{h_n^2}\sum_{k=1}^{d}\sum_{\ell=1}^{d} \left[\Delta_{h_n,e_k} f(\theta_{n-1}+h_n e_\ell;\zeta_n) - \Delta_{h_n,e_k} f(\theta_{n-1};\zeta_n)\right] e_k e_\ell^\top, \qquad (5.38)\end{aligned}$$

This construction can be viewed as a multi-query (with d^2+1 queries of function values at each iteration) `(KW)` scheme with the `(I)` choice of the random directions. Other choices of the search directions can be used as well. Each additional function-value query beyond the first one provides an estimate $\widetilde{G}_{n,kl}$ for the (k,l)-th entry of the matrix $\widetilde{G}_n$. To reduce the computational cost in $\widetilde{G}_n$, at each iteration, the algorithm may compute a random subset of entries of $\widetilde{G}_n$ and partially inhere the remaining entries from the previous estimator $\widetilde{G}_{n-1}$. For example, each entry $\widetilde{G}_{n,k\ell}$ is updated with probability $p \in (0,1]$. The procedure thus requires $\mathrm{O}(pd^2)$ function-value queries at each step. If we set $p = \mathrm{O}(1/d^2)$, then the query complexity is reduced to O(1) per step. Since the construction of (5.38) does not guarantee symmetry, an additional symmetrization step needs to be conducted, as

$$\widetilde{H}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{\widetilde{G}_i + \widetilde{G}_i^\top}{2}. \tag{5.39}$$

The next lemma quantifies the estimation error of the Hessian estimator $\widetilde{H}_n$ in (5.39).

Lemma 5.4.1. *Under Assumptions 5.3.1, 5.3.3, 5.3.4, and 5.3.5, we have*

$$\mathbb{E}\|\widetilde{H}_n - H\|^2 \leq C_1 n^{-\alpha} + C_2 p^{-1} n^{-1}. \tag{5.40}$$

Before we come to the proof of the Hessian estimator (5.39) in Lemma 5.4.1, we first introduce a naive method to estimate Hessian matrix H which we omit in the main text.

Inspired by the previous gradient estimator, we can estimate the Hessian matrix H by the following

$$\widehat{G}_n = \frac{1}{mh_n^2}\sum_{j=1}^{m}\left[\Delta_{h_n v_n^{(j)}} f(\theta_{n-1} + h_n u_n^{(j)}; \zeta_n) - \Delta_{h_n v_n^{(j)}} f(\theta_{n-1}; \zeta_n)\right] u_n^{(j)} v_n^{(j)\top},$$

where $\{u_n^{(j)}\}_{j=1}^m$ and $\{v_n^{(j)}\}_{j=1}^m$ are *i.i.d.* random vectors and $m > 0$ is a parameter (which might be different from m in the previous section). Therefore, our naive Hessian estimator is,

$$\widetilde{H}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{G}_i + \widehat{G}_i^\top}{2}. \tag{5.41}$$

where the $(\widehat{G}_i + \widehat{G}_i^\top)/2$ term ensures the symmetry of $\widetilde{H}_n$. The function query complexity is O(m) per step for this Hessian estimation.

Now we restate our Lemma 5.4.1 for the both estimators (5.41) and (5.39).

Lemma 5.4.2. *Under the assumptions in Theorem 5.3.10, we have the following result for the Hessian estimator* (5.41),

$$\mathbb{E}\|\widetilde{H}_n - H\|^2 \le C_1 n^{-\alpha} + C_2 \left(1 + \frac{1}{m}\right) n^{-1}. \tag{5.42}$$

The Hessian estimator (5.39) *satisfies,*

$$\mathbb{E}\|\widetilde{H}_n - H\|^2 \le C_1 n^{-\alpha} + C_2 p^{-1} n^{-1}. \tag{5.43}$$

Proof. In the case of naive Hessian estimator (5.41), we decompose $\widetilde{H}_n - H$ as follows,

$$\begin{aligned}
&\widetilde{H}_n - H \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{G}_i + \widehat{G}_n^\top}{2} - H \\
&= \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\widehat{G}_i + \widehat{G}_i^\top}{2} - \left(\frac{1}{m}\sum_{j=1}^{m} u_i^{(j)} u_i^{(j)\top}\right) \nabla^2 f(\theta_{n-1};\zeta_n) \left(\frac{1}{m}\sum_{j=1}^{m} v_i^{(j)} v_i^{(j)\top}\right)\right) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \left(\left(\frac{1}{m}\sum_{j=1}^{m} u_i^{(j)} u_i^{(j)\top}\right) \nabla^2 f(\theta_{n-1};\zeta_n) \left(\frac{1}{m}\sum_{j=1}^{m} v_i^{(j)} v_i^{(j)\top}\right) - \nabla^2 f(\theta_{i-1};\zeta_i)\right) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \left[\nabla^2 f(\theta_{i-1};\zeta_i) - \nabla^2 f(0;\zeta_i)\right] + \frac{1}{n}\sum_{i=1}^{n} \left(\nabla^2 f(0;\zeta_i) - H\right).
\end{aligned} \tag{5.44}$$

For the first term in the decomposition (5.44),

$$
\begin{aligned}
&\mathbb{E}_{n-1}\Bigg[\|\frac{1}{h_n^2}\big[f(\theta_{n-1}+h_nu+h_nv;\zeta_n)-f(\theta_{n-1}+h_nu;\zeta_n)-f(\theta_{n-1}+h_nv;\zeta_n)\\
&\quad+\ f(\theta_{n-1};\zeta_n)\big]uv^\top-uu^\top\nabla^2 f(\theta_{n-1};\zeta_n)vv^\top\Big\|^2\Bigg|u,v\Bigg]\\
&\le \mathbb{E}_{n-1}[\|\frac{1}{h_n^2}uu^\top\int_0^{h_n}\int_0^{h_n}\nabla^2 f(\theta_{n-1}+s_1u+s_2v;\zeta_n)\\
&\quad-\nabla^2 f(\theta_{n-1};\zeta_n)ds_1ds_2vv^\top\|^2\Big|u,v]\\
&\le \frac{1}{h_n^2}\|u\|^2\|v\|^2\int_0^{h_n}\int_0^{h_n}\mathbb{E}_{n-1}[\|\nabla^2 f(\theta_{n-1}+s_1u+s_2v;\zeta_n)\\
&\quad-\nabla^2 f(\theta_{n-1};\zeta_n)\|^2\big|u,v]ds_1ds_2\\
&\le \frac{C}{h_n^2}\|u\|^2\|v\|^2\int_0^{h_n}\int_0^{h_n}\|s_1u+s_2v\|^2\,ds_1ds_2\le Ch_n^2\|u\|^2\|v\|^2(\|u\|^2+\|v\|^2).
\end{aligned}
$$

The above derivation implies that

$$
\mathbb{E}\|\widehat{G}_n-\left(\frac{1}{m}\sum_{j=1}^m u_i^{(j)}u_i^{(j)\top}\right)\nabla^2 f(\theta_{n-1};\zeta_n)\left(\frac{1}{m}\sum_{j=1}^m v_i^{(j)}v_i^{(j)\top}\right)\|\le Ch_n^2.
$$

Therefore, we can show that

$$
\begin{aligned}
&\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n\left(\frac{\widehat{G}_i+\widehat{G}_i^\top}{2}-\left(\frac{1}{m}\sum_{j=1}^m u_i^{(j)}u_i^{(j)\top}\right)\nabla^2 f(\theta_{i-1};\zeta_i)\left(\frac{1}{m}\sum_{j=1}^m v_i^{(j)}v_i^{(j)\top}\right)\right)\right\|^2\\
&\le \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n\left(\widehat{G}_i-\left(\frac{1}{m}\sum_{j=1}^m u_i^{(j)}u_i^{(j)\top}\right)\nabla^2 f(\theta_{i-1};\zeta_i)\left(\frac{1}{m}\sum_{j=1}^m v_i^{(j)}v_i^{(j)\top}\right)\right)\right\|^2\\
&\le C\frac{1}{n}\sum_{i=1}^n h_i^2\le Cn^{-2\gamma},
\end{aligned}
\tag{5.45}
$$

where in the first inequality, we use the fact that, $\widehat{G}_i$ and $\widehat{G}_i^\top$ has the same distribution.

For the second term, notice that

$$\begin{aligned}
&\mathbb{E}_{n-1}\left\|\left(\frac{1}{m}\sum_{j=1}^m u_j u_j^\top\right)\nabla^2 f(\theta_{n-1};\zeta_n)\left(\frac{1}{m}\sum_{j=1}^m v_j v_j^\top\right)-\nabla^2 f(\theta_{n-1};\zeta_n)\right\|^2\\
&\leq \mathbb{E}_{n-1}\left\|\frac{1}{m}u_i u_i^\top - I_d\right\|^2\left\|\nabla^2 f(\theta_{n-1};\zeta_n)\right\|^2\left\|\frac{1}{m}vv^\top - I_d\right\|^2\\
&\quad+\mathbb{E}_{n-1}\left\|\frac{1}{m}u_i u_i^\top - I_d\right\|^2\left\|\nabla^2 f(\theta_{n-1};\zeta_n)\right\|^2+\mathbb{E}_{n-1}\left\|\nabla^2 f(\theta_{n-1};\zeta_n)\right\|^2\left\|\frac{1}{m}vv^\top - I_d\right\|^2\\
&\leq \frac{C}{m}\left(1+\|\theta_{n-1}\|^2\right).
\end{aligned}$$

Furthermore, the second term is a sum of martingale difference sequence and we have

$$\begin{aligned}
&\mathbb{E}\|\frac{1}{n}\sum_{i=1}^n\left(\left(\frac{1}{m}\sum_{j=1}^m u_i^{(j)}u_i^{(j)\top}\right)\nabla^2 f(\theta_{n-1};\zeta_n)\left(\frac{1}{m}\sum_{j=1}^m v_i^{(j)}v_i^{(j)\top}\right)-\nabla^2 f(\theta_{i-1};\zeta_i)\right)\|^2\\
&=\frac{1}{n}\sum_{i=1}^n\mathbb{E}\|\left(\left(\frac{1}{m}\sum_{j=1}^m u_i^{(j)}u_i^{(j)\top}\right)\nabla^2 f(\theta_{n-1};\zeta_n)\left(\frac{1}{m}\sum_{j=1}^m v_i^{(j)}v_i^{(j)\top}\right)-\nabla^2 f(\theta_{i-1};\zeta_i)\right)\|^2\\
&\leq C\frac{1}{n^2}\sum_{i=1}^n\frac{1}{m}\left(1+\mathbb{E}\|\theta_{n-1}\|^2\right)\leq C\frac{1}{mn}. \qquad (5.46)
\end{aligned}$$

For the third term in (5.44), we have

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n\nabla^2 f(\theta_{i-1};\zeta_i)-\nabla^2 f(0;\zeta_i)\right\|^2 &\leq \frac{1}{n}\sum_{i=1}^n\mathbb{E}\left\|\nabla^2 f(\theta_{i-1};\zeta_i)-\nabla^2 f(0;\zeta_i)\right\|^2\\
&\leq \frac{C}{n}\sum_{i=1}^n\mathbb{E}\|\theta_i\|^2\leq Cn^{-\alpha}. \qquad (5.47)
\end{aligned}$$

For the final term, we have

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla^2 f(0;\zeta_i)-H\right\|^2 &\le \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla^2 f(0;\zeta_i)-H\right\|^2\\ &\le \frac{C}{n^2}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla^2 f(0;\zeta_i)^2-H^2\right\|\le Cn^{-1}, \qquad (5.48)\end{aligned}$$

where the second inequality is due to the fact that it is an equality in Frobenius norm.

Combine the previous estimates (5.45), (5.46), (5.47) and (5.48), our naive Hessian estimator satisfies,

$$\mathbb{E}\left\|\widetilde{H}_n-H\right\|^2\le Cn^{-\alpha}+C(1+\frac{1}{m})n^{-1}.$$

Similarly, for the Hessian estimator (5.39), we have the following decomposition,

$$\begin{aligned}\widetilde{H}_n-H =&\frac{1}{n}\sum_{i=1}^{n}\frac{\widetilde{G}_i+\widetilde{G}_i^\top}{2}-H\\ =&\frac{1}{n}\sum_{i=1}^{n}\frac{\widetilde{G}_i+\widetilde{G}_i^\top}{2}-\frac{\widehat{G}_i+\widehat{G}_i^\top}{2}+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\widehat{G}_i+\widehat{G}_i^\top}{2}-\nabla^2 f(\theta_{i-1};\zeta_i)\right)\\ &+\frac{1}{n}\sum_{i=1}^{n}\left[\nabla^2 f(\theta_{i-1};\zeta_i)-\nabla^2 f(0;\zeta_i)\right]+\frac{1}{n}\sum_{i=1}^{n}\nabla^2 f(0;\zeta_i)-H. \qquad (5.49)\end{aligned}$$

Given $\widehat{G}_n$, our Bernoulli sampling Hessian estimator $\widetilde{G}_n$ satisfies,

$$\begin{aligned}\mathbb{E}\left\|\widetilde{G}_n-\widehat{G}_n\right\|_{\text{Fro}}^2 &=\mathbb{E}\left[\sum_{j=1}^{d}\sum_{k=1}^{d}\frac{1}{p}\left(\widehat{G}_n^{(jk)}B_n^{(jk)}-\widehat{G}_n^{(jk)}\right)^2\right]\\ &=\sum_{j=1}^{d}\sum_{k=1}^{d}\mathbb{E}\left(\frac{1}{p}B_n^{(jk)}-1\right)^2\left(\widehat{G}_n^{(jk)}\right)^2\\ &=\frac{1-p}{p}\sum_{j=1}^{d}\sum_{k=1}^{d}\mathbb{E}\left(\widehat{G}_n^{(jk)}\right)^2=\frac{1-p}{p}\|\widehat{G}_n\|_{\text{Fro}}^2,\end{aligned}$$

where the entries of B_n are *i.i.d.* and follow a Bernoulli distribution, $B_n^{(k\ell)} \sim \text{Bernoulli}(p)$, for some fixed $p \in (0,1)$. Here the second equality uses the fact that $B_i^{(jk)}$ are independent from each other. Therefore,

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{G}_i - \widehat{G}_i\right\|^2 \leq \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{G}_i - \widehat{G}_i\right\|_{\text{Fro}}^2 \leq C\frac{1-p}{p}n^{-2}\sum_{i=1}^{n}\mathbb{E}\left\|\widehat{G}_i\right\|^2.$$

With $1/t\sum_{i=1}^{n}\mathbb{E}\|\widehat{G}_i\|^2 \leq C + Cn^{-\alpha}$, the first term in decomposition (5.49) satisfies,

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{G}_i - \widehat{G}_i\right\|^2 \leq C\frac{1-p}{p}n^{-1}. \tag{5.50}$$

Other terms can be bounded similarly as in the first case:

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{G}_i + \widehat{G}_i^\top}{2} - \nabla^2 f(\theta_{i-1};\zeta_i)\right\|^2 \leq Cn^{-2\gamma}, \tag{5.51}$$

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla^2 f(\theta_{i-1};\zeta_i) - \nabla^2 f(0;\zeta_i)\right\|^2 \leq Cn^{-\alpha}, \tag{5.52}$$

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla^2 f(0;\zeta_i) - H\right\|^2 \leq Cn^{-1}. \tag{5.53}$$

Combine inequality (5.50), (5.51), (5.52) and (5.53), we obtain the desired result for Hessian estimator (5.39). □

From Lemma 5.4.1, as $n \to \infty$, the error rate is dominated by the $C_1 n^{-\alpha}$ term, where α is the parameter of the decaying step sizes.

Remark 5.4.3. *In the construction of the estimator of the limiting covariance matrix $H^{-1}QH^{-1}$, it is necessary to avoid the possible singularity of $\widetilde{H}_n$. A common practice is to adopt a thresholding version of $\widetilde{H}_n$ in (5.39). Let $U\widetilde{\Lambda}_n U^\top$ be the eigenvalue decomposition* *of $\widetilde{H}_n$, and define*

$$\widehat{H}_n = U\widehat{\Lambda}_n U^\top, \;\; \widehat{\Lambda}_{n,kk} = \max\left\{\kappa_1, \widetilde{\Lambda}_{n,kk}\right\}, \; k = 1, 2, \ldots, d, \tag{5.54}$$

for any positive constant $\kappa_1 < \lambda$ where λ is defined in Assumption 5.3.1. It is guaranteed by construction that $\widehat{H}_n$ is strictly positive definite and thus invertible.

On the other hand, the estimator of Gram matrix Q can be naturally constructed as

$$\widehat{Q}_n := \frac{1}{n}\sum_{i=1}^{n} \widehat{g}_{h_i,v_i}(\theta_{i-1}; \zeta_i)\, \widehat{g}_{h_i,v_i}(\theta_{i-1}; \zeta_i)^\top, \tag{5.55}$$

where $\widehat{g}_{h_i,v_i}(\theta_{i-1}; \zeta_i)$ is the (KW) update in the i-th iteration obtained by (5.10). As both $\widehat{H}_n$ in (5.54) and $\widehat{Q}_n$ in (5.55) can be constructed sequentially without storing historical data[2], the final plug-in estimator $\widehat{H}_n^{-1}\widehat{Q}_n\widehat{H}_n^{-1}$ can also be constructed in an online fashion. Based on Lemma 5.4.1, we obtain the following consistency result of the covariance matrix estimator $\widehat{H}_n^{-1}\widehat{Q}_n\widehat{H}_n^{-1}$.

Theorem 5.4.4. *Assume Assumptions 5.3.1, 5.3.3, 5.3.4, and 5.3.5 hold for $\delta = 2$. Set the step size as $\eta_n = \eta_0 n^{-\alpha}$ for some constant $\eta_0 > 0$ and $\alpha \in \left(\frac{1}{2}, 1\right)$, and the spacing parameter as $h_n = h_0 n^{-\gamma}$ for some constant $h_0 > 0$, and $\gamma \in \left(\frac{1}{2}, 1\right)$. We have*

$$\mathbb{E}\left\|\widehat{H}_n^{-1}\widehat{Q}_n\widehat{H}_n^{-1} - H^{-1}QH^{-1}\right\| \le Cn^{-\alpha/2}.$$

To prove Theorem 5.4.4, we first present the following lemma on the error rate of $\widehat{Q}_n$.

Lemma 5.4.5. *Under conditions in Theorem 5.4.4, our online Gram matrix estimate $\widehat{Q}_n$*

[2] The sequence $\widehat{Q}_n := \frac{1}{n}\sum_{i=1}^{n} Q_i$ with $Q_i = \widehat{g}_{h_i,v_i}(\theta_{i-1}; \zeta_i)\, \widehat{g}_{h_i,v_i}(\theta_{i-1}; \zeta_i)^\top$ can be constructed only with one-pass over the sequential data. In particular, we could compute $\widehat{Q}_n$ sequentially as $\widehat{Q}_n = \frac{1}{n}((n-1)\widehat{Q}_{n-1} + Q_i)$.

has the following convergence rate,

$$\mathbb{E}\|\widehat{Q}_n - Q\| \le Cn^{-\alpha/2}.$$

Proof. Recall the update rule,

$$\theta_n = \theta_{n-1} - \eta_n \nabla F(\theta_{n-1}) + \eta_n(\xi_n + \gamma_n + \varepsilon_n),$$

and our Gram matrix estimate $\widehat{Q}_n$ is,

$$\widehat{Q}_n = \frac{1}{n}\sum_{i=1}^{n}(\nabla F(\theta_{i-1}) - \xi_i - \gamma_i - \varepsilon_i)(\nabla F(\theta_{i-1}) - \xi_i - \gamma_i - \varepsilon_i)^\top.$$

It can be seen that we have the following estimates,

$$\begin{aligned}
\mathbb{E}_{n-1}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F(\theta_{i-1})\nabla F(\theta_{i-1})^\top\right\| &\le C\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{n-1}\|\theta_{i-1}\|^2 \le Cn^{-\alpha},\\
\mathbb{E}_{n-1}\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\xi_i^\top\right\| &\le C\frac{1}{n}\sum_{i=1}^{n}h_n^2 \le Cn^{-2\gamma},\\
\mathbb{E}_{n-1}\left\|\frac{1}{n}\sum_{i=1}^{n}\gamma_i\gamma_i^\top\right\| &\le C\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}_{n-1}\|\theta_{i-1}\|^2 + h_n^2) \le Cn^{-\alpha},\\
\mathbb{E}_{n-1}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\varepsilon_i^\top\right\| &\le C\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}_{n-1}\|\theta_{i-1}\|^2 + h_n^2 + 1) \le C.
\end{aligned}$$

The crossing terms between them can be bounded by Cauchy-Schwarz inequality. Therefore, we can find that all terms in $\widehat{Q}_n$ except $\sum_{i=1}^{n}\varepsilon_i\varepsilon_i^\top/t$ can be bounded by $Cn^{-\alpha/2}$. So it suffices to prove,

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\varepsilon_i^\top - Q\right\| \le Cn^{-\alpha/2}. \tag{5.56}$$

Define a new sequence $z_n := \varepsilon_n\varepsilon_n^\top - \mathbb{E}_{n-1}\varepsilon_n\varepsilon_n^\top$. Then z_n is a martingale difference sequence and we have

$$\begin{aligned}\left\|\varepsilon_n\varepsilon_n^\top - Q\right\| &\le \|z_n\| + \left\|\mathbb{E}_{n-1}\varepsilon_n\varepsilon_n^\top - Q\right\| \\ &\le \|z_n\| + C\left(\|\theta_{n-1}\| + \|\theta_{n-1}\|^4 + h_n + h_n^4\right),\end{aligned}$$

where the last inequality leverages inequality (5.33). Now we have,

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\varepsilon_i^\top - Q\right\| &\le \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n z_i\right\| + C\mathbb{E}\left(\|\theta_{n-1}\| + \|\theta_{n-1}\|^4 + h_n + h_n^4\right) \\ &\le \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n z_i\right\| + Cn^{-\alpha/2}.\end{aligned}$$

Thus we turn the proof of (5.56) into,

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n z_i\right\| \le Cn^{-1/2}. \tag{5.57}$$

By Hölder's inequality, it can be derived that,

$$\mathbb{E}_{n-1}\|z_n\|^2 \le \mathbb{E}_{n-1}\|\varepsilon_n\|^4 \le C(\|\theta_{n-1}\|^4 + h_n^4 + 1).$$

Combine Lemma 5.3.8 with Lemma 5.3.6, we have

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n z_i\right\|^2 \le \frac{1}{n^2}\sum_{i=1}^n C\mathbb{E}\left(\|\theta_{i-1}\|^4 + h_i^4 + 1\right) \le Cn^{-1}.$$

Therefore, condition (5.57) is satisfied through Jensen's inequality. □

We now provide a matrix perturbation inequality from [40].

Lemma 5.4.6. *If a matrix $B = A + E$ where A and B are invertible, we have,*

$$\left\| B^{-1} - A^{-1} \right\| \leq \|A^{-1}\|^2 \|E\| \frac{1}{1 - \|A^{-1}E\|}.$$

Proof. Notice that

$$\begin{aligned} B^{-1} = (A+E)^{-1} &= A^{-1} - A^{-1}\left(A^{-1} + E^{-1}\right)^{-1} A^{-1} \\ &= A^{-1} - A^{-1}E\left(A^{-1}E + I\right)^{-1} A^{-1}. \end{aligned}$$

Therefore, the inversion error is,

$$\begin{aligned} \|B^{-1} - A^{-1}\| &= \left\| A^{-1}E\left(A^{-1}E + I\right)^{-1} A^{-1} \right\| \\ &\leq \|A^{-1}\|^2 \|E\| \|(A^{-1}E + I)^{-1}\| \\ &\leq \|A^{-1}\|^2 \|E\| \frac{1}{\lambda_{\min}(A^{-1}E + I)} \\ &\leq \|A^{-1}\|^2 \|E\| \frac{1}{1 - \|A^{-1}E\|}, \end{aligned}$$

where we use Weyl's inequality in the last inequality. □

We now come back to the main proof of Theorem 5.4.4.

Proof. For the thresholding estimator $\widehat{H}_n$, since $\|\widehat{H}_n - \widetilde{H}_n\| \leq \|\widetilde{H}_n - H\|$ by construction, it is consistent with the rate below,

$$\mathbb{E}\|\widehat{H}_n - H\|^2 \leq 2\mathbb{E}\|\widetilde{H}_n - H\|^2 + 2\mathbb{E}\|\widehat{H}_n - \widetilde{H}_n\|^2 \leq 4\mathbb{E}\|\widetilde{H}_n - H\|^2 \leq Cn^{-\alpha}, \tag{5.58}$$

where the last inequality from Lemma 5.4.1.

By Lemma 5.4.6, the inverse matrix error satisfies,

$$
\begin{aligned}
&\mathbb{E}\|\widehat{H}_n^{-1} - H^{-1}\|^2 \\
\leq\ & \mathbb{E}\left[1_{\|H^{-1}(\widehat{H}_n - H)\| \leq 1/2} 2\|\widehat{H}_n - H\|\|H^{-1}\|^2 + 1_{\|H^{-1}(\widehat{H}_n - H)\| \geq 1/2}\|\widehat{H}_n^{-1} - H^{-1}\|\right]^2 \\
\leq\ & 8\|H^{-1}\|^4\mathbb{E}\|\widehat{H}_n - H\|^2 + 2(\kappa_1^{-1} + \lambda_{\min}^{-1}(H))^2\mathbb{P}\left(\|H^{-1}(\widehat{H}_n - H)\| \geq \frac{1}{2}\right) \\
\leq\ & 8\|H^{-1}\|^4\mathbb{E}\|\widehat{H}_n - H\|^2 + \frac{1}{2\lambda^2}(\kappa_1^{-1} + \lambda_{\min}^{-1}(H))^2\mathbb{E}\|\widehat{H}_n - H\|^2 \\
\leq\ & C\, n^{-\alpha}, && (5.59)
\end{aligned}
$$

where the third inequality follows from Markov's inequality and the last one from (5.58).

We now consider our target term. With previous results (5.58), (5.59), and Lemma 5.4.5, we can obtain that,

$$
\begin{aligned}
&\mathbb{E}\left\|\widehat{H}_n^{-1}\widehat{Q}_n\widehat{H}_n^{-1} - H^{-1}QH^{-1}\right\| \\
=\ & \mathbb{E}\left\|\widehat{H}_n^{-1}(\widehat{Q}_n - Q)\widehat{H}_n^{-1} + (H^{-1} + \widehat{H}_n^{-1} - H^{-1})Q(H^{-1} + \widehat{H}_n^{-1} - H^{-1}) - H^{-1}QH^{-1}\right\| \\
\leq\ & \mathbb{E}\left\|\widehat{H}_n^{-1}(\widehat{Q}_n - Q)\widehat{H}_n^{-1}\right\| + \mathbb{E}\left\|H^{-1}Q(\widehat{H}_n^{-1} - H^{-1})\right\| + \mathbb{E}\left\|(\widehat{H}_n^{-1} - H^{-1})QH^{-1}\right\| \\
& + \mathbb{E}\left\|(\widehat{H}_n^{-1} - H^{-1})Q(\widehat{H}_n^{-1} - H^{-1})\right\| \\
\leq\ & \kappa_1^{-2}\mathbb{E}\left\|\widehat{Q}_n - Q\right\| + 2\lambda^{-1}\|Q\|\mathbb{E}\left\|\widehat{H}_n^{-1} - H^{-1}\right\| + \|Q\|\mathbb{E}\left\|\widehat{H}_n^{-1} - H^{-1}\right\|^2 \\
\leq\ & Cn^{-\alpha/2},
\end{aligned}
$$

which completes the proof. □

Theorem 5.4.4 establishes the consistency and the rate of the convergence of our proposed covariance matrix estimator $\widehat{H}_n^{-1}\widehat{Q}_n\widehat{H}_n^{-1}$. Given Theorems 5.3.10 and 5.4.4, a confidence interval of the projected true parameter $w^\top\theta^\star$ for any $w \in \mathbb{R}^d$ can be constructed via a projection of $\overline{\theta}_n$ and $\widehat{H}_n^{-1}\widehat{Q}_n\widehat{H}_n^{-1}$ onto w. Specifically, for a pre-specified confidence level q and the corresponding z-score $z_{q/2}$, we can obtain an asymptotic exact confidence interval as $n \to \infty$,

$$\mathbb{P}\left\{ w^\top \theta^\star \in \left[w^\top \overline{\theta}_n - \frac{z_{q/2}}{\sqrt{n}} \sqrt{w^\top \widehat{H}_n^{-1} \widehat{Q}_n \widehat{H}_n^{-1} w}, \quad w^\top \overline{\theta}_n + \frac{z_{q/2}}{\sqrt{n}} \sqrt{w^\top \widehat{H}_n^{-1} \widehat{Q}_n \widehat{H}_n^{-1} w} \right] \right\}$$
$$\to 1 - q.$$

5.4.1 Online inference without additional function-value queries

Despite the simplicity of the plug-in approach, the proposed estimator $\widehat{H}_n^{-1} \widehat{Q}_n \widehat{H}_n^{-1}$ incurs additional computational and storage cost as it requires additional function-value queries for constructing $\widehat{H}_n$. It raises a natural question: *is it possible to conduct inference only based on* (KW) *iterates* $\{\theta_i\}_{i=1,2,...}$ *without additional function-value queries*?

In this section, we provide an affirmative answer to this question, and propose an alternative online statistical inference procedure using the intermediate (KW) iterates only, without requiring any additional function-value query. Intuitively, the (AKW) estimator in (5.1) is constructed as the average of all intermediate (KW) iterates $\{\theta_i\}_{i=1}^n$. If all iterates were independent and identically distributed, the asymptotic covariance could have been directly estimated by the sample covariance matrix of the iterates $\frac{1}{n}\sum_{i=1}^n (\theta_i - \overline{\theta})(\theta_i - \overline{\theta})^\top$. Unfortunately, the (KW) iterates are far from independent and indeed highly correlated. Nevertheless, the autocorrelation structure of the iterates can be carefully analyzed and utilized to construct the estimator of $H^{-1} Q H^{-1}$.

In this paper, we adopt an alternative approach to take more advantage of the autocorrelation structure by leveraging the techniques from robust testing literature [1, 124, 136]. Such an estimator is often referred to as the Fixed Bandwidth Heteroskedasticity and Autocorrelation Robust estimator (*fixed-b* HAR) in the econometrics literature. The *fixed-b* HAR estimator is able to overcome the series correlation and heteroskedasticity in the error terms for the OLS estimates of the linear regression (e.g. [124]). For the (RM) scheme, [136] utilized and generalized this technique to construct an online statistical inference procedure, and refer to this method as the *random scaling* method.

In particular, we present the following theorem based on a functional extension of the distributional analysis of the intermediate (KW) iterates $\{\theta_t\}$ as a stochastic process.

Theorem 5.4.7. *For any $w \in \mathbb{R}^d$, under the assumptions in Theorem 5.3.10, we have*

$$\sqrt{n}\frac{w^\top(\overline{\theta}_n - \theta^\star)}{\sqrt{w^\top V_n w}} \Longrightarrow \frac{W_1}{\sqrt{\int_0^1 (W_r - rW_1)^2\, dr}}, \tag{5.60}$$

where $V_n = \frac{1}{n^2}\sum_{i=1}^n i^2(\overline{\theta}_i - \overline{\theta}_n)(\overline{\theta}_i - \overline{\theta}_n)^\top$, and $\overline{\theta}_i = \frac{1}{i}\sum_{\ell=1}^i \theta_\ell$ is the average of iterates up to the i-th iteration, and $\{W_t\}_{t\geq 0}$ is the standard one-dimensional Brownian motion.

Proof. We first show that we can extend our result in Theorem 5.3.10 to the following form,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nr\rfloor} \theta_i \Longrightarrow \Sigma^{1/2}\boldsymbol{W}_r, \quad r \in [0,1].$$

where $\Sigma = H^{-1}QH^{-1}$ and $\boldsymbol{W}_r$ is a d-dimensional vector of independent standard Brownian motions on $[0,1]$. For any $r \in [0,1]$, we consider the following partial summation process,

$$\overline{B}_n(r) = \frac{1}{n}\sum_{i=1}^{\lfloor nr\rfloor} \Delta_i,$$

where $\Delta_i = \theta_i - \theta^\star = \theta_i$. Now consider the following alternative partial summation process,

$$\overline{B}_n'(r) = \frac{1}{n}\sum_{i=1}^{\lfloor nr\rfloor} \Delta_i',$$

where

$$\Delta_i' = \Delta_{i-1}' - \eta_i H\Delta_{i-1}' + \eta_n(\xi_n + \gamma_n + \varepsilon_n), \quad \Delta_0' = \Delta_0 = \theta_0.$$

From Theorem 2 in [162], we know that $\sqrt{n}\sup_r |\overline{B}'_n(r) - \overline{B}_n(r)| = o_p(1)$. Now we consider the weak convergence of $\overline{B}'_n(r)$ instead. Using the decomposition below,

$$\sqrt{n}\overline{B}'_n(r) = \frac{1}{\sqrt{n}\lfloor nr\rfloor\eta_{\lfloor nr\rfloor}}\theta_0 + \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nr\rfloor} H^{-1}(\xi_n + \gamma_n + \varepsilon_n) + \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nr\rfloor} w_i^{\lfloor nr\rfloor}(\xi_n + \gamma_n + \varepsilon_n),$$

where $1/\sqrt{n}\sum_{i=1}^{n}\|w_i^n\| \to 0$. Using the result from Lemma 5.3.6, the first and the third terms on the RHS are $o_p(1)$. Combining Theorem 4.2 from [91] and Equation (5.30), we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nr\rfloor} H^{-1}(\xi_n + \gamma_n + \varepsilon_n) \Longrightarrow \Sigma^{1/2}\boldsymbol{W}_r.$$

Therefore, for any $w \in \mathbb{R}^d$, we have

$$C_n(r) = \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nr\rfloor} w^\top\theta_i \Rightarrow w^\top(w^\top\Sigma w)^{1/2}W_r, \quad r \in [0,1].$$

Here W_r is the standard one dimensional Brownian motion. In addition,

$$w^\top V_n w = \frac{1}{n}\sum_{i=1}^{n}\left[C_n\left(\frac{i}{n}\right) - \frac{i}{n}C_n(1)\right]\left[C_n\left(\frac{i}{n}\right) - \frac{i}{n}C_n(1)\right]^\top.$$

Notice that $w^\top(\overline{\theta}_n) = \frac{1}{\sqrt{n}}C_n(1)$, and

$$n\frac{(w^\top\overline{\theta}_n)^2}{w^\top V_n w} \Rightarrow \frac{W_1^2}{\int_0^1 (W_r - rW_1)^2 dr},$$

using the continuous mapping theorem. □

As an important special case, when $w = e_k$ for $k = 1, 2, \ldots, d$, we have the convergence

Quantile	90%	95%	97.5%	99%
[1] Table 1	3.875	5.323	6.747	8.613

Table 5.1: Cumulative probability table of the limiting distribution.

in each coordinate to the following pivotal limiting distribution,

$$\frac{\sqrt{n}(\overline{\theta}_{n,k} - \theta^\star_k)}{\sqrt{V_{n,kk}}} \Longrightarrow \frac{W_1}{\sqrt{\int_0^1 (W_r - rW_1)^2 \, dr}}, \tag{5.61}$$

For the asymptotic distribution defined on the right hand side in (5.61), we repeat the quantiles of the distribution published by [1] in Table 5.1[3]. Combining the asymptotic results in (5.61) and Table 5.1, we can construct coordinate-wise confidence intervals for the true parameter $\theta^\star$. In addition, as

$$V_n = \frac{1}{n^2}\sum_{i=1}^n i^2(\overline{\theta}_i - \overline{\theta}_n)(\overline{\theta}_i - \overline{\theta}_n)^\top = \frac{1}{n^2}\sum_{i=1}^n i^2\overline{\theta}_i\overline{\theta}_i^\top - \frac{2}{n^2}\overline{\theta}_n\sum_{i=1}^n i^2\overline{\theta}_i^\top + \frac{1}{n^2}\sum_{i=1}^n i^2\overline{\theta}_n\overline{\theta}_n^\top \tag{5.62}$$

can be constructed in an online fashion via the iterative updates of the matrix $\sum_{i=1}^n i^2\overline{\theta}_i\overline{\theta}_i^\top$ and the vector $\sum_{i=1}^n i^2\overline{\theta}_i$, the proposed online inference procedure only requires one pass over the data.

5.4.2 Finite-difference stochastic Newton method

As a by-product and an application, the online finite-difference estimator of Hessian in (5.54) enables us to develop the (KW) version of the stochastic Newton's method. Existing literature that handles the (RM) version of the stochastic Newton's method traces back to [171]. Given

[3] Since the distribution on the right hand side of (5.60) is symmetric, we only provide one-side quantiles in the table.

an initial point θ_0, the `(KW)` stochastic Newton's method has the following updating rule,

$$\theta_n = \theta_{n-1} - \frac{1}{n} \widehat{H}_{n-1}^{-1} \widehat{g}_{h_n, v_n}(\theta_{n-1}; \zeta_n), \tag{5.63}$$

Here $\widehat{H}_n^{-1}$ a recursive estimator of H^{-1}. We modify the thresholding Hessian estimator $\widehat{H}_n$ in (5.54) as follows. Let $U\widetilde{\Lambda}_n U^\top$ be the eigenvalue decomposition of $\widetilde{H}_n$ in (5.39), and define

$$\widehat{H}_n = U\widehat{\Lambda}_n U^\top, \;\; \widehat{\Lambda}_{n,kk} = \max\left\{\kappa_1, \min\left\{\kappa_2, \widetilde{\Lambda}_{n,kk}\right\}\right\}, \; k = 1, 2, \ldots, d, \tag{5.64}$$

for some constants $0 < \kappa_1 < \lambda < L_f < \kappa_2$, where λ, L_f are defined in Assumption 5.3.1.

Theorem 5.4.8. *Under the assumptions in Theorem 5.3.10, the Hessian estimator $\widehat{H}_n$ in* (5.64) *converges in probability to the empirical Hessian matrix H. The stochastic Newton estimator θ_n in* (5.63) *converges to $\theta^\star$ almost surely and has the following limiting distribution,*

$$\sqrt{n}\left(\theta_n - \theta^\star\right) \Longrightarrow \mathcal{N}\left(0, H^{-1} Q H^{-1}\right), \tag{5.65}$$

for the same Q as in Theorem 5.3.10.

Theorem 5.4.8 states that the final iterate of the `(KW)` stochastic Newton method (5.63) entails the same asymptotic distribution as the averaged `(AKW)` estimator (5.1). In contrast to `(AKW)`, (5.63) leverages additional Hessian information to achieve the asymptotic normality and efficiency. Nevertheless, the numerical implementation of the `(KW)` stochastic Newton's method requires to update a Hessian estimator $\widehat{H}_n$ in all iterations, which demands significant additional computation unless such an estimator is yet computed and maintained along the procedure for other purposes.

Proof. Notice that

$$\theta_n = \theta_{n-1} - \frac{1}{n} H_{n-1}^{-1} \nabla F(\theta_{n-1}) + \frac{1}{n} \overline{H}_{n-1}^{-1} \left(\xi_n + \gamma_n + \varepsilon_n \right).$$

We now show that Lemma 5.3.6 holds under $\alpha = 1$. Following from the same logic in Lemma 5.3.6, we can show that there exists some universal constant $n_0 > 0$, such that for all $n > n_0$, and some constants C_1, C_2,

$$\mathbb{E}_{n-1} \|\theta_n\|^2 \le \left(1 - \frac{C_1}{n} \right) \|\theta_{n-1}\|^2 + C_2 n^{-2}. \tag{5.66}$$

Therefore, $\theta_n \to 0$ almost surely by martingale convergence theorem [168]. Now we consider the convergence rate of θ_n.

Using the proof in Lemma 5.3.6, we can show that

$$\mathbb{E}_{n-1} \|\theta_n\|^2 \le C \left(n^{-C_1/2} + n^{-1} \right). \tag{5.67}$$

Similarly,

$$\mathbb{E}_{n-1} \|\theta_n\|^{2+\delta} \le C \left(n^{-C_1/2} + n^{-(1+\delta)} \right). \tag{5.68}$$

Now we consider the limiting distribution.

$$\begin{aligned} \theta_n &= \theta_{n-1} - \frac{1}{n} H^{-1} \nabla F(\theta_{n-1}) - \frac{1}{n} \left(H_{n-1}^{-1} - H^{-1} \right) \nabla F(\theta_{n-1}) + \frac{1}{n} H_{n-1}^{-1} \left(\xi_n + \gamma_n + \varepsilon_n \right) \\ &= \left(1 - \frac{1}{n} \right) \theta_{n-1} - \frac{1}{n} H^{-1} \delta_n - \frac{1}{n} \left(H_{n-1}^{-1} - H^{-1} \right) \nabla F(\theta_{n-1}) + \frac{1}{n} H_{n-1}^{-1} \left(\xi_n + \gamma_n + \varepsilon_n \right), \end{aligned}$$

where $\delta_n = \nabla F(\theta_{n-1}) - H\theta_{n-1}$. By induction, we can find that

$$\begin{aligned}\theta_n = &\frac{1}{n}\sum_{k=0}^{n-1} H_k^{-1}\varepsilon_{k+1} + \frac{1}{n}\sum_{k=0}^{n-1} H_k^{-1}\left(\xi_{k+1}+\gamma_{k+1}\right)\\ &-\frac{1}{n}H^{-1}\sum_{k=0}^{n-1}\delta_{k+1} - \frac{1}{n}\sum_{k=0}^{n-1}\left(H_k^{-1}-H^{-1}\right)\nabla F(\theta_k).\end{aligned}$$

The last three terms in the RHS above all converge to zero due to Assumption 5.3.5. Now we only need to show that $\frac{1}{\sqrt{n}}\sum_{k=0}^{n-1} H_k^{-1}\varepsilon_{k+1}$ converges to a normal distribution. Consider

$$\mathbb{E}_k\left[H_k^{-1}\varepsilon_{k+1}\varepsilon_{k+1}^\top H_k^{-1}\right] = H_k^{-1}\mathbb{E}_k\left[\varepsilon_{k+1}\varepsilon_{k+1}^\top\right]H_k^{-1},$$

recall that in (5.33) we have shown that $\mathbb{E}_k\left[\varepsilon_{k+1}\varepsilon_{k+1}^\top\right]$ converges almost surely to Q. Therefore, by Assumption 5.3.5, $\mathbb{E}_k\left[H_k^{-1}\varepsilon_{k+1}\varepsilon_{k+1}^\top H_k^{-1}\right]$ converges in probability to $H^{-1}QH^{-1}$.

Obviously, we can get the tail bound similar to (5.34) and by martingale central limit theorem [63, Theorem 2.1.9],

$$\frac{1}{\sqrt{n}}\sum_{k=0}^{n-1} H_k^{-1}\varepsilon_{k+1} \Longrightarrow \mathcal{N}\left(0, H^{-1}QH^{-1}\right).$$

□

5.5 Numerical experiments

In this numerical section, we first investigate the empirical performance of the proposed inference procedures and their corresponding coverage rates. We consider linear regression and logistic regression models (Examples 5.2.1–5.2.2) where $\{x_i, y_i\}_{i=1}^n$ is an *i.i.d.* sample with the covariate $x \sim \mathcal{N}(0,\Sigma)$ and the response $y \in \mathbb{R}$. The true model parameter $\theta^\star \in \mathbb{R}^d$ is selected uniformly from the unit sphere before the experiments. For both models, we consider two different structures of the covariance matrices Σ: identity matrix I_d and equicorrelation covariance matrix (Equicorr in the tables), i.e., $\Sigma_{k\ell} = 0.2$ for all $k \neq \ell$ and $\Sigma_{kk} = 1$. The parameter α in the step size is specified to $\alpha = 0.501$. The variance of noise ϵ in the linear regression model (Example 5.2.1) is set to $\sigma^2 = 0.2$.

d	Σ	Estimation error		Average coverage rate		
		Parameter	Plug-in Cov.	Plug-in	Fixed-b	Oracle
Linear regression						
5	Identity	0.0031 (0.0010)	0.0384 (0.0106)	0.9448 (0.1035)	0.9464 (0.1174)	0.9436 (0.1040)
	Equicorr	0.0035 (0.0012)	0.0342 (0.0092)	0.9428 (0.1096)	0.9488 (0.1195)	0.9412 (0.1102)
20	Identity	0.0135 (0.0023)	0.1126 (0.0190)	0.9319 (0.0594)	0.9039 (0.0657)	0.9288 (0.0616)
	Equicorr	0.0172 (0.0029)	0.1124 (0.0199)	0.9194 (0.0644)	0.9014 (0.0674)	0.9170 (0.0656)
100	Identity	0.0748 (0.0062)	0.5707 (0.0648)	0.9309 (0.0261)	0.7501 (0.0397)	0.9012 (0.0336)
	Equicorr	0.0921 (0.0076)	0.5615 (0.0647)	0.9331 (0.0250)	0.7435 (0.0418)	0.9044 (0.0320)
Logistic regression						
5	Identity	0.0265 (0.0115)	0.0587 (0.0434)	0.9432 (0.1219)	0.9360 (0.1685)	0.9440 (0.1148)
	Equicorr	0.0299 (0.0131)	0.0697 (0.0514)	0.9440 (0.1196)	0.9364 (0.1566)	0.9464 (0.1207)
20	Identity	0.0728 (0.0124)	0.1030 (0.0250)	0.9418 (0.0532)	0.8956 (0.1156)	0.9403 (0.0540)
	Equicorr	0.0799 (0.0146)	0.1213 (0.0359)	0.9383 (0.0577)	0.8949 (0.1106)	0.9369 (0.0561)
100	Identity	0.2440 (0.0211)	0.5236 (0.1646)	0.9673 (0.0193)	0.7022 (0.0838)	0.9082 (0.0295)
	Equicorr	0.2867 (0.0253)	0.7685 (0.2933)	0.9608 (0.0185)	0.6950 (0.0728)	0.9041 (0.0314)

Table 5.2: Estimation errors and averaged coverage rates of the proposed algorithm with search direction (I) and two function queries ($m = 1$).

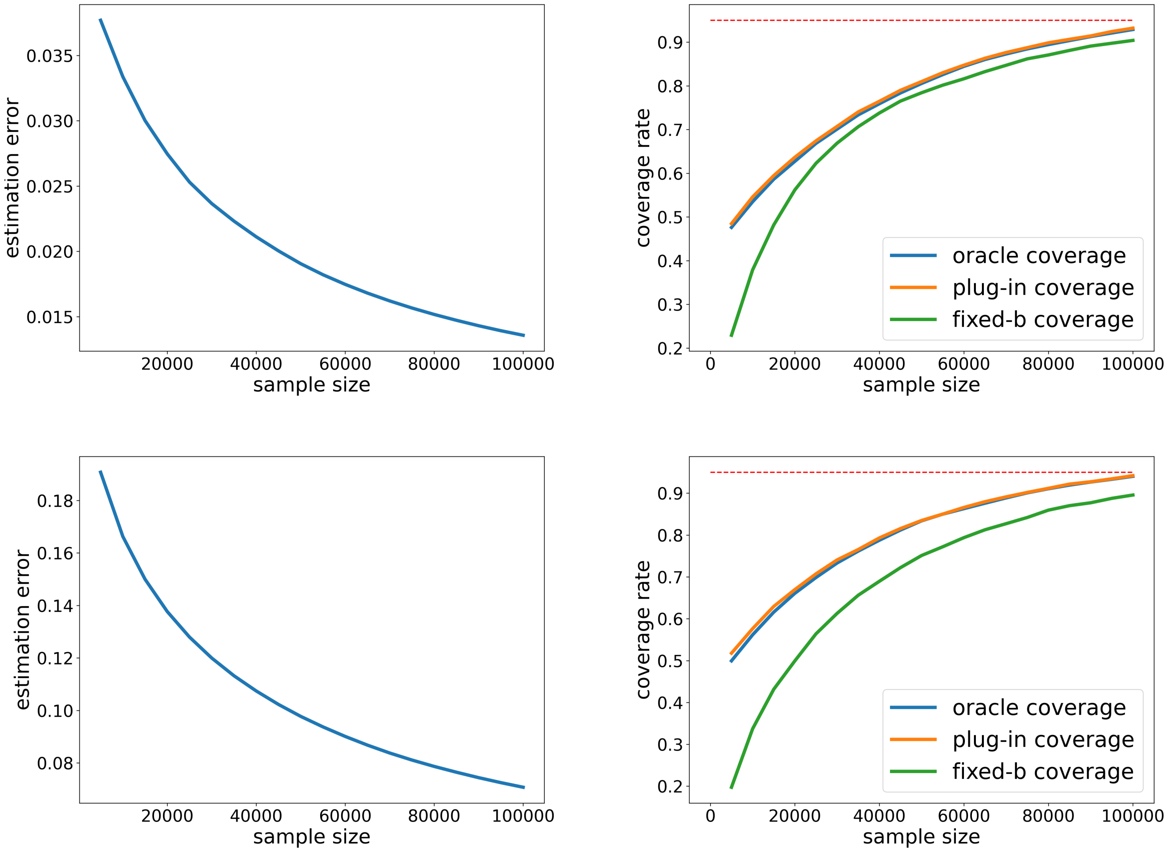


Figure 5.2: Convergence of the parameter estimation error $\|\overline{\theta}_n - \theta^{\star}\|$ and coverage rates v.s. the sample size n.

5.5.1 *Estimation errors of* (AKW) *and the performance of inference procedures*

We set the sample size $n = 10^5$ and the parameter dimension $d = 5, 20, 100$. We first report the performance of (AKW) with the search direction uniformly sampled from the natural basis, referred to as (I) in Section 5.3.1. In Table 5.2, we present the estimation error for the parameter $\theta^\star$ in the Euclidean norm and the relative error of the plug-in covariance estimator in the spectral norm (see the first two columns), with 100 Monte-Carlo simulations,

$$\frac{\|\overline{\theta}_n - \theta^\star\|}{\|\theta^\star\|}, \quad \frac{\|\widehat{H}_n^{-1}\widehat{Q}_n\widehat{H}_n^{-1} - H^{-1}QH^{-1}\|}{\|H^{-1}QH^{-1}\|}. \tag{5.69}$$

Corresponding standard errors are reported in the brackets. We compare the plug-in covariance estimator (plug-in) based inference (5.39) and fixed-b HAR (fixed-b) based inference (5.61).

Next, we set the nominal coverage probability as 95% and we project $\theta \in \mathbb{R}^d$ onto $w = (1, 1, \ldots, 1)^\top/\sqrt{d}$ to construct confidence intervals. In particular, we report the performance of the confidence interval with the average coverage rate and the average length of the intervals for (1) the plug-in covariance matrix estimator[4] (5.38) and (2) the fixed-b HAR procedure in (5.62). As an oracle benchmark, we also report the length of the confidence interval with respect to the true covariance matrix $H^{-1}QH^{-1}$ of the plug-in approach and the corresponding coverage rate. As shown from Table 5.2, the coverage rate of the plug-in covariance estimator and the oracle coverage rates are very close to the desired 95% coverage, while the fixed-b HAR approach is comparable in small dimension $d = 5, 20$ but has lower coverage rates for the large dimension $d = 100$. The average lengths of both methods are comparable to the lengths derived by the true limiting covariance.

[4] Here we use updating probability $p = 1$ for the plug-in estimation. In other words, $d^2 + 1$ queries of function values are obtained at each iteration. In Section 5.5.2 below, we extend the comparison for different p.

Then, we fix $d = 20$ and the identity design matrix $\Sigma = I$. We present in Figure 5.2 the parameter estimation error $\|\overline{\theta}_n - \theta^\star\|$ and the coverage rates as the sample size n grows. Plots (a) to (b) show the cases of linear regression and plots (c) to (d) show the cases of logistic regression. Dashed lines in plots (b) and (d) correspond to the nominal 95% coverage. In subplots (b) and (d) of Figure 5.2, we show the coverage rates for the plug-in and fixed-b HAR approaches as the sample size n increases. As one can see, coverage rates of the plug-in approach almost match the oracle case using the true asymptotic covariance matrix $H^{-1}QH^{-1}$. For the linear regression case, the plug-in and fixed-b HAR approaches are comparable. For the logistic regression case, the coverage rate of the fixed-b HAR inference procedure is slightly inferior than that of the plug-in method. On the other hand, the fixed-b HAR approach does not require additional function queries for the explicit estimation of the Hessian matrix.

Estimator	Comp.	Estimation error		Average coverage rate		Average length	
	time	Hessian	Cov.	Estimator	Oracle	Estimator	Oracle
Identity							
Plug-in $p = 1/400$	4.74s	0.1780	0.3179	0.8965	0.9288	3.6570	3.5065
		(0.0115)	(0.0423)	(0.0696)	(0.0616)	(0.0195)	-
Plug-in $p = 1/20$	25.42s	0.0393	0.1503	0.9244	0.9288	3.5511	3.5065
		(0.0043)	(0.0282)	(0.0665)	(0.0616)	(0.0169)	-
Plug-in $p = 1$	510.53s	0.0271	0.1126	0.9319	0.9288	3.5337	3.5065
		(0.0021)	(0.0190)	(0.0594)	(0.0616)	(0.0164)	-
Fixed-b	2.82s	-	-	0.9039	0.9288	3.7424	3.5065
		-	-	(0.0657)	(0.0616)	(0.4292)	-
Equicorr							
Plug-in $p = 1/400$	4.78s	0.0381	0.4211	0.8815	0.9170	4.4547	4.2753
		(0.0043)	(0.0421)	(0.0753)	(0.0656)	(0.0304)	-
Plug-in $p = 1/20$	25.60s	0.0117	0.1540	0.9122	0.9170	4.3489	4.2753
		(0.0025)	(0.0271)	(0.0691)	(0.0656)	(0.0293)	-
Plug-in $p = 1$	512.07s	0.0082	0.1124	0.9194	0.9170	4.3140	4.2753
		(0.0018)	(0.0199)	(0.0644)	(0.0656)	(0.0207)	-
Fixed-b	2.85s	-	-	0.9014	0.9170	4.5582	4.2753
		-	-	(0.0674)	(0.0656)	(0.5681)	-

Table 5.3: Results for the linear regression model.

Estimator	Comp.	Estimation error		Average coverage rate		Average length	
	time	Hessian	Cov.	Estimator	Oracle	Estimator	Oracle
Identity							
Plug-in $p = 1/400$	5.70s	0.1812	0.3293	0.9039	0.9403	5.0715	4.8374
		(0.0281)	(0.0792)	(0.0501)	(0.0540)	(0.2071)	-
Plug-in $p = 1/20$	32.32s	0.0737	0.1636	0.9330	0.9403	4.9599	4.8374
		(0.0114)	(0.0393)	(0.0593)	(0.0540)	(0.1833)	-
Plug-in $p = 1$	643.86s	0.0597	0.1030	0.9418	0.9403	4.8751	4.8374
		(0.0093)	(0.0250)	(0.0532)	(0.0540)	(0.1973)	-
Fixed-b	3.13s	-	-	0.8956	0.9403	5.1763	4.8374
		-	-	(0.1156)	(0.0540)	(0.4362)	-
Equicorr							
Plug-in $p = 1/400$	5.75s	0.0993	0.3620	0.8880	0.9369	5.9456	5.6356
		(0.0287)	(0.0992)	(0.0540)	(0.0561)	(0.1716)	-
Plug-in $p = 1/20$	32.53s	0.0356	0.1441	0.9288	0.9369	5.7766	5.6356
		(0.0120)	(0.0440)	(0.0599)	(0.0561)	(0.1556)	-
Plug-in $p = 1$	645.56s	0.0240	0.1213	0.9383	0.9369	5.6873	5.6356
		(0.0101)	(0.0359)	(0.0577)	(0.0561)	(0.1715)	-
Fixed-b	3.17s	-	-	0.8949	0.9369	5.7532	5.6356
		-	-	(0.1106)	(0.0561)	(0.4064)	-

Table 5.4: Results for the logistic regression model.

5.5.2 Comparison of the inference procedures

In this subsection, we provide detailed comparisons of different inference procedures. Specifically, we fix dimension $d = 20$, and compare the performance for the plug-in and fixed-b HAR schemes. For plug-in estimators, at each iteration, we update the Hessian estimator $\widehat{H}_n$ in (5.54) using (5.38) with probability p chosen from 1, d^{-1}, d^{-2}. The fixed-b scheme is updated by (5.62). We report the computation time, the estimation error of the Hessian matrix, and the average coverage rate and length of these candidates based on 100 replications. The computation time is recorded in a simulation environment running Python 3.8 with a single 10-core Apple M1 Max chip.

The simulation for linear regression and logistic regression is given below in Tables 5.3–5.4, respectively. Corresponding standard errors are reported in the brackets. We compare the plug-in covariance estimator (plug-in) based inference (5.54) using $p = 1, 1/20, 1/400$ and fixed-b HAR (fixed-b) based inference (5.62). As can be referred from the two tables, the fixed-b HAR approach gives the fastest execution, due to the fact that no additional function queries are required for Hessian matrix computation. The fixed-b HAR method is even faster than the case of $p = d^{-2}$ where we update only one entry (in expectation) for the Hessian matrix in each `(KW)` step. Among plug-in cases, the performance of inference improves as p increases, and it achieves a relatively more reliable coverage for $p \geq d^{-1}$, (i.e., at least d entries (in expectation) receive updates for the Hessian estimator per `(KW)` step), with a significant cost of the computation time. In practice, we would recommend the fixed-b HAR method for those computation-sensitive tasks, and the plug-in method with Hessian sampling probability $p \geq d^{-1}$ in less computation-sensitive tasks.

d	$\mathcal{P}_v$	Estimation error		Average coverage rate		Average length	
		Parameter	Plug-in Cov.	Plug-in	Oracle	Plug-in	Oracle
5	`(I)`	0.0265	0.0587	0.9432	0.9440	3.1136	3.1078
		(0.0115)	(0.0434)	(0.1219)	(0.1148)	(0.8648)	-
	`(S)`	0.0264	0.0599	0.9396	0.9376	3.0639	3.0625
		(0.0124)	(0.0453)	(0.1276)	(0.1250)	(0.8211)	-
	`(G)`	0.0312	0.0718	0.9412	0.9420	3.6304	3.6237
		(0.0139)	(0.0498)	(0.1193)	(0.1176)	(0.9770)	-
20	`(I)`	0.0728	0.1030	0.9418	0.9403	4.8751	4.8374
		(0.0124)	(0.0250)	(0.0532)	(0.0540)	(0.6441)	-
	`(S)`	0.0711	0.1017	0.9438	0.9419	4.8414	4.8156
		(0.0116)	(0.0246)	(0.0523)	(0.0524)	(0.6322)	-
	`(G)`	0.0749	0.1054	0.9427	0.9423	5.0873	5.0507
		(0.0121)	(0.0248)	(0.0563)	(0.0523)	(0.6654)	-
100	`(I)`	0.2440	0.5236	0.9673	0.9082	12.0661	10.3175
		(0.0211)	(0.1646)	(0.0193)	(0.0295)	(1.0106)	-
	`(S)`	0.2353	0.5122	0.9605	0.9145	13.1366	11.1788
		(0.0205)	(0.1530)	(0.0201)	(0.0358)	(1.0891)	-
	`(G)`	0.2357	0.5147	0.9614	0.9161	13.2836	11.2901
		(0.0202)	(0.1531)	(0.0205)	(0.0380)	(1.0929)	-

Table 5.5: Comparison among different direction distributions $\mathcal{P}_v$.

5.5.3 Choices of the search direction distribution

In this subsection, we compare the results for different directions $\mathcal{P}_v$. We report the results for the logistic regression model with the identity design matrix $\Sigma = I$ in Table 5.5. Detailed

$m;\Sigma$	$\mathcal{P}_v$	Estimation error		Average coverage rate		Average length	
		Parameter	Plug-in Cov.	Plug-in	Oracle	Plug-in	Oracle
10; Identity	`(I+WOR)`	0.0916	0.1972	0.9547	0.9342	3.7013	3.4794
		(0.0103)	(0.1053)	(0.0225)	(0.0330)	(0.2970)	-
	`(I+WR)`	0.0947	0.2004	0.9551	0.9353	3.8800	3.6383
		(0.0106)	(0.1025)	(0.0215)	(0.0310)	(0.3053)	-
	`(S)`	0.0958	0.2134	0.9552	0.9320	3.8893	3.6352
		(0.0118)	(0.1172)	(0.0219)	(0.0368)	(0.3054)	-
10; Equicorr	`(I+WOR)`	0.1184	0.2581	0.9404	0.9126	3.6432	3.3700
		(0.0122)	(0.1278)	(0.0252)	(0.0382)	(0.2240)	-
	`(I+WR)`	0.1235	0.2828	0.9431	0.9125	3.8352	3.5234
		(0.0145)	(0.1573)	(0.0266)	(0.0437)	(0.2498)	-
	`(S)`	0.1224	0.2753	0.9435	0.9135	3.8225	3.5165
		(0.0144)	(0.1501)	(0.0259)	(0.0422)	(0.2614)	-
100; Identity	`(I+WOR)`	0.0261	0.0531	0.9455	0.9438	0.8978	0.8938
		(0.0022)	(0.0135)	(0.0297)	(0.0305)	(0.0290)	-
	`(I+WR)`	0.0333	0.0568	0.9455	0.9441	1.4037	1.3948
		(0.0030)	(0.0196)	(0.0253)	(0.0262)	(0.0803)	-
	`(S)`	0.0334	0.0556	0.9458	0.9439	1.4034	1.3941
		(0.0028)	(0.0199)	(0.0231)	(0.0247)	(0.0816)	-
100; Equicorr	`(I+WOR)`	0.0328	0.0664	0.9490	0.9441	0.9056	0.8971
		(0.0035)	(0.0199)	(0.0339)	(0.0356)	(0.0493)	-
	`(I+WR)`	0.0453	0.0823	0.9494	0.9444	1.4183	1.3946
		(0.0046)	(0.0322)	(0.0240)	(0.0270)	(0.0766)	-
	`(S)`	0.0451	0.0821	0.9497	0.9449	1.4157	1.3930
		(0.0048)	(0.0321)	(0.0249)	(0.0270)	(0.0777)	-

Table 5.6: Comparison among different sampling schemes.

specification of (I),(S),(G) can be referred to Section 5.3.1. We consider the logistic regression model with design matrix $\Sigma = I$, and the (AKW) estimators are computed based on the case of two function queries ($m = 1$). Corresponding standard errors are reported in the brackets. Table 5.5 suggests the (AKW) algorithms with search directions (I), (S), (G) achieve similar performance for parameter estimation error and average coverage rates, while the average confidence intervals of (G) are generally larger. The observations in the numerical experiments match our Proposition 5.3.12.

5.5.4 Multi-query (AKW) *estimator*

We further conduct experiments for the (KW) algorithm with multiple function-value queries ($m > 1$) and compare the performance of $m = 10, 100$ using different search directions with sampling schemes (I+WR), (I+WOR), and (S). We note that (I+WR) and (I+WOR) refer to the uniform sampling from natural basis with and without replacement, respectively; and (S) refers to the uniform sampling from the sphere. We report the results of the logistic regression model with dimension $d = 100$ in Table 5.6. Corresponding standard errors are reported in the brackets.

When $m = 10$, the (KW) algorithm using all three sampling schemes achieves similar performance in both estimation and inference. When $m = 100$, the algorithm with (I+WOR) achieves better performance than the other two sampling schemes by constructing around 30% shorter confidence intervals on average while achieving comparable coverage rates.

We further present in Figure 5.3 the estimation error of the parameters and covariance matrices when we increase the function-query complexity m. The numerical results matches the magnitudes of Q with regard to different m in Theorems 5.3.17–5.3.18, which could help practitioners choose an appropriate m to balance the accuracy and computational cost. We report the logistic regression results with with $n = 10^5$, $d = 100$, and the identity design matrix $\Sigma = I$ in Figure 5.3. The x-axis is the number of function evaluations per step (i.e.,

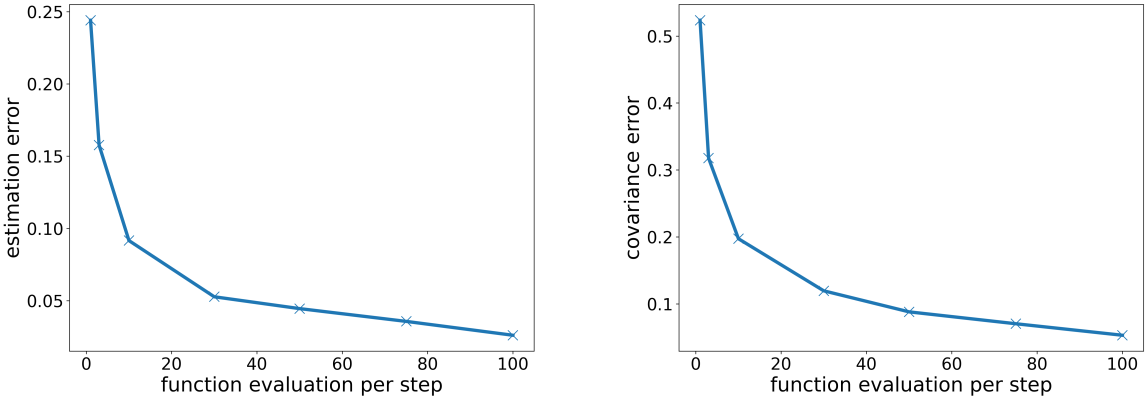


Figure 5.3: The parameter estimation error and the relative covariance estimation error (see (5.69)) for multiple function-value evaluations.

$m + 1)$.

5.6 Conclusion and future work

In this paper, we investigate the statistical inference problem for the Kiefer-Wolfowitz stochastic optimization algorithm with random search directions. We show the asymptotic normality for the (KW)-type estimators and provide consistent estimators of the asymptotic covariance matrix to facilitate the inference. Our theoretical analysis provides a comprehensive comparison on the impact of different random search directions, the number of multi-query evaluations, and sampling schemes. Our findings are validated by numerical experiments.

For future works, our results and estimation methods may be potentially useful to understand asymptotic behaviors of other gradient-free variants of stochastic optimization algorithms, e.g. moment-adjusted stochastic gradients [139], stochastic optimization under constraints [62], high dimensional stochastic algorithms [35, 177], and SGD in contextual bandit settings [40].

CHAPTER 6

ONLINE STATISTICAL INFERENCE FOR CONTEXTUAL BANDITS VIA STOCHASTIC GRADIENT DESCENT

This is a joint work with He Li, Yichen Zhang, and Xi Chen.

6.1 Introduction

Following the seminal work of [167], the stochastic multi-armed bandit problem has been studied extensively in the literature, where an agent aims to make optimal decisions sequentially among multiple arms and only the selected arm reveals rewards consequently. As the agent's choice is often influenced by additional covariates, also referred to as contexts, contextual bandit problems have gained renewed attention in the past decades [201, 132, etc.]. With the development of internet and data technology, contextual bandit algorithms play an important role in sequential decision-making applications, such as online advertisement [138], precision medicine [125], e-commence [163, 45], and public policy [120]. Such decisions are often referred to as recommendations, treatments, interventions, and public orders, while the rewards can be healthcare outcomes, welfare utility, revenue as well as any measure of satisfaction of decisions.

Most contextual bandit algorithms are built with the goal of learning the best action under different contexts. In sequential settings, it is often formulated as minimizing the expected cumulative regret that the practitioner would have received if she knows the optimal action. While the importance of this regret minimization is undisputed, *reliable uncertainty quantification* of the learned decision rule is evidently important in many featured applications. For example, in a personalized medicine application where the intervention decision is to choose t''he best medical treatment to optimize some health outcome, the risk for the selected treatment plays a critical and even sometimes life-threatening role in decision-making.

Such examples call for the crucial need for a valid and reliable statistical inference procedure accompanying the decision-making process to provide guidance on policy interventions. Inferential studies help not only prompt risk alerts in recommendations, but also gain scientific knowledge of questions such as the effectiveness of medicines.

Particularly, consider a linear contextual bandit environment where the observed data at each decision point t is a triplet $\zeta_t = (X_t, A_t, Y_t)$ for all $t \geq 1$, consisting of covariate X_t, action A_t, and reward $Y_t = X_t^\top \theta^*_{A_t} + \epsilon_t$ where $\theta^*_{A_t} \in \mathbb{R}^d$ is unknown parameters of interest governed by a finite set of actions $\mathcal{A}$, and $\epsilon_t \in \mathbb{R}$ is the noise under certain modeling assumptions. For illustrative simplicity, we consider a binary action space $\mathcal{A} = \{0, 1\}$ corresponding to a duplet of underlying model parameters $(\theta^*_0, \theta^*_1) \in \mathbb{R}^d \times \mathbb{R}^d$, and actions $A_t \in \mathcal{A}$ are selected according to a policy $A_t \sim \pi(X_t, \mathcal{H}_{t-1})$ where $\mathcal{H}_{t-1}$ denotes the trajectory of observations until time $t-1$. At the time t, a typical policy π prefers the action with a higher mean reward $X_t^\top \theta^*_a$ for $a \in \mathcal{A}$, while reserving a small probability to explore a random action to avoid potential myopic short-sighted exploitation. For example, in the widely-used ε-greedy policy,

$$\Pr\big(A_t = a \mid X_t, \theta_{0,t-1}, \theta_{1,t-1}\big) = (1-\varepsilon)\mathbb{I}\big\{a = \text{argmax}_{a\in\mathcal{A}} X_t^\top \theta_{a,t-1}\big\} + \frac{\varepsilon}{2}, \tag{6.1}$$

This procedure heavily relies on a series of estimators $(\theta_{0,t-1}, \theta_{1,t-1})$ on-the-fly, of the underlying model parameters. Despite that a return-oriented policy would undoubtedly favor the action with a higher reward, it is often as crucial to obtain the confidence of decisions, i.e., conducting statistical inference for (θ^*_0, θ^*_1) in the prescribed applications. This model of statistical inference of model parameters in decision-making problems appears recently in literature (See e.g., 39, 205, and a brief survey in Section 6.1.1 below). A typical inferential task provides a confidence interval of the underlying parameters (θ^*_0, θ^*_1) or significance levels when testing hypotheses of parameters, or its margin $\theta^*_1 - \theta^*_0$.

Since the sequential decision-making problem relies on updating the estimator for every

t throughout the horizon, it is important to provide a computationally efficient fully-online algorithm for both estimation and inference purposes. The existing literature of sequential decision-making mostly focuses on the convergence rate and efficiency, while computational efficiency and storage applicability of the estimation algorithm is often optimistically neglected. As such, they often provide online decision-making procedures governed by an offline scheme of parameter estimation. At each iteration t, an "offline" M-estimator $(\theta_{0,t}, \theta_{1,t})$ is often obtained using the sample path $\{(X_1, y_1),\ (X_1, y_2), \ldots, (X_t, y_t)\}$ up to time t. For example, when using the linear estimator, the computation cost accumulates in a non-scalable manner to at least $\mathcal{O}(T^3)$ over the entire horizon T.

To facilitate computationally efficient online inference, we adopt the stochastic gradient descent (SGD) algorithms in conducting statistical inference in fully-online decision-making. SGD, dated back to [168], has been widely used in large-scale stochastic optimization thanks to its computational and storage efficiency. Let θ_0 denote an initial estimation, the SGD iteratively updates the parameter as follows,

$$\theta_t = \theta_{t-1} - \eta_t \nabla \ell(\theta_{t-1}; \zeta_t), \tag{6.2}$$

where η_t is a positive non-increasing sequence referred to as the step-size sequence and $\nabla \ell$ is the gradient for smooth individual loss function ℓ. For the SGD update above, under the *i.i.d.* setting where $\zeta_t = (X_t, Y_t)$, the classical result by [162] uses the average $\bar{\theta}_t^{(\texttt{SGD})} = t^{-1} \sum_{s=0}^{t-1} \theta_s$ as the final estimator to accelerate the estimation. They characterize the limiting distribution and statistical efficiency of the averaged SGD, i.e.,

$$\sqrt{t}\big(\bar{\theta}_t^{(\texttt{SGD})} - \theta^*\big) \Longrightarrow \mathcal{N}\big(0, H^{(\texttt{SGD})-1} S^{(\texttt{SGD})} H^{(\texttt{SGD})-1}\big),$$

given a series of predetermined learning rates $\eta_t = \eta_0 t^{-\alpha}$ for $\eta_0 >$ and $0.5 < \alpha < 1$. Here $H^{(\texttt{SGD})}$ and $S^{(\texttt{SGD})}$ are the Hessian and Gram matrix at $\theta = \theta^*$ for some population loss function under *i.i.d.* settings. For model well-specified settings, this asymptotic covariance matrix matches the inverse Fisher information matrix and thus the resulting averaged estimator $\bar{\theta}_t^{(\mathtt{SGD})}$ is asymptotically efficient.

SGD fits well into the online decision-making scheme, as the underlying parameter (θ_0^*, θ_1^*) is the solution to the following stochastic optimization under certain modeling assumptions,

$$\theta_a^* \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \, \mathbb{E}\left[\ell\big(\theta; (X_t, Y_t)\big) \mid X_t, A_t = a\right], \quad a \in \mathcal{A}, \tag{6.3}$$

where the function $\ell \in \mathbb{R}^d \to \mathbb{R}$ will be constructed accordingly. For example, in a linear contextual bandit $Y_t = X_t^\top \theta_{A_t}^* + \epsilon_t$ with *i.i.d.* covariates X_t and mean-zero noise $\{\epsilon_t\}$, a natural choice of $\ell\big(\theta; (X_t, Y_t)\big) = \big(y_t - X_t^\top \theta\big)^2$ is the squared loss. As the outcome Y_t at every time t is adaptively collected upon the decision of action A_t, only one of the $\big(\theta_{0,t}, \theta_{1,t}\big)$ is updated. To compensate missing updates, a generalized SGD updates

$$\theta_t = \theta_{t-1} - \eta_t w_t \nabla \ell\big(\theta_{t-1}; (X_t, Y_t)\big). \tag{6.4}$$

with a weighting parameter w_t determined by the decision policy $\pi(X_t, \mathcal{H}_{t-1})$. This procedure first appeared in [41] where they used the inverse probability weighting (IPW) for an ε-greedy policy (See (`IPW`) below for the explicit form of the weight). As a consequence, the weighted stochastic gradient $w_t \nabla \ell\big(\theta_{t-1}; (X_t, Y_t)\big)$ is proved to be an unbiased estimator of a weighted population loss function where the weight is independent to the entire the historical information. While the unbiasedness property of stochastic gradient and its independence from the prior trajectory clear the technical difficulty of theoretical analysis of the asymptotic normality of the IPW-weighted ASGD estimator, IPW increases its asymptotic variance by a factor of $1/\varepsilon$. With such a factor, the proposed algorithm leads to a highly-volatile estimator in practice and entails an overly wide confidence interval while making inferential calls. Designing ameliorate decision-making algorithms to enhance the asymptotic efficiency of the estimator remains challenging yet important.

In this paper, we allow a general choice of the weighting parameter w_t in (6.4), which admits the IPW weights as a special case, and derive the explicit formula for the asymptotic distribution of the generalized-weighting ASGD algorithm, thus provides us a way to compare different choices of w_t and even optimize over w_t for some simple models. Our proposed estimator greatly improves the asymptotic efficiency over IPW-ASGD and achieves comparable efficiency as if the practitioner picks one arm steadily. This estimator helps construct narrow yet reliable confidence intervals for the underlying parameter of interest. The analysis also reveals a recommendation of optimal choices of weights w_t in certain policies. To overcome the technical challenge raised in dependent weighting parameters, we propose a new definition of the loss function, which is different from the loss function used in classical SGD literature (e.g., 44) and adaptive SGD literature [41]. We use two parameters θ and θ' to separate the effect of weighting parameters in SGD and that of decision-making procedures in the local geometric landscape of the loss function.

As a separate interest, our framework allows non-smooth loss functions such as quantile loss. In contrast to linear regression, quantile regression provides estimates of a range of conditional quantiles of the reward Y_t. Since contextual bandit problems often appear in an interactive environment, the underlying reward model is more likely to differ across the distribution of the rewards and contexts or involves outliers. Linear regression methods estimate only the mean effects which is usually an incomplete summary of the effect of exposures for certain outcomes. For example, when recommending health care interventions, associations between health care and health outcomes can be highly different among individuals at high-, median-, and low-level utilization of health care. Quantile regression finds ubiquitous applications in many fields such as operations management of business inventory and risk management of financial assets [169, 13]. Therefore, it is worth exploring the use of quantile-based objective functions in sequential decision-making problems. In this paper, we establish a general framework that allows certain nonsmooth objective functions including quantile regression.

We emphasize the technical challenges and summarize the methodology contribution and theoretical advances in the following facets.

- We study the online statistical inference of model parameters in a contextual bandit framework of sequential decision-making. We adopt the existing fully-online reweighting algorithm for SGD but extend it in two directions: for a general choice of weights and handling non-smooth loss functions via stochastic subgradient. An important example is the quantile loss functions with applications in newsvendor problems and risk management. Moreover, this example provides robustness due to the fact that the objective function is globally Lipschitz. We establish the asymptotic normality result and characterize how the asymptotic covariance depends on the weight choice.

- We show that SGD under ε-greedy policies with inverse probability weighting (IPW) in [41] suffers from an unbounded asymptotic variance when the exploration rate, ε is close to 0, i.e., the relative efficiency of adaptive models versus non-adaptive models diverges to infinity. Our proposed algorithm features a general policy with a flexible specification of the weights to avoid such deficiency and obtain a bounded relative efficiency. We further provide some practical insights into the optimal weight specification.

- Beyond the asymptotic normality of the proposed estimator, we further establish an analysis of the higher-order remainder term in its Bahadur representation. In classical *i.i.d.* SGD settings, the remainder term has the rate of $\mathcal{O}_p\big(t^{-\alpha+\frac{1}{2}} + t^{\alpha-1}\big)$. On the contrary, under the adaptive decision-making environment, the reminder term has a slower rate of $\mathcal{O}_p\big(t^{-\alpha+\frac{1}{2}} + t^{-\frac{\alpha}{4}} + t^{\alpha-1}\big)$. This slower rate can be considered as the effect of the discontinuous indicator function for the ε-greedy policy.

6.1.1 Related works

Online statistical inference for model parameters in SGD The asymptotic distribution of averaged stochastic gradient descent (ASGD) is first given in [172] in [162]. Since then, there has been a rapid growth of interest recently in conducting statistical inference for model parameters in stochastic gradient algorithms. [44] proposed two online estimators (plug-in and batch-means) in constructing estimators of limiting covariance matrix of ASGD, of which [211] extended the batch-means to overlapped batches. [72] proposed a perturbation-based resampling procedure to conduct inference for ASGD. [189] proposed a tree-structured inference scheme to construct confidence intervals. [136, 137] generalized the results in [162] to a functional central limit theorem and proposed an online inference procedure called random-scaling for smooth objectives and quantile regression, respectively.

Statistical inference in online decision-making problems [39] studied the asymptotic distribution of the parameters under a linear contextual bandit framework. [56, 122] considered adaptive linear regression where the vector contexts are correlated over time. [205, 206] conducted statistical inference for M-estimators in contextual bandit and non-Markovian environments. [94] used multiplier bootstrap to offer uncertainty quantification for exploration in the bandit settings. [41] conducted statistical inference under the contextual bandit settings via SGD. There also exists related statistical inference literature in reinforcement learning as a well-known online decision-making setting. [164] conducted statistical inference for TD (and GTD) learning. [178] constructed the confidence interval for policy values in Markov decision processes. [179] conducted statistical inference for confounded Markov decision processes. [37] developed the confidence interval for heterogeneous Markov decision processes.

6.1.2 Notations

We first introduce some notations in our paper. For any pair of positive integers $m < n$, we use $[m : n]$ as a shorthand for the discrete set of $\{m, m+1, \ldots, n\}$. For any vector $\theta \in \mathbb{R}^d$, we use $\theta_{[m:n]}$ to denote the vector consisting of the m-th to n-th coordinates of θ. Similarly, $\theta_{[m:n],t}$ is the corresponding subvector of θ_t. For a set of random variables X_n and a corresponding set of constants a_n, $X_n = \mathcal{O}_p(a_n)$ means that X_n/a_n is stochastically bounded and $X_n = o_p(a_n)$ means that X_n/a_n converges to zero in probability as n goes to infinity. We denote $\overset{p}{\to}$, and $\overset{d}{\to}$ as convergence in probability and convergence in distribution, respectively.

For convenience, let $\|\cdot\|$ denote the standard Euclidean norm for vectors and the spectral norm for matrices. We use the standard Loewner order notation $\Sigma \succeq 0$ if a matrix Σ is positive semi-definite. Denote I_d as the identity matrix in $\mathbb{R}^{d\times d}$. For any square matrix Σ, $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ represent the smallest and the largest eigenvalues, respectively. We also introduce $\mathbb{I}(\cdot)$ for the indicator function, and $\lesssim$ is used for inequalities with omitted constants.

The remainder of the paper is organized as follows. In Section 6.2, we consider the environment where we collect data adaptively. We describe the weighted version SGD under this setting and give two illustrative examples of the classical regression problems. In Section 6.4, we first introduce the technical assumptions before we present the asymptotic distribution for general weighted SGD under this adaptive data collection scheme, along with a comparison on the statistical efficiency to the previous result proposed in [41]. We further justify our assumptions under the two illustrative regression examples and show the asymptotic normality for these two cases. Section 6.4.2 gives the finite-sample rate for our SGD update under adaptive environment. We compare our result with the classical SGD rate, where the slower rate is due to the adaptively collected data. Simulation studies and real data analyses in Section 6.5 lend numerical support to the theoretical claims in this paper, which also provides hands-on guidelines to practitioners.

6.2 Problem setup

We consider a contextual bandit environment where the observed data at each decision point t is a triplet $\zeta_t = (X_t, A_t, Y_t)$ for all $t \geq 1$, consisting of covariate X_t, action A_t, and reward Y_t. In this paper, we consider a finite action space, i.e., $A_t \in \mathcal{A}$ and $|\mathcal{A}| < \infty$. We assume a stochastic contextual bandit environment in which $\{X_t, Y_t(a) : a \in \mathcal{A}\} \overset{i.i.d}{\sim} \mathbb{P} \in \mathbf{P}$ for all $t \geq 1$. The contextual bandit environment distribution $\mathbb{P}$ is in a space of possible environment distributions $\mathbf{P}$. Here $Y_t(a)$ corresponds to the (heuristic) reward Y_t given a fixed action a regardless of the realized action A_t. Note that $Y_t(a)$ is observed for $a = A_t$ only, but not observed for any other $a \in \mathcal{A}, a \neq A_t$.

We define the trajectory until time t as $\mathcal{H}_t := \{X_s, A_s, Y_s\}_{s=1}^{t}$ for $t \geq 1$ and $\mathcal{H}_0 := \emptyset$. Actions $A_t \in \mathcal{A}$ are selected according to some policy $A_t \sim \pi\left(X_t, \mathcal{H}_{t-1}\right)$, which defines action distribution. Even though the covariate reward tuples are *i.i.d.*, the observed data $\{X_t, A_t, Y_t\}_{t \geq 1}$ are not *i.i.d.*, since the actions are selected using policies $\pi\left(X_t, \mathcal{H}_{t-1}\right)$ which is a function of past data, $\mathcal{H}_{t-1}$. Non-independence of observations is a key property of adaptively collected data.

We are interested in constructing confidence regions for some unknown $\theta^* \in \mathbb{R}^d$. Under the finite action space where $|\mathcal{A}| < \infty$, we can use θ^* as the concatenated vector of θ_a^* for all $a \in \mathcal{A}$, where we assume that θ_a^* is a conditionally maximizing value of some loss function $\ell(\theta; \zeta)$ for $\mathbb{P} \in \mathbf{P}$,

$$\theta_a^*(\mathbb{P}) \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \mathbb{E}_{\mathbb{P}_{Y|X}}\left[\ell\left(\theta; \zeta\right) \mid X, A = a\right]. \tag{6.5}$$

Note that (6.5) represents an implicit modeling assumption that such an underlying θ_a^* does not depend on X for a given loss $\ell(\theta; \zeta)$. This assumption is generally satisfied in many statistical applications. For example, in a classical regression setting, a natural choice for $\ell(\theta; \zeta)$ is as follows,

$$\ell(\theta; \zeta_t) = \rho\left(Y_t - X_t^\top \theta_{A_t}\right), \tag{6.6}$$

where $\theta \in \mathbb{R}^d$ is the concatenated vector of $\theta_{A_t} \in \mathbb{R}^p$ for all possible choices of $A_t \in \mathcal{A}$ and $d = p|\mathcal{A}|$. Here $\rho(\cdot)$ is some convex loss function. Note that ℓ can be non-smooth as long as $\nabla \ell$ exists almost surely. We illustrate several examples of popular statistical models, and we will refer to these examples throughout the paper.

Example 6.2.1 (Linear Regression)**.** *Consider a two-arm linear contextual bandit problem where*

$$\mathbb{E}[Y_t \mid A_t, X_t] = X_t^\top \theta^*_{A_t} = (1 - A_t)\big(X_t^\top \theta^*_{[1:p]}\big) + A_t\big(X_t^\top \theta^*_{[p+1:2p]}\big),$$

where $\theta^* \in \mathbb{R}^d$ is the concatenated vector of $\theta^*_{[1:p]}$ and $\theta^*_{[p+1:2p]}$ and $\theta^*_{[1:p]} \neq \theta^*_{[p+1:2p]}$, $\{X_t, Y_t(a) : a \in \mathcal{A}\} \overset{i.i.d}{\sim} \mathbb{P} \in \mathbf{P}$ for all $t \geq 1$, and $\mathcal{A} = \{0, 1\}$. The true reward Y_t is generated by $\mathbb{E}[Y_t \mid A_t, X_t] + \mathcal{E}_t$ where $\{\mathcal{E}_t\}$ are i.i.d. *random error with mean zero and variance σ^2. Under the linear regression model, our loss function ℓ is defined as*

$$\ell(\theta; \zeta_t) = \frac{1}{2}(1 - A_t)\left(Y_t - X_t^\top \theta_{[1:p]}\right)^2 + \frac{1}{2}A_t\left(Y_t - X_t^\top \theta_{[p+1:2p]}\right)^2.$$

Example 6.2.2 (Logistic Regression)**.** *Consider a two-arm contextual bandit problem under the logistic model with binary rewards where $A_t \in \mathcal{A} = \{0, 1\}$, $Y_t \in \{-1, 1\}$, and*

$$\mathbb{P}(Y_t \mid A_t, X_t) = \left(1 + \exp\left(-(1 - A_t)Y_t X_t^\top \theta^*_{[1:p]} + -A_t Y_t X_t^\top \theta^*_{[p+1:2p]}\right)\right)^{-1}.$$

We consider the entropy loss

$$\ell(\theta;\zeta_t) = (1-A_t)\log\left(1+\exp\left(-Y_t X_t^\top \theta_{[1:p]}\right)\right) + A_t \log\left(1+\exp\left(-Y_t X_t^\top \theta_{[p:2p]}\right)\right).$$

Example 6.2.3 (Quantile Regression)**.** *Consider a two-arm linear contextual bandit problem where*

$$Y_t = (1-A_t)X_t^\top \theta^*_{[1:p]} + A_t X_t^\top \theta^*_{[p+1:2p]} + \mathcal{E}_t,$$

and $\{\mathcal{E}_t\}$ *are* i.i.d. *random error such that,* $\Pr(\mathcal{E}_t \le 0) = \tau$ *for some given quantile level* $\tau \in (0,1)$*. Consider a quantile loss such that*

$$\ell(\theta;\zeta_t) = (1-A_t)\rho_\tau\left(Y_t - X_t^\top \theta_{[1:p]}\right) + A_t \rho_\tau\left(Y_t - X_t^\top \theta_{[p+1:2p]}\right),$$

where $\rho_\tau(u) = u(\tau - \mathbb{I}(u<0))$*.*

6.3 SGD with weighted stochastic gradients

Under the adaptive data collection scheme, we now consider a generalized version of the classical SGD (6.2) with weights w_t depends only on the triplet (A_t, X_t, θ_{t-1}), as follows,

$$\theta_t = \theta_{t-1} - \eta_t w_t \nabla \ell(\theta_{t-1};\zeta_t). \tag{6.7}$$

We consider the following three popular choices of weight w_t as examples throughout the paper.

(IPW) Inverse probability weighting: $w_t(A_t, X_t, \theta_{t-1}) = \dfrac{1}{2\Pr(A_t \mid X_t, \theta_{t-1})}$.

(sqrt-IPW) Square-root importance weights: $w_t(A_t, X_t, \theta_{t-1}) = \sqrt{\dfrac{1}{2\Pr(A_t \mid X_t, \theta_{t-1})}}$.

`(vanilla)` No weights applied to the SGD updates. $w_t(A_t, X_t, \theta_{t-1}) = 1$.

These weighting schemes are well-rooted in literature, such as `(IPW)` by [41] that corrects the action distribution to some deterministic stable policy, and `(sqrt-IPW)`in [93] and [205]. The proposed method is not limited to the analysis of these three weights but is applied to general weight specifications.Before we present our main result, we revisit the three aforementioned motivating examples and illustrate the weighted SGD algorithm for the three models.

Example 6.3.1 (6.2.1 continued.). *Under the linear regression model, the weighted SGD* (6.7) *writes as*

$$\theta_{[1:p],t} = \theta_{[1:p],t-1} - \eta_t w_t X_t \left(X_t^\top \theta_{[1:p],t-1} - Y_t \right), \quad A_t = 0;$$
$$\theta_{[p+1:2p],t} = \theta_{[p+1:2p],t-1} - \eta_t w_t X_t \left(X_t^\top \theta_{[p+1:2p],t-1} - Y_t \right), \quad A_t = 1.$$

Example 6.3.2 (6.2.2 continued.). *Under a logistic regression model, the weighted SGD* (6.7) *writes as*

$$\theta_{[1:p],t} = \theta_{[1:p],t-1} + \eta_t w_t \left(1 + \exp\left(Y_t X_t^\top \theta_{[1:p],t-1} \right) \right)^{-1} Y_t X_t, \quad A_t = 0;$$
$$\theta_{[p+1:2p],t} = \theta_{[p+1:2p],t-1} + \eta_t w_t \left(1 + \exp\left(Y_t X_t^\top \theta_{[p+1:2p],t-1} \right) \right)^{-1} Y_t X_t, \quad A_t = 1.$$

Example 6.3.3 (6.2.3 continued.). *Under the quantile regression model, the weighted SGD* (6.7) *writes as*

$$\theta_{[1:p],t} = \theta_{[1:p],t-1} + \eta_t w_t \left(\tau - \mathbb{I}(Y_t - X_t^\top \theta_{[1:p],t-1} < 0) \right) X_t, \quad A_t = 0;$$
$$\theta_{[p+1:2p],t} = \theta_{[p+1:2p],t-1} + \eta_t w_t \left(\tau - \mathbb{I}(Y_t - X_t^\top \theta_{[p+1:2p],t-1} < 0) \right) X_t, \quad A_t = 1.$$

Given our path of $\{\theta_t\}_{t\geq 1}$, we assume the policy $\pi\left(X_t, \mathcal{H}_{t-1}\right)$ depend on the history

$\mathcal{H}_{t-1}$ only through θ_{t-1}, our estimator from the latest step, i.e., $A_t \sim \pi(X_t, \theta_{t-1})$. We consider a common policy, ε-greedy, to address the exploration-and-exploitation dilemma, where the probability of action A_t is defined as,

$$\Pr(A_t = 0 \mid X_t, \theta_{t-1}) = (1-\varepsilon)\mathbb{I}\left\{X_t^\top \theta_{[1:p],t-1} > X_t^\top \theta_{[p+1:2p],t-1}\right\} + \frac{\varepsilon}{2}, \tag{6.8}$$

for some constant $\varepsilon \in (0,1)$. In practice, the ε is often set as some small constant close to zero. Note that this setting can be relaxed to a deterministic sequence $\{\varepsilon_t\}$ which converges to some constant $\varepsilon_\infty \in (0,1)$. Note that this may be relaxed to $A_t \sim \pi(X_t, \Phi_{t-1})$ for some other statistic Φ_{t-1} relies on the history $\theta_0, \theta_1, \cdots, \theta_{t-1}$, for example, the running average of the $\{\theta_s\}_{s=0}^{t-1}$.

We now use the linear regression model in Example 6.2.1 with random design as a special case of our main result that will be presented in Theorem 6.4.7 below. We specify w_t as a function of $\Pr(A_t \mid X_t, \theta_{t-1})$, i.e., $w_t(A_t, X_t, \theta_{t-1}) = \varphi(\Pr(A_t \mid X_t, \theta_{t-1}))$. The following Theorem 6.3.4 provides a new way to further determine the optimal weighting scheme to minimize the asymptotic variance of the average SGD estimator.

Theorem 6.3.4. *In the linear regression setting in Example 6.2.1, assume that $X_t \sim \mathcal{N}(\mu, I_p)$, and $\varphi(\cdot) : (0,1) \mapsto \mathbb{R}^+$ is continuous. The averaged SGD estimator $\bar{\theta}_t$ converges* *to θ^* almost surely as $t \to \infty$ and*

$$\sqrt{t}(\bar{\theta}_t - \theta^*) \xrightarrow{d} \mathcal{N}(0, H^{-1}SH^{-1}), \quad \textit{where} \qquad S = \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix}, \ H = \begin{bmatrix} H_1 & 0 \\ 0 & H_2 \end{bmatrix},$$

$$S_1 = \sigma^2 \left((1-\frac{\varepsilon}{2})\varphi^2(1-\frac{\varepsilon}{2})G_1^* + \frac{\varepsilon}{2}\varphi^2(\frac{\varepsilon}{2})G_2^* \right),$$

$$S_2 = \sigma^2 \left(\frac{\varepsilon}{2}\varphi^2(\frac{\varepsilon}{2})G_1^* + (1-\frac{\varepsilon}{2})\varphi^2(1-\frac{\varepsilon}{2})G_2^* \right),$$

$$H_1 = (1-\frac{\varepsilon}{2})\varphi(1-\frac{\varepsilon}{2})G_1^* + \frac{\varepsilon}{2}\varphi(\frac{\varepsilon}{2})G_2^*, \qquad H_2 = \frac{\varepsilon}{2}\varphi(\frac{\varepsilon}{2})G_1^* + (1-\frac{\varepsilon}{2})\varphi(1-\frac{\varepsilon}{2})G_2^*,$$

$$G_1^* = \Phi\left(a^*\right) I_p + \frac{1}{\sqrt{2\pi}} a^* e^{\frac{a^{*2}}{2}} \nu^* \nu^{*\top}, \qquad G_2^* = \left(1 - \Phi\left(a^*\right)\right) I_p - \frac{1}{\sqrt{2\pi}} a^* e^{\frac{a^{*2}}{2}} \nu^* \nu^{*\top},$$

and $\nu^* = (\theta^*_{[1:p]} - \theta^*_{[p+1:2p]})/\|\theta^*_{[1:p]} - \theta^*_{[p+1:2p]}\|$, $a^* = \frac{\mu^\top \nu^*}{\sqrt{1+(\mu^\top \nu^*)^2}}$, and Φ is the cumulative distribution function of standard normal distribution.

Theorem 6.3.4 can be considered as a special case of our main result which will be presented in 6.4. It follows directly from Remark 6.4.1 and Corollary 6.4.8. Here the definition of ν^* and consequently the existence of the asymptotic covariance matrix are assured by the implicit non-degenerate model assumption such that $\theta^*_{[1:p]} \neq \theta^*_{[p+1:2p]}$.

In light of Theorem 6.3.4, we are ready to conduct a comparison among the three popular weighting schemes. Here we specify $\varphi_\gamma(\varepsilon) = \varepsilon^\gamma$ as a class of power functions parameterized by γ. This class of weights covers the following three popular weighting schemes: `(IPW)` as $\gamma = -1$, `(sqrt-IPW)` as $\gamma = -1/2$, and `(vanilla)` as $\gamma = 0$, up to some constants. We can write $H^{-1}SH^{-1}$ as follows,

$$H^{-1}SH^{-1} = \sigma^2 \begin{bmatrix} c_1 I + c_2 \nu^* \nu^{*\top} & 0 \\ 0 & c_3 I + c_4 \nu^* \nu^{*\top} \end{bmatrix}, \tag{6.9}$$

where

$$
\begin{aligned}
c_1 &= \frac{(1-\frac{\varepsilon}{2})^{1+2\gamma}\Phi(a) + (\frac{\varepsilon}{2})^{1+2\gamma}(1-\Phi(a))}{((1-\frac{\varepsilon}{2})^{1+\gamma}\Phi(a) + (\frac{\varepsilon}{2})^{1+\gamma}(1-\Phi(a)))^2}, \\
c_2 &= \frac{(1-\frac{\varepsilon}{2})^{1+2\gamma}(\Phi(a) + \frac{1}{\sqrt{2\pi}}ae^{\frac{a^2}{2}}) + (\frac{\varepsilon}{2})^{1+2\gamma}(1-\Phi(a) - \frac{1}{\sqrt{2\pi}}ae^{\frac{a^2}{2}})}{((1-\frac{\varepsilon}{2})^{1+\gamma}(\Phi(a) + \frac{1}{\sqrt{2\pi}}ae^{\frac{a^2}{2}}) + (\frac{\varepsilon}{2})^{1+\gamma}(1-\Phi(a) - \frac{1}{\sqrt{2\pi}}ae^{\frac{a^2}{2}}))^2} - c_1, \\
c_3 &= \frac{(1-\frac{\varepsilon}{2})^{1+2\gamma}(1-\Phi(a)) + (\frac{\varepsilon}{2})^{1+2\gamma}\Phi(a)}{((1-\frac{\varepsilon}{2})^{1+\gamma}(1-\Phi(a)) + (\frac{\varepsilon}{2})^{1+\gamma}\Phi(a))^2}, \\
c_4 &= \frac{(1-\frac{\varepsilon}{2})^{1+2\gamma}(1-\Phi(a) - \frac{1}{\sqrt{2\pi}}ae^{\frac{a^2}{2}}) + (\frac{\varepsilon}{2})^{1+2\gamma}(\Phi(a) + \frac{1}{\sqrt{2\pi}}ae^{\frac{a^2}{2}})}{((1-\frac{\varepsilon}{2})^{1+\gamma}(1-\Phi(a) - \frac{1}{\sqrt{2\pi}}ae^{\frac{a^2}{2}}) + (\frac{\varepsilon}{2})^{1+\gamma}(\Phi(a) + \frac{1}{\sqrt{2\pi}}ae^{\frac{a^2}{2}}))^2} - c_3.
\end{aligned}
$$

The eigenvalues of the asymptotic covariance matrix are $c_1, c_1+c_2, c_3, c_3+c_4$ in the above equations. Since H_1, H_2, G_1, G_2 all have the form $\tilde{b}I + \tilde{c}\nu^*\nu^{*\top}$ for some constants $\tilde{b}$ and $\tilde{c}$, they can be simultaneously diagonalized. When varying γ, the eigenvectors stay fixed and each eigenvalue changes in the following form with some $b \in (0,1)$,

$$
g(\gamma) = \frac{(1-\varepsilon/2)^{1+2\gamma}b + (\varepsilon/2)^{1+2\gamma}(1-b)}{\big((1-\varepsilon/2)^{1+\gamma}b + (\varepsilon/2)^{1+\gamma}(1-b)\big)^2}. \tag{6.10}
$$

In practice, for the ε-greedy policy, the parameter ε is usually taken as some small constant. When ε gets close to 0, it can be inferred from (6.10) that $\gamma \geq -1/2$ leads to a finite covariance matrix when ε goes to zero. This includes `(vanilla)` and `(sqrt-IPW)` but excludes `(IPW)`. Furthermore, the minimum of (6.10) is obtained at $\gamma = 0$ for all $b \in (0,1)$. Therefore, under the settings in Theorem 6.3.4, `(vanilla)` has an asymptotic covariance matrix that is dominated by any other asymptotic covariance matrix obtained from a power-law weighted scheme, $\varphi(\varepsilon) = \varepsilon^\gamma$. The following Corollary 6.3.5 concludes the above discussion and further extends to general weighting schemes w.

Corollary 6.3.5 (Optimal weights in linear regression). *Under Assumption 6.4.2 to Assumption 6.4.6, the* `(vanilla)`*SGD has the optimal asymptotic covariance matrix in the* *linear regression setting, i.e.,* $\Sigma_{\mathrm{vnl}} \preceq \tilde{\Sigma}$*, where* Σ_{vnl} *is the asymptotic covariance matrix of vanilla SGD and* $\tilde{\Sigma}$ *is the asymptotic covariance matrix under any other weighting function* φ *where* $w_t(A_t, X_t, \theta_{t-1}) = \varphi(\Pr(A_t \mid X_t, \theta_{t-1}))$*.*

Proof. Consider a k-arm bandit linear regression setting. Our central limit theorem gives the covariance of the form

$$\Sigma = H^{-1} S H^{-1},$$

where

$$H = \nabla^2 \mathcal{L}_{\theta^*}(\theta^*) = \mathbb{E}[w(\theta^*)\nabla^2 \ell(\theta^*)],$$

and

$$S = \mathbb{E}[\xi_{\theta^*}(\theta^*; \zeta)\xi_{\theta^*}(\theta^*; \zeta)^\top] = \mathbb{E}[(w(\theta^*)^2 \nabla \ell(\theta^*) \nabla \ell(\theta^*)^\top].$$

One special property of linear regression is that

$$\nabla^2 \ell(\theta^*)^2 = (X_t X_t^\top) \otimes E_i,$$

where E_k is the matrix with 1 on the (k,k) entry and other entries to be 0, i means we choose the i-th bandit, and

$$(\nabla \ell(\theta^*))(\nabla \ell(\theta^*))^\top = (X_t X_t^\top) \otimes E_i \mathbb{E}\sigma^2.$$

So they differ only by a constant factor. If w is constant instead of stochastic, we denote the corresponding H, S, Σ as H_c, S_c, Σ_c. We claim that $\Sigma \succeq \Sigma_c$. Equal weight is optimal for linear regression. This is equivalent to

$$(\mathbb{E}[w(\theta^*)\nabla^2\ell(\theta^*)])^{-1}(\mathbb{E}[w(\theta^*)^2\nabla^2\ell(\theta^*)])(\mathbb{E}[w(\theta^*)\nabla^2\ell(\theta^*)])^{-1} \succeq (\mathbb{E}[\nabla^2\ell(\theta^*)])^{-1},$$

which is the same as

$$\mathbb{E}[w(\theta^*)^2\nabla^2\ell(\theta^*)] - (\mathbb{E}[w(\theta^*)\nabla^2\ell(\theta^*)])(\mathbb{E}[\nabla^2\ell(\theta^*)])^{-1}(\mathbb{E}[w(\theta^*)\nabla^2\ell(\theta^*)]) \succeq 0.$$

By Schur complement, this is equivalent to

$$\begin{bmatrix} \mathbb{E}[w(\theta^*)^2\nabla^2\ell(\theta^*)] & \mathbb{E}[w(\theta^*)\nabla^2\ell(\theta^*)] \\ \mathbb{E}[w(\theta^*)\nabla^2\ell(\theta^*)] & \mathbb{E}[\nabla^2\ell(\theta^*)] \end{bmatrix} \succeq 0.$$

Note that

$$\begin{bmatrix} w(\theta^*)^2\nabla^2\ell(\theta^*) & w(\theta^*)\nabla^2\ell(\theta^*) \\ w(\theta^*)\nabla^2\ell(\theta^*) & \nabla^2\ell(\theta^*) \end{bmatrix} \succeq 0.$$

Therefore, our conclusion holds. □

6.4 Asymptotic normality under general models

We provide the main theoretical results in this section. To facilitate the analysis of the asymptotic behavior of SGD update (6.7) in a general model, we define the function $\mathcal{L}_{\theta'}(\theta)$ as follows,

$$\mathcal{L}_{\theta'}(\theta) = \mathbb{E}_{\mathbb{P}}\left[\mathbb{E}_{\pi(X,\theta')}\left(w(\theta'; X, A)\ell(\theta; X, A, Y) \mid X\right)\right], \tag{6.11}$$

where $A \sim \pi(X, \theta')$, $\theta', \theta \in \mathbb{R}^d$, and gradient weight w depending on θ', action A and covariate X. Note that the objective $\mathcal{L}_{\theta'}(\theta)$ is a function of θ with a parameter θ' that corresponds to the current estimate being used to select the action. Typically we use $\theta' = \theta_{t-1}$ at iteration t. Below we will always use the expression $\nabla \mathcal{L}_{\theta'}(\theta)$ to represent the partial gradient of $\mathcal{L}_{\theta'}(\theta)$ with respect to the variable θ, i.e.,

$$\nabla \mathcal{L}_{\theta'}(\theta) = \frac{\partial}{\partial \theta} \mathcal{L}_{\theta'}(\theta) \in \mathbb{R}^d, \ \ \nabla^2 \mathcal{L}_{\theta'}(\theta) = \frac{\partial^2}{\partial \theta^2} \mathcal{L}_{\theta'}(\theta) \in \mathbb{R}^{d \times d}.$$

We note that for quantile regression in Example 6.2.3, even though the individual objective $\ell(\theta; X, Y)$ is nonsmooth, the population objective $\mathcal{L}_{\theta'}(\theta)$ is second-order differentiable with some smoothness conditions on the error distribution. Finally, we denote $\xi_{\theta'}(\theta; \zeta)$ as the difference between the stochastic gradient and population gradient of the loss defined in (6.11), i.e.,

$$\xi_{\theta'}(\theta; \zeta) = w(\theta'; X, A) \nabla \ell(\theta; \zeta) - \nabla \mathcal{L}_{\theta'}(\theta), \tag{6.12}$$

By definition, we can easily verify that $w(\theta'; X, A) \nabla \ell(\theta; \zeta)$ is an unbiased estimator of $\nabla \mathcal{L}_{\theta'}(\theta)$, $\mathbb{E}[\xi_{\theta'}(\theta; \zeta)] = 0$.

In the previous work of [41], the loss function is defined with respect to some predetermined stable policy π_{stable}, i.e.,

$$\tilde{\mathcal{L}}(\theta) = \mathbb{E}_{\mathbb{P}} \left[\mathbb{E}_{\pi_{\text{stable}}} \left(\ell(\theta; X, A, Y) \mid X \right) \right], \tag{6.13}$$

where $A \sim \pi_{\text{stable}}$ and π_{stable} is a Bernoulli($1/|\mathcal{A}|$), uniformly distributed on the action space $\mathcal{A}$. To match the SGD update with the loss function $\tilde{\mathcal{L}}(\cdot)$, they choose the (IPW)-weighted SGD such that $w_t = \frac{\pi_{\text{stable}}}{\pi(X, \theta)}$. This weighting scheme corrects the sampling distribution of the action A_t towards the Bernoulli distribution under the stable policy. However, this definition cannot be extended to a general weighting scheme and the resulting asymptotic covariance matrix could be extremely large as we will see in the discussion after Theorem 6.4.7. Our framework allows a broader class of weighting schemes, and our theoretical analysis relies heavily on our definition of the loss function $\mathcal{L}_{\theta'}(\theta)$ in (6.11). By expressing the loss using two different variables θ and θ', we separate the loss $\ell(\theta;\zeta)$ from the policy $\pi(X,\theta')$ and the weight $w(\theta';X,A)$, as we have a focus on the local geometry of $\ell(\theta;\zeta)$ instead of the local geometry of $\pi(X,\theta')$ and $w(\theta';X,A)$.

In the Remark 6.4.1 below, we illustrate our definition of $\mathcal{L}_{\theta'}(\theta)$ and our assumptions above using a special case where the covariate X follows a normal distribution.

Remark 6.4.1. *Under the linear setting in Theorem 6.3.4. We have for any ε,*

$$\mathcal{L}_{\theta'}(\theta) = (\theta^* - \theta)^\top G(\theta^* - \theta) + \frac{\sigma^2}{2}\left((1-\frac{\varepsilon}{2})\varphi(1-\frac{\varepsilon}{2}) + \frac{\varepsilon}{2}\varphi(\frac{\varepsilon}{2})\right),$$

where we denote $\Phi(\cdot)$ as the c.d.f. for standard normal distribution and

$$G = \begin{bmatrix} (1-\frac{\varepsilon}{2})\varphi(1-\frac{\varepsilon}{2})G_1 + \frac{\varepsilon}{2}\varphi(\frac{\varepsilon}{2})G_2 & 0 \\ 0 & (1-\frac{\varepsilon}{2})\varphi(1-\frac{\varepsilon}{2})G_2 + \frac{\varepsilon}{2}\varphi(\frac{\varepsilon}{2})G_1 \end{bmatrix}.$$

$$G_1 = \Phi\left(a\right) I_p + \frac{1}{\sqrt{2\pi}} a e^{\frac{a^2}{2}} \nu' \nu'^\top, \quad G_2 = \left(1 - \Phi\left(a\right)\right) I_p - \frac{1}{\sqrt{2\pi}} a e^{\frac{a^2}{2}} \nu' \nu'^\top,$$

where ν' is the normalized margin between the two arms of θ',

$$\nu' = (\theta'_{[1:p]} - \theta'_{[p+1:2p]}) / \left\|\theta'_{[1:p]} - \theta'_{[p+1:2p]}\right\|, \quad \textit{and} \quad a = \frac{\mu^\top \nu'}{\sqrt{1 + (\mu^\top \nu')^2}}.$$

6.4.1 Asymptotic normality

We first introduce some regularity assumptions on the population loss function $\mathcal{L}_{\theta'}(\theta)$, the individual loss function $\ell(\theta;\zeta)$, and the gradient weight $w(\theta'; X, A)$.

Assumption 6.4.2. *There exists some constants $\underline{w}, \overline{w}$, such that $0 < \underline{w} < w_t < \overline{w}$ for all $t \geq 1$.*

Assumption 6.4.3. *The loss function $\mathcal{L}_{\theta'}(\theta)$ is convex with respect to $\theta \in \mathbb{R}^d$, continuously differentiable with respect to $\theta \in \mathbb{R}^d$, and twice continuously differentiable with respect to θ at θ^*. Moreover, there exists some constants $\delta, \lambda > 0$, such that $\langle \nabla \mathcal{L}_{\theta}(\theta), \theta - \theta^* \rangle > 0,\ \forall \theta \neq \theta^*$ and*

$$\langle \nabla \mathcal{L}_{\theta}(\theta), \theta - \theta^* \rangle \geq \lambda \|\theta - \theta^*\|^2, \ \ \forall\, \theta \in \{\theta : \|\theta - \theta^*\| \leq \delta\}.$$

Assumption 6.4.4. *The Hessian matrix $\nabla^2 \mathcal{L}_{\theta'}(\theta) \in \mathbb{R}^{d\times d}$ exists for all $(\theta; \theta') \in \mathbb{R}^d \times \mathbb{R}^d$ and the Hessian matrix at $(\theta^*; \theta^*)$ is positive definite, i.e., $H \triangleq \nabla^2 \mathcal{L}_{\theta^*}(\theta^*) \succ 0$. Moreover, the Hessian matrix $\nabla^2 \mathcal{L}_{\theta'}(\theta)$ is K-Lipschitz continuous at (θ^*, θ^*), i.e.,*

$$\left\| \nabla^2 \mathcal{L}_{\theta'}(\theta) - \nabla^2 \mathcal{L}_{\theta^*}(\theta^*) \right\| \leq K\|\theta - \theta^*\| + K\|\theta' - \theta^*\|,$$

for all (θ, θ') such that $\|\theta - \theta^*\| + \|\theta' - \theta^*\| \leq 2\delta$.

Assumption 6.4.5. *For any action $A \in \mathcal{A}$ and covariate X, we further assume,*

$$\mathbb{E}_{\mathbb{P}_{Y|X}} \left(\|\nabla \ell(\theta; \zeta)\|^2 \mid X, A \right) \leq \phi(X)(1 + \|\theta - \theta^*\|^2),$$

for some function $\phi(\cdot)$ such that $\mathbb{E}[\phi(X)] = \kappa$ for some constant $\kappa > 0$. We also assume the Gram matrix of $\xi_{\theta'}(\theta; \zeta)$ at $(\theta^*; \theta^*)$, $S \triangleq \mathbb{E}[\xi_{\theta^*}(\theta^*; \zeta)\xi_{\theta^*}(\theta^*; \zeta)^\top]$, exists.

Assumption 6.4.6. *Let $\Delta(X,\theta) = d_{TV}(\pi(X,\theta), \pi(X,\theta^*))$ be the total variation distance of $\pi(X,\theta)$ and $\pi(X,\theta^*)$. For function $\phi(X)$ defined in Assumption 6.4.5, we have*

$$\begin{aligned}
&\lim_{\theta\to\theta^*} \mathbb{E}_{\mathbb{P}_X}[\Delta(X,\theta)\phi(X)] = 0,\\
&\lim_{\theta\to\theta^*} \mathbb{E}_{\mathbb{P}_{Y|X}}\left[\|\nabla\ell(\theta;\zeta) - \nabla\ell(\theta^*;\zeta)\|^2 \mid X, A\right] = 0,\\
&\lim_{\theta\to\theta^*} \mathbb{E}_{\mathbb{P}_X}\left[|w(\theta;X,A) - w(\theta^*;X,A)|^2\phi(X) \mid A\right] = 0.
\end{aligned}$$

Assumption 6.4.2 is a common assumption on the weights applied to the stochastic gradient, which is used in many adaptive setting literature, e.g., [39], [41], and [205]. The convexity and continuity on the population loss $\mathcal{L}$ in Assumption 6.4.3 is a standard requirement in classical SGD literature [162, 44, 41, 62]. We can also find similar arguments in the SGD literature mentioned above for Assumption 6.4.3 to Assumption 6.4.5, whereas we generalize the previous assumptions on our loss function $\mathcal{L}_\theta(\theta)$ with an extra variable θ'. Assumption 6.4.6 further gives some regularity on the function $\phi(\cdot)$ defined in Assumption 6.4.5. Later we will further verify our assumptions on the two examples we mentioned above, i.e., the linear regression and the quantile regression. It is noteworthy to mention that, in Assumption 6.4.5 and Assumption 6.4.6, we only implicitly assume $\nabla\ell$ exists almost surely under $\mathbb{P}_{Y|X}$. Therefore, our assumption is not restricted to smooth loss function ℓ, it also covers many non-smooth statistical problems like quantile regression and robust regression. We now state our main result that characterizes the limiting distribution of the averaged weighted SGD iterates defined in (6.7) under general models.

Theorem 6.4.7. *Under Assumption 6.4.2 to Assumption 6.4.6, the averaged SGD estimator $\bar{\theta}_t$ converges to θ^* almost surely when $t \to \infty$ and*

$$\sqrt{t}(\bar{\theta}_t - \theta^*) \xrightarrow{d} \mathcal{N}(0, H^{-1}SH^{-1}),$$

where $H = \nabla^2 \mathcal{L}_{\theta^*}(\theta^*)$ *and* $S = \mathbb{E}[\xi_{\theta^*}(\theta^*; \zeta)\xi_{\theta^*}(\theta^*; \zeta)^\top]$.

Proof. By definition, the loss function can be written as

$$\begin{aligned}\mathcal{L}_{\theta'}(\theta) &= \mathbb{E}[w(\theta'; X, A')\ell(\theta; Y, X, A')] \\ &= \mathbb{E}_{\mathbb{P}_X, \pi(X, \theta')}\left\{\mathbb{E}_{\mathbb{P}_{Y|X}}[w(\theta'; X, A')\ell(\theta; Y, X, A')] \mid X, A'\right\}.\end{aligned}$$

By Equation (6.5), θ^* is the minimizer, i.e., $\theta^* \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \mathcal{L}_{\theta'}(\theta)$. Because $\mathcal{L}$ is differentiable, We have $\nabla \mathcal{L}_{\theta'}(\theta^*) = 0$. Moreover, we have

$$\begin{aligned}\|\nabla \mathcal{L}_\theta(\theta) - H(\theta - \theta^*)\| &= \|\nabla \mathcal{L}_\theta(\theta) - \nabla \mathcal{L}_\theta(\theta^*) - H(\theta - \theta^*)\| \\ &= \left\| \int_0^1 \left(\nabla^2 \mathcal{L}_\theta(\theta^* + s(\theta - \theta^*)) - H\right)(\theta - \theta^*) ds \right\| \\ &\leq K\|\theta - \theta^*\|^2, \end{aligned} \tag{6.14}$$

for $\|\theta - \theta^*\| < \delta$ as stated in Assumption 6.4.4.

By Equation (6.12), we have $w(\theta'; X, A)\nabla \ell(\theta; \zeta) = \xi_{\theta'}(\theta; \zeta) + \nabla \mathcal{L}_{\theta'}(\theta)$. Notice that we have the following inequality,

$$\mathbb{E}[\|w(\theta; X, A)\nabla \ell(\theta; \zeta)\|^2] \leq \overline{w}^2 \mathbb{E}[\nabla \ell(\theta; \zeta)^2] \leq \overline{w}^2 \kappa(1 + \|\theta - \theta^*\|^2).$$

Therefore, by the fact that $\mathbb{E}[\xi_\theta(\theta; \zeta)] = 0$, the two terms can be bounded by

$$\begin{aligned}\|\nabla \mathcal{L}_\theta(\theta)\|^2 &\leq \overline{w}^2 \kappa(1 + \|\theta - \theta^*\|^2), \\ \mathbb{E}\left[\|\xi_\theta(\theta; \zeta)\|^2\right] &\leq \overline{w}^2 \kappa(1 + \|\theta - \theta^*\|^2).\end{aligned}$$

The above bounds can already guarantee the almost surely convergence of θ_t by Theorem 2 of [162]. Now we need to quantify the difference between $\xi_{\theta_{t-1}}(\theta_{t-1}; \zeta)$ and $\xi_{\theta^*}(\theta^*; \zeta)$. Using the coupling we defined in Equation (6.20), it can be bounded by

$$
\begin{aligned}
&\mathbb{E}_{\mathbb{P},\nu}\|\xi_{\theta_{t-1}}(\theta_{t-1};\zeta) - \xi_{\theta^*}(\theta^*;\zeta)\|^2 \\
\leq\ & 2\|\nabla\mathcal{L}_{\theta_{t-1}}(\theta_{t-1}) - \nabla\mathcal{L}_{\theta^*}(\theta^*)\|^2 \\
&+ 2\mathbb{E}_{\mathbb{P},\nu}\left[\|w(\theta_{t-1};X_t,A_t)\nabla\ell(\theta_{t-1};\zeta_t) - w(\theta^*;X_t,A^*)\nabla\ell(\theta^*;X_t,A^*,Y_t)\|^2\right] \\
:=\ & 2M_1 + 2M_2.
\end{aligned}
$$

From (6.14), we have the following bound for M_1,

$$M_1 \leq 2K^2\|\theta_{t-1} - \theta^*\|^4 + 2\|H\|_2^2\|\theta_{t-1} - \theta^*\|^2. \tag{6.15}$$

Therefore, as θ_{t-1} converges to θ^*, M_1 converges to 0.

The second term has the following inequality,

$$M_2 \leq \mathbb{E}_{\mathbb{P}_X}\left[\Delta(X_t,\theta_{t-1})M_3 + (1-\Delta(X_t,\theta_{t-1}))M_4\right], \tag{6.16}$$

where

$$
\begin{aligned}
M_3 := \mathbb{E}_{\mathbb{P}_{Y|X},\nu}\Bigg[\Bigg\| & w(\theta_{t-1};X_t,A_t)\nabla\ell(\theta_{t-1};X_t,A_t,Y_t) \\
& - w(\theta^*;X_t,A^*)\nabla\ell(\theta^*;X_t,A^*,Y_t)\Bigg\|^2 \mid X_t, A_t \neq A^*\Bigg], \\
M_4 := \mathbb{E}_{\mathbb{P}_{Y|X},\nu}\Bigg[\Bigg\| & w(\theta_{t-1};X_t,A_t)\nabla\ell(\theta_{t-1};X_t,A_t,Y_t) \\
& - w(\theta^*;X_t,A^*)\nabla\ell(\theta^*;X_t,A^*,Y_t)\Bigg\|^2 \mid X_t, A_t = A^*\Bigg].
\end{aligned}
$$

The third term M_3 can be bounded as,

$$\begin{aligned} M_3 &\leq 2\overline{w}^2 \mathbb{E}_{\mathbb{P}_{Y|X},\nu} \left[\|\nabla \ell(\theta_{t-1}; X_t, A_t, Y_t)\|^2 + \|\nabla \ell(\theta^*; X_t, A^*, Y_t)\|^2 \mid X_t, A_t \neq A^* \right] \\ &\leq 4\overline{w}^2 (1 + \|\theta_{t-1} - \theta^*\|^2) \phi(X_t). \end{aligned} \tag{6.17}$$

Finally, we have

$$\begin{aligned} M_4 \leq & \max_{A \in \mathcal{A}} \mathbb{E}_{\mathbb{P}_{Y|X}} [\|w(\theta_{t-1}; X_t, A) \nabla \ell(\theta_{t-1}; X_t, A, Y_t) \\ & \quad - w(\theta^*; X_t, A) \nabla \ell(\theta^*; X_t, A, Y_t)\|^2 \mid X_t, A] \\ & \leq 2 \max_{A \in \mathcal{A}} \mathbb{E}_{\mathbb{P}_{Y|X}} [\|w(\theta_{t-1}; X_t, A) \nabla \ell(\theta_{t-1}; X_t, A, Y_t) \\ & \quad - w(\theta_{t-1}; X_t, A) \nabla \ell(\theta^*; X_t, A, Y_t)\|^2 \mid X_t, A] \\ & \quad + 2 \max_{A \in \mathcal{A}} \mathbb{E}_{\mathbb{P}_{Y|X}} [\|w(\theta_{t-1}; X_t, A) \nabla \ell(\theta^*; X_t, A, Y_t) \\ & \quad - w(\theta^*; X_t, A) \nabla \ell(\theta^*; X_t, A, Y_t)\|^2 \mid X_t, A] \\ & \leq 2\overline{w}^2 \max_{A \in \mathcal{A}} \mathbb{E}_{\mathbb{P}_{Y|X}} \left[\|\nabla \ell(\theta_{t-1}; X_t, A, Y_t) - \nabla \ell(\theta^*; X_t, A, Y_t)\|^2 \mid X_t, A \right] \\ & \quad + 2 \max_{A \in \mathcal{A}} |w(\theta_{t-1}; X_t, A) - w(\theta^*; X_t, A)|^2 \phi(X_t). \end{aligned} \tag{6.18}$$

Combining (6.17) and (6.18) into (6.16), we have

$$\begin{aligned} M_2 \leq & \mathbb{E}_{\mathbb{P}_X} \left[\Delta(X_t, \theta_{t-1}) 4\overline{w}^2 (1 + \|\theta_{t-1} - \theta^*\|^2) \phi(X_t) \right] \\ & + \mathbb{E}_{\mathbb{P}_X} \left[2\overline{w}^2 \max_{A \in \mathcal{A}} \mathbb{E}_{\mathbb{P}_{Y|X}} [\|\nabla \ell(\theta_{t-1}; X_t, A, Y_t) - \nabla \ell(\theta^*; X_t, A, Y_t)\|^2 \mid X_t, A] \right] \\ & + \mathbb{E}_{\mathbb{P}_X} \left[2 \max_{A \in \mathcal{A}} |w(\theta_{t-1}; X_t, A) - w(\theta^*; X_t, A)|^2 \phi(X_t) \right]. \end{aligned} \tag{6.19}$$

Using Assumption 6.4.6, when $\theta_{t-1} \to \theta^*$, we have M_2 converges to 0.

We can now conclude from our above results that

$$\lim_{\theta_t \to \theta^*} \mathbb{E}\|\xi_{\theta_{t-1}}(\theta_{t-1};\zeta) - \xi_{\theta^*}(\theta^*;\zeta)\|^2 = 0.$$

Note that all three conditions in Theorem 2 of [162] are verified under our assumptions, we can conclude that the asymptotic normality result holds in Theorem 6.4.7. □

To emphasize the technical challenge in the theoretical analysis, our loss function $\mathcal{L}$ in (6.11) is not defined by the stable policy as in the prior works [41]. The action $A_t \sim \pi(X_t, \theta_{t-1})$ and $A^* \sim \pi(X_t, \theta^*)$ are no longer in the same probability space, and therefore we specify a coupling between A_t and A^* to compare them. A natural choice is the coupling such that

$$\Delta(X,\theta) = d_{TV}(\pi(X,\theta), \pi(X,\theta^*)) = \frac{1}{2}\sum_{i=1}^{|\mathcal{A}|} |p_i - q_i| = \mathbb{P}(A \neq A^*), \tag{6.20}$$

where $p_i = \Pr(A = A_i), q_i = \Pr(A^* = A_i)$.

To further illustrate our assumptions and central limit theorem result in Theorem 6.4.7, we validate them under two examples we mentioned above, i.e., linear regression (Example 6.2.1) and quantile regression (Example 6.2.3). Under ε-greedy policy defined in Equation (6.8), Theorem 6.4.7 holds for these two cases. In Corollary 6.4.8 below, we demonstrate that Assumptions 6.4.2–6.4.6 are quite natural and can be satisfied by the linear regression example we discussed in Example 6.2.1.

Corollary 6.4.8. *Consider the linear setting defined in Proposition 6.3.4, that the covariate X has finite $\mathbb{E}_{\mathbb{P}_X}\|X\|^4$ and $\mathbb{E}_{\mathbb{P}_X}[XX^\top] \succ 0$. Further assume that the probability density function of X, $p(x)$, is smooth and $\int_{x^\top\theta^*_{[1:p]}=x^\top\theta^*_{[p+1:2p]}} x\otimes x\otimes x p(x)dx$ exists, and the function $\varphi(\cdot) : (0,1) \mapsto \mathbb{R}^+$ is continuous. Assumptions 6.4.2–6.4.6 are satisfied and Theorem 6.4.7 holds.*

Proof. Now let's compute $\mathcal{L}_{\theta'}(\theta)$, under ε-greedy policy defined in (6.8),

$$\begin{aligned}
\mathcal{L}_{\theta'}(\theta) = & \frac{1}{2}\mathbb{E}_{\mathbb{P}}\{\mathbb{E}_{\pi(X,\theta')}[\varphi(\Pr(A \mid X, \theta'))((1-A)(Y - X^\top\theta_{[1:p]})^2 \\
& + A(Y - X^\top\theta_{[p+1:2p]})^2) \mid X]\} \\
= & \frac{1}{2}(1-\frac{\varepsilon}{2})\varphi(1-\frac{\varepsilon}{2})\mathbb{E}_{\mathbb{P}_X}\left[\mathbb{I}\{X^\top\theta'_{[1:p]} > X^\top\theta'_{[p+1:2p]}\}\left(X^\top\theta^*_{[1:p]} - X^\top\theta_{[1:p]}\right)^2\right] \\
& + \frac{\varepsilon}{4}\varphi(\frac{\varepsilon}{2})\mathbb{E}_{\mathbb{P}_X}\left[\mathbb{I}\{X^\top\theta'_{[1:p]} > X^\top\theta'_{[p+1:2p]}\}\left(X^\top\theta^*_{[p+1:2p]} - X^\top\theta_{[p+1:2p]}\right)^2\right] \\
& + \frac{\varepsilon}{4}\varphi(\frac{\varepsilon}{2})\mathbb{E}_{\mathbb{P}_X}\left[\mathbb{I}\{X^\top\theta'_{[1:p]} < X^\top\theta'_{[p+1:2p]}\}\left(X^\top\theta^*_{[1:p]} - X^\top\theta_{[1:p]}\right)^2\right] \\
& + \frac{1}{2}(1-\frac{\varepsilon}{2})\varphi(1-\frac{\varepsilon}{2})\mathbb{E}_{\mathbb{P}_X}[\mathbb{I}\{X^\top\theta'_{[1:p]} < X^\top\theta'_{[p+1:2p]}\} \\
& \left(X^\top\theta^*_{[p+1:2p]} - X^\top\theta_{[p+1:2p]}\right)^2] \\
& + \sigma^2\left[\frac{1}{2}(1-\frac{\varepsilon}{2})\varphi(1-\frac{\varepsilon}{2}) + \frac{\varepsilon}{4}\varphi(\frac{\varepsilon}{2})\right]. \qquad (6.21)
\end{aligned}$$

Obviously, the first part of Assumption 6.4.3 is satisfied under this form of loss function $\mathcal{L}$.

Also, we can calculate the gradient of $\mathcal{L}$ with respect to θ as follows,

$$\begin{aligned}
\nabla_{[1:p]}\mathcal{L}_{\theta'}(\theta) = & (1-\frac{\varepsilon}{2})\varphi(1-\frac{\varepsilon}{2})\mathbb{E}_{\mathbb{P}_X}\left[\mathbb{I}\{X^\top\theta'_{[1:p]} > X^\top\theta'_{[p+1:2p]}\}XX^\top\left(\theta_{[1:p]} - \theta^*_{[1:p]}\right)\right] \\
& + \frac{\varepsilon}{2}\varphi(\frac{\varepsilon}{2})\mathbb{E}_{\mathbb{P}_X}\left[\mathbb{I}\{X^\top\theta'_{[1:p]} < X^\top\theta'_{[p+1:2p]}\}XX^\top\left(\theta_{[1:p]} - \theta^*_{[1:p]}\right)\right] \\
\nabla_{[p+1:2p]}\mathcal{L}_{\theta'}(\theta) = & \frac{\varepsilon}{2}\varphi(\frac{\varepsilon}{2})\mathbb{E}_{\mathbb{P}_X}\left[\mathbb{I}\{X^\top\theta'_{[1:p]} > X^\top\theta'_{[p+1:2p]}\}XX^\top\left(\theta_{[p+1:2p]} - \theta^*_{[p+1:2p]}\right)\right] \\
& + (1-\frac{\varepsilon}{2})\varphi(1-\frac{\varepsilon}{2})\mathbb{E}_{\mathbb{P}_X}[\mathbb{I}\{X^\top\theta'_{[1:p]} < X^\top\theta'_{[p+1:2p]}\} \\
& XX^\top\left(\theta_{[p+1:2p]} - \theta^*_{[p+1:2p]}\right)].
\end{aligned}$$

Therefore, the second part of Assumption 6.4.3 is naturally satisfied since we have the following,

$$\langle \nabla \mathcal{L}_{\theta}(\theta), \theta - \theta^* \rangle \geq \min\left\{(1-\frac{\varepsilon}{2})\varphi(1-\frac{\varepsilon}{2}), \frac{\varepsilon}{2}\varphi(\frac{\varepsilon}{2})\right\} \mathbb{E}_{\mathbb{P}_X}[XX^\top] \\ \left[\left(\theta_{[1:p]} - \theta^*_{[1:p]}\right)^2 + \left(\theta_{[p+1:2p]} - \theta^*_{[p+1:2p]}\right)^2\right].$$

We now consider the Hessian matrix, we have

$$\nabla^2 \mathcal{L}_{\theta'}(\theta) = \begin{bmatrix} H_1 & 0 \\ 0 & H_2 \end{bmatrix},$$

where

$$\begin{aligned} H_1 &= (1-\frac{\varepsilon}{2})\varphi(1-\frac{\varepsilon}{2})\mathbb{E}_{\mathbb{P}_X}\left[\mathbb{I}\{X^\top \theta'_{[1:p]} > X^\top \theta'_{[p+1:2p]}\}XX^\top\right] \\ &\quad + \frac{\varepsilon}{2}\varphi(\frac{\varepsilon}{2})\mathbb{E}_{\mathbb{P}_X}\left[\mathbb{I}\{X^\top \theta'_{[1:p]} < X^\top \theta'_{[p+1:2p]}\}XX^\top\right] \\ H_2 &= \frac{\varepsilon}{2}\varphi(\frac{\varepsilon}{2})\mathbb{E}_{\mathbb{P}_X}\left[\mathbb{I}\{X^\top \theta'_{[1:p]} > X^\top \theta'_{[p+1:2p]}\}XX^\top\right] \\ &\quad + (1-\frac{\varepsilon}{2})\varphi(1-\frac{\varepsilon}{2})\mathbb{E}_{\mathbb{P}_X}\left[\mathbb{I}\{X^\top \theta'_{[1:p]} < X^\top \theta'_{[p+1:2p]}\}XX^\top\right]. \end{aligned}$$

Obviously, the Hessian matrix exists for all $(\theta, \theta') \in \mathbb{R}^d \times \mathbb{R}^d$, and the Hessian matrix at $(\theta^*; \theta^*)$ is positive definite since $\lambda_{\min}\mathbb{E}_{\mathbb{P}_X}[XX^\top] > 0$. We now check the Lipschitz continuity of $\nabla^2 \mathcal{L}_{\theta'}(\theta)$ at (θ^*, θ^*). It is a constant function with respect to θ, so we only need to consider its Lipschitz continuity with respect to θ'.

For a smooth integrable function $p(x)$, define the function $J(\theta) = \int \mathbb{I}(\theta^\top x > c)p(x)dx$. It is easy to see that

$$\nabla J(\theta) = \int_{\theta^\top x = c} p(x)x dx.$$

Apply this formula to H_1 and H_2, we get

$$\frac{\partial}{\partial \theta'_{[1:p]}} H_1 = -\frac{\partial}{\partial \theta'_{[p+1:2p]}} H_1 = -\frac{\partial}{\partial \theta'_{[1:p]}} H_2 = \frac{\partial}{\partial \theta'_{[p+1:2p]}} H_2 \tag{6.22}$$

$$= \Big((1-\frac{\varepsilon}{2})p(1-\frac{\varepsilon}{2}) - \frac{\varepsilon}{2}p(\frac{\varepsilon}{2})\Big) \int_{x^\top \theta'_{[1:p]} = x^\top \theta'_{[p+1:2p]}} x \otimes x \otimes x p(x) dx. \tag{6.23}$$

So $\nabla^2 \mathcal{L}_{\theta'}(\theta)$ is Lipschitz continuous at (θ^*, θ^*) as long as $\int_{x^\top \theta^*_{[1:p]} = x^\top \theta^*_{[p+1:2p]}} x \otimes x \otimes x p(x) dx$ exists. Therefore, we verify Assumption 6.4.4.

For any A, X, we can bound

$$\mathbb{E}_{\mathbb{P}_{Y|X}}(\|\nabla \ell(\theta;\zeta)\|^2 \mid X, A) \leq \|X\|^2\sigma^2 + \|XX^\top XX^\top\| \|\theta - \theta^*\|^2.$$

The matrix S can be computed by

$$S = \mathbb{E}[\nabla \ell(\theta;\zeta)\nabla \ell(\theta;\zeta)^\top] := \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix},$$

where

$$\begin{aligned} S_1 &= (1-\frac{\varepsilon}{2})\varphi^2(1-\frac{\varepsilon}{2})\sigma^2 \mathbb{E}_{\mathbb{P}_X}\left[\mathbb{I}\{X^\top \theta^*_{[1:p]} > X^\top \theta^*_{[p+1:2p]}\} XX^\top\right] \\ &\quad + \frac{\varepsilon}{2}\varphi^2(\frac{\varepsilon}{2})\sigma^2 \mathbb{E}_{\mathbb{P}_X}\left[\mathbb{I}\{X^\top \theta^*_{[1:p]} < X^\top \theta^*_{[p+1:2p]}\} XX^\top\right], \\ S_2 &= (1-\frac{\varepsilon}{2})\varphi^2(1-\frac{\varepsilon}{2})\sigma^2 \mathbb{E}_{\mathbb{P}_X}\left[\mathbb{I}\{X^\top \theta^*_{[1:p]} < X^\top \theta^*_{[p+1:2p]}\} XX^\top\right] \\ &\quad + \frac{\varepsilon}{2}\varphi^2(\frac{\varepsilon}{2})\sigma^2 \mathbb{E}_{\mathbb{P}_X}\left[\mathbb{I}\{X^\top \theta^*_{[1:p]} > X^\top \theta^*_{[p+1:2p]}\} XX^\top\right]. \end{aligned}$$

So Assumption 6.4.5 is satisfied with $\phi(X) = \|X\|^2\sigma^2 + \|XX^\top\|^2$.

By definition

$$\Delta(X, \theta) = (1-\varepsilon)\left|\mathbb{I}\left(X^\top \theta^*_{[1:p]} > X^\top \theta^*_{[p+1:2p]}\right) - \mathbb{I}\left(X^\top \theta_{[1:p]} > X^\top \theta_{[p+1:2p]}\right)\right|.$$

Take any convergent sequence $\theta_n \to \theta^*$. It is clear that $\Delta(X, \theta_n)\phi(X)$ converges to 0 almost surely. Furthermore, $|\Delta(X, \theta_n)\phi(X)| \leq \|X\|^4$ and $\mathbb{E}\|X\|^4 \leq \infty$. By dominated convergence theorem, $\lim_{n\to\infty} \mathbb{E}[\Delta(X, \theta_n)\phi(X)] = 0$.

$$
\begin{aligned}
&|w(\theta_{t-1}; X_t, A) - w(\theta^*; X_t, A)| \\
&= |\varphi(\Pr(A_t|X_t; \theta_{t-1})) - \varphi(\Pr(A|X_t; \theta^*))| \\
&= |\varphi(1 - \frac{\varepsilon}{2}) - \varphi(\frac{\varepsilon}{2})| \left| \mathbb{I}\left(X^\top \theta^*_{[1:p]} > X^\top \theta^*_{[p+1:2p]}\right) - \mathbb{I}\left(X^\top \theta_{[1:p]} > X^\top \theta_{[p+1:2p]}\right) \right|.
\end{aligned}
$$

Using the same argument as above, we can derive that

$$
\lim_{\theta\to\theta^*} \mathbb{E}_{\mathbb{P}_X}\left[|w(\theta; X, A) - w(\theta^*; X, A)|^2 \phi(X) \mid A\right] = 0.
$$

Finally, we have the following inequality,

$$
\mathbb{E}_{\mathbb{P}}\left[\|\nabla \ell(\theta; \zeta) - \nabla \ell(\theta^*; \zeta)\|^2 \mid A\right] \leq \mathbb{E}_{\mathbb{P}_X}\left\|XX^\top(\theta - \theta^*)\right\|^2.
$$

Therefore, Assumption 6.4.6 is satisfied since

$$
\lim_{\theta\to\theta^*} \mathbb{E}_{\mathbb{P}}\left[\|\nabla \ell(\theta; \zeta) - \nabla \ell(\theta^*; \zeta)\|^2 \mid A\right] = 0 \square
$$

As discussed earlier, our assumption allows a much broader setting than the class of smooth individual loss functions. Under our assumptions, the individual loss function $\ell(\theta; \zeta)$ can be non-smooth. We will justify this argument in the quantile regression example below.

Corollary 6.4.9. *Consider the quantile regression setting defined in Example 6.2.3, assume that*

- *The covariate X has finite $\mathbb{E}_{\mathbb{P}_X}[XX^\top]$ and $\mathbb{E}_{\mathbb{P}_X}[XX^\top] \succ 0$;*
- *The p.d.f. of X, denoted as $p(x)$, is smooth and $\int_{x^\top \theta^*_{[1:p]} = x^\top \theta^*_{[p+1:2p]}} x \otimes x \otimes x p(x) dx$*

exists;

- *The p.d.f. of $\mathcal{E}$, denoted as $q(x)$, is smooth and bounded. Also, $q(0) > 0$ and $q'(x)$ is bounded;*

- *Assume $w_t(A_t, X_t, \theta_{t-1}) = \varphi(\Pr(A_t \mid X_t, \theta_{t-1}))$ for some continuous function $\varphi(\cdot) : (0, 1) \mapsto \mathbb{R}^+$.*

Under the above conditions, the Assumption 6.4.2 to Assumption 6.4.6 are satisfied and Theorem 6.4.7 holds.

Proof. We first define $\psi_\tau(u)$ as follows,

$$\psi_\tau(u) = \mathbb{E}_{\mathcal{E}}\rho_\tau(u + \mathcal{E}) = \int_{-\infty}^{-u} (u + x)(\tau - 1)q(x)dx + \int_{-u}^{\infty} (u + x)\tau q(x)dx.$$

The first and second order derivative of $\psi_\tau(u)$ can be computed as

$$\begin{aligned}
\psi'_\tau(u) &= \mathbb{E}_{\mathcal{E}}\rho'_\tau(u + \mathcal{E}) = \int_{-\infty}^{-u} (\tau - 1)q(x)dx + \int_{-u}^{\infty} \tau q(x)dx, \\
\psi''_\tau(u) &= -(\tau - 1)q(-u) + \tau q(-u) = q(-u).
\end{aligned}$$

Because $\psi'_\tau(0) = 0, \psi''_\tau(0) = q(0) > 0$, there exists $\delta > 0$ such that for all $|u| < \delta$,

$$\psi'_\tau(u)u \geq \frac{1}{2}q(0)u^2.$$

Now let's compute $\nabla \mathcal{L}_{\theta'}(\theta)$, under ε-greedy policy defined in (6.8).

$$\begin{aligned}
\nabla_{[1:p]} \mathcal{L}_{\theta'}(\theta) = & (1-\frac{\varepsilon}{2})\varphi(1-\frac{\varepsilon}{2})\mathbb{E}_{\mathbb{P}_X}[\mathbb{I}\{X^\top \theta'_{[1:p]} > X^\top \theta'_{[p+1:2p]}\} \\
& X\psi'_\tau\left(X^\top\left(\theta_{[1:p]} - \theta^*_{[1:p]}\right)\right)] \\
& + \frac{\varepsilon}{2}\varphi(\frac{\varepsilon}{2})\mathbb{E}_{\mathbb{P}_X}\left[\mathbb{I}\{X^\top \theta'_{[1:p]} < X^\top \theta'_{[p+1:2p]}\} X\psi'_\tau\left(X^\top\left(\theta_{[1:p]} - \theta^*_{[1:p]}\right)\right)\right] \\
\nabla_{[p+1:2p]} \mathcal{L}_{\theta'}(\theta) = & \frac{\varepsilon}{2}\varphi(\frac{\varepsilon}{2})\mathbb{E}_{\mathbb{P}_X}[\mathbb{I}\{X^\top \theta'_{[1:p]} > X^\top \theta'_{[p+1:2p]}\} \\
& X\psi'_\tau\left(X^\top\left(\theta_{[p+1:2p]} - \theta^*_{[p+1:2p]}\right)\right)] \\
& + (1-\frac{\varepsilon}{2})\varphi(1-\frac{\varepsilon}{2})\mathbb{E}_{\mathbb{P}_X}[\mathbb{I}\{X^\top \theta'_{[1:p]} < X^\top \theta'_{[p+1:2p]}\} \\
& X\psi'_\tau\left(X^\top\left(\theta_{[p+1:2p]} - \theta^*_{[p+1:2p]}\right)\right)].
\end{aligned}$$

Because $\mathbb{E}[XX^\top]$ is positive definite, there exists a constant $C > 0$ such that $\mathbb{E}[\mathbb{I}\{\|X\| < C\}XX^\top]$ is positive definite. For any $\|\theta_{[1:p]} - \theta^*_{[1:p]}\| \leq \delta/C$, we have the following,

$$\begin{aligned}
& \langle \nabla_{[1:p]} \mathcal{L}_\theta(\theta), \theta_{[1:p]} - \theta^*_{[1:p]} \rangle \\
\geq & \frac{1}{2} q(0) \min\left\{(1-\frac{\varepsilon}{2})\varphi(1-\frac{\varepsilon}{2}), \frac{\varepsilon}{2}\varphi(\frac{\varepsilon}{2})\right\} \mathbb{E}_{\mathbb{P}_X}[X^\top\left(\theta_{[1:p]} - \theta^*_{[1:p]}\right) \\
& \psi'_\tau\left(X^\top\left(\theta_{[1:p]} - \theta^*_{[1:p]}\right)\right)] \\
\geq & C'\mathbb{E}_{\mathbb{P}_X}\left[\mathbb{I}\{|X^\top \theta_{[1:p]} - X^\top \theta^*_{[p+1:2p]}| < \delta\} \left|X^\top\left(\theta_{[1:p]} - \theta^*_{[1:p]}\right)\right|^2\right] \\
\geq & C'\mathbb{E}_{\mathbb{P}_X}\left[\mathbb{I}\{\|X\| < C\}XX^\top\right]\left\|\theta_{[1:p]} - \theta^*_{[1:p]}\right\|^2 \\
\geq & C'\left\|\theta_{[1:p]} - \theta^*_{[1:p]}\right\|^2,
\end{aligned}$$

for some constant $C' > 0$. So the second part of Assumption 6.4.3 is satisfied.

We now consider the Hessian matrix, we have

$$\nabla^2 \mathcal{L}_{\theta'}(\theta) = \begin{bmatrix} H_1 & 0 \\ 0 & H_2 \end{bmatrix},$$

where

$$\begin{aligned}
H_1 &= (1-\frac{\varepsilon}{2})\varphi(1-\frac{\varepsilon}{2})\mathbb{E}_{\mathbb{P}_X}\left\{\mathbb{I}\{X^\top\theta'_{[1:p]} > X^\top\theta'_{[p+1:2p]}\}XX^\top p\left[X^\top\left(\theta_{[1:p]}-\theta^*_{[1:p]}\right)\right]\right\} \\
&\quad + \frac{\varepsilon}{2}\varphi(\frac{\varepsilon}{2})\mathbb{E}_{\mathbb{P}_X}\left\{\mathbb{I}\{X^\top\theta'_{[1:p]} < X^\top\theta'_{[p+1:2p]}\}XX^\top p\left[X^\top\left(\theta_{[1:p]}-\theta^*_{[1:p]}\right)\right]\right\} \\
H_2 &= \frac{\varepsilon}{2}\varphi(\frac{\varepsilon}{2})\mathbb{E}_{\mathbb{P}_X}\left\{\mathbb{I}\{X^\top\theta'_{[1:p]} > X^\top\theta'_{[p+1:2p]}\}XX^\top p\left[X^\top\left(\theta_{[p+1:2p]}-\theta^*_{[p+1:2p]}\right)\right]\right\} \\
&\quad + (1-\frac{\varepsilon}{2})\varphi(1-\frac{\varepsilon}{2})\mathbb{E}_{\mathbb{P}_X}\{\mathbb{I}\{X^\top\theta'_{[1:p]} < X^\top\theta'_{[p+1:2p]}\} \\
&\quad XX^\top p\left[X^\top\left(\theta_{[p+1:2p]}-\theta^*_{[p+1:2p]}\right)\right]\}.
\end{aligned}$$

Obviously, the Hessian matrix exists for all $(\theta,\theta')\in\mathbb{R}^d\times\mathbb{R}^d$, and the Hessian matrix at $(\theta^*;\theta^*)$ is positive definite since $\lambda_{\min}\mathbb{E}_{\mathbb{P}_X}[XX^\top]>0$. We now check the Lipschitz continuity of $\nabla^2\mathcal{L}_{\theta'}(\theta)$ at (θ^*,θ^*). Its Lipschitz continuity with respect to θ' can be checked by the same argument as in the linear case. It is clear differentiable with respect to θ, so it is also Lipschitz continuous with respect to θ. Therefore, we verify Assumption 6.4.4.

For any A, X, we can bound

$$\mathbb{E}_{\mathbb{P}_{Y|X}}(\|\nabla\ell(\theta;\zeta)\|^2 \mid X, A) \le \|X\|^2.$$

The matrix S can be computed by

$$S=\mathbb{E}[\nabla\ell(\theta;\zeta)\nabla\ell(\theta;\zeta)^\top] := \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix},$$

where

$$\begin{aligned}
S_1 &= (1-\frac{\varepsilon}{2})\varphi^2(1-\frac{\varepsilon}{2})\tau(1-\tau)\mathbb{E}_{\mathbb{P}_X}\left[\mathbb{I}\{X^\top\theta^*_{[1:p]} > X^\top\theta^*_{[p+1:2p]}\}XX^\top\right] \\
&\quad + \frac{\varepsilon}{2}\varphi^2(\frac{\varepsilon}{2})\tau(1-\tau)\mathbb{E}_{\mathbb{P}_X}\left[\mathbb{I}\{X^\top\theta^*_{[1:p]} < X^\top\theta^*_{[p+1:2p]}\}XX^\top\right], \\
S_2 &= (1-\frac{\varepsilon}{2})\varphi^2(1-\frac{\varepsilon}{2})\tau(1-\tau)\mathbb{E}_{\mathbb{P}_X}\left[\mathbb{I}\{X^\top\theta^*_{[1:p]} < X^\top\theta^*_{[p+1:2p]}\}XX^\top\right] \\
&\quad + \frac{\varepsilon}{2}\varphi^2(\frac{\varepsilon}{2})\tau(1-\tau)\mathbb{E}_{\mathbb{P}_X}\left[\mathbb{I}\{X^\top\theta^*_{[1:p]} > X^\top\theta^*_{[p+1:2p]}\}XX^\top\right].
\end{aligned}$$

So Assumption 6.4.5 is satisfied with $\phi(X) = \|X\|^2$.

Using the same argument as above, we can derive that

$$\begin{aligned}
&\lim_{\theta\to\theta^*}\mathbb{E}_{\mathbb{P}_X}[\Delta(X,\theta)\phi(X)] = 0, \\
&\lim_{\theta\to\theta^*}\mathbb{E}_{\mathbb{P}_X}\left[|w(\theta;X,A) - w(\theta^*;X,A)|^2\phi(X) \mid A\right] = 0.
\end{aligned}$$

Finally, we have the following inequality,

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_{\mathcal{E}}}\left(\rho'_\tau(u+\mathcal{E}) - \rho'_\tau(\mathcal{E})\right)^2 &= \mathbb{E}_{\mathbb{P}_{\mathcal{E}}}\left(\mathbb{I}\{u+\mathcal{E}<0\} - \mathbb{I}\{\mathcal{E}<0\}\right)^2 \\
&\le \Pr(|\mathcal{E}| < |u|).
\end{aligned}$$

So we can bound

$$\mathbb{E}_{\mathbb{P}}\left[\|\nabla\ell(\theta;\zeta) - \nabla\ell(\theta^*;\zeta)\|^2 \mid A\right] \le \mathbb{E}_{\mathbb{P}_X}\left[\|X\|^2\Pr(|\mathcal{E}| < |X^\top(\theta-\theta^*)|)\right].$$

Again, we can use dominated convergence theorem to prove this term converges to 0 as $\theta\to\theta^*$. Therefore, Assumption 6.4.6 is satisfied. □

Corollary 6.4.9 states that we can also obtain the limiting distribution for some non-smooth loss functions like a quantile loss.

In Corollary 6.4.8 and Corollary 6.4.9 above, we use the ε-greedy policy with fixed constant $\varepsilon \in (0,1)$ throughout the whole SGD process. This policy can be relaxed to a general ε_t-greedy policy, for some deterministic sequence $\{\varepsilon_t\}$ varying with respect to time t, such that $\varepsilon_t \in (0,1)$ and $\varepsilon_t \to \varepsilon_\infty$. The asymptotic normality result also holds under this setting. To illustrate the method, we work under the linear regression setting with assumptions in Corollary 6.4.8. The new policy $A_t \sim \pi_t$ is defined by

$$\Pr(A_t = 0 \mid X_t, \theta_{t-1}) = (1-\varepsilon_t)\mathbb{I}\{X_t^\top \theta_{[1:p],t-1} > X_t^\top \theta_{[p+1:2p],t-1}\} + \frac{\varepsilon_t}{2},$$

instead of (6.8). The weight w_t is again defined as some functions of $\Pr(A_t = 0 \mid X_t, \theta_{t-1})$. Assume $\lim_{t\to\infty} \varepsilon_t = \varepsilon_\infty$, for some constant $\varepsilon_\infty \in (0,1)$. Notice that ε_t is a deterministic sequence, meaning it does not change with respect to X, A, Y and θ, θ'.

The definition of $\mathcal{L}_{\theta'}(\theta)$ should be change accordingly, i.e.,

$$\mathcal{L}_{t,\theta'}(\theta) = \mathbb{E}_{\mathbb{P}}\left[\mathbb{E}_{\pi_t(X,\theta')}\left(w_t(\theta'; X, A)\ell(\theta; X, A, Y) \mid X\right)\right],$$
$$\mathcal{L}_{\infty,\theta'}(\theta) = \mathbb{E}_{\mathbb{P}}\left[\mathbb{E}_{\pi_\infty(X,\theta')}\left(w_\infty(\theta'; X, A)\ell(\theta; X, A, Y) \mid X\right)\right].$$

Furthermore, the matrix H, S should be defined with respect to $\mathcal{L}_\infty$.

Theorem 6.4.10. *Under the ε_t-greedy policy we discussed above, with the same conditions as Corollary 6.4.8, the asymptotic normality also holds for averaged SGD estimator $\bar{\theta}_t$, i.e.,*

$$\sqrt{t}(\bar{\theta}_t - \theta^*) \to N(0, H^{-1}SH^{-1}).$$

Proof. We will follow the steps in the proof of Theorem 6.4.7. To simplify the notation, we denote $R(\theta)$ as $\nabla\mathcal{L}_{\infty,\theta}(\theta)$, ξ_t as $w_t(\theta_{t-1}; X, A)\nabla\ell(\theta_{t-1}; \zeta) - \nabla\mathcal{L}_{\infty,\theta_{t-1}}(\theta_{t-1})$, $\xi_t(0)$ as $w_\infty(\theta^*; X, A)\nabla\ell(\theta^*; \zeta) - \nabla\mathcal{L}_{\infty,\theta^*}(\theta^*)$, and $\xi_t(\theta_{t-1})$ as $\xi_t - \xi_t(0)$.

Because $\varepsilon_t \to \varepsilon_\infty$, ε_t is uniformly bounded away from 0 and 1 for sufficiently large t. So we still have the following inequality,

$$\mathbb{E}[\|\xi_t\|^2 \mid \mathcal{F}_{t-1}] + R(\theta_{t-1})^2 \leq K_2(1 + \|\theta_{t-1}\|^2).$$

The only thing that remains unproved is

$$\mathbb{E}[\|\xi_t(\theta_{t-1})\|^2 \mid \mathcal{F}_{t-1}] \leq \delta(\theta_t),$$

with $\lim_{\theta \to 0} \delta(\theta) = 0$.

Similarly, $\{\xi_t(0)\}$ are *i.i.d.*, and $\xi_t(0)$ can be coupled with ξ_t so that the distance between them can be measured in TV distance between π_t and π_∞.

$$\begin{aligned}
&\mathbb{E}[\|\xi_t(\theta_{t-1})\|^2 \mid \mathcal{F}_{t-1}] \\
&= \mathbb{E}[\|w_t(\theta_{t-1}; X, A)\nabla\ell(\theta_{t-1}; \zeta) - \nabla\mathcal{L}_{\infty,\theta_{t-1}}(\theta_{t-1}) \\
&\quad - w_\infty(\theta^*; X, A^*)\nabla\ell(\theta; \zeta) + \nabla\mathcal{L}_{\infty,\theta^*}(\theta^*)\|^2 \mid \theta_{t-1}] \\
&\leq C\mathbb{E}[\|w_t(\theta_{t-1}; X, A)\nabla\ell(\theta_{t-1}; \zeta) - w_\infty(\theta^*; X, A^*)\nabla\ell(\theta; \zeta)\|^2 \mid \theta_{t-1}] \\
&\quad + C\mathbb{E}[\|\nabla\mathcal{L}_{\infty,\theta_{t-1}}(\theta_{t-1}) - \nabla\mathcal{L}_{\infty,\theta^*}(\theta^*)\|^2 \mid \theta_{t-1}].
\end{aligned}$$

The second term has been bounded by (6.15), the first term can be further decompose as

$$\begin{aligned}
&\mathbb{E}[\|w_t(\theta_{t-1}; X, A)\nabla\ell(\theta_{t-1}; \zeta) - w_\infty(\theta^*; X, A^*)\nabla\ell(\theta^*; \zeta)\|^2 \mid \theta_{t-1}] \\
&\leq C\mathbb{E}_{\mathbb{P}_X}\left[\Delta_t(X_t, \theta_{t-1})(1 + \|\theta_{t-1} - \theta^*\|^2)\phi(X_t)\right] \\
&\quad + C\mathbb{E}\left[\max_{A \in \mathcal{A}} \mathbb{E}_{\mathbb{P}_{Y|X}}[\|\nabla\ell(\theta_{t-1}; X, A, Y) - \nabla\ell(\theta^*; X, A, Y)\|^2 \mid X, A]\right] \\
&\quad + C\mathbb{E}_{\mathbb{P}_X}\left[\max_{A \in \mathcal{A}} |w_t(\theta_{t-1}; X_t, A) - w_\infty(\theta^*; X_t, A)|^2(1 + \|\theta_{t-1} - \theta^*\|^2)\phi(X_t)\right],
\end{aligned}$$

where $\Delta_t(X, \theta) = d_{TV}(\pi_t(X, \theta), \pi_\infty(X, \theta^*))$. This decomposition is similar to (6.19). We now have

$$d_{TV}(\pi_t(X,\theta),\pi_\infty(X,\theta^*)) \leq C d_{TV}(\pi_\infty(X,\theta),\pi_\infty(X,\theta^*)) + C|\varepsilon_t - \varepsilon_\infty|$$

Similarly, we have the following upper bound as well,

$$\begin{aligned}
&|w_t(\theta_{t-1};X_t,A) - w_\infty(\theta^*;X_t,A)| \\
\leq\, & C|\Pr_{\pi_t}(A_t|X_t;\theta_{t-1}) - \Pr_{\pi_\infty}(A|X_t;\theta^*)| \\
\leq\, & C|\varepsilon_t - \varepsilon_\infty| + \left|\mathbb{I}\left(X^\top\theta^*_{[1:p]} > X^\top\theta^*_{[p+1:2p]}\right) - \mathbb{I}\left(X^\top\theta_{[1:p]} > X^\top\theta_{[p+1:2p]}\right)\right|.
\end{aligned}$$

Combining these bounds above and Theorem 6.4.7, it is sufficient to guarantee the validity of the central limit theorem result.

Also notice that, from the above proof, it is easy to see that as long as $|\varepsilon_t - \varepsilon_\infty| = \mathcal{O}(t^{-\alpha/2})$, we can still obtain the same result as Theorem 6.4.11. □

In order to provide statistical inference for the model parameter, we need to estimate the variance of $\hat{\theta}_t$, $H^{-1}SH^{-1}$, as we established in Theorem 6.4.7, in a fully online fashion. A few options have been provided from SGD inference literature, e.g., the plug-in estimator [44, 41], the batch-means estimator [44, 211], the bootstrap estimator [72], the random scaling estimator [136]. Among the above, the plug-in estimator is expected to achieve a very good numerical behavior as evident from classical SGD approaches. In this paper, we use the plug-in estimator [44] for smooth loss functions ℓ, and leave the other methods as an interesting future work. In adaptive settings, the online plugin estimators for S and H are given by,

$$\hat{S}_n = \frac{1}{n}\sum_{t=1}^{n} w_t^2 \nabla\ell(\theta_{t-1};\zeta_t)\nabla\ell(\theta_{t-1};\zeta_t)^\top, \quad \hat{H}_n = \frac{1}{n}\sum_{t=1}^{n} w_t \nabla^2\ell(\theta_{t-1};\zeta_t).$$

With the plug-in estimators $(\hat{S}_t, \hat{H}_t)$, an online plug-in inference procedure can be provided by replacing S and H in the asymptotic covariance matrix in Theorem 6.4.7 to $(\hat{S}_t, \hat{H}_t)$. We defer the detailed procedure to Section 6.5.2 below.

6.4.2 Bahadur representations

In this section, we further present the Bahadur representation of our weighted SGD update (6.7) under the adaptive data collection environment. Aside from the asymptotic normality result in Theorem 6.4.7, the Bahadur representation characterizes the remainder term beyond the normal approximation, which helps conduct a finer convergence analysis of the proposed estimator. The Bahadur representation was first studied in [12] for quantile regression, and generalized to M-estimators by [33, 97] and many others. For the SGD estimator under classical non-adaptive settings (6.2), the Bahadur representation can be inferred by the proof of Theorem 2 in [162] as,

$$\sqrt{t}\Sigma^{-1/2}(\bar{\theta}_t^{(\mathtt{SGD})} - \theta^*) = W + \mathcal{O}_p\big(t^{-\alpha+\frac{1}{2}} + t^{-\frac{\alpha}{2}} + t^{\alpha-1}\big), \tag{6.24}$$

where $\Sigma = H^{(\mathtt{SGD})-1}S^{(\mathtt{SGD})}H^{(\mathtt{SGD})-1}$, and W is the leading term as a sum of independent variables that converges to a standard normal distribution as $t \to \infty$. The other term on the right-hand side is a higher-order remainder term that converges faster than the leading term W under common regularity conditions. In the following theorem, we provide the Bahadur representation of the proposed weighted SGD (6.7) under adaptive settings.

Theorem 6.4.11. *Under the conditions in Theorem 6.4.7 and ε-greedy algorithm defined in* (6.8)*, we further assume:*

- *There exists constant $C_1 > 0$, such that $\int_{x^\top\theta_{[1:p]}=x^\top\theta_{[p+1:2p]}} x \otimes x \otimes x p(x)dx \leq C_1$ for all θ;*

- *Given θ, θ^*, the following inequality holds for some constant $C_2 > 0$,*

$$\mathbb{E}\left[\left|\mathbb{I}\left(X^\top \theta^*_{[1:p]} > X^\top \theta^*_{[p+1:2p]}\right) - \mathbb{I}\left(X^\top \theta_{[1:p]} > X^\top \theta_{[p+1:2p]}\right)\right|(1+\|X\|^4)\right]$$
$$\leq C_2\|\theta - \theta^*\|.$$

We have the following decomposition

$$\sqrt{t}\Sigma^{-1/2}(\bar{\theta}_t - \theta^*) = \underbrace{\frac{1}{\sqrt{t}}\sum_{i=1}^{t-1}\Sigma_t^{-1/2}Q_i^t\xi_{\theta^*}(\theta^*;\zeta_i)}_{W} \tag{6.25}$$
$$+ \underbrace{\frac{1}{\sqrt{t}}\sum_{i=1}^{t-1}\Sigma^{-1/2}Q_i^t(\xi_{\theta_{i-1}}(\theta_{i-1};\zeta_i) - \xi_{\theta^*}(\theta^*;\zeta_i))}_{R_1}$$
$$+ \underbrace{\frac{1}{\sqrt{t}\eta_0}\Sigma^{-1/2}Q_0^t(\theta_0 - \theta^*)}_{R_2} + \underbrace{\frac{1}{\sqrt{t}}\sum_{i=1}^{t-1}\Sigma^{-1/2}Q_i^t(\mathcal{L}_{\theta_i}(\theta_i) - H(\theta_i - \theta^*))}_{R_3}$$
$$+ \underbrace{\frac{1}{\sqrt{t}}\sum_{i=1}^{t-1}(\Sigma^{-1/2} - \Sigma_t^{-1/2})Q_i^t\xi_{\theta^*}(\theta^*;\zeta_i)}_{R_4}$$
$$= W + R_1 + R_2 + R_3 + R_4, \tag{6.26}$$

where $\mathbb{E}[W] = 0, \mathbb{E}[WW^\top] = I_d$, $\Sigma_t = \frac{1}{t}\sum_{i=1}^{t-1} Q_i^t S Q_i^t$, and $Q_i^t = \eta_i \sum_{j=i}^{t-1}\prod_{k=i+1}^{j}(I_d - \eta_k H)$ for $t > 0$. Furthermore, we have,

$$\mathbb{E}\|R_1\|^2 \lesssim t^{-\frac{\alpha}{2}},\ \ \mathbb{E}\|R_2\|^2 \lesssim t^{-1},\ \ \mathbb{E}\|R_3\| \lesssim t^{-\alpha+\frac{1}{2}},\ \ \mathbb{E}\|R_4\|^2 \lesssim t^{2\alpha-2}.$$

Proof. To address the randomness in the adaptive policy A_t, it is necessary to define a coupling for all categorical distributions with $|\mathcal{A}|$ categories simultaneously. Previously in the proof of Theorem 6.4.7, we used the total variation distance to bound $\mathbb{P}(A_t \neq A^*)$. Here we need a generalized coupling defined as follows.

Consider the $(|\mathcal{A}|-1)$-simplex $S=\{(x_1,\dots,x_{|\mathcal{A}|}) \mid x_i \geq 0, \sum x_i = 1\}$. It has $|\mathcal{A}|$ vertices given by $V_i=(0,\dots,0,1,0,\dots,0)$ where 1 is in the i-th coordinate. Take a point P uniformly from S. For any categorical distribution with probability $(p_1,\dots,p_{|\mathcal{A}|})$, define $K=(p_1,\dots,p_{|\mathcal{A}|})$. The probability that P lies in the sub-simplex with vertices $\{V_1,\dots,\hat{V_i},\dots,V_{|\mathcal{A}|},K\}$ (V_i is deleted) is exactly p_i. Thus, K gives a partition of S that has the required categorical distribution and we can use this to define the action A. Furthermore, given two different distribution K,K', it is easy to see that the quantity $\mathbb{P}(A\neq A')$ is bounded by $Cd_{TV}(K,K')$, where C is some positive constant which only depends on $|\mathcal{A}|$. So all previous bounds still holds up to a constant.

In conclusion, the probability space we have used for stochastic gradient descent can be redefined using *i.i.d.* random variables $(X_t,Y_t,P_t), t\geq 1$, where P_t obeys a uniform distribution on a $(|\mathcal{A}|-1)$-simplex. We also redefine $\zeta_t=(X_t,Y_t,P_t)$.

We would like to note that in this proof and the proofs thereafter, with a slight abuse of notation, we will use C to represent different positive constants.

To prove the bounds for R_1,R_2,R_3,R_4, we need some preliminary results. The bounds in Assumption 6.4.3 and Assumption 6.4.4 actually hold globally by our assumptions. That is,

$$\langle \nabla\mathcal{L}_\theta(\theta),\theta-\theta^*\rangle \geq \lambda\|\theta-\theta^*\|^2,$$
$$\left\|\nabla^2\mathcal{L}_{\theta'}(\theta)-\nabla^2\mathcal{L}_{\theta^*}(\theta^*)\right\| \leq K\|\theta-\theta^*\|+K\|\theta'-\theta^*\|,$$

for all θ,θ'. The second inequality comes from Assumption (a). So we can estimate

$$\|\nabla\mathcal{L}_{\theta_t}(\theta_t)-H(\theta_t-\theta^*)\| \leq C\|\theta_t-\theta^*\|^2. \tag{6.27}$$

We also have

$$
\begin{aligned}
&\mathbb{E}\left[\|\xi_{\theta_{t-1}}(\theta_{t-1};\zeta_t)+\nabla\mathcal{L}_{\theta_{t-1}}(\theta_{t-1})\|^4 \mid |\theta_{t-1}\right] \\
&=\mathbb{E}\left[\|w(\theta_{t-1};\zeta_t)\nabla\ell(\theta_{t-1};\zeta_t)\|^4 \mid |\theta_{t-1}\right] \\
&\leq \bar{w}^4\mathbb{E}\left[\|\nabla\ell(\theta_{t-1};\zeta_t)\|^4 \mid |\theta_{t-1}\right].
\end{aligned}
$$

Therefore, for some positive constant $C > 0$,

$$
\begin{aligned}
&\mathbb{E}\left[\|\xi_{\theta_{t-1}}(\theta_{t-1};\zeta_t)+\nabla\mathcal{L}_{\theta_{t-1}}(\theta_{t-1})\|^4 \mid |\theta_{t-1}\right] \\
&\leq C(\sigma^4\mathbb{E}\|X\|^4+\mathbb{E}\|XX^\top(\theta_{t-1}-\theta^*)\|^4) \\
&\leq C(1+\|\theta_{t-1}-\theta^*\|^4).
\end{aligned}
$$

The above argument implies that

$$
\mathbb{E}\left[\|\xi_{\theta_{t-1}}(\theta_{t-1};\zeta_t)\|^4 \mid |\theta_{t-1}\right] \leq C(1+\|\theta_{t-1}-\theta^*\|^4),
$$

and $\mathbb{E}\|\xi_{\theta^*}(\theta^*;\zeta_t)\|^4 \leq C$. We also utilize the following bounds from [44],

$$
\begin{aligned}
\mathbb{E}\|\theta_t-\theta^*\|^2 &\leq Ct^{-\alpha}, \\
\mathbb{E}\|\theta_t-\theta^*\|^4 &\leq Ct^{-2\alpha}.
\end{aligned}
$$

These two inequalities above are also derived as Lemma 5.12 and 5.14 in [176].

From [162], we know that $\|Q_i^t\| \leq C$. Moreover, $\|\Sigma_t^{-1/2}Q_i^t\| \leq C$ is guaranteed in [176].

In the proof of Lemma 1 of [162], it can be seen that,

$$\begin{aligned}
H^{-1} - Q_i^t &= H^{-1} - \eta_i \sum_{j=i}^{t-1} \prod_{k=i+1}^{j} (I_d - \eta_k H) \\
&= \sum_{j=i}^{t-1} (\eta_j - \eta_i) \prod_{k=i+1}^{j} (I_d - \eta_k H) + H^{-1} \prod_{k=i+1}^{t} (I_d - \eta_k H),
\end{aligned}$$

and the first term is $O(i^{\alpha-1})$. In Lemma D.2 of [44], it is proved that

$$\left\| \prod_{k=i+1}^{t} (I_d - \eta_k H) \right\| \leq e^{-C(t-i)\eta_t}.$$

So we have

$$\begin{aligned}
\|\Sigma_t - \Sigma\| &\leq \frac{C}{t} + \frac{C}{t} \sum_{i=1}^{t} (i^{\alpha-1} + e^{-C(t-i)\eta_t}) \\
&\leq Ct^{-1} + Ct^{\alpha-1} + \frac{e^{-C\eta_t}}{t(1-e^{-C\eta_t})} \\
&\leq Ct^{-1} + Ct^{\alpha-1} + Ct^{-\alpha-1} \\
&\leq Ct^{\alpha-1}. \qquad (6.28)
\end{aligned}$$

Now we proceed to the main part of the proof. Similar to (6.19), we have

$$\begin{aligned}
&\mathbb{E}\|\xi_{\theta_{t-1}}(\theta_{t-1}; \zeta_t) - \xi_{\theta^*}(\theta^*; \zeta_t)\|^2 \\
&\leq C\mathbb{E}[(\Delta(X, \theta) + \max_{A \in \mathcal{A}} |w(\theta; X, A) - w(\theta^*; X, A)|^2)(1 + \|\theta - \theta^*\|^2)\phi(X)] + C\|\theta - \theta^*\|^2.
\end{aligned}$$

We can further give the following inequality, following the steps in the proof of Theorem 6.4.7,

$$\begin{aligned}
&|w(\theta; X, A) - w(\theta^*; X, A)|^2 \\
&\leq (\underline{w} - \overline{w})^2 \left| \mathbb{I}\left(X^\top \theta^*_{[1:p]} > X^\top \theta^*_{[p+1:2p]}\right) - \mathbb{I}\left(X^\top \theta_{[1:p]} > X^\top \theta_{[p+1:2p]}\right) \right|,
\end{aligned}$$

and the same bound also holds for $\Delta(X, \theta)$ with a different constant. So the whole term can be estimated by

$$\begin{aligned}
&\mathbb{E}[(\Delta(X,\theta) + \max_{A \in \mathcal{A}} |w(\theta; X, A) - w(\theta^*; X, A)|^2)(1 + \|\theta - \theta^*\|^2)\phi(X)] \\
&\leq C\mathbb{E}\left[\left|\mathbb{I}\left(X^\top \theta^*_{[1:p]} > X^\top \theta^*_{[p+1:2p]}\right) - \mathbb{I}\left(X^\top \theta_{[1:p]} > X^\top \theta_{[p+1:2p]}\right)\right| (1 + \|\theta - \theta^*\|^2)\phi(X)\right] \\
&\leq C\|\theta^*\|\mathbb{E}\left|\|\theta^* - \theta\|(1 + \|\theta - \theta^*\|^2)\phi(X)\right| \\
&\leq C\|\theta^*\|\mathbb{E}\left|\|\theta^* - \theta\|(1 + \|\theta - \theta^*\|^2)\|\right| \\
&\leq Ct^{-\alpha/2}.
\end{aligned}$$

Combining the results above, we obtain that

$$\mathbb{E}\|\xi_{\theta_{t-1}}(\theta_{t-1}; \zeta_t) - \xi_{\theta^*}(\theta^*; \zeta_t)\|^2 \leq Ct^{-\alpha/2}. \tag{6.29}$$

With all these intermediate results in hand, we can proceed to the conclusion as follows. First of all, by inequality (6.29), we have the following bound for R_1,

$$\begin{aligned}
\mathbb{E}\|R_1\|^2 &\leq Ct^{-1}\mathbb{E}\|\sum_{i=1}^{n} \xi_{\theta_{i-1}}(\theta_{i-1}; \zeta_i) - \xi_{\theta^*}(\theta^*; \zeta_i)\|^2 \\
&\leq Ct^{-1}\sum_{i=1}^{n} \mathbb{E}\|\xi_{\theta_{i-1}}(\theta_{i-1}; \zeta_i) - \xi_{\theta^*}(\theta^*; \zeta_i)\|^2 \\
&\leq Ct^{-1}\sum_{i=1}^{n} i^{-\alpha/2} \leq Ct^{-\alpha/2}.
\end{aligned}$$

Also, it is easy to derive that

$$\mathbb{E}\|R_2\|^2 \leq Ct^{-1}.$$

Using the above intermediate result (6.27), the R_3 term has the convergence rate below,

$$\begin{aligned}\mathbb{E}\|R_3\| &\le t^{-1/2}\sum_{i=0}^{t-1}\mathbb{E}\|\nabla\mathcal{L}_{\theta_i}(\theta_i) - H(\theta_i - \theta^*)\| \\ &\le Ct^{-1/2}\sum_{i=0}^{t-1}\mathbb{E}\|\theta_i - \theta^*\|^2 \\ &\le Ct^{-1/2}\sum_{i=0}^{t-1} i^{-\alpha} \le Ct^{-\alpha+\frac{1}{2}}.\end{aligned}$$

Finally, we can bound R_4 using our result in (6.28),

$$\mathbb{E}\|R_4\|^2 \le \frac{C}{t}\left\|\Sigma^{-1/2} - \Sigma_t^{-1/2}\right\|^2\sum_{i=1}^{t-1}\mathbb{E}\|\xi_{\theta^*}(\theta^*;\zeta_i)\|^2 \le Ct^{2\alpha-2}.$$

□

Given the Bahadur representation of $\bar{\theta}_t$, we now emphasize the difference in the convergence rate of the adaptive SGD and the classical SGD results [162, 176]. The remainder term under adaptive settings has a rate of $\mathcal{O}_p\left(t^{-\alpha+\frac{1}{2}} + t^{-\frac{\alpha}{4}} + t^{\alpha-1}\right)$, slower than that under classical settings (6.24). If we minimizer the order of the rate over $\alpha \in (\frac{1}{2}, 1)$, we have that the optimal convergence rate of the remainder term is $\mathcal{O}(t^{-0.2})$ with $\alpha = 0.8$.

Note that to derive the above decomposition, we require a slightly stronger condition (condition (a) in the theorem statement). The second condition in Theorem 6.4.11 requires a certain level of continuity of the distribution of covariate X. These extra conditions can be easily satisfied, e.g., when X obeys a non-degenerate normal distribution. For a non-degenerated normal variable X, Assumption (a) of Theorem 6.4.11 clear holds. Assumption (b) can also be transformed into a (stronger) differentiability condition. Denote the left hand side of Assumption (b) as $F(\theta)$,

$$F(\theta) = \mathbb{E}\left[\left|\mathbb{I}\left(X^\top\theta^*_{[1:p]} > X^\top\theta^*_{[p+1:2p]}\right) - \mathbb{I}\left(X^\top\theta_{[1:p]} > X^\top\theta_{[p+1:2p]}\right)\right|(1 + \|X\|^4)\right].$$

It is not differentiable at θ^*, but directional derivatives exist. In fact, we have for some constants $\underline{C}, \overline{C}$

$$0 \leq \underline{C} \leq \frac{\partial F(\theta)}{\partial v}\big|_{\theta=\theta^*} \leq \overline{C},$$

for any v orthogonal to θ^*, and the directional derivatives with respect to θ^* is 0. Together, they imply

$$\underline{C}|(\theta-\theta^*)^\top \theta^*| \leq F(\theta) \leq \overline{C}|(\theta-\theta^*)^\top \theta^*|,$$

which implies Assumption (b). So Theorem 6.4.11 holds.

Now we furfther provide a lower bound on the remainding terms for a non-degenerated normal variable X. By same argument, we can prove that

$$0 \leq \underline{C} \leq \frac{\partial \mathbb{E}\left[\Delta(X,\theta)\right]}{\partial v}\big|_{\theta=\theta^*} \leq \overline{C},$$

where $\Delta(X,\theta) = (1-\epsilon)\left|\mathbb{I}\left(X^\top \theta^*_{[1:p]} > X^\top \theta^*_{[p+1:2p]}\right) - \mathbb{I}\left(X^\top \theta_{[1:p]} > X^\top \theta_{[p+1:2p]}\right)\right|$. So it can be bounded by

$$\underline{C}|(\theta-\theta^*)^\top \theta^*| \leq \mathbb{E}\left[\Delta(X,\theta)\right] \leq \overline{C}|(\theta-\theta^*)^\top \theta^*|.$$

For R_1, we first decompose the term $\xi_{\theta_{t-1}}(\theta_{t-1};\zeta_t) - \xi_{\theta^*}(\theta^*;\zeta_t)$ as follows

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P},\nu}[\|\xi_{\theta_{t-1}}(\theta_{t-1};\zeta_t) - \xi_{\theta^*}(\theta^*;\zeta_t)\|^2] \\
&= \mathbb{E}_{\mathbb{P},\nu}\left[\|w(\theta_{t-1};X_t,A_t)\nabla\ell(\theta_{t-1};\zeta_t) - w(\theta^*;X_t,A^*)\nabla\ell(\theta^*;X_t,A^*,Y_t)\|^2\right] \\
&\quad - \|\nabla\mathcal{L}_{\theta_{t-1}}(\theta_{t-1}) - \nabla\mathcal{L}_{\theta^*}(\theta^*)\|^2 \\
&= M_2 - M_1,
\end{aligned}$$

where M_1, M_2, M_3, M_4 has been defined in the proof of Theorem 6.4.7. From previous estimates, $M_1 \leq Ct^{-\alpha}$. Previous decomposition can also provide lower bounds,

$$\begin{aligned}
&M_2 \geq C\mathbb{E}[\Delta(X_t, \theta_{t-1})M_3],\\
&M_3 \geq C\mathbb{E}_{\mathbb{P}_{Y|X}}\left[\|\nabla\ell(\theta_{t-1}; X_t, A_t, Y_t)\|^2 + \|\nabla\ell(\theta^*; X_t, A^*, Y_t)\|^2 \mid X_t, A_t \neq A^*\right] \geq C.
\end{aligned}$$

Combining all inequalities together, we have

$$\mathbb{E}\|\xi_{\theta_{t-1}}(\theta_{t-1}; \zeta_t) - \xi_{\theta^*}(\theta^*; \zeta_t)\|^2 \geq C\mathbb{E}[\Delta(X_t, \theta_{t-1})].$$

The proof of Theorem 6.4.11 implies that $\Sigma_t^{-1/2}$ and Q_i^t are bounded from below for sufficiently large $i, t > i_0$. So

$$\begin{aligned}
\mathbb{E}\|R_1\|^2 &= \frac{1}{t}\sum_{i=1}^{t-1} \|\Sigma_t^{-1/2} Q_i^t(\xi_{\theta_{t-1}}(\theta_{t-1}; \zeta_t) - \xi_{\theta^*}(\theta^*; \zeta_t))\|^2\\
&\geq \frac{C}{t}\sum_{i>i_0}^{t-1} \mathbb{E}\|\xi_{\theta_{t-1}}(\theta_{t-1}; \zeta_t) - \xi_{\theta^*}(\theta^*; \zeta_t)\|^2\\
&\geq \frac{C}{t}\sum_{i>i_0}^{t-1} \mathbb{E}[\Delta(X_i, \theta_{i-1})]\\
&\geq \frac{C}{t}\sum_{i>i_0}^{t-1} |(\theta_{i-1} - \theta^*)^\top \theta^*|.
\end{aligned}$$

Theorem 6.4.11 implies that

$$\frac{1}{t}\sum_{i=1}^{t} \mathbb{E}|(\theta_i - \theta^*)^\top \theta^*| \geq \mathbb{E}|(\bar{\theta}_t - \theta^*)^\top \theta^*| \geq Ct^{-1/2}.$$

Therefore, we can come to the conclusion that

$$\mathbb{E}\|R_1\|^2 \geq Ct^{-1/2}.$$

Even though the lower bound does not match the upper bound established in Theorem 6.4.11, it indicates a strictly slower convergence than the classical SGD setting. This slower rate results from the reliance between the convergence of the estimator θ_t and that of the policy function $\pi(X_t, \theta_{t-1})$. Such a phenomenon is not limited to the ε-greedy policy, but also expected for a wide range of different policies.

6.5 Numerical experiments

In this section, we investigate the empirical performance of the proposed estimators on normal approximation. We further construct the confidence intervals using a plug-in estimator of the asymptotic covariance matrices and report their coverage rates. Lastly, we validate the performance of the proposed estimator and inference procedure on a logistic regression of a real dataset.

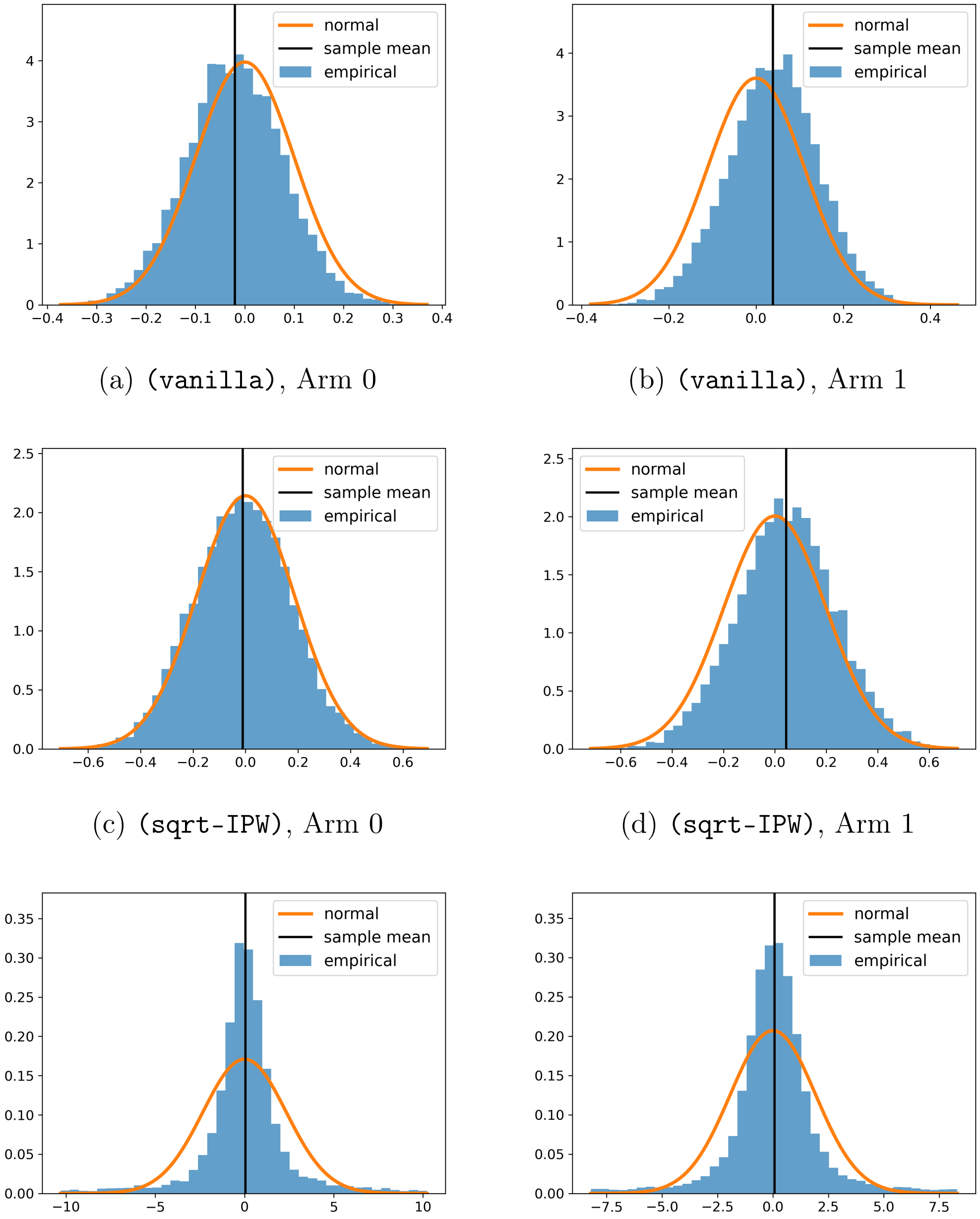
6.5.1 Normal approximation

We verify Theorem 6.4.7 under linear regression and quantile regression (Example 6.2.1 and Example 6.2.3). For both examples, the true parameter $\theta^* \in \mathbb{R}^{20}$ and

$$Y_t = (1 - A_t) X_t^\top \theta^*_{[1:10]} + A_t X_t^\top \theta^*_{[11:20]} + \mathcal{E}_t.$$

In the numerical experiments below, we fix the sample size as $80,000$. The covariate $X_t \sim \mathcal{N}(0, I_{10})$ and the noise $\{\mathcal{E}_s\}_{s=1}^t$ is *i.i.d.* with standard deviation $\sigma = 0.1$. We use ε-greedy policy (6.8) to select actions, and set $\varepsilon = 0.02$.

For the SGD update (6.7), we specify the step sizes as $\eta_t = \eta \cdot \max(t, 300)^{-\alpha}$. As indicated in Theorem 6.4.11, we set the parameter α in the step size as $\alpha = 0.8$ for both linear regression and quantile regression. We compare three weighting schemes below, `(IPW)`, `(sqrt-IPW)`, `(vanilla)`.



(a) `(vanilla)`, Arm 0

(b) `(vanilla)`, Arm 1

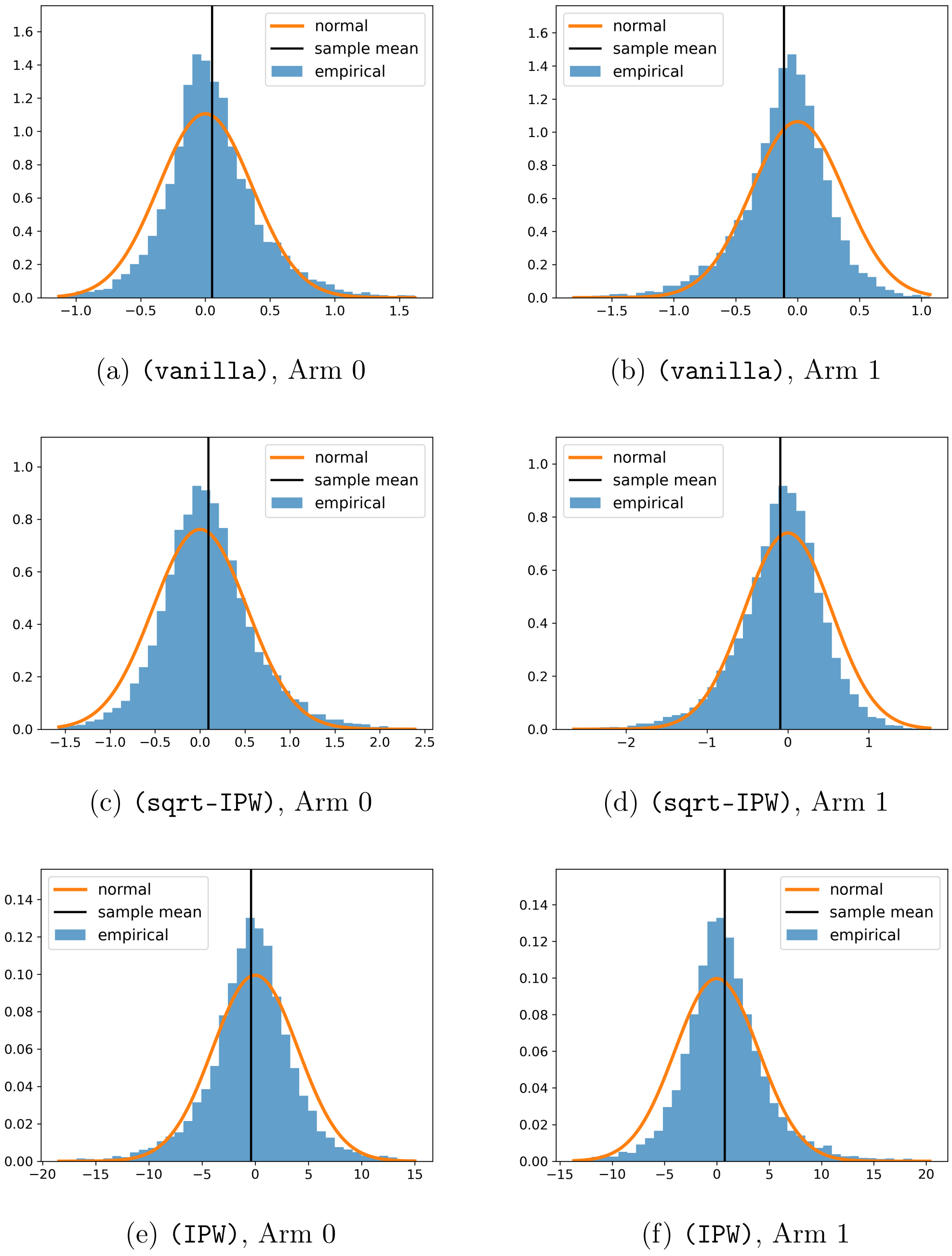
(c) `(sqrt-IPW)`, Arm 0

(d) `(sqrt-IPW)`, Arm 1

(e) `(IPW)`, Arm 0

(f) `(IPW)`, Arm 1

Figure 6.1: SGD on linear regression with different weights.



(a) (vanilla), Arm 0

(b) (vanilla), Arm 1

(c) (sqrt-IPW), Arm 0

(d) (sqrt-IPW), Arm 1

(e) (IPW), Arm 0

(f) (IPW), Arm 1

Figure 6.2: SGD on quantile regression with different weights.

We first present the results for linear regression. In Figure 6.1, we plot the empirical distribution of $\sqrt{t}(\bar{\theta}_t - \theta^*)$ using $10,000$ Monte-Carlo simulations. We also plot the density function of a zero-mean normal distribution that matches the second-order moments. As can be inferred from the plots, the vanilla SGD and the square-root importance weight SGD have much smaller standard deviation compared with `(IPW)`, this finding matches our discussion in Section 6.3. We also conduct simulations on quantile regression with quantile level $\tau = 0.75$. The empirical distribution is reported in Figure 6.2.

6.5.2 Online statistical inference

In this section, we demonstrate the online plug-in inference procedure based on the limiting distribution of our proposed estimator $\bar{\theta}_t$ in Theorem 6.4.7. As we mentioned in the previous section, the plug-in estimator constructs a pair $(\hat{S}_n, \hat{H}_n)$ to estimate (S, H) in the asymptotic covariance matrix $H^{-1}SH^{-1}$.

$$\hat{S}_n = \frac{1}{n}\sum_{t=1}^{n} w_t^2 \nabla \ell(\theta_{t-1}; \zeta_t) \nabla \ell(\theta_{t-1}; \zeta_t)^\top, \quad \hat{H}_n = \frac{1}{n}\sum_{t=1}^{n} w_t \nabla^2 \ell(\theta_{t-1}; \zeta_t).$$

Using linear regression as an example, we first establish the consistency of the plug-in estimator under the following additional assumption.

Assumption 6.5.1. *For any action $A \in \mathcal{A}$ and covariate X, we assume that $\nabla^2 \ell(\theta;\zeta)$ exists and $\mathbb{E}_{\mathbb{P}_{Y|X}}\left(\|\nabla^2 \ell(\theta;\zeta)\|^2 \mid X, A\right)$ is bounded by $\psi(X)(1 + \|\theta - \theta^*\|^2)$ for some function $\psi(\cdot)$ such that $\mathbb{E}[\psi(X)] < \infty$. In addition, we have $\lim_{\theta \to \theta^*} \mathbb{E}_{\mathbb{P}_X}[\Delta(X,\theta)\psi(X)] = 0$ where $\Delta(X,\theta)$ is defined in Assumption 6.4.5, and*

$$\lim_{\theta \to \theta^*} \mathbb{E}_{\mathbb{P}_{Y|X}}\left[\|\nabla^2 \ell(\theta;\zeta) - \nabla^2 \ell(\theta^*;\zeta)\|^2 \mid X, A\right] = 0,$$
$$\lim_{\theta \to \theta^*} \mathbb{E}_{\mathbb{P}_X}\left[|w(\theta; X, A) - w(\theta^*; X, A)|^2 \psi(X) \mid A\right] = 0.$$

Proposition 6.5.2. *Under Assumption 6.4.2 to Assumption 6.4.6, and Assumption 6.5.1, the plug-in estimators are consistent, i.e.,* $\hat{S}_n \to S$ *and* $\hat{H}_n \to H$ *in probability.*

Proof. Recall the definition of our estimators

$$\hat{S}_n = \frac{1}{n}\sum_{t=1}^{n} w_t^2 \nabla \ell(\theta_{t-1};\zeta_t)\nabla \ell(\theta_{t-1};\zeta_t)^\top, \quad \hat{H}_n = \frac{1}{n}\sum_{t=1}^{n} w_t \nabla^2 \ell(\theta_{t-1};\zeta_t).$$

In the proof of Theorem 6.4.7, the bound on M_2 (Equation (6.19)) implies the following convergence in L^2,

$$w(\theta_{t-1};X_t,A)\nabla \ell(\theta_{t-1};\zeta_t) \to w(\theta^*;X_t,A)\nabla \ell(\theta^*;\zeta_t).$$

Therefore we have the following convergence of $\hat{S}_n$ in L^1,

$$\hat{S}_n - \frac{1}{n}\sum_{t=1}^{n} w(\theta^*;X_t,A)^2 \nabla \ell(\theta^*;\zeta_t)\nabla \ell(\theta^*;\zeta_t)^\top \to 0$$

Notice that by Law of Large Numbers,

$$\frac{1}{n}\sum_{t=1}^{n} w(\theta^*;X_t,A)^2 \nabla \ell(\theta^*;\zeta_t)\nabla \ell(\theta^*;\zeta_t)^\top \to S,$$

in probability. Thus, combining our findings above, we can easily see that our plug-in estimator for gram matrix $\hat{S}_n \to S$ in probability.

Now we come to the consistency proof of $\hat{H}_n$. Notice that our assumption 6.5.1 is simply a repetition of Assumption 6.4.5 and Assumption 6.4.6 in Theorem 6.4.7 with ϕ replaced by ψ and with gradient replaced by Hessian. So our proof of Theorem 6.4.7 from bound (6.16) to bound (6.19) can be adapted here to prove the following convergence in L^2,

$$w(\theta_{t-1};X_t,A)\nabla^2 \ell(\theta_{t-1};\zeta_t) \to w(\theta^*;X_t,A)\nabla^2 \ell(\theta^*;\zeta_t).$$

Similarly, we have $\hat{H}_n \to H$ in probability. □

Under the same setting as in Section 6.5.1, we show the inference results for linear regression in Table 6.1. Averaged coverage rate and average length of the confidence intervals are reported for plug-in estimator and oracle estimator. We also include standard error in the parentheses. The comparison of the three candidate weighted-SGD schemes is clearly stated. Both `(vanilla)` and `(sqrt-IPW)` provide a valid conference interval, while `(IPW)` provides a much wider confidence interval than its oracle.

Weight & Arm	Sample size	Plug-in Cov.	Oracle Cov.	Plug-in Len.	Oracle Len.
`(vanilla)`, Arm 0	2×10^4	0.78 (0.14)	0.73 (0.15)	0.63 (0.03)	0.55
	8×10^4	0.88 (0.09)	0.86 (0.09)	0.57 (0.01)	0.55
`(vanilla)`, Arm 1	2×10^4	0.89 (0.09)	0.83 (0.12)	0.63 (0.03)	0.55
	8×10^4	0.94 (0.07)	0.93 (0.08)	0.58 (0.01)	0.55
`(sqrt-IPW)`, Arm 0	2×10^4	0.78 (0.14)	0.72 (0.15)	0.82 (0.12)	0.72
	8×10^4	0.88 (0.10)	0.87 (0.11)	0.74 (0.04)	0.72
`(sqrt-IPW)`, Arm 1	2×10^4	0.84 (0.12)	0.78 (0.14)	0.83 (0.13)	0.72
	8×10^4	0.91 (0.09)	0.90 (0.10)	0.75 (0.05)	0.72
`(IPW)`, Arm 0	2×10^4	0.81 (0.15)	0.47 (0.32)	19.18 (34.94)	2.79
	8×10^4	0.85 (0.14)	0.62 (0.33)	13.04 (28.04)	2.79
`(IPW)`, Arm 1	2×10^4	0.82 (0.15)	0.51 (0.32)	16.76 (32.12)	2.79
	8×10^4	0.86 (0.13)	0.65 (0.32)	11.47 (25.80)	2.79

Table 6.1: Inference results of linear regression with different weighting schemes.

6.5.3 *Real data analysis*

In this section, we apply our online estimation and inference framework to Yahoo! Today module user click-log dataset and conduct statistical inference for model parameters. We use the news recommendation and user response records on May 1^{st}, 2009. On this day, we consider the two most recommended (recommended $405,888$ times) articles, No.109510 and No.109520 for analysis.

We follow the experiment settings in [41]. The action A_t is specified to be 1 when Article No.109510 is recommended and $A_t = 0$ when Article No.109520 is recommended. The original user features have six covariates, where the first five sum up to one, and the sixth is a constant 1. In our experiments below, we keep the second to fifth covariates in the original features as $X_{[2:5]}$ and specify $X_{[1]} = 1$ as the intercept.

As the reward Y_t is binary, we consider a logistic regression model (Example 6.2.2) and set $Y_t = 1$ if the user clicks on the article link and $Y_t = -1$ if not. We use the ε-greedy algorithm (6.8). In order to match our online decision-making process with our offline dataset, we keep the entry if the recorded offline action matches the action given by our online ε-greedy algorithm with two specifications of $\varepsilon \in \{0.2, 0.02\}$.

We now present the online statistical inference results. For our SGD update, we use the same settings as above experiments, i.e., 300-step meltdown and $\alpha = 0.8$. We compare three weighting schemes below, vanilla SGD (6.2), square-root importance weight SGD (6.3), and IPW SGD (6.3). Table 6.2 below gives the result for $\varepsilon = 0.2$ and Table 6.3 gives the result for $\varepsilon = 0.02$. In both table, the vanilla SGD and the square-root importance SGD have smaller standard errors and smaller p-values. There are also more insignificant parameters for IPW SGD. The results of IPW SGD are worse when we decrease the value of ε, matches our findings in Theorem 6.4.7 and discussions in Section 6.3.

Weight & Arm	Parameter	Estimate	S.E.	95% LB	95% UB	t-value	p-value
(vanilla), Arm 0	θ_1	-2.56	0.04	-2.64	-2.48	-65.52	0.00
	θ_2	-0.26	0.08	-0.43	-0.10	-3.11	0.00
	θ_3	-0.48	0.07	-0.62	-0.34	-6.80	0.00
	θ_4	-0.23	0.06	-0.34	-0.12	-4.09	0.00
	θ_5	-0.90	0.07	-1.03	-0.77	-13.65	0.00
(vanilla), Arm 1	θ_6	-2.55	0.05	-2.65	-2.44	-47.77	0.00
	θ_7	-0.24	0.08	-0.40	-0.09	-3.06	0.00
	θ_8	-0.45	0.07	-0.58	-0.32	-6.76	0.00
	θ_9	-0.41	0.11	-0.62	-0.19	-3.71	0.00
	θ_{10}	-0.91	0.07	-1.05	-0.77	-12.31	0.00
(sqrt-IPW), Arm 0	θ_1	-2.52	0.05	-2.62	-2.43	-52.85	0.00
	θ_2	-0.30	0.11	-0.51	-0.09	-2.79	0.01
	θ_3	-0.49	0.09	-0.66	-0.31	-5.56	0.00
	θ_4	-0.28	0.07	-0.4	-0.15	-4.25	0.00
	θ_5	-0.80	0.09	-0.97	-0.63	-9.33	0.00
(sqrt-IPW), Arm 1	θ_6	-2.51	0.05	-2.61	-2.41	-49.35	0.00
	θ_7	-0.28	0.08	-0.43	-0.13	-3.60	0.00
	θ_8	-0.45	0.06	-0.58	-0.33	-7.10	0.00
	θ_9	-0.42	0.11	-0.63	-0.20	-3.83	0.00
	θ_{10}	-0.81	0.07	-0.94	-0.68	-12.02	0.00
(IPW), Arm 0	θ_1	-2.64	0.10	-2.85	-2.44	-25.54	0.00
	θ_2	-0.28	0.19	-0.64	0.08	-1.51	0.13
	θ_3	-0.51	0.15	-0.80	-0.23	-3.49	0.00
	θ_4	-0.24	0.16	-0.55	0.07	-1.54	0.12
	θ_5	-0.91	0.16	-1.23	-0.59	-5.64	0.00
(IPW), Arm 1	θ_6	-2.47	0.03	-2.53	-2.40	-76.6	0.00
	θ_7	-0.22	0.06	-0.33	-0.11	-3.83	0.00
	θ_8	-0.51	0.05	-0.60	-0.42	-11.08	0.00
	θ_9	-0.37	0.05	-0.47	-0.27	-7.40	0.00
	θ_{10}	-0.88	0.05	-0.98	-0.78	-17.67	0.00

Table 6.2: Real data analysis with online statistic inference with $\varepsilon = 0.2$.

Weight & Arm	Parameter	Estimate	S.E.	95% LB	95% UB	t-value	p-value
`(vanilla)`, Arm 0	θ_1	-2.55	0.04	-2.63	-2.48	-68.62	0.00
	θ_2	-0.31	0.09	-0.47	-0.14	-3.61	0.00
	θ_3	-0.45	0.07	-0.6	-0.31	-6.18	0.00
	θ_4	-0.23	0.05	-0.33	-0.12	-4.29	0.00
	θ_5	-0.88	0.07	-1.01	-0.75	-13.45	0.00
`(vanilla)`, Arm 1	θ_6	-2.54	0.06	-2.66	-2.42	-41.76	0.00
	θ_7	-0.29	0.09	-0.45	-0.12	-3.36	0.00
	θ_8	-0.42	0.07	-0.57	-0.28	-5.88	0.00
	θ_9	-0.42	0.19	-0.79	-0.04	-2.18	0.03
	θ_{10}	-0.89	0.08	-1.04	-0.73	-11.25	0.00
`(sqrt-IPW)`, Arm 0	θ_1	-2.49	0.05	-2.58	-2.40	-54.74	0.00
	θ_2	-0.31	0.13	-0.57	-0.05	-2.37	0.02
	θ_3	-0.45	0.12	-0.68	-0.21	-3.74	0.00
	θ_4	-0.29	0.06	-0.41	-0.17	-4.78	0.00
	θ_5	-0.82	0.08	-0.98	-0.66	-9.80	0.00
`(sqrt-IPW)`, Arm 1	θ_6	-2.48	0.08	-2.64	-2.33	-31.13	0.00
	θ_7	-0.29	0.10	-0.50	-0.09	-2.84	0.00
	θ_8	-0.42	0.09	-0.60	-0.25	-4.69	0.00
	θ_9	-0.4	0.25	-0.90	0.09	-1.60	0.11
	θ_{10}	-0.82	0.10	-1.01	-0.63	-8.49	0.00
`(IPW)`, Arm 0	θ_1	-2.75	0.33	-3.40	-2.11	-8.37	0.00
	θ_2	-0.22	0.57	-1.35	0.90	-0.39	0.70
	θ_3	-0.80	0.50	-1.78	0.18	-1.59	0.11
	θ_4	0.11	0.39	-0.65	0.87	0.28	0.78
	θ_5	-0.90	0.51	-1.89	0.09	-1.78	0.08
`(IPW)`, Arm 1	θ_6	-2.40	0.09	-2.57	-2.23	-27.81	0.00
	θ_7	-0.33	0.14	-0.60	-0.07	-2.46	0.01
	θ_8	-0.33	0.08	-0.48	-0.17	-4.17	0.00
	θ_9	-0.55	0.30	-1.14	0.05	-1.81	0.07
	θ_{10}	-1.14	0.20	-1.53	-0.76	-5.81	0.00

Table 6.3: Real data analysis with online statistic inference with $\varepsilon = 0.02$.

CHAPTER 7

HAAGERUP BOUND FOR QUATERNIONIC GROTHENDIECK INEQUALITY

7.1 Introduction

We will let $\mathbb{F} = \mathbb{R}$ and $\mathbb{C}$ and $\mathbb{H}$ be the fields of real, complex and the skew field of quaternions respectively in this article. In 1953, Grothendieck proved a powerful result that he called "the fundamental theorem in the metric theory of tensor products" [87]. His result can be stated as follows [140]: For $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ there exists a finite constant $K > 0$ such that for every $l, m, n \in \mathbb{N}$ and every matrix $M = (M_{ij}) \in \mathbb{F}^{m \times n}$,

$$\max_{\|x_i\|=\|y_j\|=1} \left| \sum_{i=1}^{m} \sum_{j=1}^{n} M_{ij} \langle x_i, y_j \rangle \right| \leq K \max_{|\varepsilon_i|=|\delta_j|=1} \left| \sum_{i=1}^{m} \sum_{j=1}^{n} M_{ij} \bar{\varepsilon}_i \delta_j \right| \tag{7.1}$$

where the maximum on the left is take over all $x_i, y_j \in \mathbb{F}^l$ of unit 2-norm, and the maximum on the right is taken over all $\varepsilon_i, \delta_j \in \mathbb{F}$ of unit absolute value (i.e., $\varepsilon_i = \pm 1$, $\delta_j = \pm 1$ over $\mathbb{R}$; $\varepsilon_i = e^{i\theta_i}$, $\delta_j = e^{i\phi_j}$ over $\mathbb{C}$). The inequality (7.1) has since been christened *Grothendieck inequality* and the smallest possible constant K *Grothendieck's constant.* The value of Grothendieck's constant depends on the choice of $\mathbb{F}$ and we will denote it by $K_G^{\mathbb{F}}$. In a recent paper [77] two authors of this paper extended the Grothendieck inequality to symmetric/Hermitian matrices, which we call symmetric Grothendieck inequality and referred as SGI. Namely, in the above inequality we can assume that M is symmetric/Hermitian and $x_i = y_i$. Furthermore, they considered more refined versions of SGI where the vectors x_i are in d-dimensional Hilbert space as in [26].

The aim of this paper to extend the Grothendieck inequality and SGI to quaternions $\mathbb{H}$. Since quaternions is a skew field, which is noncommutative, there are many different obstacles and problems to be solved before having the Haagerup type constant [89]. We now describe briefly the results we obtained. Let $\mathbb{F} = \mathbb{H}$. We first show that the inequality (7.1) holds, where x_i, y_j in the quaternion Hilbert space $\mathbb{H}^l$, where $l \geq m + n$, and $\varepsilon_i, \delta_j \in \mathbb{H}$ with the constant $K_G^{\mathbb{H}}$. Using the analogous results to Krivine [129] and Haagerup [89] we show that $K_G^{\mathbb{H}} \leq 1.2168$. This result is achieved by establishing the most difficult technical part of our paper. Let ${}_2F_1(a, b; c; x)$ be the classical hypergeometric function. Denote $p_\ell(x) = x {}_2F_1(\frac{1}{2}, \frac{1}{2}; \ell; x^2)$ for $\ell \in \mathbb{N}$. It was shown by Haagerup that the inverse function $p_\ell^{-1}(x)$ has first positive Taylor coefficient, while all other Taylor coefficients are nonpositive for $\ell = 2$. In this paper we prove Haagerup result for $\ell = 3$. By numerical computing we validate that the same result holds for at least $\ell = 4, 5, 6$, for the first one hundred Taylor coefficients.

Denote by $\mathbb{S}^n(\mathbb{F}) \subset \mathbb{F}^{n \times n}$ the real space of self-adjoint matrices. i.e., $A^* = A$. We show that we have two analogs of the Grothendieck inequality (7.1) on $\mathbb{S}^n(\mathbb{H})$:

$$\max_{\|x_i\|=1} \left| \Re \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \langle x_i, x_j \rangle \right| \leq K_\gamma^{\mathbb{H}} \max_{|\delta_i|=1} \left| \Re \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \bar{\delta}_i \delta_j \right|, \tag{7.2}$$

$$\max_{\|x_i\|\leq 1} \left| \Re \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \langle x_i, x_j \rangle \right| \leq K_\Gamma^{\mathbb{H}} \max_{|\delta_i|\leq 1} \left| \Re \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \bar{\delta}_i \delta_j \right|.$$

Furthermore $K_G^{\mathbb{H}} \leq K_\Gamma^{\mathbb{H}} \leq K_\gamma^{\mathbb{H}} \leq \frac{64}{9\pi} - 1 \approx 1.263537$.

We now describe briefly the "conic Grothendieck inequality" for various cones in $\mathbb{S}^n(\mathbb{H})$. Denote by $\mathbb{S}_+^n(\mathbb{H})$ the cone of positive semidefinite self-adjoint quaternionic matrices. We show that in this case (7.1) is equivalent to the inequality of the form (7.2) with the constant $32/9\pi$, which is sharp. This is a quaternionic version of Nesterov-Rietz $\pi/2$ theorem [152, 166] for the real numbers, and Nemirovski-Roos-Terlaky $4/\pi$ theorem for the complex numbers [151]. We next consider the subcone of $\mathbb{S}_+^n(\mathbb{R})$ of Laplacian matrices. In this case the constant in (7.1) can be reduced to $K \leq 1.0338$. This is the quaternionic version of the well-known Goemans-Williamson inequality [85].

7.2 Quaternions

7.2.1 Basic facts on quaternions

Recall that $\mathbb{H}$ can be viewed as $\mathbb{R}^4$. So $a \in \mathbb{H}$ is of the form $a = a_0 + a_1\mathrm{i} + a_2\mathrm{j} + a_3\mathrm{k}$. We can identify a with $a = (a_0, a_1, a_2, a_3)^\top$. We define the real part of a to be $\Re a := a_0$, and the conjugate of a to be $\bar{a} = a^* = a_0 - a_1\mathrm{i} - a_2\mathrm{j} - a_3\mathrm{k}$. The product table of $\mathrm{i}, \mathrm{j}, \mathrm{k}$ is given by

$$\mathrm{i}^2 = \mathrm{j}^2 = \mathrm{k}^2 = -1,\ \mathrm{ij} = -\mathrm{ji} = \mathrm{k},\ \mathrm{jk} = -\mathrm{kj} = \mathrm{i},\ \mathrm{ki} = -\mathrm{ik} = \mathrm{j}.$$

Hence $\mathbb{H}$ is a noncommutative ring over $\mathbb{R}$. Observe next that $a\bar{a} = \bar{a}a = a_0^2 + a_1^2 + a_2^2 + a_3^2$. Hence $|a| = \sqrt{a\bar{a}} \ge 0$ and equality holds if and only if $a = 0$. Thus for $a \ne 0$ the element $|a|^{-1}\bar{a} = \bar{a}|a|^{-1}$ is the unique inverse of a in $\mathbb{H}$. So $\mathbb{H}$ is a skew field over the field $\mathbb{R}$, where 1 is the identity element. Frobenius theorem claims that the only skew fields over $\mathbb{R}$ are $\mathbb{R}, \mathbb{C}, \mathbb{H}$. For $A \in \mathbb{F}^{m\times n}$ we denote $A^* := \bar{A}^\top \in \mathbb{F}^{n\times m}$.

There exists a standard way to present quaternions similar to the complex numbers: $z + w\mathrm{j}$, where $z, w \in \mathbb{C}$. Indeed, if $z = x + y\mathrm{i}, w = u + v\mathrm{i}$, the identity $\mathrm{ij} = \mathrm{k}$ yields $z + w\mathrm{j} = x + y\mathrm{i} + u\mathrm{j} + v\mathrm{k}$. Thus to multiply quaternions we have to remember that the product of complex numbers is commutative and

$$\overline{z + w\mathrm{j}} = \bar{z} - w\mathrm{j}, \quad w\mathrm{j} = \mathrm{j}\bar{w}. \tag{7.3}$$

By representing a complex number $z = x + y\mathrm{i}$ as 2×2 real valued matrix $\begin{pmatrix} x & y \\ -y & x \end{pmatrix}$

we can represent $C(a)$ as

$$C(a,\mathbb{R}) = \begin{pmatrix} x & y & u & v \\ -y & x & -v & u \\ -u & v & x & -y \\ -v & -u & y & x \end{pmatrix}, \quad a = (z,w), z = x + y\mathrm{i}, w = u + v\mathrm{i}.$$

Then $a \to C(a,\mathbb{R})$ is an isomorphism of $\mathbb{H}$ and the induced 4-dimensional subalgebra $\mathcal{C}(\mathbb{H},\mathbb{R}) = \{C(a,\mathbb{R}),\ a \in \mathbb{H}\} \subset \mathbb{R}^{4\times 4}$.

There exists yet another representation of $\mathbb{H}$ as a real subalgebra of 2×2 complex valued matrices $\mathbb{C}^{2\times 2}$. First observe that one can view a as $a = (z,w) \in \mathbb{C}^2$. Note that $\bar{a} = (\bar{z}, -w)$. (Warning: if one views (z,w) as a vector with complex entries then $\overline{(z,w)} = (\bar{z},\bar{w})$.) Let

$$C(a) = \begin{pmatrix} z & w \\ -\bar{w} & \bar{z} \end{pmatrix} \in \mathbb{C}^{2\times 2}, \quad a = (z,w). \tag{7.4}$$

Then the map $a \to C(a)$ is an isomorphism of $\mathbb{H}$ and the induced complex 2-dimensional subalgebra $\mathcal{C}(\mathbb{H}) = \{C(a),\ a \in \mathbb{H}\} \subset \mathbb{C}^{2\times 2}$. Note that $A(\mathbb{H}) \cap \mathbb{R}^{2\times 2}$ is a subalgebra isomorphic to $\mathbb{C}$. Observe that

$$|a|^2 = \det C(a), \quad C(\bar{a}) = C(a)^*, \Re(a) = \frac{1}{2}\operatorname{tr}(C(a)). \tag{7.5}$$

As $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ we deduce that

$$\Re(ab) = \Re(ba) = \Re(\overline{ab}) = \Re(\bar{b}\bar{a}) = \Re(\bar{a}\bar{b}),\ a, b \in \mathbb{H}. \tag{7.6}$$

7.2.2 Vector spaces

We next consider a right vector space $\mathbb{V}$ over $\mathbb{H}$. It is a commutative group with 0 element denoted as 0. We will denote in this section by the lower case bold letter vectors in $\mathbb{V}$. For the right vector space $\mathbb{V}$ the scalar vector product va satisfies the standard assumptions:

$$(v+w)a = va + wa,\ v(a+b) = va + wb,\ v(ab) = (va)b,\ v1 = v.$$

We can define similarly the left vector space over $\mathbb{H}$. In this paper, we only work with right vector space. Linear dependence, linear independence, subspace, span of a set of vectors, finitely generated subspaces, basis are defined as for the vector spaces over a field. Every finitely generated vector space over $\mathbb{H}$ has a basis of the same cardinality, denoted by dim $\mathbb{V}$. Denote $[l] = \{1, \dots, l\} \subset \mathbb{N}$. We view

$$\mathbb{H}^l = \{x = (x_1, \dots, x_l)^\top,\ x_i \in \mathbb{H}, i \in [l]\}, \quad \mathbb{H}_l = \{x = (x_1, \dots, x_l),\ x_i \in \mathbb{H}, i \in [l]\},$$

as right and left vector spaces over $\mathbb{H}$ respectively. Clearly, $\dim \mathbb{H}^l = \dim \mathbb{H}_l = l$ and $e_i = (\delta_{1i}, \dots, \delta_{li}), i \in [l]$ is the standard basis in $\mathbb{H}^l$ and $\mathbb{H}_l$.

When a basis is specified, for example, the standard basis in the right vector space $\mathbb{H}^l$, the expression av is meaningful and we will use it when necessary. However, the reader should keep in mind this expression should be considered as an additional structure related to a particular basis. Its meaning will be different if we choose a different basis.

Denote

$$\overline{x} = (\bar{x}_1, \dots, \bar{x}_l)^\top,\ x^* = (\bar{x}_1, \dots, \bar{x}_l), \quad \text{for } x = (x_1, \dots, x_l)^\top \in \mathbb{H}^l. \tag{7.7}$$

Similar notations apply for $x \in \mathbb{H}_l$. Note that $\overline{x}$ is defined with respect to the standard basis in $\mathbb{H}_l$.

Let $M = (M_{ij}) \in \mathbb{H}^{m\times n}$. Define $C(M) = (C(M_{ij})) \in \mathbb{C}^{(2m)\times(2n)}$ to be the block matrix with 2×2 blocks $C(M_{ij})$. Again, this embedding commutes with conjugate transpose, addition and multplication of matrices. For $m = n$ the matrix $M \in \mathbb{H}^{n\times n}$ is called (quaternion) self-adjoint if $M^* = M$. We denote by $\mathbb{S}^n(\mathbb{F}) \subset \mathbb{F}^{n\times n}$ the real space of self-adjoint matrices: $M^* = M$. When no ambiguity arises we will drop the dependence on $\mathbb{F}$.

It is helpful to introduce a convenient relabeling of the block matrix $C(M)$ denoted as $\hat{C}(M) = P_m C(M) P_n^\top$, where $P_m \in \{0, 1\}^{(2m)\times(2m)}$ is the following permutation matrix: The matrix P_m permutes the rows $1, 2, \dots, m + 1, \dots, 2m$ to $1, 3, \dots, 2m - 1, 2, 4, \dots, 2m$ respectively. Then $\hat{C}(M)$ has the following block structure:

$$\hat{C}(M) = \begin{pmatrix} Z & W \\ -\overline{W} & \bar{Z} \end{pmatrix}, \quad Z, W. \in \mathbb{C}^{m\times n} \tag{7.8}$$

Clearly, this partition is another isomorphism $\iota : \mathbb{H}^{m\times n} \to \mathbb{C}^{(2m)\times(2n)}$ which is preserved under multiplication and conjugate transpose of matrices. Note $M \in \mathbb{S}^n(H)$ if in the above representation of $\hat{C}(M)$, where $Z \in \mathbb{S}^n(\mathbb{C})$ and W is skew symmetric $W^\top = -W$.

7.2.3 Inner product on quaternion vector space

Assume that $\mathbb{V}$ is a right vector space over $\mathbb{H}$. A mapping $\langle \cdot, \cdot \rangle : \mathbb{V} \times \mathbb{V} \to \mathbb{H}$ is called an inner product if the following conditions hold:

$$\begin{aligned} \langle y, x\rangle &= \overline{\langle x, y\rangle}, \\ \langle xa + yb, z\rangle &= \bar{a}\langle x, z\rangle + \bar{b}\langle y, z\rangle, \\ \langle z, xa + yb\rangle &= \langle z, x\rangle a + \langle z, y\rangle b, \\ \langle x, x\rangle &> 0 \text{ for } x \neq 0. \end{aligned}$$

The norm is defined as $\|x\| = \sqrt{\langle x, x\rangle}$. Let $\mathcal{H}$ be a right vector space of $\mathbb{H}$ with an inner product. We also call $\mathcal{H}$ the Hilbert space over quaternions. All analysis in this paper is essentially finite dimensional. Incooperating the completeness in our definition does not change our result.

Lemma 7.2.1. *1. $\|xa\| = |a|\|x\|$.*

2. The Cauchy-Schwarz inequality holds for quaternion vector space,

$$|\langle x, y\rangle| \leq \|x\|\|y\|.$$

3. $\|\cdot\|$ is subadditive, i.e., satisfies the triangle inequality. Hence $\|\cdot\|$ is indeed a norm on $\mathbb{V}$.

Proof. 1. $\|xa\|^2 = \langle xa, xa\rangle = a^*\langle x, x\rangle a = \|x\|^2|a|^2$.

2. Suppose that x is not a scalar multiple of y, and that neither x nor y is 0. Then $x - ya$ is not 0 for any a. So

$$\|x - ya\|^2 = \|x\|^2 + \|y\|^2|a|^2 - \langle x, y\rangle a - a^*\langle x, y\rangle > 0$$

Let $a = t\mu$ with real t and $|\mu| = 1$ so that $\langle x, y\rangle a = |\langle x, y\rangle|t$. Then

$$\|x\|^2 + \|y\|^2 t^2 - 2|\langle x, y\rangle|t > 0$$

holds for all t. So $|\langle x, y\rangle| \leq \|x\|\|y\|$.

3. $(\|x\| + \|y\|)^2 - \|x + y\|^2 = 2\|x\|\|y\| - \langle x, y\rangle - \langle y, x\rangle \geq 0$.

□

Two vectors x, y are called orthogonal if $\langle x, y\rangle = 0$. A set of vectors $x_1, \ldots, x_l \in \mathbb{V}$ is an orthonormal system if $\langle x_i, x_j\rangle = \delta_{ij}$ for $i, j \in [l]$.

Lemma 7.2.2. *(Gram-Schmidt process) Let $x_1, \ldots, x_n$ be vectors in a right inner product space $\mathbb{V}$ over $\mathbb{H}$. Assume that $x_1 \neq 0$. Then there exists $m \in [n]$ orthonormal vectors $y_1, \ldots, y_m$ in the span of $x_1, \ldots, x_n$ with the following property. For each $i \in [n]$ there exists $j(i) \in [i]$ such that $x_1, \ldots, x_i$ are in the span of $y_1, \ldots, y_{j(i)}$. The vectors $y_1, \ldots, y_m$ are obtained by the Gram-Schmidt process.*

Proof. Let $y_1 = x_1 \|x_1\|^{-1}$. Suppose we defined the orthonormal vectors $y_1, \ldots, y_j$ such that their span, denoted as $\mathbb{V}_j$, contains the vectors $x_1, \ldots, x_i$. So $j(i) = j$. Let

$$z_{i+1} = x_{i+1} - \sum_{k=1}^{j} y_k \langle y_k, x_{i+1} \rangle.$$

Assume first that $z_{i+1} \neq 0$. A straightforward calculation shows that z_{i+1} is orthogonal on y_k for $k \in [j]$. Then let $y_{j+1} = z_{i+1} \|z_{i+1}\|^{-1}$. Assume second that $z_{i+1} = 0$. Then $j(i+1) = j$ and we replace x_{i+1} by x_{i+2}. □

In what follows we need the following lemma

Lemma 7.2.3. *Let $x, y \in \mathbb{H}^l$. Then*

$$\Re(\langle x, y \rangle) = \Re(\langle \overline{y}, \overline{x} \rangle). \tag{7.9}$$

Proof. Use (7.6) and the definitions of $\overline{x}, \overline{y}$ to deduce

$$\Re(\langle x, y \rangle) = \Re(\sum_{i=1}^{l} x_i \bar{y}_i) = \Re(\sum_{i=1}^{l} \bar{y}_i x_i) = \Re(\langle \overline{y}, \overline{x} \rangle).$$

□

Let $b_1, \ldots, b_m$ be an orthonormal basis in a right vector space $\mathbb{V}$. Then

$$x = \sum_{i=1}^{m} b_i \langle b_i, x \rangle, \; y = \sum_{i=1}^{m} b_i \langle b_i, y \rangle, \; \langle x, y \rangle = \sum_{i=1}^{n} \langle x, b_i \rangle \langle b_i, y \rangle. \tag{7.10}$$

Definition 7.2.4. *Let $\mathbb{V}$ be a right m-dimensional vector space over $\mathbb{H}$ with the inner product $\langle\cdot\rangle$. Let $\mathcal{B} = \{b_1, \ldots, b_m\}$ be an orthonormal basis in $\mathbb{V}$. For a vector $v \in \mathbb{V}$ we define*

$$\overline{v}_{\mathcal{B}} = \sum_{i=1}^{m} b_i \overline{\langle b_i, v\rangle} = \sum_{i=1}^{m} b_i \langle v, b_i\rangle.$$

7.2.4 Tensor products over quaternions

There is no natural way to define the tensor product space over $\mathbb{H}$. So the definition below is coordinate dependent and should not be confused with the universal construction often used in other settings.

Given quaternion vector spaces $\mathbb{H}^m, \mathbb{H}^n$ with standard basis, we define the tensor product $\mathbb{H}^m \otimes \mathbb{H}^n$ as the space of $\mathbb{H}^{m\times n}$ matrices and $u \otimes v$ can be identified with $uv^\top$. On matrices $\mathbb{H}^m \otimes \mathbb{H}^n$ we define the inner product as:

$$\langle A, B\rangle = \text{ trace } A^*B = \sum_{i,j=1}^{m,n} A_{ij}^* B_{ij},\ A = (A_{ij}), B = (B_{ij}) \in \mathbb{H}^m \otimes \mathbb{H}^n.$$

Therefore

$$\langle u \otimes v, x \otimes y\rangle = \sum_{i,j=1}^{m,n} v_j^* u_i^* x_i y_j.$$

Notice that

$$\langle u, x\rangle\langle v, y\rangle = \sum_{i,j=1}^{m,n} u_i^* x_i v_j^* y_j.$$

In general, the two quantities are not the same due to noncommutativity.

Lemma 7.2.5. *Let $u, x \in \mathbb{H}^m, v, y \in \mathbb{H}^n$.*

1. *If $\langle u, x\rangle \in \mathbb{R}$ then $\langle u \otimes v, x \otimes y\rangle = \langle u, x\rangle\langle v, y\rangle$.*

2. *$\|u \otimes v\| = \|u\|\|v\|$.*

3. *$\Re(\langle \overline{v} \otimes v, \overline{y} \otimes y\rangle) = |\langle v, y\rangle|^2$.*

Proof. *(1)* If $\langle u, x\rangle \in \mathbb{R}$, then

$$\sum_{i,j=1}^{m,n} v_j^* u_i^* x_i y_j = \sum_{j=1}^{n} v_j^* (\sum_{i=1}^{m} u_i^* x_i) y_j = \sum_{i,j=1}^{m,n} u_i^* x_i v_j^* y_j.$$

(2) As $\langle u, u\rangle \geq 0$ it follows that$\langle u \otimes v, u \otimes v\rangle = \langle u, u\rangle\langle v, v\rangle$. Hence $\|u \otimes v\| = \|u\|\|v\|$.

(3)

$$\Re(\langle \overline{v} \otimes v, \overline{y} \otimes y\rangle) = \Re(\sum_{i,j=1}^{n,n} v_j^* v_i y_i^* y_j) = \Re(\sum_{i=1}^{n} v_i y_i^* \sum_{i=1}^{n} y_i v_i^*) = |\langle v, y\rangle|^2.$$

□

7.2.5 Schur's theorem for quaternions

Recall $\mathbb{S}^n(\mathbb{F}) \subset \mathbb{F}^{n\times n}$ is the space of A satisfying $A^* = A$. We call such matrices self-adjoint. Assume that $A = (a_{ij}) \in \mathbb{S}^n(\mathbb{H})$. We associate with A the quaternion form $Q(x) := x^* A x = \sum_{i=1}^{n}\sum_{j=1}^{n} \bar{x}_i a_{ij} x_j$ for $x \in \mathbb{H}^n$. As $(x^*Ax)^* = x^*Ax$ it follows that $Q(x)$ is always a real number. The matrix A is called positive semidefinite if $Q(x) \geq 0$ for all x. We denote by $\mathbb{S}_+^n(\mathbb{F})$ the cone of positive semidefinite self-adjoint matrices over $\mathbb{F}$. It is easy to check that $\mathbb{S}_+^n(\mathbb{H}) \cap \mathbb{R}^{n\times n} = \mathbb{S}_+^n(\mathbb{R})$. If $Q(x) > 0$ for all $x \neq 0$, $\langle x, y\rangle = x^*Ay$ defines an inner product in $\mathbb{H}^n$.

Denote by $\mathbb{U}^n(\mathbb{F}) \subset \mathbb{F}^{n\times n}$ the group of unitary matrices $U^*U = UU^* = I$. The spectral theorem of $A \in \mathbb{S}^n(\mathbb{F})$ claims that there exists a unitary U and a real diagonal D such that $A = UDU^*$ [74]. The columns of U are the eigenvectors of A with real left eigenvalues, which are the corresponding diagonal entries of D. Thus $A \in \mathbb{S}_+^n(\mathbb{F})$ if and only if all the left real eigenvalues of A are nonnegative. In that case A has a unique square root $A^{1/2} = UD^{1/2}U^* \in \mathbb{S}_+^n(\mathbb{F})$. Hence $A = \langle x_i, x_j\rangle$, where $x_1, \ldots, x_n$ are the columns of $A^{1/2}$. In particular, $\|x_1\| = \cdots = \|x_n\| = 1$ if and only if the diagonal entries of A are 1. To get the expression $A = \langle x_i, x_j\rangle$, we can also use the Cholesky decomposition. The usual algorithm for Cholesky decomposition works for quaternions.

Lemma 7.2.6. *Assume that $M \in \mathbb{H}^{n\times n}$ has representation $\hat{C}(M) \in \mathbb{C}^{(2n)\times(2n)}$ given by (7.8). Then*

1. *$M \in \mathbb{S}^n(\mathbb{H})$ if and only if $Z \in \mathbb{S}^n(\mathbb{C})$ and W is skew symmetric: $W^\top = -W$.*

2. *$M \in \mathbb{S}^n(\mathbb{H})$ if and only if $\bar{M} \in \mathbb{S}^n(\mathbb{H})$. Furthermore $M \in \mathbb{S}^n_+(\mathbb{H})$ if and only if $\bar{M} \in \mathbb{S}^n_+(\mathbb{H})$.*

3. *$M \in \mathbb{S}_+(\mathbb{H})$ if and only if $\hat{C}(M) \in \mathbb{S}^{2n}_+(\mathbb{C})$.*

Proof. (1) Assume that $M = Z + W\mathrm{j}$, where $Z, W \in \mathbb{C}^{n\times n}$. Then $M^* = Z^* - W^\top\mathrm{j}$. Thus $M^* = M$ if and only if $Z^* = Z$ and $-W^\top = W$. This is equivalent to the statement that $\hat{C}(M) \in \mathbb{S}^{2n}(\mathbb{C})$.

(2) As $\bar{M} = \bar{Z} - W\mathrm{j}$ we deduce that M is self-adjoint if and only if $\bar{M}$ is self-adjoint. Suppose that $M \in \mathbb{S}^n_+(\mathbb{H})$. Then $M = UDU^*$ where U is unitary and D is a real diagonal with nonnegative diagonal entries. Then $\bar{M} = \overline{UDU^*} = \overline{U^*}D\bar{U} = U^\top D\bar{U}$. As $\bar{U}$ is unitary we deduce that $\bar{M} \in \mathbb{S}^n_+(\mathbb{H})$. Similarly if $\bar{M}$ is positive semidefinite, then M is positive semidefinite.

(3) Assume that M is self-adjoint. Then $\hat{C}(M)$ is positive semidefinite if and only if $x^*Zx + y^*\bar{Z}y + 2\Re x^*Wy \geq 0$ for $x, y \in \mathbb{C}^n$. Replace x, y with $x, -y$ we deduce that the above Hermitian form is nonnegative if and only if the form $x^*Zx + y^*\bar{Z}y - 2\Re x^*Wy$ is nonnegative.

Assume that $M = Z + W\mathrm{j} \in \mathbb{S}^n(\mathbb{H})$. Let $x = z + w\mathrm{j} \in \mathbb{H}^n$, where $z, w \in \mathbb{C}^n$. A straightforward calculation shows:

$$x^*Mx = (z^* - w^\top\mathrm{j})(Z + W\mathrm{j})(z + w\mathrm{j}) = z^*Zz + w^\top\bar{Z}\overline{w} - z^*W\overline{w} + w^\top\bar{W}z$$

Clearly $(w^\top\bar{W}z)^* = z^*W^\top\overline{w} = -z^*W\overline{w}$. As x^*Mx is a real number it follows that $x^*Mx = z^*Zz + w^\top\bar{Z}\overline{w} - 2\Re z^*W\overline{w}$. Set $y = \overline{w}$ to deduce the claim. $\square$

For $A = (a_{ij}), B = (b_{ij}) \in \mathbb{F}^{m \times n}$ denote by $A \circ B = (a_{ij}b_{ij})$ the Schur product of two matrices. Assume that $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. Then the Schur product of two self-adjoint matrices is self-adjoint. Furthermore, Schur's theorem claims that the Schur product of two positive semidefinite matrices is positive semidefinite. Assume that $\mathbb{F} = \mathbb{H}$. Since $\mathbb{H}$ is not commutative, the product of two quaternion self-adjoint is usually not self-adjoint. There are two simple exceptions: either $B \in \mathbb{S}^n(\mathbb{R})$ or $B = \bar{A}$.

Lemma 7.2.7 (The Schur product theorem for quaternions)**.** *For symmetric positive semidefinite real matrix M and self-adjoint positive semidefinite quaternion matrix N, their Hadamard product, defined by $(M \circ N)_{ij} := M_{ij}N_{ij}$, is self-adjoint positive semidefinite. The matrix $L_{ij} := N_{ij}N^*_{ij} = \|N_{ij}\|^2$ is also self-adjoint positive semidefinite.*

Proof. M can be written as $M_{ij} = \langle a_i, a_j \rangle$ where $a_i \in \mathbb{R}^n$. And N can be written as $N_{ij} = \langle x_i, x_j \rangle$ where $x_i \in \mathbb{H}^n$. By Lemma 7.2.5,

$$\langle a_i \otimes x_k, a_j \otimes x_l \rangle = \langle a_i, a_j \rangle \langle x_k, x_l \rangle.$$

So both the Kronecker product and Schur product of M and N are semidefinite positive.

For the second claim, again by Lemma 7.2.5, $L_{ij} = \|\langle x_i, x_j \rangle\|^2 = \mathfrak{R}(\langle \bar{x}_i \otimes x_i, \bar{x}_j \otimes x_j \rangle)$. So L is the real part of a positive semidefinite matrix and it is also positive semidefinite. □

Lemma 7.2.8. *Let $\mathcal{H}$ be a Hilbert space over quaternions. Assume that $x_1, \ldots, x_n$ be n unit vectors in $\mathcal{H}$. Then for each $m \in \mathbb{N}$, there exist m unit vectors $x_{1,m} \ldots, x_{n,m} \in \mathbb{H}^n$ such that $\langle x_{i,m}, x_{j,m} \rangle = \langle x_i, x_j \rangle |\langle x_i, x_j \rangle|^{2m}$.*

Proof. Let $A_0 = (\langle x_i, x_j \rangle)$. Then $A_0 \in \mathbb{S}^n_+$ is a correlation matrix. Denote $A_m = A_{m-1} \circ (A_0 \circ \bar{A}_0)$ for $m \in \mathbb{N}$. Use induction and Lemma 7.2.7 to deduce that A_m is a correlation matrix. Then $A_m = (\langle x_{i,m}, x_{j,m} \rangle)$. Hence $\langle x_{i,m}, x_{j,m} \rangle = \langle x_i, x_j \rangle |\langle x_i, x_j \rangle|^{2m}$. □

7.2.6 The kernel trick

We now state the kernel trick for quaternion Hilbert space [194]. This technique has been used in [78] to deduce in a unified way the Krivine-Haagerup upper bound on the Grothendieck constant $K_G^{\mathbb{F}}$ for the fields of real or complex numbers $\mathbb{F}$.

Lemma 7.2.9. *Let $\mathcal{H}$ be a Hilbert space over quaternions. Assume that $x_1, \ldots, x_n, y_1, \ldots, y_n$ are $2n$ unit vectors in $\mathcal{H}$. Suppose that $g(z)$ is an analytic function in the unit complex disk with Taylor series satisfying the following conditions.*

$$g(z) = \sum_{i=0}^{\infty} a_m z^{2i}, \quad a_i \in \mathbb{R}, \quad \sum_{i=0}^{\infty} |a_m| = 1$$

Then there exists $2n$ unit vectors $u_1, \ldots, u_n, v_1, \ldots, v_n$ in a Hilbert space $\mathcal{H}'$ such that

$$\langle x_i, y_j \rangle g(|\langle x_i, y_j \rangle|) = \langle u_i, v_j \rangle, \quad i, j = 1, \ldots, n. \tag{7.11}$$

Proof. Let $z_i = x_i, z_{n+i} = y_i$ for $i \in [n]$. Lemma 7.2.8 yields the existence of unit vectors $w_{i,m} \in \mathbb{H}^{2n}$ such that $\langle w_{i,m}, w_{j,m} \rangle = \langle z_i, z_j \rangle |\langle z_i, z_j \rangle|^{2m}$ for each $m \in \mathbb{N}$ and $i, j \in [2n]$.

Let $\mathcal{H}' = \mathcal{H} \oplus (\oplus_{m=1}^{\infty} \mathbb{H}^{2n})$ with the corresponding induced inner product. Define

$$u_i = x_i \sqrt{|a_0|} \oplus (\oplus_{m=1}^{\infty} w_{i,m} \sqrt{|a_m|}),$$
$$v_i = y_i \operatorname{sgn} a_0 \sqrt{|a_0|} \oplus (\oplus_{m=1}^{\infty} w_{n+i,m} \operatorname{sgn} a_m \sqrt{|a_m|})$$

for $i \in [n]$. Then u_i, v_i are unit vectors and (7.11) holds. □

7.2.7 Existence of the Grothendieck constant for quaternions

In this paper we view $\mathbb{H}^{m\times n}$ as a left vector space over $\mathbb{H}$. We introduce two norms on $\mathbb{H}^{m\times n}$:

$$\|M\|_{\infty,1,\mathbb{H}} = \max\{|\sum_{i,j}^{m,n} M_{ij}\bar{\varepsilon}_i\delta_j|,\ \varepsilon_i, \delta_j \in \mathbb{H}, |\varepsilon_i| = |\delta_j| = 1, i \in [m], j \in [n]\}, \tag{7.12}$$

$$\|M\|_{G,\mathbb{H}} = \max\{|\sum_{i,j}^{m,n} M_{ij}\langle x_i, y_j\rangle|, x_i, y_i \in \mathcal{H}, \|x_i\| = \|y_j\| = 1, i \in [m], j \in [n]\}. \tag{7.13}$$

Here $\mathcal{H}$ is a right Hilbert space over quaternions. If we choose $\mathcal{H}$ to be one dimensional, then the maximum in the Grothendieck norm is the maximum of $(\infty, 1)$ norm. Hence we have the inequality $\|M\|_{\infty,1} \le \|M\|_{G,\mathbb{H}}$. Thus the problem is now the following: do we have the reverse inequality independent on the dimensions m, n: $K_G^{\mathbb{H}}\|M\|_{\infty,1} \ge \|M\|_{G,\mathbb{H}}$? By multiplying each δ_j and y_j by a fixed $a \in \mathbb{H}, |a| = 1$ from the right, for $j \in [n]$, we can replace absolute values in the definitions of the norms $\|M\|_{\infty,1}, |M\|_{G,\mathbb{H}}$ by the real part

$$\|M\|_{\infty,1,\mathbb{H}} = \max\{\Re(\sum_{i,j}^{m,n} M_{ij}\bar{\varepsilon}_i\delta_j),\ \varepsilon_i, \delta_j \in \mathbb{H}, |\varepsilon_i| = |\delta_j| = 1, i \in [m], j \in [n]\}, \tag{7.14}$$

$$\|M\|_{G,\mathbb{H}} = \max\{\Re(\sum_{i,j}^{m,n} M_{ij}\langle x_i, y_j\rangle), x_i, y_j \in \mathcal{H}, \|x_i\| = \|y_j\|, i \in [m], j \in [n]\}. \tag{7.15}$$

Observe next that the maximum in the characterizations above we can replace the equalities $|\varepsilon_i| = |\delta_j| = \|x_i\| = \|y_j\| = 1$ by the inequalities $|\varepsilon_i|, |\delta_j|, \|x_i\|, \|y_j\| \le 1$ [77].

Next we now are going to replace the maximum in the above characterization for $\|M\|_{\infty,1,\mathbb{H}}$ with quaternions with matrices of sizes $(2m) \times (2n)$ with complex entries. We start with the following lemma which follows by straightforward calculation:

Lemma 7.2.10. *Assume that the quaternions $\alpha, \varepsilon, \delta$ have the following matrix representa-*

tions:

$$\alpha = \begin{pmatrix} a & b \\ -\bar{b} & \bar{a} \end{pmatrix}, \varepsilon = \begin{pmatrix} z & w \\ -\bar{w} & \bar{z} \end{pmatrix}, \delta = \begin{pmatrix} u & v \\ -\bar{v} & \bar{u} \end{pmatrix} \in \mathbb{C}^{2\times 2}.$$

Then

$$\Re(\alpha\bar{\varepsilon}\delta) = \Re((\bar{z},\bar{w}) \begin{pmatrix} a & -\bar{b} \\ b & \bar{a} \end{pmatrix} (u,v)^\top) = \Re((\bar{z},\bar{w})A(\alpha)^\top (u,v)^\top).$$

Lemma 7.2.11. *Let*

$$M = (M_{ij}) \in \mathbb{H}^{m\times n}, \varepsilon_i = (z_i, w_i), \delta_j = (u_i, v_i) \in \mathbb{H}, z_i, w_i, u_j, v_j \in \mathbb{C},$$

where $|z_i|^2 + |w_i|^2 = |u_i|^2 + |v_i|^2 = 1, i \in [m], j \in [n]$. *To each quaternion* $M_{ij} = (M_{ij,1}, M_{ij,2})$ *associate the matrix* $C(M_{ij}) = \begin{pmatrix} M_{ij,1} & M_{ij,2} \\ -\bar{M}_{ij,2} & \bar{M}_{ij,1} \end{pmatrix}$. *Let* $\tilde{M} = (C(M_{ij})^\top) \in \mathbb{C}^{(2m)\times(2n)}$ *and*

$$e = (\bar{z}_1, \bar{w}_1, \bar{z}_2, \bar{w}_2, \ldots, \bar{z}_m, \bar{w}_m)^\top \in \mathbb{C}^{2m}, \quad d = (u_1, v_1, u_2, v_2, \ldots, u_n, v_n)^\top \in \mathbb{C}^{2n}.$$

Then

$$Re(\sum_{i,j}^{m,n} M_{ij}\bar{\varepsilon}_i\delta_j) = \Re(e^\top \tilde{M} d),$$

where $\varepsilon_i, \delta_j \in \mathbb{H}, |\varepsilon_i| = |\delta_j| = |z_i|^2 + |w_i^2| = |u_j|^2 + |v_j|^2 = 1$, *and*

$$\|M\|_{1,\infty,\mathbb{H}} = \max\{\Re(e^\top \tilde{M} d), e \in \mathbb{C}^{2m}, d \in \mathbb{C}^2 n, |z_i|^2 + |w_i^2| = |u_j|^2 + |v_j|^2 = 1\}. \quad (7.16)$$

In particular

$$\|M\|_{1,\infty,\mathbb{H}} \le \|\tilde{M}\|_{1,\infty,\mathbb{C}} \le 2\|M\|_{1,\infty,\mathbb{H}}. \quad (7.17)$$

Proof. Use Lemma 7.2.11 to obtain the first equality of the lemma. Clearly, the set $|z|^2 +$

$|w|^2 = 1$ is a subset of $|z| \leq 1, |w| \leq 1$. Hence the characterizations (7.16) and the complex norm $\|\tilde{M}\|_{\infty,1,\mathbb{C}}$ yield the first inequality in (7.17). In the equality $|z|^2 + |w|^2 = 1$, choose a particular case $|z| = |w| = \frac{1}{\sqrt{2}}$. Then $|\sqrt{2}z| = |\sqrt{2}w| = 1$. Hence the maximal characterization for $2\tilde{M}$ corresponding to $\|2M\|_{\infty,1,\mathbb{H}}$ is not less than $\|\tilde{M}\|_{\infty,1,\mathbb{C}}$. □

We now recast $\|M\|_{G,\mathbb{H}}$ in terms of the above matrix $\tilde{M}$. We can first write $x, y \in \mathbb{H}^l$ as $z + w\mathrm{j}, u + v\mathrm{j}$, where $z, w, u, v \in \mathbb{C}^l$. We next observe that

$$\langle x, y\rangle = \langle z + w\mathrm{j}, u + v\mathrm{j}\rangle = (\langle z, u\rangle + \langle v, w\rangle) + (\langle z, v\rangle - \langle u, w\rangle)\mathrm{j}. \tag{7.18}$$

Introduce the following four vectors:

$$f_1 = (z^\top, \overline{w}^\top)^\top,\ f_2 = (w^\top, -\overline{z}^\top)^\top,\ g_1 = (u^\top, \overline{v}^\top)^\top,\ g_2 = (v^\top, -\overline{u}^\top)^\top \in \mathbb{C}^{2l}.$$

Then

$$\langle x, y\rangle = \langle f_1, g_1\rangle + \langle f_1, g_2\rangle\mathrm{j} = \overline{\langle f_2, g_2\rangle} - \overline{\langle f_2, g_1\rangle}\mathrm{j}.$$

Note that $\langle f, g\rangle = f^* g$ is the standard inner product on $\mathbb{C}^{2l}$. Furthermore

$$\|f_1\| = \|f_2\| = \|x\|, \quad \|g_1\| = \|g_2\| = \|y\|,$$

The next lemma relates $\Re(\alpha\langle x, y\rangle)$ and $\Re(\sum_{i,j=1}^{2}(A(\alpha)^\top)_{ij}\langle f_i, g_j\rangle)$:

Lemma 7.2.12. *Let $\alpha = a + b\mathrm{j} \in \mathbb{H}$ and $x, y \in \mathbb{H}^l$. Then*

$$\Re(\alpha\langle x, y\rangle) = \frac{1}{2}\Re(\sum_{p,q=1}^{2}(A(\alpha)^\top)_{pq}\langle f_p, g_q\rangle).$$

Proof. The above formulas yield

$$\Re(\alpha\langle x,y\rangle) = \Re(a(\langle z,u\rangle + \langle v,w\rangle) - b(\overline{\langle z,v\rangle - \langle u,w\rangle})$$

$$= \Re(a(\langle u,z\rangle + \langle w,v\rangle) - b(\langle v,z\rangle - \langle w,u\rangle)$$

$$= \Re(a\langle f_1,g_1\rangle + b\langle f_2,g_1\rangle)$$

$$= \Re(\bar{a}\langle f_2,g_2\rangle - \bar{b}\langle f_1,g_2\rangle).$$

As $A(\alpha) = \begin{pmatrix} a & b \\ -\bar{b} & \bar{a} \end{pmatrix}$ we get

$$\Re(\sum_{p,q=1}^{2} (A(\alpha)^\top)_{pq}\langle f_p,g_q\rangle) = \Re(a\langle f_1,g_1\rangle - \bar{b}\langle f_1,g_2\rangle + b\langle f_2,g_1\rangle + \bar{a}\langle f_2,g_2\rangle).$$

Compare the two expressions to deduce the lemma. □

The following result is an improvement of the Grothendieck's result [87] that $K_G^{\mathbb{C}} \le 2K_G^{\mathbb{R}}$, and an analog of result in [77].

Theorem 7.2.13. *$K_G^{\mathbb{H}} \le K_G^{\mathbb{C}}$.*

Proof. Let $M = (M_{ij}) \in \mathbb{H}^{m\times n}$. Let $\tilde{M} \in \mathbb{C}^{(2m)\times(2n)}$ be defined as in the proof of Lemma 7.2.11. Lemma 7.2.11 claims that $\|\tilde{M}\|_{\infty,1,\mathbb{C}} \le 2\|\tilde{M}\|_{\infty,1,\mathbb{H}}$.

Assume that the entries of $\tilde{M}$ are $\hat{M}_{ij} \in \mathbb{C}$ where $i \in [2m], j \in [2n]$. Let $x_1,\dots,x_m$, $y_1,\dots,y_n \in \mathcal{H}$, be vectors of norm one, where $\mathcal{H}$ is an inner product space over quaternions. As one has Gram-Schmidt process in $\mathcal{H}$ we can assume that $x_1,\dots,x_m,y_1,\dots,y_n \in \mathbb{H}^{n+m}$. Consider the Grothendieck norm $\|M\|_{G,\mathbb{H}}$. For each x_i, y_j define $f_{2i-1}, f_{2i}, g_{2j-1}, f_{2j}$ as before the proof of Lemma 7.2.12. The proof of Lemma 7.2.12 that $\|M\|_{G,\mathbb{H}}$ is maximum on all $f_1, \ldots, f_{2m}, g_1, \ldots, g_{2n} \in \mathbb{C}^{2(n+m)}$ of the expression

$$\frac{1}{2}\Re(\sum_{i,j=1}^{m,n}\sum_{p,q=1}^{2}(C(M_{i,j})^\top)_{pq}\langle f_{2(i-1)+p}, g_{2(j-1)+q}\rangle).$$

Clearly, the above maximum is not more then the maximum for $\frac{1}{2}\|\tilde{M}\|_{G,\mathbb{C}}$, since the vectors $f_{2i-1}, f_{2i}, g_{2j-1}, f_{2j}$ are of a special form. Hence

$$\|M\|_{G,\mathbb{H}} \le \frac{1}{2}\|\tilde{M}\|_{G,\mathbb{C}} \le \frac{K_G^{\mathbb{C}}}{2}\|\tilde{M}\|_{\infty,1,\mathbb{C}} \le K_G^{\mathbb{C}}\|M\|_{\infty,1,\mathbb{H}}.$$

(The right hand side inequality follows from (7.17).) □

7.2.8 Semidefinite programming for computing quaternion Grothendieck norm

In this subsection we state the computation of $\|M\|_{G,\mathbb{H}}$ as an SDP problem on positive semidefinite Hermitian matrices. We use the characterization (7.15). Let $z_i = x_i, z_{n+j} = y_j$ for $i \in [m], j \in [n]$. Denote by $\mathbb{G}^p(\mathbb{H}) \subset \mathbb{S}_+^p(\mathbb{H})$ the convex set of quaternion correlation matrices, i.e., all positive semidefinite quaternionic matrices whose diagonal entries are 1. Assume that $G = G(z_1, \ldots, z_{m+n}) \in \mathbb{G}^{m+n}(\mathbb{H})$. Let $H = \hat{C}(G)$. Then $H \in \mathbb{S}^{2(m+n)}(\mathbb{C})$ is a complex correlation matrix of the form (7.8), where $Z, W \in \mathbb{C}^{m+n}$ and $W^\top = -W$. Let us denote this real subspace of complex correlation matrices by $\mathcal{C}^{2(m+n)}$. Define as in [77]

$$A(M) = \begin{bmatrix} 0 & M \\ M^* & 0 \end{bmatrix} \in \mathbb{S}^{m+n}(\mathbb{H})$$

Lemma 7.2.14. *Assume that $M \in \mathbb{H}^{m\times n}$. Then*

$$\|M\|_{G,\mathbb{H}} = \frac{1}{2}\max\{\Re\operatorname{tr} A(M)G, G \in \mathbb{G}^{m+n}(\mathbb{H})\} = \frac{1}{4}\max\{\operatorname{tr}\hat{C}(A(M))H, H \in \mathcal{C}^{2(m+n)}\}.$$

Proof. Let $G = G(z_1, \ldots, z_{m+n})$ be defined as above. Then

$$\begin{aligned}
\operatorname{tr} A(M)\bar{G} &= \sum_{i=1}^{m}\sum_{j=1}^{n} M_{ij}\overline{\langle y_j, x_i\rangle} + \overline{M_{ji}}\overline{\langle x_i, y_j\rangle} = \sum_{i=1}^{m}\sum_{j=1}^{n} M_{ij}\langle x_i, y_j\rangle + \overline{M_{ji}}\overline{\langle x_i, y_j\rangle},\\
\Re \operatorname{tr} A(M)\bar{G} &= \sum_{i=1}^{m}\sum_{j=1}^{n} \Re M_{ij}\langle x_i, y_j\rangle + \Re\overline{M_{ij}}\overline{\langle x_i, y_j\rangle} = 2\Re\sum_{i=1}^{m}\sum_{j=1}^{n} M_{ij}\langle x_i, y_j\rangle
\end{aligned}$$

(To deduce the last equality we used (7.6).) Lemma 7.2.6 yields that $\bar{G} \in \mathbb{S}_+^{m+n}$ if and only if $\bar{G} \in \mathbb{S}_+^{m+n}$. As the diagonal entries of $\bar{G}$ are 1 we deduce that $\bar{G} \in \mathbb{G}^{m+n}(\mathbb{H})$ if and only if $G \in \mathbb{G}^{m+n}(\mathbb{H})$. Use (7.15) to deduce the first part of the characterization of $\|M\|_{G,\mathbb{H}}$. Since $A(M), G \in \mathbb{S}^{m+n}(\mathbb{H})$ we have

$$\begin{aligned}
&M = M_1 + M_2\mathrm{j},\ A(M) = A_1 + A_2\mathrm{j} = \begin{bmatrix} 0 & M_1 \\ M_1^* & 0 \end{bmatrix} + \begin{bmatrix} 0 & M_2 \\ -M_2^\top & 0 \end{bmatrix}\mathrm{j},\ G = G_1 + G_2\mathrm{j},\\
&M_1, M_2 \in \mathbb{C}^{m\times n}, \quad G_1 \in \mathbb{S}^{m+n}(\mathbb{C}), \quad G_2 \in \mathbb{C}^{(m+n)\times(m+n)}, \quad G_2^\top = -G_2.
\end{aligned}$$

Hence

$$\bar{G} = \bar{G}_1 - G_2\mathrm{j}, \quad \Re \operatorname{tr} A(M)\bar{G} = \Re(A_1\bar{G}_1 + A_2\bar{G}_2).$$

Observe next

$$\begin{aligned}
\hat{C}(A(M)) &= \begin{bmatrix} A_1 & A_2 \\ -\bar{A}_2 & \bar{A}_1 \end{bmatrix}, \quad \hat{C}(\bar{G}) = \begin{bmatrix} \bar{G}_1 & -G_2 \\ \bar{G}_2 & G_1 \end{bmatrix},\\
\operatorname{tr} \hat{C}(A(M))\hat{C}(\bar{G})) &= \operatorname{tr}(A_1\bar{G}_1 + A_2\bar{G}_2 + \bar{A}_2 G_2 + \bar{A}_1 G_1) = 2\Re(A_1\bar{G}_1 + A_2\bar{G}_2).
\end{aligned}$$

As $\bar{G} \in \mathbb{G}^{m+n}(\mathbb{H})$ we deduce that $\hat{C}(\bar{G}) \in \mathcal{C}^{2(m+n)}$. Vice versa, if $H \in \mathcal{C}^{2(m+n)}$ then $H = \widehat{C(\bar{G})}$ for some $\bar{G} \in \mathbb{G}^{m+n}(\mathbb{H})$. Hence G is a quaternion correlation matrix. This proves the lemma. □

7.2.9 The sign function for quaternions

We define our sign function over $\mathbb{F}$

$$\operatorname{sgn} z = \begin{cases} z/|z| & z \neq 0, \\ 0 & z = 0. \end{cases} \tag{7.19}$$

Denote by $\mathrm{S}^3 = \{a \in \mathbb{H},\ |a| = 1\}$, the 3-dimensional sphere in $\mathbb{R}^4$. Note that multiplication by $\phi_a(b) = ab$, and $\psi_a(b) = ba$ are orientation preserving orthogonal transformation on $\mathbb{H}$ for a fixed $a \in \mathrm{S}^3$ and $b \in \mathbb{H}$. In particular $\phi_a(\mathrm{S}^3) = \psi_a(\mathrm{S}^3) = \mathrm{S}^3$.

On S^3 let $d\sigma$ be the Haar measure on S^3, which is invariant under the action of ϕ_a, ψ_b. We now give the following generalization of Haagerup formula [90] for $\operatorname{sgn}(z)$ for quaternions:

Lemma 7.2.15. *Let $z \in \mathbb{H}$. Then*

$$\operatorname{sgn}(z) = \frac{3}{8\pi} \int_{w \in \mathrm{S}^3} \operatorname{sgn}(\Re(\bar{w}z)) w d\sigma(w). \tag{7.20}$$

Proof. Clearly for $z = 0$ (7.20) trivially holds. We next assume that $z = 1$. Hence the left hand side of (7.20) is 1. Let $w = w_0 + w_1\mathrm{i} + w_2\mathrm{j} + w_3\mathrm{k} \in \mathrm{S}^3$. Then $\Re(w^*) = w_0$ and $\operatorname{sgn}(\Re(\bar{w})) = \operatorname{sgn}(w_0)$. We first observe that

$$\int_{w \in \mathrm{S}^3} \operatorname{sgn}(w_0) w_j d\sigma(w) = 0 \text{ for } j \in [3].$$

This follows from the observation that the transformation $w \mapsto \bar{w}$ is Haar measure preserving on S^3. Hence

$$\int_{w \in \mathrm{S}^3} \operatorname{sgn}(w_0) w_j d\sigma(w) = -\int_{w \in \mathrm{S}^3} \operatorname{sgn}(w_0) w_j d\sigma(w), \quad j \in [3].$$

Observe next that $\mathrm{sgn}(\Re(\bar{w}))w_0 = \mathrm{sgn}(w_0)w_0 = |w_0|$. Thus we need to show that

$$\int_{w\in \mathrm{S}^3} |w_0| d\sigma(w) = \frac{8\pi}{3}.$$

We now introduce the spherical coordinates on $\mathbb{R}^4$ as follows:

$$w_0 = \cos(\phi_0), w_1 = \sin(\phi_0)\cos(\phi_1), w_2 = \sin(\phi_0)\sin(\phi_1)\cos(\phi_2),$$
$$w_3 = \sin(\phi_0)\sin(\phi_1)\sin(\phi_2), d\sigma = \sin^2(\phi_0)\sin(\phi_1)d\phi_0 d\phi_1 d\phi_2,$$
$$\phi_0, \phi_1 \in [0,\pi], \phi_2 \in [0, 2\pi].$$

Hence

$$\int_{w\in \mathrm{S}^3} |w_0| d\sigma(w) = \big(\int_0^\pi |\cos(\phi_0)|\sin^2(\phi_0)d\phi_0\big)\big(\int_0^\pi \sin(\phi_1)d\phi_1\big)\big(\int_0^{2\pi} d\phi_2\big) = \frac{8\pi}{3}.$$

Hence (7.20) holds for $z = 1$. For a general $z \neq 0$ we recall that $\mathrm{sgn}(z) = \mathrm{sgn}(tz)$ for any $t > 0$. Hence it is ehough to show (7.20) for $z \in \mathrm{S}^3$. That is we need to show the equality

$$z = \frac{3}{8\pi}\int_{w\in \mathrm{S}^3} \mathrm{sgn}(\Re(\bar{w}z))w d\sigma(w), \quad z \in \mathrm{S}^3$$

By multiplying by $\bar{z}$ from the left it is sufficient to show that

$$1 = \frac{3}{8\pi}\int_{w\in \mathrm{S}^3} \mathrm{sgn}(\Re(\bar{w}z))\bar{z}w d\sigma(w).$$

Now introduce a new variable $u = \phi_{\bar{z}}(w) = \bar{z}w$ on S^3 . Note that $\bar{u} = \bar{w}z$. Since the Haar measure on S^3 is invariant under $\phi_{\bar{z}}$ we get that $d\sigma(w) = d\sigma(u)$. Hence the above equality is equivalent to (7.20) for $z = 1$, which was proved. □

7.2.10 Quaternion Gaussian

Recall the distribution of $G_n^{\mathbb{R}}(z)$ and $G_n^{\mathbb{C}}(z)$ given [90]. The Gaussian quaternions $\mathbb{H}^n$ has the distribution

$$G_n^{\mathbb{H}}(z) = \big(\frac{\pi}{2}\big)^{-2n} \exp(-2\|z\|_2^2).$$

The variance chosen here is totally arbitrary.

Theorem 7.2.16. *Assume that $u, v \in \mathcal{H}$ are of norm one, where $\mathcal{H}$ is a right vector space over $\mathbb{H}$. $m(z)$ is the Lebesgue measure in $\mathbb{H}^n$, then*

$$\begin{aligned}\int_{\mathbb{H}^n} \operatorname{sgn}\langle u, z\rangle \operatorname{sgn}\langle z, v\rangle G_n^{\mathbb{H}}(z) dm(z) &= \langle u, v\rangle f_{\mathbb{H}}(|\langle u, v\rangle|) \\ &= \langle u, v\rangle \frac{3}{2} \int_0^{\frac{\pi}{2}} \frac{\cos^4 t}{\sqrt{1 - |\langle u, v\rangle|^2 \sin^2 t}} dt.\end{aligned}$$

If $\langle u, v\rangle$ is real, then

$$\begin{aligned}\int_{\mathbb{H}^n} \operatorname{sgn}\langle z, u\rangle \operatorname{sgn}\langle v, z\rangle G_n^{\mathbb{H}}(z) dm(z) &= \langle u, v\rangle f_{\mathbb{H}}(|\langle u, v\rangle|) \\ &= \langle u, v\rangle \frac{3}{2} \int_0^{\frac{\pi}{2}} \frac{\cos^4 t}{\sqrt{1 - |\langle u, v\rangle|^2 \sin^2 t}} dt.\end{aligned}$$

Proof. We will prove the first formula, the proof of second formula is the same. First of all we point out that it is sufficient to prove this formula for $n = 2$. Indeed since span (u, v) is at most two dimensional, by performing Gram-Schmidt orthogonalization process we can find an orthonormal basis in $\mathbb{F}^n$ such that $u, v \in$ span (e_1, e_2). $\mathbb{F}^n$ can be decomposed as span (e_1, e_2) and its complement. By the corresponding decomposition of the Gaussian variable, its integration on the complement dimension is simply 1.

Recall the famous Grothendieck inequality: for any fixed real vectors u, v of norm 1, we have

$$\int_{\mathbb{R}^n} \operatorname{sgn}\langle x, u\rangle \operatorname{sgn}\langle x, v\rangle G_n^{\mathbb{R}}(x)dx = \frac{2}{\pi}\arcsin\langle u, v\rangle.$$

For $u, v, z \in \mathbb{H}^n$, denote their real vector form as $u, v, z \in \mathbb{R}^{4n}$. Then $\langle u, v\rangle = \Re\langle u, v\rangle$ and the Grothendieck inequality becomes:

$$\int_{\mathbb{H}^n} \operatorname{sgn}\Re\langle u, z\rangle \operatorname{sgn}\Re\langle z, v\rangle G_n^{\mathbb{H}}(z)dm(z) = \frac{2}{\pi}\arcsin\Re\langle u, v\rangle.$$

By Lemma 7.2.15, we have

$$\begin{aligned}
&\int_{\mathbb{H}^n} \operatorname{sgn}\langle u, z\rangle \operatorname{sgn}\langle z, v\rangle G_n^{\mathbb{H}}(z)dm(z)\\
=&\frac{9}{64\pi^2}\int_{w_1,w_2\in \mathrm{S}^3}\int_{z\in\mathbb{H}^n} \operatorname{sgn}(\Re(\bar{w}_1\langle u, z\rangle))w_1 \operatorname{sgn}(\Re(\bar{w}_2\langle z, v\rangle))w_2 d\sigma(w_1)d\sigma(w_2)G_n^{\mathbb{H}}(z)dm(z)\\
=&\frac{9}{32\pi^3}\int_{w_1,w_2\in \mathrm{S}^3} \arcsin(\Re(\langle uw_1, v\bar{w}_2\rangle))w_1w_2 d\sigma(w_1)d\sigma(w_2)
\end{aligned}$$

(1) Assume now $\langle u, v\rangle = a \in \mathbb{R}$, then $\Re\langle uw_1, v\bar{w}_2\rangle = a\Re(\bar{w}_1\bar{w}_2) = a\Re(w_1w_2)$. Thus we deduce

$$\begin{aligned}
\langle u, v\rangle f_{\mathbb{H}}(|\langle u, v\rangle|) &= \frac{9}{32\pi^3}\int_{w_1,w_2\in \mathrm{S}^3} \arcsin(|\langle u, v\rangle|\Re(w_1w_2))w_1w_2 d\sigma(w_1)d\sigma(w_2)\\
&= \frac{9}{32\pi^3}\int_{w_1,w_2\in \mathrm{S}^3} \arcsin(|\langle u, v\rangle|\Re(w_1(w_1^{-1}w_2))w_1(w_1^{-1}w_2)d\sigma(w_1)d\sigma(w_2)\\
&= \frac{9}{16\pi}\int_{q\in \mathrm{S}^3} \arcsin(|\langle u, v\rangle|\Re(q))q d\sigma(q)
\end{aligned}$$

The second equality is due to the fact that $d\sigma(w_2)$ is a Haar measure. The third equality is due to the fact that the volume of S^3 is equal to $2\pi^2$.

Observe that the left hand side of this equality is real. Hence

$$\langle u, v\rangle f_{\mathbb{H}}(|\langle u, v\rangle|) = \frac{9}{16\pi}\int_{q\in \mathrm{S}^3} \arcsin(|\langle u, v\rangle|\Re(q))\Re(q)d\sigma(q).$$

Now use the spherical coordinates as before to deduce that

$$\begin{aligned}\langle u,v\rangle f_{\mathbb{H}}(|\langle u,v\rangle|) &= \frac{9}{4}\int_0^{\pi} \arcsin(|\langle u,v\rangle|\cos\phi_0)\cos\phi_0\sin^2\phi_0 d\phi_0\\ &= \frac{9}{2}\int_0^{\frac{\pi}{2}} \arcsin(|\langle u,v\rangle|\cos\phi_0)\cos\phi_0\sin^2\phi_0 d\phi_0\\ &= \frac{9}{2}\int_0^{\frac{\pi}{2}} \arcsin(|\langle u,v\rangle|\sin t)\sin t\cos^2 t\, dt.\end{aligned}$$

Finally do the integration by part to get

$$\langle u,v\rangle f_{\mathbb{H}}(|\langle u,v\rangle|) = \frac{3|\langle u,v\rangle|}{2}\int_0^{\frac{\pi}{2}} \frac{\cos^4 t}{\sqrt{1-|\langle u,v\rangle|^2\sin^2 t}}dt.$$

Thus

$$f_{\mathbb{H}}(|\langle u,v\rangle|) = \frac{3}{2}\int_0^{\frac{\pi}{2}} \frac{\cos^4 t}{\sqrt{1-|\langle u,v\rangle|^2\sin^2 t}}dt. \tag{7.21}$$

(2) If $\langle u,v\rangle$ is not real, then there is a norm 1 quaternion c such that $\langle uc,v\rangle \in \mathbb{R}$.

$$\begin{aligned}&\int_{\mathbb{H}^n} \operatorname{sgn}\langle u,z\rangle \operatorname{sgn}\langle z,v\rangle G_n^{\mathbb{H}}(z)dm(z)\\ =&\int_{\mathbb{H}^n} c\operatorname{sgn}\langle uc,z\rangle \operatorname{sgn}\langle z,v\rangle G_n^{\mathbb{H}}(z)dm(z)\\ =&c\frac{3\langle cu,v\rangle}{2}\int_0^{\frac{\pi}{2}} \frac{\cos^4 t}{\sqrt{1-|\langle cu,v\rangle|^2\sin^2 t}}dt\\ =&\frac{3\langle u,v\rangle}{2}\int_0^{\frac{\pi}{2}} \frac{\cos^4 t}{\sqrt{1-|\langle u,v\rangle|^2\sin^2 t}}dt\end{aligned}$$

□

7.2.11 The function p(x)

Define a function over the open *quaternion* unit disk $D_{\mathbb{H}} = \{z \in \mathbb{H}, |z| < 1\}$ by

$$P(z) := \frac{3z}{2}\int_0^{\pi/2} \frac{\cos^4 t}{(1-|z|^2\sin^2 t)^{1/2}}\, dt, \qquad z \in D_{\mathbb{H}}, \tag{7.22}$$

and the function $p(x)$ as the restriction of P to $(-1,1) \subseteq \mathbb{R}$. Note that $p'(x) > 0$ on $(-1,1)$, $p(-1) = -1$ and $p(1) = 1$. Hence $p : [-1,1] \to [-1,1]$ is a strictly increasing continuous bijection. Since $[-1,1]$ is compact, p is a homeomorphism of $[-1,1]$ onto itself. By Taylor expansion, we get

$$(1 - x^2 \sin^2 t)^{-1/2} = \sum_{k=0}^{\infty} \frac{(2k-1)!!}{(2k)!!} x^{2k} \sin^{2k} t, \qquad |x| \le 1,\ 0 \le t < \pi/2,$$

and

$$\int_0^{\pi/2} \cos^4 t \sin^{2k} t \, dt = \frac{3\pi}{2} \cdot \frac{(2k-1)!!}{(2k+4)!!},$$

we get

$$p(x) = \sum_{k=0}^{\infty} \frac{9\pi}{16(k+1)(k+2)} \left[\frac{(2k-1)!!}{(2k)!!}\right]^2 x^{2k+1}, \qquad x \in [-1,1]. \tag{7.23}$$

Let $p_\ell(x) = x_2F_1(\frac{1}{2}, \frac{1}{2}; \ell; x^2)$ be the function introduced in the introduction. A straightforward calculation shows that $p(x) = \frac{9\pi}{32} p_3(x)$.

This calculation coincides with the formula [27, (3.2)], namely $\mathcal{E}_4(z) = p(x)$. More generally, we have $\mathcal{E}_{2d}(z) = C_{2d} p_{d+1}(x)$, where $C_{2d} = (1/d)\big(\Gamma((2d+1)/2)/\Gamma(d)\big)^2$. Thus, whenever the inverse function of $p_\ell(x)$ has first Taylor coefficient positive and all other nonpositive one can improve the value of the Grothendieck constants $K_{G,2(\ell-1)}^{\mathbf{R}}$ as in [77].

Compare above $p(x)$ with the real Haagerup function

$$h(x) = \sum_{k=0}^{\infty} \frac{\pi}{4(k+1)} \left[\frac{(2k-1)!!}{(2k)!!}\right]^2 x^{2k+1}, \qquad x \in [-1,1]. \tag{7.24}$$

Observe that $h(x) = \frac{\pi}{4} p_2(x)$. Note that

$$(x^3 p(x))' = \frac{3}{2} x^2 h(x).$$

As $\left[\frac{(2k-1)!!}{(2k)!!}\right] < 1$ It follows that

$$\sum_{k=0}^{\infty} \frac{9\pi}{16(k+1)(k+2)} \left[\frac{(2k-1)!!}{(2k)!!}\right] < \sum_{k=0}^{\infty} \frac{9\pi}{16(k+1)(k+2)} = \frac{9\pi}{16} \sum_{k=0}^{\infty} (\frac{1}{k+1} - \frac{1}{k+2}) = \frac{9\pi}{16}.$$

Therefore the power series for $p(x)$ (or $P(z)$) converge uniformly to a continuous function on the closed quaternion unit disk $\bar{D}_{\mathbb{H}}$. Note that $p(z)$ is analytic in the open complex unit disk $D = \{z \in \mathbb{C}, |z| < 1\}$. Use the ratio test for the coefficients to see that the radius of convergence of the series for $p(z)$ is $r = 1$. Since the Taylor coefficients of $p(z)$ are nonnegative, the Vivanti-Pringsheim theorem yields that $z = 1$ is a singular point. As $p(-z) = -p(z)$ it follows that $z = -1$ is also singular point. As $p'(0) > 0$ it follows that $p(z)$ has an inverse analytic function in some disc $D(r) = \{z \in \mathbb{C}, |z| < r\}$. So

$$p^{-1}(z) = \sum_{k=0}^{\infty} c_{2k+1} z^{2k+1}, \qquad z \in \mathbb{C}, |z| < r, 0 < r \leq 1. \tag{7.25}$$

The reason that $r \leq 1$ is because 1 is the singular point of $p(z)$. The coefficients c_{2k+1} are given by the Lagrange inversion formula:

$$c_{2k+1} = \frac{1}{(2k+1)!} \lim_{t \to 0} \left[\frac{d^{2k}}{dt^{2k}} \left(\frac{t}{p(t)}\right)^{2k+1}\right]. \tag{7.26}$$

Since $p'(x) > 0$ for $x \in (-1, 1)$ the function $p^{-1}(z)$ is an analytic function in some simply connected domain containing $(-1, 1)$.

Assume that $p(z) = w$. Then $z = p^{-1}(w)$. As $p_3(z) = \frac{32}{9\pi} w$ it follows that $z = p_3^{-1}(\frac{32}{9\pi} w) = p^{-1}(w)$. Hence the Taylor coefficients of $p^{-1}(w)$ and $p_3^{-1}(w)$ have the same signs.

7.2.12 Haagerup's method

In this subsection we will try to apply methods in [90] to show that $c_{2k+1} < 0$ for $k > 0$. We first show that as in [90, Lemma 2.2] that $p(z)$ can be extended to a continuous function $p^+(z)$ in the closed upper half plane $\mathbb{C}^+ = \{z \in \mathbb{C}, \Im(z) \geq 0\}$ which is analytic in the open upper half plane $\mathbb{C}_o^+ = \{z \in \mathbb{C}, \Im(z) > 0\}$. Recall that

$$p(x) = \frac{9}{2} \int_0^{\frac{\pi}{2}} \sin t \cos^2 t \arcsin(x \sin t) dt, \quad -1 \leq x \leq 1. \tag{7.27}$$

The analytic function $\sin z$ is a bijection of $[-\frac{\pi}{2}, \frac{\pi}{2}] \times [0, \infty)$ onto the closed upper half plane. Let $\arcsin^+ z : \mathbb{C}^+ \to [-\frac{\pi}{2}, \frac{\pi}{2}] \times [0, \infty)$. Note that

$$\begin{aligned}
&\arcsin^+ x = \arcsin x \text{ for } -1 \leq x \leq 1, \\
&\arcsin^+ x = \frac{\pi}{2} + \mathbf{i} \operatorname{arccosh} x \text{ for } x \geq 1, \\
&\arcsin^+ x = -\frac{\pi}{2} + \mathbf{i} \operatorname{arccosh}(-x) \text{ for } x \leq -1.
\end{aligned}$$

Furthermore $\arcsin^+ z$ is analytic in $\mathbb{C}_o^+$. Hence we can define

$$p^+(z) = \frac{9}{2} \int_0^{\frac{\pi}{2}} \sin t \cos^2 t \arcsin^+(z \sin t) dt. \tag{7.28}$$

Lemma 7.2.17. *The function $p(x)$ given by* (7.27) *has an analytic extension to $\mathbb{C}_o^+$ given*

(7.28). *Furthermore $p^+(z)$ is continuous on $\mathbb{C}^+$. Its value for $x > 1$ is given by the formulas:*

$$\begin{aligned}
\Re(p^+(x)) &= \frac{9}{2}\Big(\int_0^{\sin t = \frac{1}{x}} \sin t \cos^2 t \arcsin(x \sin t)\, dt + \int_{\sin t = \frac{1}{x}}^{\frac{\pi}{2}} \sin t \cos^2 t \frac{\pi}{2} dt\Big) \\
&= \frac{3}{2}\int_0^{\sin t = \frac{1}{x}} \frac{x \cos^4 t\, dt}{\sqrt{1 - x^2 \sin^2 t}}, \\
\Im(p^+(x)) &= \frac{9}{2}\int_{\sin t = \frac{1}{x}}^{\frac{\pi}{2}} \sin t \cos^2 t \arccos(x \sin t)\, dt \\
&= \frac{3}{2}\int_{\sin t = \frac{1}{x}}^{\frac{\pi}{2}} \frac{x \cos^4 t\, dt}{\sqrt{x^2 \sin^2 t - 1}}.
\end{aligned}$$

Similar equalities hold for $x < -1$. Furthermore

$$\begin{aligned}
\Re(p^+(x)) &= \frac{3}{2}\int_0^{\frac{\pi}{2}} (1 - x^{-2}\sin^2 u)^{\frac{3}{2}}\, du,\ \Im(p^+(x)) \\
&= \frac{3}{2}(1 - x^{-2})^2 \int_0^{\frac{\pi}{2}} \frac{\sin^4 v}{\sqrt{1 - (1 - x^{-2})\sin^2 v}}\, dv \qquad (7.29)
\end{aligned}$$

Proof. The contents of the proof are exactly the same as those of Haagerup with the following modification. First recall that $(\cos^3 t)' = -3 \sin t \cos^2 t$. Hence

$$\begin{aligned}
&3\int_0^{\sin t = \frac{1}{x}} \sin t \cos^2 t \arcsin(x \sin t) dt \\
&= -\cos^3 t \arcsin(x \sin t)|_{t=0}^{\sin t = \frac{1}{x}} + \int_0^{\sin t = \frac{1}{x}} \frac{x \cos^4 t\, dt}{\sqrt{1 - x^2 \sin^2 t}},
\end{aligned}$$

and

$$3\int_{\sin t = \frac{1}{x}}^{\frac{\pi}{2}} \sin t \cos^2 t \frac{\pi}{2} dt = -\frac{\pi}{2}\cos^3 t|_{\sin t = \frac{1}{x}}^{\frac{\pi}{2}}.$$

These equalities show the first identity for $\Re(p^+(x))$. Use the same integration by part for the first identity for $\Im(p^+(x))$.

For identities (7.29) use the same substitutions as in [90, page 205]. For the expression $\Re(p^+(x))$ use $\sin u = x \sin t$. Since in the integrant we have $\cos^4 t = \cos^2 t \cos^2 t$ we need to multiply the integrant of Haagerup by $\cos^2 t = 1 - \sin^2 t = 1 - x^{-2}\sin^2 u$, which gives the factor $(1 - x^{-2}\sin^2 u)^{\frac{3}{2}}$.

For the expression $\Im(p^+(x))$ use the substitution $\sin v = \frac{\cos t}{\sqrt{1-x^{-2}}}$. Again $\cos^2 t = (1 - x^{-2})\sin^2 v$. □

Lemma 7.2.18. *We have the following series expansions for $x \geq 1$:*

$$\psi_1(x) = \Re(p^+(x)) = \frac{9\pi}{16}\Big(\frac{4}{3} - x^{-2} + \sum_{k=2}^{\infty} \frac{(2k-5)!!(2k-1)!!}{2^{2k-2}(k!)^2} x^{-2k}\Big), \tag{7.30}$$

$$\psi_2(x) = \Im(p^+(x)) = \frac{3\pi}{16}\sum_{k=0}^{\infty} \frac{(2k-1)!!(2k+3)!!}{2^{2k}k!(k+2)!}(1 - x^{-2})^{k+2}. \tag{7.31}$$

Furthermore the functions ψ_1 and ψ_2 strictly increase for $x \geq 1$.

Proof. First, we use the following Taylor expansions:

$$(1-t)^{\frac{3}{2}} = 1 - \frac{3}{2}t + 3\sum_{k=2}^{\infty} \frac{(2k-5)!!}{2^k k!} t^k, \quad (1-t)^{-\frac{1}{2}} = \sum_{k=0}^{\infty} \frac{(2k-1)!!}{2^k k!} t^k.$$

Second, we use the formula $\int_0^{\frac{\pi}{2}} \sin^{2n} u du = \frac{(2n-1)!!}{2^{n+1}n!}\pi$, where $(-1)!! = 1$.

To show that $\psi_1(x)$ strictly increases observe that $1 - x^{-2}\sin^2 u$ is strictly increasing for $x \geq 1$. To show that $\psi_2(x)$ striclty increases for $x \geq 1$ observe that the functions $(1 - x^{-2})$ and $\frac{1}{1-(1-x^{-2})\sin^2 u}$ strictly increase for $x \geq 1$. □

Use (7.28) and the arguments of [90][Lemmas 2.3 and 2.4] to deduce:

Lemma 7.2.19. *1. $\Im(p^+(z)) \geq \Im(p^+(|z|))$ for $|z| \geq 1$, $\Im(z) \geq 0$.*

2. $p^+(z)$ has no zero in $\mathbb{C}^+$ except $z = 0$.

Lemma 7.2.20. *Assume that the Taylor series of $p^{-1}(x)$ are given by (7.25). Let $\alpha > 1$. Then for any odd positive integer n we have*

$$c_n = \frac{2}{\pi n}\int_1^{\alpha} \Im(p^+(x)^{-n})\, dx + r_n(\alpha), \text{ where } |r_n(\alpha)| \leq \frac{\alpha}{n}(\Im p^+(\alpha))^{-n}.$$

We now imitate the steps in the proof of Haagerup for nonpositivity of Taylor series of $h(z)$ for $k \geq 2$. We will do that without relying on the complete elliptic integrals, only using Lemmas 7.2.17 and 7.2.18. We first start with the following Lemma:

Lemma 7.2.21. *The ratio* $\frac{\psi_2(x)}{\psi_1(x)}$ *strictly increases for* $x \geq 1$.

Proof. Let

$$\frac{\psi_2(x)}{\psi_1(x)} = \frac{(1-x^{-2})^2 \int_0^{\frac{\pi}{2}} \frac{\sin^4 v}{\sqrt{1-(1-x^{-2})\sin^2 v}}}{\int_0^{\frac{\pi}{2}} (1-x^{-2}\sin^2 u)^{\frac{3}{2}}\, du} = \frac{(1-x^{-2})^{\frac{1}{2}} \int_0^{\frac{\pi}{2}} \frac{\sin^4 v}{\sqrt{1-(1-x^{-2})\sin^2 v}}}{\int_0^{\frac{\pi}{2}} [(1-x^{-2})^{-1}(1-x^{-2}\sin^2 u)]^{\frac{3}{2}}\, du}$$

The proof of Lemma 7.2.18 yields that the numerator of the last expression strictly increases for $x \geq 1$. Thus it is enough to show that the denominator of the last expression strictly decreases for $x \geq 1$. This would follow from the claim that

$$\frac{1-(1-x^{-2})\sin^2 u}{1-x^{-2}} = \cos^2 u + \frac{1}{x^2-1}$$

is strictly decreasing. This is obvious from the last expression. □

Clearly

$$\psi_1(1) = 1,\ \psi_1(\infty) = \frac{3\pi}{4}, \quad \psi_2(1) = 0,\ \psi_2(\infty) = \infty.$$

Corollary 7.2.22. *The complex function* p^+ *has the following expression for* $x \geq 1$

$$\theta(x) = \arctan\frac{\psi_2(x)}{\psi_1(x)},\ p^+(x) = \psi_1(x) + \mathrm{i}\psi_2(x) = |p^+(x)|\, e^{\mathrm{i}\theta(x)} = \sqrt{\psi_1^2(x)+\psi_2^2(x)}\, e^{\mathrm{i}\theta(x)}.$$

The function $\theta(x)$ *strictly increases for* $x \geq 1$, *where* $\theta(1) = 0$ *and* $\theta(\infty) = \frac{\pi}{2}$.

The above corollary is the analog of [90][Lemma 2.7].

Next we need an analog of Lemma 2.8. We first start with the following lemma .

Lemma 7.2.23. *Let $\chi(x) : [0, \frac{\pi}{2}) \to [1, \infty)$ be the inverse function of $\theta(x)$. Then the substitution $x = \chi(y)$ yields:*

$$\frac{d\chi}{dy} = \frac{\psi_1^2(\chi(y)) + \psi_2^2(\chi(y))}{\psi_2'(\chi(y))\psi_1(\chi(y)) - \psi_1'(\chi(y))\psi_2(\chi(y))},$$
$$|p^+(x)|^{-n}dx = |p^+(\chi(y))|^{-n}\frac{\psi_1^2(\chi(y)) + \psi_2^2(\chi(y))}{\psi_2'(\chi(y))\psi_1(\chi(y)) - \psi_1'(\chi(y))\psi_2(\chi(y))}dy =$$
$$\big(|p^+|^{n-2}(\chi(y))(\psi_2'(\chi(y))\psi_1(\chi(y)) - \psi_1'(\chi(y))\psi_2(\chi(y)))\big)^{-1}.$$

Proof. As $y = \theta(x)$ is strictly increasing on $[1, \infty)$ it follows that $x = \chi(y)$ is strictly increasing on $[0, \frac{\pi}{2})$. Clearly

$$\frac{dy}{dx} = (\arctan\frac{\psi_2(x)}{\psi_1(x)})' = \frac{\psi_2'(x)\psi_1(x) - \psi_1'(x)\psi_2(x)}{\psi_1^2(x) + \psi_2^2(x)}.$$

This equality implies straightforward the lemma. □

The following proposition follows from (7.29), the numerical calculations 7.3.3 and the last part of Lemma 7.2.18.

Proposition 7.2.24. *Let $\omega(x) = \psi_2'(x)\psi_1(x) - \psi_1'(x)\psi_2(x)$. Then $\omega(x)$ strictly increases on $[0, \tau]$, where*

$$\tau \approx 1.732,\ \omega(\tau) \approx 1.360,\ \omega(1) = \omega'(\tau) = 0.$$

Assume that $\omega(x)|p^+(x)|^m$ is strictly increasing on $[1, \alpha]$ for some $m > 0$. Then for each integer $k \geq 0$ the function $\omega(x)|p^+(x)|^{m+k}$ strictly increases on the interval $[1, \alpha]$.

Let us choose $\alpha = 5$ and

$$\theta_0 = \theta(5) = \arctan\frac{\psi_2(5)}{\psi_1(5)} \approx 0.8097.$$

By Proposition 7.3.4, $\omega(x)|p^+(x)|^7$ increases in $[1, 5]$. We now give an analog of [90][Lemma 2.8].

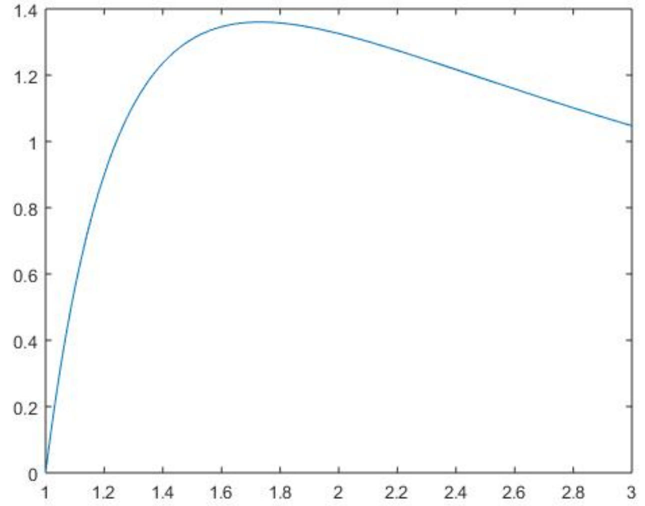


Figure 7.1: Graph of function $\omega(x)$.

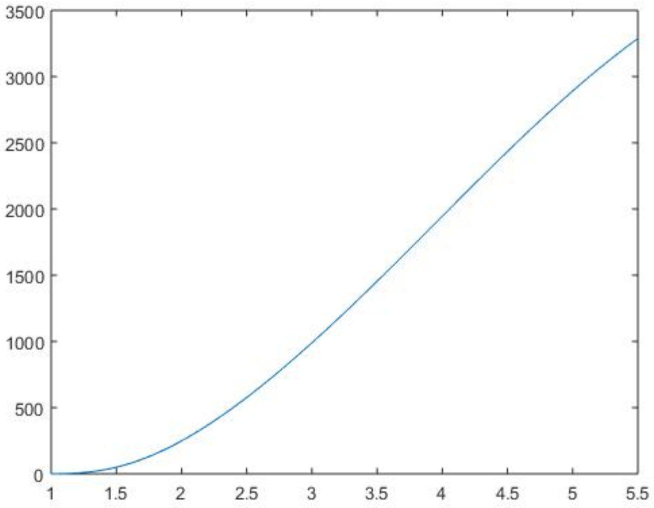


Figure 7.2: Graph of function $\omega(x)|p^+(x)|^7$.

Lemma 7.2.25. *Let $\alpha = 5$ and $\theta_0 = \theta(\alpha)$. For a fixed $n \in \mathbb{N}$ let $p = \lfloor \frac{n\theta_0}{\pi} \rfloor$. Set*

$$I_r = \frac{2}{\pi n} \int_{\theta(x)=\frac{\pi}{n}(r-1)}^{\theta(x)=\frac{\pi}{n}r} |p^+(x)|^{-n} |\sin n\theta(x)| dx$$

for $r = 1, \ldots, p$. Put

$$I' = \frac{2}{\pi n} \int_{\theta(x)=\frac{\pi}{n}p}^{\alpha} |p^+(x)|^{-n} |\sin n\theta(x)| dx$$

Then:

1.

$$\frac{2}{\pi n} \int_1^{\alpha} \Im(p^+(x)^{-n}) dx = -I_1 + I_2 - \cdots + (-1)^p I_p + (-1)^{p+1} I'.$$

2. For $n \geq 9$ one has $p \geq 2$ and $I_1 > I_2 > \cdots > I_p > I'$.

Proof. (*1*) Observe that $\Im(p^+(x)^{-n}) = |p^+(x)|^{-n} \sin(-n\theta(x)) = -|p^+(x)|^{-n} \sin(n\theta(x))$. Hence

$$\begin{aligned} \int_1^{\alpha} \Im(p^+(x)^{-n}) dx &= -\int_1^{\alpha} |p^+(x)|^{-n} \sin(n\theta(x)) dx \\ &= \frac{\pi n}{2}(-I_1 + I_2 - \cdots + (-1)^p I_p + (-1)^{p+1} I'). \end{aligned}$$

(*2*) Let $x = \chi(y)$ for $y \in [0, \theta_0]$. Use Lemma 7.2.23 and the definition of $\omega(x)$ in Proposition (7.2.24) to deduce

$$\begin{aligned} I_r &= \frac{2}{\pi n} \int_{\frac{\pi}{n}(r-1)}^{\frac{\pi}{n}r} \big(\omega(\chi(y)) |p^+(\chi(y))|^{n-2}\big)^{-1} |\sin ny| dy, \qquad (7.32) \\ I' &= \frac{2}{\pi n} \int_{\frac{\pi}{n}p}^{\theta_0} \big(\omega(\chi(y)) |p^+(\chi(y))|^{n-2}\big)^{-1} |\sin ny| dy. \end{aligned}$$

Recall that the function $\omega(x)|p^+(x)|^7$ strictly increases on $[1, 5]$. Use Proposition (7.2.24) to deduce that $\omega(x)|p^+(x)|^{n-2}$ is strictly increasing on $[1, \alpha]$ for $n \geq 9$.

Since $|\sin(ny)|$ is periodic with period π/n, it follows that

$$I_1 > I_2 > \cdots > I_p.$$

Additionally,

$$\begin{aligned} I' &= \frac{2}{\pi n}\int_{\frac{\pi}{n}p}^{\theta_0} \big(\omega(\chi(y))|p^+(\chi(y))|^{n-2}\big)^{-1}|\sin ny|\,dy \\ &\le \frac{2}{\pi n}\int_{\frac{\pi}{n}(p-1)}^{\theta_0-\pi/n} \big(\omega(\chi(y))|p^+(\chi(y))|^{n-2}\big)^{-1}|\sin ny|\,dy \\ &< I_p. \end{aligned}$$

□

We now give the analog of q in [90]:

Lemma 7.2.26. *Let*

$$\mu(x) = \frac{\psi_2'(x)\psi_2(x) + \psi_1'(x)\psi_1(x)}{\psi_2'(x)\psi_1(x) - \psi_1'(x)\psi_2(x)}, \qquad x \in (1,2].$$

Then $\mu(x)$ strictly decreases on the interval $(1, 1.732]$. Furthermore

$$(\log|p^+(\chi(y))|)' = \mu(\chi(y)), \qquad y \in (0, \frac{\pi}{2}).$$

Proof. The claim that $\mu(x)$ strictly decreases on $(1, 1.732]$ follows from the Proposition 7.3.5. Clearly

$$\log|p^+(x)| = \frac{1}{2}\log|p^+(x)|^2 = \frac{1}{2}\log(\psi_2^2(x) + \psi_1^2(x)).$$

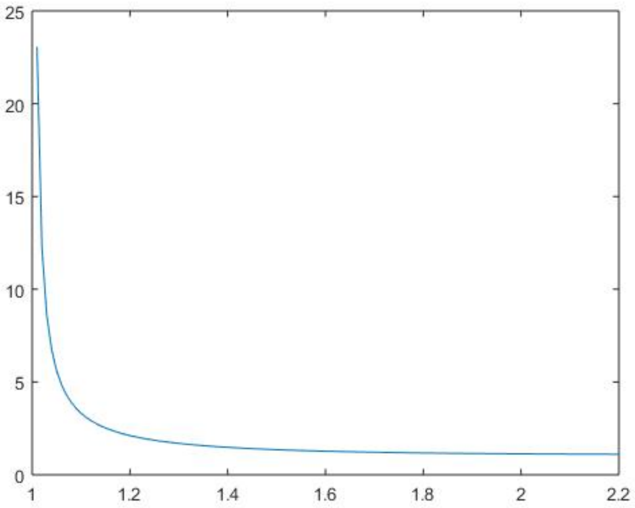


Figure 7.3: Graph of function $\mu(x)$.

Hence

$$(\log|p^+(\chi(y))|)' = \frac{\psi_2'(\chi(y))\psi_2(\chi(y)) + \psi_1'(\chi(y))\psi_1(\chi(y))}{\psi_2^2(\chi(y)) + \psi_1^2(\chi(y))}\chi'(y) = \mu(\chi(y)).$$

□

We now give the analog of [90][Lemma 2.9]:

Lemma 7.2.27. *Let $n \geq 21$. $I_1, I_2, \ldots$ are defined as in Lemma 7.2.25 and $q = \mu(\tau) \approx$ 1.2020. Put*

$$c = |p^+(\tau)|e^{-q\theta(\tau)} \approx 1.2923.$$

Then

1. $I_1 > \frac{0.326}{n^2}c^{-n}$,

2. $I_2 < 0.033 I_1$.

Proof. Recall that $(\log|p^+(\chi(y))|)' = \mu(\chi(y))$ and $\mu(x)$ is strictly decreasing on $(1, 1.732]$.

Hence $\mu(\chi(y))$ strictly decreasing on $(0, \theta(1.732)]$. In particular, for $y \in (0, \theta(\tau)]$

$$\mu(\chi(y)) \geq \mu(\chi(\theta(\tau))) = \mu(\tau) = q \approx 1.2020.$$

Therefore

$$\log|p^+(\chi(u))| - \log|p^+(\chi(y))| = \int_y^u \mu(\chi(t))dt \geq q(u-y), \text{ for } 0 \leq y \leq u \leq \theta(\tau). \quad (7.33)$$

Choose $u = \theta(\tau)$ to obtain

$$|p^+(\chi(y))| \leq ce^{qy} \text{ for } y \in [0, \theta(\tau)]. \quad (7.34)$$

(*1*) We first use (7.32) for $r = 1$:

$$I_1 = \frac{2}{\pi n} \int_0^{\frac{\pi}{n}} \left(\omega(\chi(y))|p^+(\chi(y))|^{n-2}\right)^{-1} |\sin ny| dy.$$

Since $n \geq 21$ it follows that

$$\frac{\pi}{n} \leq \frac{\pi}{21} \approx 0.1496 < \theta(\tau) \approx 0.3224.$$

As $\omega(x)$ is increasing on the interval $[1, \tau]$ it follows that $(\omega(\chi(y)))^{-1} > \omega^{-1}(\tau)$ on $(0, \frac{\pi}{n}]$. Hence

$$I_1 > \frac{2}{\omega(\tau)\pi n} \int_0^{\frac{\pi}{n}} |p^+(\chi(y))|^{2-n} \sin ny \, dy.$$

Apply inequality (7.34) to deduce that

$$\begin{aligned} I_1 &> \frac{2}{\omega(\tau)\pi n}\int_0^{\frac{\pi}{n}} (ce^{qy})^{2-n}\sin ny\,dy \\ &= \frac{2c^{2-n}}{\omega(\tau)\pi n^2}\int_0^{\pi} e^{-\frac{(n-2)qy}{n}}\sin y\,dy \\ &\geq \frac{2c^{2-n}}{\omega(\tau)\pi n^2}\int_0^{\pi} e^{-qy}\sin y\,dy \\ &= \frac{2c^2}{\omega(\tau)\pi n^2}\frac{1+e^{-q\pi}}{1+q^2}c^{-n}. \end{aligned}$$

Since

$$\frac{2c^2}{\omega(\tau)\pi}\frac{1+e^{-q\pi}}{1+q^2} \approx 0.32697,$$

this completes the proof of (1).

(*2*) We now use (7.32) for $r = 2$:

$$I_2 = \frac{2}{\pi n}\int_{\frac{\pi}{n}}^{\frac{2\pi}{n}} \left(\omega(\chi(y))|p^+(\chi(y))|^{n-2}\right)^{-1}|\sin ny|dy.$$

We now make a substitution $y = t + \frac{\pi}{n}$ in the integral formula for I_2:

$$I_2 = \frac{2}{\pi n}\int_0^{\frac{\pi}{n}} \left(\omega(\chi(t+\frac{\pi}{n}))|p^+(\chi(t+\frac{\pi}{n}))|^{n-2}\right)^{-1}\sin nt\,dt.$$

Let

$$\hat{I}_2 = \frac{2}{\pi n}\int_0^{\frac{\pi}{n}} \left(\omega(\chi(t))|p^+(\chi(t+\frac{\pi}{n}))|^{n-2}\right)^{-1}\sin nt\,dt.$$

Since $n \geq 21$ it follows that

$$\frac{2\pi}{n} \leq \frac{2\pi}{21} \approx 0.2992 < \theta(\tau) \approx 0.3224.$$

As $\omega(x)$ is strictly increasing on the interval $[1,\tau]$, we know that $(\omega(\chi(y)))^{-1}$ strictly decreases on $(0, \frac{2\pi}{n}]$. Then one gets $I_2 < \hat{I}_2$. Next use the inequality (7.33) for $y = t$ and $u = t + \frac{\pi}{n}$ which yields

$$e^{-\frac{(n-2)q\pi}{n}} |p^+(\chi(t))|^{-(n-2)} \geq |p^+(\chi(t + \frac{\pi}{n}))|^{-(n-2)}, \qquad t \in [0, \frac{\pi}{n}].$$

Hence for $n \geq 21$

$$\hat{I}_2 \leq e^{-\frac{(n-2)q\pi}{n}} I_1 \leq e^{-\frac{19q\pi}{21}} I_1 < 0.033 I_1.$$

□

Theorem 7.2.28. *$c_1 = \frac{32}{9\pi}$ and $c_{2k+1} < 0$ for $k \geq 1$.*

Proof. Let $k \geq 10$. Applying Lemma 7.2.20 with $\alpha = 5$ and Lemma 7.2.25, we have

$$\begin{aligned} -c_{2k+1} &= I_1 - I_2 + \cdots + (-1)^{p-1} I_p + (-1)^p I' - r_{2k+1}(5) \\ &> I_1 - I_2 - r_{2k+1}(5). \end{aligned}$$

Using Lemma 7.2.27, we get

$$(I_1 - I_2) > \frac{0.315}{(2k+1)^2} 1.293^{-(2k+1)} \qquad |r_{2k+1}(5)| \leq \frac{5}{2k+1} 2.4^{-(2k+1)}.$$

Applying the aforementioned formulas to $-c_{2k+1} > I_1 - I_2 - r_{2k+1}(5)$, it follows that for $k \geq 10$

$$\begin{aligned} -c_{2k+1} &> (I_1 - I_2) - r_{2k+1}(5) \\ &> \frac{0.315}{(2k+1)^2} 1.293^{-(2k+1)} - \frac{5}{2k+1} 2.4^{-(2k+1)} \\ &= \frac{0.315}{(2k+1)^2} 1.293^{-(2k+1)} \left[1 - \frac{5(2k+1)}{0.315} \left(\frac{1.293}{2.4} \right)^{2k+1} \right] \\ &> 0. \end{aligned}$$

By directly using (7.26), we can obtain following approximations (rounded to two decimal places).

n	1	3	5	7	9
$-c_n$	$-32/9\pi$	0.12	$4.84 \cdot 10^{-3}$	$2.58 \cdot 10^{-3}$	$1.22 \cdot 10^{-3}$

n	11	13	15	17	19
$-c_n$	$6.76 \cdot 10^{-4}$	$4.15 \cdot 10^{-4}$	$2.74 \cdot 10^{-4}$	$1.91 \cdot 10^{-4}$	$1.39 \cdot 10^{-4}$

□

Since $c_{2k+1} \leq 0$ for $k \geq 1$, $h_1(z) = c_1 z - h^{-1}(z)$ has nonnnegative Taylor coefficients. The Vivanti-Pringsheim theorem states that if the radius of convergence of Taylor series of $h_1(z)$ is r then $h_1(z)$, and hence $h^{-1}(z)$, has a singular point at r. As $h'(t) > 0$ on $(0,1)$, and $h(1) = 1$, it follows that $r \geq 1$. Clearly $h^{-1}(t) \leq \sum_{k=0}^{N} c_{2k+1} t^{2k+1}$ for $t \in (0,1)$. That is $\sum_{k=1}^{N} |c_{2k+1}| t^{2k+1} \leq c_1 t - h^{-1}(t)$ for $t \in (0,1)$. In particular $\sum_{k=1}^{N} |c_{2k+1}| \leq c_1 - 1$. Thus $\sum_{k=0}^{\infty} |c_{2k+1}| \leq 2c_1 - 1$.

As $c_1 \neq 0$, clearly, the function

$$\psi(x) \coloneqq \sum_{k=0}^{\infty} |c_{2k+1}| x^{2k+1} \tag{7.35}$$

is a strictly increasing and continuous on $[0, r)$. Recall that if the series (7.25) converge for $x_0 \in \mathbb{R}$ (pointwise) then $r \geq |x_0|$. Hence $\psi(x) = +\infty$ for $x > r$. $\psi(1) = \sum_{k=0}^{\infty} |c_{2k+1}| \geq c_1 = 32/(9\pi) > 1$. Thus there exists a unique $c_0 \in (0,1)$ such that $\psi(c_0) = 1$.

Proposition 7.2.29. *The following equality holds*

$$\sum_{k=1}^{\infty} |c_{2k+1}| = -\sum_{k=1}^{\infty} c_{2k+1} = c_1 - p^{-1}(1) = \frac{32}{9\pi} - 1.$$

Hence the equation $\psi(x) = 1$ has a unique solution $c_0 \in (0, \frac{9\pi}{32})$ given by the equation $c_0 = p(2c_1c_0 - 1)$. Equivalently, let x_0 be the unique solution of

$$p(x_0) = \frac{9\pi(1 + x_0)}{64}. \tag{7.36}$$

Then

$$c_0 = \frac{9\pi(1 + x_0)}{64}. \tag{7.37}$$

Proof. Observe that $\psi(x) = 2c_1x - p^{-1}(x)$. Hence $\psi(c_0) = 1$ is equivalent to $p^{-1}(c_0) = 2c_1c_0 - 1$, which implies that $c_0 = p(2c_1c_0 - 1)$. Set $x_0 = 2c_1c_0 - 1$ and use $c_0 = \frac{9\pi(1+x_0)}{64}$ to deduce the Proposition. □

7.2.13 An upper bound on $K_G^{\mathbb{H}}$

Theorem 7.2.30. *Let x_0 be the unique solution of* (7.36)*. Then*

$$K_G^{\mathbb{H}} \le \frac{64}{9\pi(x_0 + 1)} \approx 1.2168. \tag{7.38}$$

The same value of an upper bound on the constant $K_{G,4}^{\mathbb{R}}$ was calculated in [27, Table 1, p. 81]. The authors explain that the computation results in Table 1 are “just numerical, and do not yield a formal proof”.

Proof. Recall previous definition of $P(z)$ and $\psi(x)$

$$P^{-1}(z) = \arg z \, p^{-1}(|z|) = \sum_{k=0}^{\infty} c_{2k+1} z |z|^{2k}.$$

$$\psi(c_0) = \sum_{k=0}^{\infty} |c_{2k+1}| c_0^{2k+1} = 1.$$

Recall Lemma (7.2.9). Let $g(z) = \sum_{k=0}^{\infty} c_{2k+1} c_0^{2k+1} z^{2k}$. Then $p^{-1}(c_0|z|) = |z| g(|z|)$. Given unit vectors $x_1, \ldots, x_m, y_1, \ldots y_n$ in a quaternionic Hilbert space $\mathcal{H}$, then there exist unit vectors $u_1, \ldots, u_m, v_1, \ldots, v_n$ in a quaternionic Hilbert space $\mathcal{H}'$ such that $P^{-1}(c_0\langle x_i, y_j\rangle) = \langle u_i, v_j\rangle$. Let $\mathcal{H}_1$ be an l-dimensional subspace of $\mathcal{H}'$ spanned by $u_1, \ldots, u_m, v_1, \ldots, v_n$. Thus we can assume that $\mathcal{H}_1 = \mathbb{H}^l$, where $l \leq m+n$. Assume that

$$\max_{|\varepsilon_i|=|\delta_j|=1}\left|\sum_{i=1}^{m}\sum_{j=1}^{n} M_{ij}\bar{\varepsilon}_i\delta_j\right| \leq 1.$$

Then

$$\left|\sum_{i=1}^{m}\sum_{j=1}^{n} M_{ij}\operatorname{sgn}\langle u, z\rangle \operatorname{sgn}\langle z, v_j\rangle G_l^{\mathbb{H}}(z)\right| \leq G_l^{\mathbb{H}}(z), \quad z \in \mathbb{H}^l.$$

Integrate over the Lebesgue measure of $\mathbb{H}^l$ to get $\left|\sum_{i=1}^{m}\sum_{j=1}^{n} M_{ij}P(\langle u_i, v_j\rangle)\right| \leq 1$. Now

$$1 \geq \left|\sum_{i=1}^{m}\sum_{j=1}^{n} M_{ij}P(\langle u_i, v_j\rangle)\right| = \left|\sum_{i=1}^{m}\sum_{j=1}^{n} M_{ij}c_0\langle x_i, y_j\rangle\right|.$$

Take maximum on unit vectors $x_1, \ldots, x_m, y_1, \ldots, y_n$ to deduce that $\|M\|_G \leq 1/c_0$. □

7.2.14 The function $p(x)$ (7.23) and the constant $K_G^{\mathbb{H}}$

We claim that the function $p(x)$ (7.23) is the function $\varphi_4^{\mathbb{R}}$ [77, (40)]. The proof is very similar to part (ii) of Lemma 3.1 in [77]. Hence we have inequalities: $K_{G,4}^{\mathbb{R}} \leq K_{G,2}^{\mathbb{C}} \leq K_G^{\mathbb{H}}$.

7.3 Increasing properties of three functions

Recall the definiton of our functions:

$$\psi_1(x) = \Re(p^+(x)) = \frac{3}{2}\int_0^{\frac{\pi}{2}} (1 - x^{-2}\sin^2 u)^{\frac{3}{2}}\, du, \tag{7.39}$$

$$\psi_2(x) = \Im(p^+(x)) = \frac{3}{2}(1 - x^{-2})^2 \int_0^{\frac{\pi}{2}} \frac{\sin^4 v}{\sqrt{1-(1-x^{-2})\sin^2 v}}\, dv, \tag{7.40}$$

$$\omega(x) = \psi_2'(x)\psi_1(x) - \psi_2(x)\psi_1'(x), \tag{7.41}$$

$$\psi_1(1) = 1, \quad \psi_1(\infty) = \frac{3\pi}{4}, \quad \psi_2(1) = 0, \quad \psi_2(\infty) = \infty \tag{7.42}$$

It is easy to show from the definitions that $\psi_1(x)$ and $\psi_2(x)$ are increasing functions on $[1, \infty)$.We are interested in some properties of related functions on the interval $[1, 5]$. For the function $\psi_2(x)$, since it is a function of $y = \sqrt{1 - x^{-2}}$ for $x \in [1, \infty)$, $y \in [0, \sqrt{24}/5]$, we can differentiate this function as many times as needed on the interval $[1, 5]$. However, for the function $\psi_1(x)$, one can show that the second derivative does not exists at 1, i.e., it value at 1 is ∞.

7.3.1 Basic properties

Set

$$\phi_1 = \frac{3}{4}\left(\frac{4}{3} - x^{-2} + \sum_{k=2}^{\infty} \frac{(2k-5)!!(2k-1)!!}{2^{2k-2}(k!)^2} x^{-2k}\right), \tag{7.43}$$

$$\phi_2 = \sum_{k=0}^{\infty} \frac{(2k-1)!!(2k+3)!!}{2^{2k}k!(k+2)!}(1 - x^{-2})^{k+2}. \tag{7.44}$$

Thus

$$\psi_1 = \frac{3\pi}{4}\phi_1, \; \psi_2(x) = \frac{3\pi}{16}\phi_2, \; \omega = \frac{9\pi^2}{64}\tilde{\omega}, \; \tilde{\omega} = \phi_2'\phi_1 - \phi_2\phi_1'.$$

Furthermore

$$\phi_1' = \frac{3}{2}\big(x^{-3} - \sum_{k=2}^{\infty} \frac{(2k-5)!!(2k-1)!!}{2^{2k-2}(k!)(k-1)!} x^{-2k-1}\big), \tag{7.45}$$

$$\phi_1'' = \frac{3}{2}\big(-3x^{-4} + \sum_{k=2}^{\infty} \frac{(2k-5)!!(2k+1)!!}{2^{2k-2}(k!)(k-1)!} x^{-2k-2}\big), \tag{7.46}$$

$$\phi_2' = 2x^{-3} \sum_{k=2}^{\infty} \frac{(2k-1)!!(2k+3)!!}{2^{2k}k!(k+1)!} (1-x^{-2})^{k+1}, \tag{7.47}$$

$$\phi_2'' = 2x^{-6} \sum_{k=2}^{\infty} \frac{(2k-1)!!(2k+3)!!}{2^{2k}k!(k+1)!} (1-x^{-2})^{k}(2k+5-3x^2). \tag{7.48}$$

We introduce following finite series, where m will be specified later. All those functions are rational functions.

$$\phi_{1,m}(x) = \frac{3}{4}\big(\frac{4}{3} - x^{-2} + \sum_{k=2}^{m} \frac{(2k-5)!!(2k-1)!!}{2^{2k-2}(k!)^2} x^{-2k}\big)$$

$$
\begin{aligned}
\hat{\phi}_{1,m}(x) &= \frac{3}{4}\big(\frac{4}{3} - x^{-2} + (\sum_{k=2}^{m-1} \frac{(2k-5)!!(2k-1)!!}{2^{2k-2}(k!)^2} x^{-2k}) \\
&\quad + \frac{(2m-5)!!(2m-1)!!}{2^{2m-2}(m!)^2} x^{-2(m-1)} \frac{1}{x^2-1}\big), \\
\phi'_{1,m}(x) &= \frac{3}{2}\big(x^{-6} - \sum_{k=2}^{m} \frac{(2k-5)!!(2k-1)!!}{2^{2k-2}k!(k-1)!} x^{-2k-1}\big), \\
\bar{\phi}_{1,m}(x) &= \frac{3}{2}\big(x^{-6} - (\sum_{k=2}^{m-1} \frac{(2k-5)!!(2k-1)!!}{2^{2k-2}k!(k-1)!} x^{-2k-1}) \\
&\quad - \frac{(2m-5)!!(2m-1)!!}{2^{2m-2}m!(m-1)!} x^{-2m+1} \frac{1}{x^2-1}\big), \\
\phi''_{1,m}(x) &= \frac{3}{2}\big(- 3x^{-4} + \sum_{k=2}^{m} \frac{(2k-5)!!(2k+1)!!}{2^{2k-2}k!(k-1)!} x^{-2k-2}\big), \\
\tilde{\phi}_{1,m}(x) &= \frac{3}{2}\big(- 3x^{-4} + \sum_{k=2}^{m-1} \frac{(2k-5)!!(2k+1)!!}{2^{2k-2}k!(k-1)!} x^{-2k-2} \\
&\quad + \frac{(2m-5)!!(2m+1)!!}{2^{2m-2}m!(m-1)!} x^{-2m} \frac{1}{x^2-1}\big). \\
\phi_{2,m}(x) &= \sum_{k=0}^{m} \frac{(2k-1)!!(2k+3)!!}{2^{2k}k!(k+2)!} (1-x^{-2})^{k+2}, \\
\hat{\phi}_{2,m}(x) &= \sum_{k=0}^{m-1} \frac{(2k-1)!!(2k+3)!!}{2^{2k}k!(k+2)!} (1-x^{-2})^{k+2} \\
&\quad + \frac{(2m-1)!!(2m+3)!!}{2^{2m}m!(m+2)!} (1-x^{-2})^{m+2} x^2, \\
\phi'_{2,m}(x) &= 2x^{-3} \sum_{k=0}^{m} \frac{(2k-1)!!(2k+3)!!}{2^{2k}k!(k+1)!} (1-x^{-2})^{k+1}, \\
\bar{\phi}_{2,m}(x) &= 2x^{-3} \sum_{k=0}^{m-1} \frac{(2k-1)!!(2k+3)!!}{2^{2k}k!(k+1)!} (1-x^{-2})^{k+1} \\
&\quad + \frac{(2m-1)!!(2m+3)!!}{2^{2m}m!(m+1)!} (1-x^{-2})^{m+1} x^2, \\
\phi''_{2,m}(x) &= 2x^{-6} \sum_{k=0}^{m} \frac{(2k-1)!!(2k+3)!!}{2^{2k}k!(k+1)!} (1-x^{-2})^{k} \big(2k+5-3x^2\big).
\end{aligned}
$$

We claim that we have the following inequalities:

Proposition 7.3.1. *For $x > 1$,*

$$\begin{aligned}
&0 < \phi_{1,m}(x) < \phi_1(x) < \hat{\phi}_{1,m}(x),\\
&\bar{\phi}_{1,m}(x) < \phi_1'(x) < \phi_{1,m}'(x), 0 < \phi_1'(x),\\
&\phi_{1,m}''(x) < \phi_1''(x) < \tilde{\phi}_{1,m}(x),\\
&0 < \phi_{2,m}(x) < \phi_2(x) < \hat{\phi}_{2,m}(x),\\
&0 < \phi_{2,m}'(x) < \phi_2'(x) < \bar{\phi}_{2,m}(x).
\end{aligned}$$

When $m \geq 35, 1 \leq x \leq 5$,

$$\phi_{2,m}''(x) < \phi_2''(x)$$

Proof. All the above inequalities are clear except the five inequalities

$$\begin{aligned}
&\phi_1(x) < \hat{\phi}_{1,m}(x),\ \bar{\phi}_{1,m}(x) < \phi_1'(x),\ \phi_1''(x) < \tilde{\phi}_{1,m}(x),\\
&\phi_2(x) < \hat{\phi}_{2,m}(x), \phi_2'(x) < \bar{\phi}_{2,m}(x).
\end{aligned}$$

To show the inequality $\phi_1(x) < \hat{\phi}_{1,m}(x)$, we argue as follows. First observe that coefficients of $\phi_1(x)$ is strictly decreasing for $k \geq 2$. This follows from the observation that the ratio between two terms

$$\frac{(2k-3)(2k+1)}{4(k+1)^2} < 1$$

for $k \geq 2$. Now in the infinite series of $\phi_1(x)$ we replace the coefficient $\frac{(2k-5)!!(2k-1)!!}{2^{2k-2}(k!)^2}$ by $\frac{(2k-m)!!(2m-1)!!}{2^{2m-2}(k!)^2}$ for $k \geq m$. This will increase the value of $\phi_1(x)$. The infinite sum for $k \geq m$ can be summed to

$$\frac{(2m-5)!!(2m-1)!!}{2^{2m-2}(m!)^2} x^{-2m} \frac{1}{1-x^{-2}} = \frac{(2m-5)!!(2m-1)!!}{2^{2m-2}(m!)^2} x^{-2(m-1)} \frac{1}{x^2-1}.$$

The other inequalities can be shown similarly. □

To prove certain poperties of those function, we use Mathematica. Essentially, we only use `CountRoots` to calculate the number of roots of a rational function in a given interval. If the number is 0, we can conclude that the function stays positive or negative in this interval. `CountRoots` applies an exact algorithm so that we can prove our results rigorously.

To bound infinite series by finite series, we need this trivial lemma:

Lemma 7.3.2. *For two real numbers a, b, if $0 \le a_1 \le a \le a_2$, $b_1 \le b \le b_2$, then $\min\{a_1b_1, a_2b_1\} \le ab \le \max\{a_1b_2, a_2b_2\}$. If furthermore $b \ge 0$, then $a_1b_1 \le ab \le a_2b_2$.*

Proof. Consider $\min ab$, where a, b are variables that satisfy $a_1 \le a \le a_2$ and $b_1 \le b \le b_2$. Then this minimum is $\min\{a_ib_j, i, j \in [2]\}$. We now use the assumption that $0 \le a_1$. Hence $a_1b_1 \le a_1b_2, a_2b_1 \le a_2b_2$. Hence $ab \ge \min\{a_1b_1, a_2b_1\}$. Similarly

$$\max ab = \max\{a_ib_j, i, j \in [2]\} = \max\{a_1b_2, a_2b_2\}.$$

Assume now that $b_2 \ge b \ge 0$. Then

$$\min\{ab, a \in [a_1, a_2]\} = \min\{a_1b, a_2b\} = a_1b,$$
$$\max\{ab, a \in [a_1, a_2]\} = \max\{a_1b, a_2b\} = a_2b \le a_2b_2. \qquad \square$$

7.3.2 The function $\omega(x)$

In this section, $m = 50$.

Proposition 7.3.3. *Assume that $\tau \in (1, 2)$ is the smallest value of $x \ge 1$ such that $\tilde{\omega}'(x) = 0$. Then $\tau > 1.732$.*

Proof.

$$\tilde{\omega}'(x) = \phi_2''(x)\phi_1(x) - \phi_2(x)\phi_1''(x).$$

We now apply Lemma 7.3.2 as follows. First set

$$a_1 = \phi_{1,m}(x), a = \phi_1(x), a_2 = \hat{\phi}_{1,m}(x), b_1 = \phi''_{2,m}(x) < \phi''_2(x), \quad x \in [1,5], m = 50,$$

to deduce

$$\phi''_2(x)\phi_1(x) \geq \min\{\phi''_{2,m}(x)\phi_{1,m}(x), \phi''_{2,m}(x)\hat{\phi}_{1,m}(x)\}.$$

Second set

$$a_1 = \phi_{2,m}(x), a = \phi_2(x), a_2 = \hat{\phi}_{2,m}(x), b_2 = \tilde{\phi}_{1,m}(x), \quad x \in [1,5], m = 50,$$

to deduce

$$\phi_2(x)\phi''_1(x) \leq \max\{\phi_{2,m}(x)\tilde{\phi}_{1,m}(x), \hat{\phi}_{2,m}(x)\tilde{\phi}_{1,m}(x)\},$$

which is equivalent to

$$-\phi_2(x)\phi''_1(x) \geq \min\{-\phi_{2,m}(x)\tilde{\phi}_{1,m}(x), -\hat{\phi}_{2,m}(x)\tilde{\phi}_{1,m}(x)\}.$$

So $\tilde{\omega}'(x)$ is larger than the minimum of $2 \times 2 = 4$ functions. Using `CountRoots`, we found that those 4 functions are all positive in $[1, 1.732]$, thus $\tau > 1.732$. □

7.3.3 The function $\omega(x)p_+^l$

In this section, $m = 50$. Recall that $p_+(x)$ can be written as

$$p_+(x) = \sqrt{\psi_1^2(x) + \psi_2^2(x)} = \frac{3\pi}{16}\sqrt{\tilde{p}}, \quad \tilde{p}(x) = 16\phi_1^2 + \phi_2^2.$$

Proposition 7.3.4. *The function $\omega(x)p_+^l$ increases in the interval $[1,5]$ for $l = 7$.*

Proof. Note that since ω is increasing in the interval $[1,\tau]$ we automatically have that for all $l \geq 1$ the function $\omega(x)p_+^l$ increases on $[1,\tau]$. So now we have to verified that $\omega(x)p_+^l$ increases on $[1.732, 5]$.

Set $s_l(x) = \tilde{\omega}(x)\tilde{p}^{l/2}(x)$. Then

$$s_l' = \tilde{p}^{(l-2)/2}\big((l/2)\tilde{\omega}\tilde{p}' + \tilde{\omega}'\tilde{p}\big). \tag{7.49}$$

Let

$$\rho = (l/2)\tilde{\omega}\tilde{p}' + \tilde{\omega}'\tilde{p} = l\tilde{\omega}(16\phi_1\phi_1' + \phi_2\phi_2') + \tilde{\omega}'(16\phi_1^2 + \phi_2^2)$$

We need to prove $\rho \geq 0$ in $[\tau, 5]$. When x is in $[1.732, 5]$, by `CountRoots`, we have $\bar{\phi}_{1,m} > 0, \tilde{\phi}_{1,m} < 0, \phi''_{2,m} < 0$. In view of Proposition 7.3.1 we deduce that

$$0 < \bar{\phi}_{1,m}(x) < \phi_1'(x), \quad 0 < -\tilde{\phi}_{1,m}(x) < -\phi_1''(x) \text{ for } x \in [1.732, 5].$$

Furthermore, Proposition 7.3.1 yields $\tilde{\omega} = \phi_2'\phi_1 - \phi_1'\phi_2 \geq \phi'_{2,m}\phi_{1,m} - \phi'_{1,m}\hat{\phi}_{2,m}$. Again, when x in $[1.732, 5]$, by `CountRoots`, $\phi'_{2,m}\phi_{1,m} - \phi'_{1,m}\hat{\phi}_{2,m}$ is positive. So we can bound from below the whole term $(l/2)\tilde{\omega}\tilde{p}'$ by $l(\phi'_{2,m}\phi_{1,m} - \phi'_{1,m}\hat{\phi}_{2,m})(16\phi_{1,m}\bar{\phi}_{1,m} + \phi_{2,m}\phi'_{2,m})$. This term is positive.

For the second term, if $\tilde{\omega}' \geq 0$, we have nothing to prove. So we can assume $\tilde{\omega}' < 0$. Because $\tilde{\omega}' = \phi_2''\phi_1 - \phi_2\phi_1''$, and $-\phi_2\phi_1''$ is positive, so we can assume ϕ_2'' is negative. Recall that for $m = 50$ and $x \in [1,5]$ we have the inequality $\phi''_{2,m}(x) < \phi_2''(x)$. Hence for $x \in [1.732, 5]$ we have the inequality $\phi''_{2.m}(x) < \phi_2''(x) < 0$. Therefore

$$\tilde{\omega}'(16\phi_1^2 + \phi_2^2) > (\phi''_{2,m}\hat{\phi}_{1,m} - \phi_{2,m}\tilde{\phi}_{1,m})(16\hat{\phi}_{1,m}^2 + \hat{\phi}_{2,m}^2), \quad x \in [1.732, 5], m = 50.$$

Finally, use `CountRoots` to test whether $l(\phi'_{2,m}\phi_{1,m} - \phi'_{1,m}\hat{\phi}_{2,m})(16\phi_{1,m}\bar{\phi}_{1,m} + \phi_{2,m}\phi'_{2,m}) +$

$(\phi''_{2,m}\hat{\phi}_{1,m} - \phi_{2,m}\tilde{\phi}_{1,m})(16\hat{\phi}^2_{1,m} + \hat{\phi}^2_{2,m})$ is positive in $[1.732, 5]$. The answer is yes. □

7.3.4 The function $\mu(x)$

In this section, $m = 40$. Recall the function

$$\mu(x) = \frac{\psi'_2(x)\psi_2(x) + \psi'_1(x)\psi_1(x)}{\psi'_2(x)\psi_1(x) - \psi'_1(x)\psi_2(x)}.$$

Note that

$$\frac{1}{4\mu(x)} = \frac{\phi'_2(x)\phi_1(x) - \phi'_1(x)\phi_2(x)}{\phi'_2(x)\phi_2(x) + 16\phi'_1(x)\phi_1(x)}.$$

Define $\nu(x)$ as

$$\begin{aligned}(\frac{1}{4\mu(x)})' &= \frac{\nu(x)}{(\phi'_2(x)\phi_2(x) + 16\phi'_1(x)\phi_1(x))^2},\\ \nu(x) &= (\phi'_2(x)\phi_1(x) - \phi'_1(x)\phi_2(x))'(\phi'_2(x)\phi_2(x) + 16\phi'_1(x)\phi_1(x)) -\\ &\quad (\phi'_2(x)\phi_1(x) - \phi'_1(x)\phi_2(x))(\phi'_2(x)\phi_2(x) + 16\phi'_1(x)\phi_1(x))'.\end{aligned}$$

Thus

$$\nu = \phi''_2\phi'_1(16\phi_1^2 + \phi_2^2) - \phi'_2\phi''_1(16\phi_1^2 + \phi_2^2) - \tilde{\omega}(16\phi'^2_1 + \phi'^2_2).$$

Proposition 7.3.5. *The function $\nu(x)$ is positive in $[1, 1.732]$, so $\mu(x)$ is decreasing in this interval.*

Proof. We have to consider two different intervals: $[1.01, 1.732]$ and $[1, 1.01]$.

Assume that $x \in [1.01, 1.732]$. By `CountRoots`, $\bar{\phi}_{1,m}(x) > 0$ in $[1.01, 1.732]$, so $\phi'_1(x) > \bar{\phi}_{1,m}(x) > 0$ in $[1.01, 1.732]$. Recall Proposition 7.3.1 and Lemma 7.3.2, for $x \in [1.01, 1.732]$

and $m = 40$:

$$\begin{aligned}
&a_1 = \bar{\phi}_{1,m}(x)(16\phi_{1,m}^2(x) + \phi_{2,m}^2(x)), a = \phi_1'(x)(16\phi_1^2(x) + \phi_2^2(x)),\\
&a_2 = \phi_{1,m}'(x)(16\hat{\phi}_{1,m}^2(x) + \hat{\phi}_{2,m}^2(x)),\\
&b_1 = \phi_{2,m}''(x), b = \phi_2''(x), b_2 = \bar{\phi}_{2,m}(x).
\end{aligned}$$

Then we have

$$\phi_2''\phi_1'(16\phi_1^2 + \phi_2^2) \geq \min\{\phi_{2,m}''\bar{\phi}_{1,m}(16\phi_{1,m}^2 + \phi_{2,m}^2), \phi_{2,m}''\phi_{1,m}'(16\hat{\phi}_{1,m}^2 + \hat{\phi}_{2,m}^2)\}.$$

For the second term, $-\phi_2'\phi_1''(16\phi_1^2 + \phi_2^2)$, by Proposition 7.3.1:

$$\begin{aligned}
&a_1 = \phi_{2,m}'(x)(16\phi_{1,m}^2(x) + \phi_{2,m}^2(x)), a = \phi_2'(x)(16\phi_1^2(x) + \phi_2^2(x)),\\
&a_2 = \bar{\phi}_{2,m}(x)(16\hat{\phi}_{1,m}^2(x) + \hat{\phi}_{2,m}^2(x)),\\
&b_1 = \phi_{1,m}''(x), b = \phi_1''(x), b_2 = \tilde{\phi}_{1,m}(x).
\end{aligned}$$

Use the inequality $ab \leq \max\{a_1b_2, a_2b_2\}$ in Lemma 7.3.2 to deduce

$$-\phi_2'\phi_1''(16\phi_1^2 + \phi_2^2) \geq \min\{-\bar{\phi}_{2,m}\tilde{\phi}_{1,m}(16\hat{\phi}_{1,m}^2 + \hat{\phi}_{2,m}^2), -\phi_{2,m}'\tilde{\phi}_{1,m}(16\phi_{1,m}^2 + \phi_{2,m}^2)\}.$$

For the last term $-\tilde{\omega}(16\phi_1'^2 + \phi_2'^2)$ we proceed as follows: First recall that $\tilde{\omega} > 0$ for $x > 1$. Use Proposition 7.3.1 and Lemma 7.3.2 to deduce $\tilde{\omega}_2(x) < \bar{\phi}_{2,m}\hat{\phi}_{1,m}(x) - \phi_{2,m}\bar{\phi}_{1,m}(x)$ for $x > 1$. Hence

$$\begin{aligned}
&\tilde{\omega}(x)(16\phi_1'^2(x) + \phi_2'^2(x)) < (\bar{\phi}_{2,m}\hat{\phi}_{1,m}(x) - \phi_{2,m}\bar{\phi}_{1,m}(x))((\phi_{1,m}'(x))^2 + (\bar{\phi}_{2,m}(x))^2) \Rightarrow\\
&-\tilde{\omega}(x)(16\phi_1'^2(x) + \phi_2'^2(x)) > -(\bar{\phi}_{2,m}\hat{\phi}_{1,m}(x) - \phi_{2,m}\bar{\phi}_{1,m}(x))((\phi_{1,m}'(x))^2 + (\bar{\phi}_{2,m}(x))^2)
\end{aligned}$$

There are $2 \times 2 = 4$ cases, in each case, their sum is positive in $[1.01, 1.732]$. So $\nu(x)$ is positive when $x \in [1.01, 1.732]$.

Assume that $x \in [1, 1.01]$. For $\phi_1'(x)$, we need a better lower bound than $\bar{\phi}_{1,m}(x)$ because when $x \to 1$, $\bar{\phi}_{1,m}(x)$ diverges to $-\infty$ while $\phi_1'(x)$ remains finite.

Recall the definition:

$$\phi_1(x) = \frac{4}{3\pi}\psi_1(x) = \frac{2}{\pi}\int_0^{\frac{\pi}{2}} (1 - x^{-2}\sin^2 u)^{\frac{3}{2}}\, du.$$

Take the derivative and note that $\pi < 22/7$,

$$\begin{aligned} \phi_1(x) &= \frac{6}{\pi}x^{-3}\int_0^{\frac{\pi}{2}} \sin^2 u(1 - x^{-2}\sin^2 u)^{\frac{1}{2}}\, du \\ &\geq \frac{6}{\pi}x^{-3}\int_0^{\frac{\pi}{2}} \sin^2 u(1 - \sin^2 u)^{\frac{1}{2}}\, du \\ &= \frac{6}{\pi}x^{-3}\int_0^{\frac{\pi}{2}} \sin^2 u \cos u\, du = \frac{2}{\pi x^3} \geq \frac{7}{11x^3} \end{aligned}$$

For the first term, Proposition 7.3.1 yields $\phi_2''(x) \geq \phi_{2,m}''(x), \phi_1''(x) \geq \phi_{1,m}''(x)$.

By `CountRoots`, $\phi_{2,m}''(x) > 0, \phi_{1,m}''(x) > 0$. Hence

$$\phi_2''\phi_1'(16\phi_1^2 + \phi_2^2) \geq \phi_{2,m}''\frac{7}{11x^3}(16\phi_{1,m}^2 + \phi_{2,m}^2).$$

For the second term, because ϕ_1, ϕ_2 are increasing functions, we can bound them above by value at 1.01. So the whole term can be bounded below using Proposition 7.3.1 by $-\bar{\phi}_{2,m}\tilde{\phi}_{1,m}(16\hat{\phi}_{1,m}(1.01)^2 + \hat{\phi}_{2,m}(1.01)^2)$.

For the last term, we use the same bound as for the interval $[1.01, 1.732]$. Finaly, the sum of three terms is positive in $[1, 1.01]$ using `CountRoots`. So $\nu(x)$ is positive when $x \in [1, 1.01]$. □

7.4 Symmetric Versions of Grothendieck Inequality for quaternions

In this section, we address the symmetric version of Grothendieck inequality for quaternions. Note that $\mathbb{S}^n(\mathbb{H})$ is the set of $n \times n$ quaternion matrices A satisfying $A^* = A$ and $\mathbb{S}^n_+(\mathbb{H})$ is the set of positive semidefinite matrices A satisfying $u^* A u \geq 0$ for all $u \in \mathbb{H}^n$. There is no natural left or right action of quaternion scalar on the space $\mathbb{S}^n(\mathbb{H})$, so we should view it as a real vector space. Recall that on $\mathbb{S}^n(\mathbb{C})$ the inner product is given by $\langle A, B\rangle = \operatorname{tr} A^* B = \operatorname{tr} AB \in \mathbb{R}$. Unfortunately, for $A, B \in \mathbb{S}^n(\mathbb{H})$, where $n \geq 2$ the $\operatorname{tr} AB$ does not to have to be real. Since we consider $\mathbb{S}^n(\mathbb{H})$ as a real vector space, we define an inner product on $\mathbb{S}^n(\mathbb{F})$ as $\Re \operatorname{tr} AB = \frac{1}{2} \operatorname{tr}(AB + BA)$. This definition is identical to the standard definition of the inner product for $\mathbb{F} \in \{\mathbb{R}, \mathbb{C})$ and gives the right definition for quaternions.

We consider the quantities:

$$\Re \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \langle x_i, x_j \rangle, \quad \Re \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \bar{\delta}_i \delta_j.$$

As in Section 7.2.8 we can compute the maximum of the first term using SDP. Also, as in [77], in this section we modify the definition of $\operatorname{sgn} z$ for $z \in \mathbb{H}$:

$$\operatorname{sgn} z = \begin{cases} z/|z| & z \neq 0, \\ 1 & z = 0. \end{cases}$$

This will yield the equality $|\operatorname{sgn} z| = 1$ for all z. Clearly Theorem 7.2.16 will still hold for this definition of $\operatorname{sgn} z$.

7.4.1 *Symmetric Grothendieck inequality*

Denote by $\mathbb{S}_o^n(\mathbb{H})$ the real subspace of all quaternion self-adjoint matrices with zero diagonal. Then $A \in \mathbb{S}_o^n(\mathbb{H})$ is of the form $A = D + A_0$ where $A_0 = (a_{ij,0}) \in \mathbb{S}_o^n(\mathbb{H})$ and D is a real diagonal matrix. So $\operatorname{tr} A = \operatorname{tr} D$. Observe that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\langle x_i, x_j\rangle = \operatorname{tr} A + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij,0}\langle x_i, x_j\rangle \text{ if } \|x_i\| = 1.$$

For $A \in \mathbb{S}^n(\mathbb{H})$, we consider the following quantities, which are analogous to those introduced in [77]:

$$\|A\|_\theta = \max\nolimits_{|\delta_i|=1}\left|\Re \textstyle\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\bar{\delta}_i\delta_j\right|, \qquad \|A\|_\Theta = \max_{|\delta_i|\le 1}\left|\Re \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\bar{\delta}_i\delta_j\right|,$$

$$\|A\|_\gamma = \max\nolimits_{\|x_i\|=1}\left|\Re \textstyle\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\langle x_i, x_j\rangle\right|, \qquad \|A\|_\Gamma = \max_{\|x_i\|\le 1}\left|\Re \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\langle x_i, x_j\rangle\right|.$$

Clearly $\|A\|_\theta \le \|A\|_\gamma$ and $\|A\|_\Theta \le \|A\|_\Gamma$. As in [77] we observe that $\|A\|_\theta = 0$ if and only if A is a diagonal matrix, and $\operatorname{tr} A = 0$. Indeed, observe that

$$\sum_{\delta_i \in \{-1,1\}} \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\bar{\delta}_i\delta_j = 2^n \operatorname{tr} A.$$

Assume that $\|A\|_\theta = 0$. Then we obtain that $\operatorname{tr} A = 0$. Hence $\|A\|_\theta = \|A_0\|_\theta = 0$. Observe next that

$$\sum_{\delta_i \in \{-1,1\}, i \ge 3} \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} = 2^{n-2}(a_{12}\bar{\delta}_1\delta_2 + a_{21}\bar{\delta}_2\delta_1) = 2^{n-2}(a_{12}\bar{\delta}_1\delta_2 + \bar{a}_{12}\bar{\delta}_2\delta_1).$$

Suppose that $a_{12} \ne 0$. Choose $\delta_1 = \operatorname{sgn} a_{12}$ and $\delta_2 = 1$. Then $(a_{12}\bar{\delta}_1\delta_2 + \bar{a}_{12}\bar{\delta}_2\delta_1) = 2|a_{12}|$ which is a real number. Hence the assumption that $\|A_0\|_\theta = 0$ yields that $a_{12} = 0$. Similarly we deduce that $A_0 = 0$. Hence $\|A\|_\gamma = 0$ if and only if A is a diagonal matrix with zero trace.

Denote by $\mathbb{S}^n_=(\mathbb{H})$ the real subspace of self-adjoint matrices $A = (a_{ij})$ where $a_{11} = \cdots = a_{nn}$. Then $\|A\|_\theta$ and $\|A\|_\gamma$ are norms on $\mathbb{S}^n_=(\mathbb{H})$.

We now claim that $\|A\|_\Gamma$ is a norm on $\mathbb{S}^n(\mathbb{H})$. Clearly, $\|A\|_\Gamma$ is a seminorm. Choose $\delta_i = 1$ and $\delta_j = 0$ for $i \neq j$. Then $|\Re \sum_{i=1}^n \sum_{j=1}^n a_{ij}\bar{\delta}_i\delta_j| = |a_{ii}| = 0$. Hence $A = A_0$. As $\|A_0\|_\theta \leq \|A\|_\Theta = 0$ we deduce that $A_0 = 0$. Thus $A = 0$ and $\|\cdot\|_\Theta$ is a norm. As $\|A\|_\Theta \leq \|A\|_\Gamma$ it follows that $\|\cdot\|_\Gamma$ is a norm on $\mathbb{S}^n(\mathbb{H})$.

Let $\mathcal{D}^n \subset \mathbb{S}^n_+(\mathbb{R})$ be the convex subset of all positive semidefinite diagonal matrices whose diagonal entries are in $[0, 1]$. As in [77] it is straightforward to show that

$$\|A\|_\Theta = \max\{\|DAD\|_\theta, D \in \mathcal{D}^n\}, \quad \|A\|_\Gamma = \max\{\|DAD\|_\gamma, D \in \mathcal{D}^n\}. \tag{7.50}$$

Let $K^{\mathbb{H}}_\gamma, K^{\mathbb{H}}_\Gamma$ the smallest possible constant, (which in principle may equal to ∞), for which one has the inequalities

$$\|A\|_\gamma \leq K^{\mathbb{H}}_\gamma \|A\|_\theta, \qquad \|A\|_\Gamma \leq K^{\mathbb{H}}_\Gamma \|A\|_\Theta. \tag{7.51}$$

Theorem 7.4.1 (Symmetric Grothendieck inequality)**.** *The symmetric Grothendieck constants satisfy the following relations*

$$K^{\mathbb{H}}_G \leq K^{\mathbb{H}}_\Gamma \leq K^{\mathbb{H}}_\gamma \leq \frac{64}{9\pi} - 1 \approx 1.263537. \tag{7.52}$$

Proof. The first inequality follows from the proof of Lemma 7.2.14. The second inequality follows from the characterization (7.50). We now show the third inequality. Assume that $\|A\|_\theta \leq 1$: $\left|\Re \sum_{i=1}^n \sum_{j=1}^n a_{ij}\bar{\delta}_i\delta_j\right| \leq 1$ for $|\delta_i| = 1$. Let $x_1, \ldots, x_n$ be unit vectors in a right Hilbert space $\mathcal{H}$. Hence the span of $x_1, \ldots, x_n$ is contained in a subspace of dimension at most n. Thus we shall assume that $x_1, \ldots, x_n \in \mathbb{H}^n$. We claim that

$$\left| \Re \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} P(\langle x_i, x_j \rangle) |P(\langle x_i, x_j \rangle)|^{2k} \right| \leq 1, \quad k+1 \in \mathbb{N}.$$

Let

$$\Phi(x, z_1, \ldots, z_{2k+1}) = \operatorname{sgn}\langle z_1, x \rangle \operatorname{sgn}\langle x, z_2 \rangle \operatorname{sgn}\langle z_3, x \rangle \ldots \operatorname{sgn}\langle z_{2k+1}, x \rangle.$$

We claim that

$$\int_{(\mathbb{H}^n)^{2k+1}} \overline{\Phi(x, z_1, \ldots, z_{2k+1})} \Phi(y, z_1, \ldots, z_{2k+1}) \prod_{p=1}^{2k+1} G^{\mathbb{H}}(z_p) dm(z_p)$$
$$= P(\langle x, y \rangle) |P(\langle x, y \rangle)|^{2k}.$$

Assume first that $\langle x, y \rangle$ is a real number. Then the above equality follows by applying Theorem 7.2.16 $2k+1$ times. Observe next that for $a \in \mathbb{H}, |a| = 1$ one has the equality $\Phi(xa, z_1, \ldots, z_{2k+1}) = \Phi(x, z_1, \ldots, z_{2k+1})a$. Hence replacing x by $x \operatorname{sgn}\langle x, y \rangle$, we deduce the above equality.

Clearly

$$\left| \Re \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \overline{\Phi(x_i, z_1, \ldots, z_{2k+1})} \Phi(x_j, z_1, \ldots, z_{2k+1}) \right| \leq 1.$$

Hence

$$
\begin{aligned}
&\left|\Re\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}P(\langle x_i,x_j\rangle)|P(\langle x_i,x_j\rangle)|^{2k}\right| \\
&= \left|\Re\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}(\int_{(\mathbb{H}^n)^{2k+1}}\overline{\Phi(x_i,z_1,\dots,z_{2k+1})}\Phi(x_j,z_1,\dots,z_{2k+1})\prod_{p=1}^{2k+1}G^{\mathbb{H}}(z_p)dm(z_p))\right| \\
&\le \int_{(\mathbb{H}^n)^{2k+1}}\left|\Re\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}\overline{\Phi(x_i,z_1,\dots,z_{2k+1})}\Phi(x_j,z_1,\dots,z_{2k+1})\right|\prod_{p=1}^{2k+1}G^{\mathbb{H}}(z_p)dm(z_p) \\
&\le 1.
\end{aligned}
$$

Finally

$$
\begin{aligned}
&\left|\Re\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}\langle x_i,x_j\rangle\right| = \left|\Re\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}P^{-1}(P(\langle x_i,x_j\rangle))\right| \\
&\le \sum_{k=0}^{\infty}|c_{2k+1}|\left|\Re\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}P(\langle x_i,x_j\rangle)|P(\langle x_i,x_j\rangle)|^{2k}\right| \le \sum_{k=0}^{\infty}|c_{2k+1}| = \frac{64}{9\pi}-1.
\end{aligned}
$$

□

7.4.2 Cones of positive semidefinite matrices

Assume that $A \in \mathbb{S}^n(\mathbb{R}), B \in \mathbb{S}^n(\mathbb{H})$. Then $\operatorname{tr} AB = \operatorname{tr} BA$. Hence $\Re \operatorname{tr} AB = \operatorname{tr} AB$. For any real positive definite matrix, the quantities considered here have a special relation:

Lemma 7.4.2. *For $A \in \mathbb{S}^n_+(\mathbb{R})$,*

$$
\max_{\|x_i\|=1}\left|\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}\langle x_i,x_j\rangle\right| = \max_{\|x_i\|\le 1}\left|\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}\langle x_i,x_j\rangle\right| = \max_{\|x_i\|=1,\|y_j\|=1}\left|\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}\langle x_i,y_j\rangle\right|
$$

where $x_1, \ldots, x_n, y_1, \ldots, y_n \in \mathbb{H}^l$ *and*

$$\max_{|\delta_i|=1}\left|\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\langle\delta_i, \delta_j\rangle\right| = \max_{|\delta_i|\le 1}\left|\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\langle\delta_i, \delta_j\rangle\right| = \max_{|\delta_i|=1, |\epsilon_j|=1}\left|\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\langle\delta_i, \epsilon_j\rangle\right|$$

where $\delta_1, \ldots, \delta_n, \epsilon_1, \ldots, \epsilon_n \in \mathbb{H}$.

Proof. Observe that by definition

$$\max_{\|x_i\|=1}\left|\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\langle x_i, x_j\rangle\right| \le \max_{\|x_i\|\le 1}\left|\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\langle x_i, x_j\rangle\right| \le \max_{\|x_i\|=1, \|y_j\|=1}\left|\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\langle x_i, y_j\rangle\right|.$$

Define the matrices $X = [x_1, \ldots, x_n], Y = [y_1, \ldots, y_n]$. Because A is real positive definite. Then $A = B^2$, where $B = (b_{ij}) = A^{1/2}$.

$$\langle X, Y\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\langle x_i, y_j\rangle = \sum_{p=1}^{n}\langle\sum_{i=1}^{n} x_i b_{ip}, \sum_{j=1}^{n} y_j b_{jp}\rangle.$$

is an inner product of the space $\mathbb{H}^{l\times n}$. By Cauchy-Schwarz, we have

$$|\langle X, Y\rangle| \le \sqrt{\langle X, X\rangle\langle Y, Y\rangle} \le \max_{\|x_i\|=1}\left|\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\langle x_i, x_j\rangle\right|.$$

So

$$\max_{\|x_i\|=1, \|y_j\|=1}\left|\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\langle x_i, y_j\rangle\right| \le \max_{\|x_i\|=1}\left|\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\langle x_i, x_j\rangle\right|$$

and the first equality holds. By taking $l = 1$ in the previous argument, we see that the second equality holds. □

The next theorem is a generalization of the Nesterov $\pi/2$-Theorem.

Theorem 7.4.3. *Let* $A = (a_{ij}) \in \mathbb{S}^n_+(\mathbb{R})$ *is a symmetric positive semidefinite matrices.*

Then

$$\max_{\|x_i\|=1}\left|\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\langle x_i, x_j\rangle\right| \leq \frac{32}{9\pi}\max_{|\delta_i|=1}\left|\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\langle \delta_i, \delta_j\rangle\right|. \tag{7.53}$$

where $x_1, \ldots, x_n \in \mathbb{H}^l$, $\delta_1, \ldots, \delta_n \in \mathbb{H}$. By Lemma 7.4.2, we can change the expression on both sides of the inequality.

Proof. Assume

$$\max_{|\delta_i|=1}\left|\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\langle \delta_i, \delta_j\rangle\right| = 1.$$

Then for any $x_1, \ldots, x_n \in \mathbb{H}^l$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} P(\langle x_i, x_j\rangle) \leq 1.$$

Consider the matrix $G_{ij} = \langle x_i, x_j\rangle$. The first coefficient of P is $\frac{9\pi}{32}$. By the Schur product theorem for quaternions,

$$A \circ P(G) - A \circ \frac{9\pi}{32} G \succeq 0.$$

So

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} P(\langle x_i, x_j\rangle) \geq \frac{9\pi}{32}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\langle x_i, x_j\rangle.$$

So the inequality holds. □

The constant $\frac{9\pi}{32}$ is sharp. In fact, the proof of Theorem 5.1 in [77] can be repeated with only minor difference.

7.4.3 Cones of weighted Laplacians

For any matrix $A = (a_{ij}) \in \mathbb{S}^n(\mathbb{R})$ with zero diagonal and positive off-diagonal elements, we can define L_A by

$$L_A := \operatorname{diag}(A\mathbf{1}) - A,$$

where $\operatorname{diag}(x)$ denotes the diagonal matrix whose diagonal is x and $\mathbf{1}$ is the vector of all ones. And we can define the set of all weighted Laplacians:

$$\begin{aligned} \mathbb{L}^n &:= \{L_A : A \in \mathbb{S}^n(\mathbb{R}), a_{ii} = 0, a_{ij} \geq 0 \text{ for all } i, j\} \\ &= \{L \in \mathbb{S}^n(\mathbb{R}) : L\mathbf{1} = 0, l_{ij} \leq 0 \text{ for all } i \neq j\}. \end{aligned}$$

We have $\mathbb{L}^n \subseteq \mathbb{S}^n_+(\mathbb{R})$.

Theorem 7.4.4. *Define the quaternionic Goemans-Williamson constant:*

$$\alpha^{\mathbb{H}}_{GW} := \inf_{0 \leq x \leq 1} \frac{1 + P(x)}{1 + x}.$$

Then for all $L \in \mathbb{L}^n$,

$$\max_{\|x_i\|=1} \left| \sum_{i=1}^n \sum_{j=1}^n l_{ij} \langle x_i, x_j \rangle \right| \leq \frac{1}{\alpha^{\mathbb{H}}_{GW}} \max_{|\delta_i|=1} \left| \sum_{i=1}^n \sum_{j=1}^n l_{ij} \langle \delta_i, \delta_j \rangle \right|. \tag{7.54}$$

where $x_1, \ldots, x_n \in \mathbb{H}^l$, $\delta_1, \ldots, \delta_n \in \mathbb{H}$. The value of the constant is approximate 0.967337*, so* $K \leq 1.0338$*. By Lemma 7.4.2, we can change the expression on both sides of the inequality.*

Proof. We will show that for $h \in \mathbb{H}$, we have

$$\alpha^{\mathbb{H}}_{GW} = \inf_{0 \leq x \leq 1} \frac{1 + P(x)}{1 + x} = \inf_{|h|<1} \frac{1 - \Re[P(h)]}{1 - \Re(h)}.$$

By definition,

$$\inf_{|h|<1} \frac{1 - \Re[P(h)]}{1 - \Re(h)} \leq \inf_{0 \leq x \leq 1} \frac{1 + P(x)}{1 + x}.$$

On the other hand, if $\Re(h) \geq 0$, let $h = x + y$, $x = \Re(h)$. The Taylor expansion of $P(x)$ is given in formula 7.23, $P(x) = \sum_{i=0} b_{2i+1} x^{2i+1}$. $b_{2i+1} \geq 0$ and $P(1) = 1$.

$$\Re[P(h)] = \sum_{i=0} b_{2i+1} x(x^2 + |y|^2)^i \leq \sum_{i=0} b_{2i+1} x = x.$$

So

$$\inf_{|h|<1, \Re(h) \geq 0} \frac{1 - \Re[P(h)]}{1 - \Re(h)} \geq 1 \geq \inf_{0 \leq x \leq 1} \frac{1 + P(x)}{1 + x}.$$

If $\Re(h) \leq 0$, let $h = -x + y$, $x = -\Re(h)$. Then

$$1 - \Re[P(h)] = 1 + \sum_{i=0} b_{2i+1} x(x^2 + |y|^2)^i \geq 1 + \sum_{i=0} b_{2i+1} x^{2i+1} = 1 + P(x).$$

So

$$\inf_{|h|<1, \Re(h) \leq 0} \frac{1 - \Re[P(h)]}{1 - \Re(h)} \geq \inf_{0 \leq x \leq 1} \frac{1 + P(x)}{1 + x}.$$

So we have

$$\inf_{|h|<1} \frac{1 - \Re[P(h)]}{1 - \Re(h)} \geq \inf_{0 \leq x \leq 1} \frac{1 + P(x)}{1 + x}$$

and the equality holds.

Let $L_A \in \mathbb{L}^n$. Assume

$$\max_{|\delta_i|=1} \left| \sum_{i=1}^{n} \sum_{j=1}^{n} l_{ij} \langle \delta_i, \delta_j \rangle \right| = 1.$$

For any unit vectors $x_1, \ldots, x_n \in \mathbb{H}^l$,

$$\int_{z \in \mathbb{H}^l} \sum_{i=1}^{n} \sum_{j=1}^{n} l_{ij} \operatorname{sgn}\langle x_i, z \rangle \operatorname{sgn}\langle z, x_j \rangle G_n(z) dm(z) \leq 1.$$

Because L is a real symmetric matrix, we have

$$\begin{aligned}
&\int_{z\in\mathbb{H}^l}\sum_{i=1}^{n}\sum_{j=1}^{n} l_{ij}\operatorname{sgn}\langle x_i, z\rangle \operatorname{sgn}\langle z, x_j\rangle G_n(z)dm(z)\\
&=\sum_{i=1}^{n}\sum_{j=1}^{n} l_{ij}P(\langle x_i, x_j\rangle)\\
&=\sum_{i=1}^{n}\sum_{j=1}^{n} l_{ij}\Re[P(\langle x_i, x_j\rangle)]\\
&=\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(1-\Re[P(\langle x_i, x_j\rangle)])\\
&\geq \alpha_{GW}^{\mathbb{H}}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(1-\Re[\langle x_i, x_j\rangle])\\
&= \alpha_{GW}^{\mathbb{H}}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(1-\langle x_i, x_j\rangle)\\
&= \alpha_{GW}^{\mathbb{H}}\sum_{i=1}^{n}\sum_{j=1}^{n} l_{ij}\langle x_i, x_j\rangle.
\end{aligned}$$

Hence

$$\max_{\|x_i\|=1}\left|\sum_{i=1}^{n}\sum_{j=1}^{n} l_{ij}\langle x_i, x_j\rangle\right| \leq \frac{1}{\alpha_{GW}^{\mathbb{H}}}.$$

□

7.4.4 Cones of diagonally dominant matrices

As in last section, all matrices considered in this section will be real. Let $\mathbb{S}_{dd}^n := \{A \in \mathbb{S}^n(\mathbb{R}) : a_{ii} \geq \sum_{i\neq j}|a_{ij}|\}$ be the cone of symmetric diagonally dominant matrices. Let $\mathbb{S}_{dd}^n(\mathbb{R}_+)$ be the subcone of diagonally dominant matrices with nonnegative entries.

Lemma 7.4.5. *Every $A \in \mathbb{S}_{dd}^n$ has a unique decomposition $A = P + L$ such that $P \in \mathbb{S}_{dd}^n(\mathbb{R}_+), L \in \mathbb{L}^n$ and $l_{ij}p_{ij} = 0$ whenever $i \neq j$.*

Proof. Let B be defined by $b_{ij} = -a_{ij}$ if $a_{ij} < 0$ and $i \neq j$, $b_{ij} = 0$ otherwise. Then B has zero diagonal and positive off-diagonal elements and $L_B \in \mathbb{L}^n$. Let $P := A - L_B$. Then $p_{ij} \geq 0$ and $l_{ij}p_{ij} = 0$ whenever $i \neq j$. Since $a_{ii} \geq \sum_{i \neq j} |a_{ij}| = \sum_{i \neq j}(l_{ij} + p_{ij})$, we have $p_{ii} = a_{ii} - \sum_{i \neq j} l_{ij} \geq \sum_{i \neq j} p_{ij}$, so $P \in \mathbb{S}^n_{dd}$. Uniqueness follows by the definition. □

Clearly,

$$\mathbb{L}^n \subseteq \mathbb{S}^n_{dd} \subseteq \mathbb{S}^n_+(\mathbb{R}).$$

By definition, we expect the constant of Grothendieck inequality for $\mathbb{S}^n_{dd}$ lies between those of $\mathbb{L}^n$ and $\mathbb{S}^n_+(\mathbb{R})$. The following theorem gives a bound for the constant.

Theorem 7.4.6. *Let $a_0 = \frac{9\pi}{32}$ be the constant given in Theorem 7.4.3 and $\alpha^{\mathbb{H}}_{GW}$ be the quaternionic Goemans-Williamson constant. Then for any $A \in \mathbb{S}^n_{dd}$,*

$$\max_{\|x_i\|=1} \left| \sum_{i=1}^n \sum_{j=1}^n a_{ij} \langle x_i, x_j \rangle \right| \leq \left(1 + \frac{1 - a_0}{\alpha^{\mathbb{H}}_{GW}} \right) \max_{|\delta_i|=1} \left| \sum_{i=1}^n \sum_{j=1}^n a_{ij} \langle \delta_i, \delta_j \rangle \right|. \tag{7.55}$$

where $x_1, \ldots, x_n \in \mathbb{H}^l$, $\delta_1, \ldots, \delta_n \in \mathbb{H}$. The value of the constant is approximate 1.1204*. By Lemma 7.4.2, we can change the expression on both sides of the inequality.*

Proof. Let $A = P + L$ be the decomposition given by the lemma. Assume that

$$\max_{|\delta_i|=1} \left| \sum_{i=1}^n \sum_{j=1}^n a_{ij} \langle \delta_i, \delta_j \rangle \right| = 1.$$

We have

$$\begin{aligned}1 &\geq \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}P(\langle x_i, x_j\rangle)\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} l_{ij}P(\langle x_i, x_j\rangle) + \sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}P(\langle x_i, x_j\rangle)\\ &\geq \alpha_{GW}^{\mathbb{H}} \sum_{j=1}^{n} l_{ij}\langle x_i, x_j\rangle + a_0 \sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}\langle x_i, x_j\rangle\\ &= \alpha_{GW}^{\mathbb{H}} \sum_{j=1}^{n} a_{ij}\langle x_i, x_j\rangle - (\alpha_{GW}^{\mathbb{H}} - a_0) \sum_{j=1}^{n} p_{ij}\langle x_i, x_j\rangle.\end{aligned}$$

Since $P \in \mathbb{S}_{dd}^{n}(\mathbb{R}_+)$, we have

$$\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}\langle x_i, x_j\rangle \leq \sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} \leq 1$$

So

$$\alpha_{GW}^{\mathbb{H}} \sum_{j=1}^{n} a_{ij}\langle x_i, x_j\rangle \leq 1 + \alpha_{GW}^{\mathbb{H}} - a_0$$

and the inequality follows. □

7.5 Lower bounds for Grothendieck inequality

This section use the method in [52] to give lower bounds for Grothendieck inequality. We first start for simplicity of the exposition with the real case. Denote by

$$\mathrm{S}^{n-1} = \{x \in \mathbb{R}^n, \|x\| = 1\},$$

the $n-1$ dimensional sphere. Let σ be the unique probability Haar measure on S^{n-1} invariant under the orthogonal group $O(n) \subset \mathbb{R}^{n\times n}$. Denote by $\mathrm{L}_\infty(\mathrm{S}^{n-1}, \sigma)$ the space of bounded measurable functions, with the norm

$$\|f\|_\infty = \sup_{x\in \mathrm{S}^{n-1}} |f(x)| := \lim_{p\to\infty} \big(\int_{\mathrm{S}^{n-1}} |f(x)|^p d\sigma\big)^{1/p},$$

where sup means essential support by abuse of notation. Assume that $J : \mathrm{S}^{n-1} \to \mathbb{R}$ and $K : \mathrm{S}^{n-1} \times \mathrm{S}^{n-1} \to \mathbb{R}$ are continous functions. We associate the following bilinear form on $\mathrm{L}_\infty(\mathrm{S}^{-1}, \sigma) \times \mathrm{L}_\infty(\mathrm{S}^{-1}, \sigma)$:

$$\begin{gathered} B(f, g) = \int_{\mathrm{S}^{n-1}} J(x)f(x)g(x)d\sigma(x)+ \\ \int_{\mathrm{S}^{n-1}\times\mathrm{S}^{n-1}} K(x, y)f(x)g(y)d\sigma(x)d\sigma(y). \end{gathered} \tag{7.56}$$

We now show that we can approximate $B(f, g)$ by matrix binear form.

Recall the well known fact that every $f, g \in \mathrm{L}_\infty(\mathrm{S}^{n-1}, \sigma)$ can be approximated by a piece-wise constant function $\tilde{f}, \tilde{g}$, induced by the range of f and g. That is, for a given $\varepsilon > 0$, there exist piecewise constant functions $\tilde{f}, \tilde{g}$, such that $\|f - \tilde{f}\|_\infty, \|g - \tilde{g}\|_\infty \le \varepsilon$. Furthermore, the values of $\tilde{f}$ and $\tilde{g}$ can be chosen to be in $f(S)$ and $g(T)$ respectively for some measurable sets $S, T \subset \mathrm{S}^{n-1}$, with $\sigma(S) = \sigma(T) = 1$. As $\sigma(S \cap T) = 1$ we can asume that $S = T$. Thus for given $f, g \in \mathrm{L}_\infty(\mathrm{S}^{n-1}, \sigma)$ there exists a measurable partition $\{T_1, \ldots, T_M\}$ of S^{n-1} to pairwise disjoint measurable sets: $\sigma(\cup_{i=1}^M T_i) = 1$, such that each $\tilde{f}, \tilde{g}$ is a constant function on each T_i.

Denote by $\mathrm{L}_\infty(S)$ the finite dimensional space of piece-wise constant functions, whose value is fixed on each S_i. We can view these piecewise function f as a vector $\mathbf{f} \in \mathbb{R}^n$. For these piecewise functions we can approximate $B(f, g)$ as a bilinear form

$$B_N(\mathbf{f}, \mathbf{g}) = \sum_{i=1}^N J(x_i)\sigma(S_i)f_i g_i + \sum_{i=j=1}^N K(x_i, x_j)\sigma(S_i)\sigma(S_j)f_i g_j. \tag{7.57}$$

Thus (7.57) is approximation to (7.56) Note that if $\|\mathbf{f}\|_\infty \le 1$ then $|f_i| \le 1$ for each $i \in [N]$.

Moreover, assume that $|f(x)| = |g(x)| = 1$ a.e.. Then $|f_i| = |g_i| = 1$ for $i \in [N]$. For more precise statement see Lemma 7.5.1

7.5.1 The $(\infty, 1)$ and the Grothendieck norms

We now state the $(\infty, 1)$ and the Grothendieck norms for $B(\cdot, \cdot)$. Let $D(\mathbb{k}^d) = \{x \in \mathbb{k}^d, \|x\| \leq 1\}$. Denote by $\Phi_d^{\mathbb{k}}$ the space of all maps continuous maps $\phi : S(\mathbb{k}^d) \to D(\mathbb{k}^d)$. For $\phi, \psi \in \Phi_d^{\mathbb{k}}$ let

$$\begin{aligned} B(\phi, \psi) = \int_{S(\mathbb{k}^d)\times S(\mathbb{k}^d)} & K(x, y)\langle \phi(y), \psi(x)\rangle d\sigma(x)d\sigma(y) \\ & + \int_{S(\mathbb{k}^d)} J(x)\langle \phi(x), \psi(x)\rangle d\sigma(x). \end{aligned} \tag{7.58}$$

Then for a partition $\mathcal{S}_N$, we let

$$\begin{aligned} B_N(\phi, \psi) = \sum_{i=j=1}^{N} & K(\tilde{x}_i, \tilde{x}_j)\sigma(S_i)\sigma(S_j)\langle \phi(\tilde{x}_i), \psi(\tilde{x}_j)\rangle + \\ & \sum_{i=1}^{N} J(\tilde{x}_i)\sigma(S_i)\langle \phi(\tilde{x}_i), \psi(\tilde{x}_j)\rangle, \end{aligned} \tag{7.59}$$

to be an approximation of $B(\phi, \psi)$. We define the continuous analogs of the matrix norms $\| \cdot \|_{\infty,1}^{\mathbb{k}}, \| \cdot \|_{G,d}^{\mathbb{k}}$:

$$\begin{aligned} \|B\|_{\infty,1}^{\mathbb{k}} = \sup\{|B(f, g)|, f, g \in L_\infty(S(\mathbb{k}^d)), \|f\|_\infty, \|g\|_\infty \leq 1\}, \\ \|B\|_{G,d}^{\mathbb{k}} = \sup\{|B(\phi, \psi)|, \phi, \psi \in \Phi_d^{\mathbb{k}}\}. \end{aligned} \tag{7.60}$$

We need the following approximation lemma, which follows straightforward from previous discussion:

Lemma 7.5.1. *Let $K(\cdot, \cdot) : S(\mathbb{k}^d) \times S(\mathbb{k}^d) \to \mathbb{k}$ and $J(\cdot) : S(\mathbb{k}^d) \to \mathbb{k}$ be continuous. Assume that $f, g \in L_\infty(S(\mathbb{k}^d))$ and $\phi, \psi \in \Phi_d$ are given. Then for a given $\varepsilon > 0$ there is a measurable* *partition $\mathcal{S}_N = \{S_1, \ldots, S_N\}$ of $S(\mathbb{k}^d)$ such that*

$$|B(f,g) - B_N(\mathbf{f},\mathbf{g})| \le \varepsilon,$$
$$|B(\phi,\psi) - B_N(\phi,\psi)| \le \varepsilon.$$

We now show that the matrix Grothendieck constant $K_{G,1,d}$ applies to the kernel B:

Lemma 7.5.2. *Let $K^{\mathbb{k}}_{G,1,d}$ be the Grothendieck constant over $\mathbb{k}^d$. Then for each $B(\cdot,\cdot)$ on $S(\mathbb{k}^d) \times S(\mathbb{k}^d)$ the following inequality holds:*

$$K^{\mathbb{k}}_{G,1,d}\|B\|^{\mathbb{k}}_{\infty,1} \ge \|B\|^{\mathbb{k}}_{G,d}. \tag{7.61}$$

Proof. Let $\phi_p, \psi_p \in \Phi^{\mathbb{k}}_d, p \in \mathbb{N}$ be a sequence of continuous maps, such that

$$\lim_{p\to\infty} |B(\phi_p,\psi_p)| = \|B\|^{\mathbb{k}}_{G,d}.$$

Thus is enough to show that

$$K^{\mathbb{k}}_{G,1,d}\|B\|^{\mathbb{k}}_{\infty,1} \ge |B(\phi_p,\psi_p)|$$

for each $p \in \mathbb{N}$. This follows from Lemma 7.5.1 and the definition of $\|B\|^{\mathbb{k}}_{\infty,1}$. □

7.5.2 The norms $\theta, \Theta, \gamma, \Gamma$

Assume that $K(x,y)$ and $J(x)$ are continuous hermitian and real valued functions repsectively:

$$K(y,x) = \overline{K(x,y)}, \quad x,y \in S(\mathbb{k}^d),$$
$$J : S(\mathbb{k}^d) \to \mathbb{R}.$$

Then the corresponding form B is called a hermitian form.

Let $\Psi_d^{\mathbb{k}}$ be the set of continuous maps $\phi : S(\mathbb{k}^d) \to S(\mathbb{k}^d)$. Define the following analogs of the norms $\theta, \Theta, \gamma, \Gamma$ for hermitian B:

$$\begin{aligned}
\|B\|_\theta^{\mathbb{k}} &= \sup\{|\Re B(f,f)|, f \in L_\infty(S(\mathbb{k}^d)), \|f\|_\infty = 1\},\\
\|B\|_\Theta^{\mathbb{k}} &= \sup\{|\Re B(f,f)|, f \in L_\infty(S(\mathbb{k}^d)), \|f\|_\infty \le 1\},\\
\|B\|_\gamma^{\mathbb{k}} &= \sup\{|\Re B(\phi,\phi)|, \phi \in \Psi_d^{\mathbb{k}}\},\\
\|B\|_\Gamma^{\mathbb{k}} &= \sup\{|\Re B(\phi,\phi)|, \phi \in \Phi_d^{\mathbb{k}}\}
\end{aligned} \tag{7.62}$$

The following analogs of Lemma 7.5.2 is proves as Lemma 7.5.2 :

Lemma 7.5.3. *For each hermitian B on $S(\mathbb{k}^d) \times S(\mathbb{k}^d)$ the following inequality holds:*

$$\begin{aligned}
K_{\gamma,1,d}^{\mathbb{k}} \|B\|_\theta^{\mathbb{k}} &\ge \|B\|_{\gamma,d}^{\mathbb{k}},\\
K_{\Gamma,1,d}^{\mathbb{k}} \|B\|_\Theta^{\mathbb{k}} &\ge \|B\|_{\Gamma,d}^{\mathbb{k}}.
\end{aligned} \tag{7.63}$$

7.5.3 Davie's estimate for the real case

On S^{d-1} consider $K(x,y) = d\langle y,x\rangle$ and $J(x) = -\rho$. By choosing $\phi(x) = \psi(x) = x$ for $x \in \mathrm{S}^{d-1}$ we get

$$\|B\|_{G,n} \ge |n \int_{\mathrm{S}^{d-1}\times \mathrm{S}^{d-1}} \langle y,x\rangle\langle x,y\rangle d\sigma(x) d\sigma(y) - \rho|. \tag{7.64}$$

As σ is invariant for the transformation $x \mapsto Qx$ for any orthogonal Q, we can assume that $x = e_i = (\delta_{1i} \ldots, \delta_{id})^\top$. Hence

$$\int_{\mathrm{S}^{d-1}} \langle y,x\rangle\langle x,y\rangle d\sigma(y) = \int_{\mathrm{S}^{d-1}} y_i^2 d\sigma(y), \quad i \in [n]$$

Add these equalites for all $i \in [n]$ to deduce

$$\int_{\mathrm{S}^{d-1}} \langle y,x\rangle\langle x,y\rangle d\sigma(y) = \frac{1}{d},$$

for any $x \in \mathrm{S}^{d-1}$ and

$$\|B\|_{G,d} \geq |1-\rho|.$$

Denote $M_{\rho,d} = \|B\|_{\infty,1}$. Then

$$M_{\rho,d} = \sup\{d \int_{\|f\|_\infty,\|g\|_\infty \leq 1} \langle x,y\rangle f(x)g(y)d\sigma(x)d\sigma(y) - \rho \int f(x)g(x)d\sigma(x)\}. \tag{7.65}$$

Then

$$K_G^{\mathbb{R}} \geq K_{G,1,d}^{\mathbb{R}} \geq \sup_{0<\rho<1} \frac{1-\rho}{M_{\rho,d}}. \tag{7.66}$$

7.5.4 A lower bound for $K_G^{\mathbb{H}}$

Now we will use S^{d-1} to denote the unit sphere over $\mathbb{H}^d$. Now view $\langle x,y\rangle$ as an inner product over quaternions. Again, we can use the same function $K(x,y) = d\langle y,x\rangle, J(x) = -\rho, \phi(x) = \psi(x) = x$. We get

$$\|B\|_{G,d} \geq |d \int_{\mathrm{S}^{n-1}\times\mathrm{S}^{n-1}} \langle y,x\rangle\langle x,y\rangle d\sigma(x)d\sigma(y) - \rho| = |1-\rho|, \tag{7.67}$$

and $K_G^{\mathbb{H}} \geq \sup_{0<\rho<1} \frac{1-\rho}{M_{\rho,d}}$. We only need to find the bounds for $M_{\rho,n}$. Note that

$$M_{\rho,d} = \sup \Re\{d \int_{\|f\|_\infty,\|g\|_\infty \leq 1} \langle y,x\rangle f(x)g(y)d\sigma(x)d\sigma(y) - \rho \int f(x)g(x)d\sigma(x)\}.$$

Fix $f, g \in B$. Let $h(y) = \int \langle y, x\rangle f(x) d\sigma(x)$, then $h(y) = \langle y, z\rangle \lambda$ for a scalar λ and $z \in S$. Furthermore $\lambda = h(z) = \int \langle z, x\rangle f(x) d\sigma(x)$. Define $\mu = \int \langle g(y) \langle y, z\rangle d\sigma(y)$,

$$\begin{aligned} &\Re \int \int \langle y, x\rangle f(x) g(y) d\sigma(x) d\sigma(y) \\ &= \int \Re(\langle y, z\rangle \lambda g(y)) d\sigma(y) \\ &= \int \Re(g(y) \langle y, z\rangle \lambda) d\sigma(y) \\ &= \Re(\mu\lambda) \leq |\lambda + \bar{\mu}|^2/4. \end{aligned}$$

We know that $\lambda + \bar{\mu} = \int \langle z, x\rangle (f(x) + \bar{g}(x)) d\sigma(x)$. So

$$\begin{aligned} M_{\rho,d} &\leq \sup_{\|f\|_\infty, \|g\|_\infty \leq 1} \frac{d}{4} \left| \int \langle z, x\rangle (f(x) + \bar{g}(x)) d\sigma(x) \right|^2 - \rho \Re \int f(x) g(x) d\sigma(x) \\ &\leq \sup_{\|f\|_\infty, \|g\|_\infty \leq 1, \|z\|=1} \frac{d}{4} \left| \int \langle z, x\rangle (f(x) + \bar{g}(x)) d\sigma(x) \right|^2 - \rho \Re \int f(x) g(x) d\sigma(x) \\ &= \sup_{\|f\|_\infty, \|g\|_\infty \leq 1} \frac{d}{4} \left| \int x_1 (f(x) + \bar{g}(x)) d\sigma(x) \right|^2 - \rho \Re \int f(x) g(x) d\sigma(x). \end{aligned}$$

Now use the inequality $|f(x) + \bar{g}(x)|^2 \leq \Re(2 + 2f(x)g(x))$ and write

$$\psi(x) = \sqrt{\Re(1 + f(x)g(x))/2},$$

obtaining

$$M_{\rho,d} \leq \sup_{0 \leq \psi \leq 1} d \left| \int |x_1| \psi(x) d\sigma(x) \right|^2 + \rho(1 - 2 \int \psi(x)^2 d\sigma(x)).$$

By a simple variational argument, the maximum will be attained when $\psi(x) = \min(1, |x_1|/\lambda)$

for some $\lambda \geq 0$. So

$$\begin{aligned} M_{\rho,d} \leq d\Big| \int_{|x_1|>\lambda} |x_1| d\sigma(x) + \int_{|x_1|<\lambda} |x_1|^2/\lambda d\sigma(x)\Big|^2 \\ + \rho(1 - 2\int_{|x_1|>\lambda} 1 d\sigma(x) - 2\int_{|x_1|<\lambda} |x_1|^2/\lambda^2 d\sigma(x)). \end{aligned}$$

When $d \to \infty$, the distribution of $\sqrt{4d}x_1$ approaches quaternionic (4-dimensional) standard normal, so the distribution of $r = \sqrt{d}|x_1|$ tends to $8r^3 e^{-2r^2} dr$. We change λ to $\lambda/\sqrt{d}$ in the integral and obtain

$$\begin{aligned} M_{\rho,n} \leq & \Big| \int_{\sqrt{d}|x_1|>\lambda} \sqrt{d}|x_1| d\sigma(x) + \int_{\sqrt{d}|x_1|<\lambda} d|x_1|^2/\lambda d\sigma(x)\Big|^2 \\ & + \rho(1 - 2\int_{\sqrt{d}|x_1|>\lambda} 1 d\sigma(x) - 2\int_{\sqrt{d}|x_1|<\lambda} d|x_1|^2/\lambda^2 d\sigma(x)) \\ & \to \Big| \int_\lambda^\infty 8r^4 e^{-2r^2} dr + \frac{1}{\lambda}\int_0^\lambda 8r^5 e^{-2r^2} dr\Big|^2 \\ & + \rho(1 - 2\int_\lambda^\infty 8r^3 e^{-2r^2} dr - \frac{2}{\lambda^2}\int_0^\lambda 8r^5 e^{-2r^2} dr) \\ & := F(\lambda, \rho). \end{aligned}$$

We can take derivative with respect to λ and find that

$$\frac{\partial}{\partial\lambda} F(\lambda, \rho) = h(\lambda)(\rho - \theta(\lambda)),$$

where $h(\lambda) := (1 - (2\lambda^4 + 2\lambda^2 + 1)e^{-2\lambda^2})/\lambda^3 \geq 0$ and $\theta(\lambda) := \frac{1}{4}(2 - e^{-2\lambda^2}(\lambda^2 + 2) + 3\lambda \int_\lambda^\infty e^{-2r^2} dr)$. Furthermore, $\theta'(\lambda) = \frac{1}{4}(e^{-2\lambda^2}(4\lambda^3 + 3\lambda) + 3\int_\lambda^\infty e^{-2r^2} dr) > 0$. So $\theta(\lambda)$ is an increaing function with $\theta(0) = 0$ and $\lim_{\lambda\to\infty} \theta(\lambda) = 1/2$. For fixed $\rho \geq 1/2$, $\frac{\partial}{\partial\lambda} F(\lambda, \rho)$ is nonnegative, so $F(\lambda, \rho) \leq \lim_{\lambda\to\infty} F(\lambda, \rho) \leq \rho$. For fixed $\rho < 1/2$, $h(\lambda)(\rho - \theta(\lambda))$ changes sign once as λ increasing from 0 to ∞, so $F(\lambda, \rho) \leq F(\theta^{-1}(\rho), \rho)$.

In conclusion, we have found that $M_{\rho,d} \leq \rho$ if $\rho \geq \frac{1}{2}$, $M_{\rho,d} \leq F(\theta^{-1}(\rho), \rho)$ when $\rho < \frac{1}{2}$, so

$$\sup_{0<\rho<1} \frac{1-\rho}{M_{\rho,d}} \geq \sup_{\rho} \frac{1-\rho}{F(\theta^{-1}(\rho),\rho)} = \sup_{\lambda} \frac{1-\theta(\lambda)}{F(\lambda,\theta(\lambda))} \approx 1.17849,$$

where $\lambda \approx 0.64056, \rho \approx 0.295034$ are computed using Mathematica numerically. So $K_G^{\mathbb{H}} \geq 1.17849$.

7.5.5 Lower bounds for $K_\gamma^{\mathbb{H}}$ and $K_\Gamma^{\mathbb{H}}$

The computation of $K_\gamma^{\mathbb{H}}$ and $K_\Gamma^{\mathbb{H}}$ are similar and easier than $K_G^{\mathbb{H}}$. But first we need a technical lemma.

Lemma 7.5.4. *Let U_n, V_n be independent random vectors uniformly distributed on S^n sphere. Let $T_d = |\langle U_n, V_n\rangle|$. Then the probability density function of T_n is given by*

$$f_{T_n}(t) = 4n(2n+1)(1-t^2)^{2n-1}t^3.$$

For any $\alpha > 0$,

$$\mathbb{E}(|\langle U_n, V_n\rangle|^\alpha) = \frac{\Gamma(2n+2)\Gamma(2+\alpha/2)}{\Gamma(2n+2+\alpha/2)}.$$

Proof. We identify $\mathbb{H}^{n+1} = \mathbb{R}^{4n+4}$ and introduce the following coordinates for $x \in \mathbb{R}^{4n+4}$:

$$\begin{aligned}
x_1 &= \sqrt{r^2 - \rho^2}\cos(\vartheta_1),\\
x_2 &= \sqrt{r^2 - \rho^2}\sin(\vartheta_1)\cos(\vartheta_2),\\
x_3 &= \sqrt{r^2 - \rho^2}\sin(\vartheta_1)\sin(\vartheta_2)\cos(\vartheta_3),\\
&\ \ \vdots\\
x_{4n-1} &= \sqrt{r^2 - \rho^2}\sin(\vartheta_1)\cdots\sin(\vartheta_{4n-2})\cos(\vartheta_{4n-1}),\\
x_{4n} &= \sqrt{r^2 - \rho^2}\sin(\vartheta_1)\cdots\sin(\vartheta_{4n-2})\sin(\vartheta_{4n-1}),\\
x_{4n+1} &= \rho\cos\phi_1,\\
x_{4n+2} &= \rho\sin\phi_1\cos\phi_2,\\
x_{4n+3} &= \rho\sin\phi_1\sin\phi_2\cos\phi_3,\\
x_{4n+4} &= \rho\sin\phi_1\sin\phi_2\sin\phi_3,
\end{aligned}$$

where $r = \|x\|$, $\rho \in [0, r]$, $\vartheta_1, \dots, \vartheta_{4n-2}, \phi_1, \phi_2 \in [0, \pi]$, and $\vartheta_{4n-1}, \phi_3 \in [0, 2\pi)$.

The Jacobian $J = \partial(x_1, \dots, x_{4n+4})/\partial(r, \vartheta_1, \dots, \vartheta_{4n-1}, \rho, \phi_1, \phi_2, \phi_3)$ has determinant

$$\begin{aligned}
\det J = {} & r(r^2 - \rho^2)^{2n-1}\rho^3 \sin^{4n-2}(\vartheta_1)\sin^{4n-3}(\vartheta_2)\cdots\sin(\vartheta_{4n-2})\,drd\vartheta_1\,d\vartheta_2\cdots d\vartheta_{4n-1}\\
& \sin^2\phi_1\sin\phi_2 d\rho d\phi_1 d\phi_2 d\phi_3;
\end{aligned}$$

and since $r = 1$, we get

$$\begin{aligned}
d\sigma_n = {} & (1 - \rho^2)^{2n-1}\rho^3 \sin^{4n-2}(\vartheta_1)\sin^{4n-3}(\vartheta_2)\cdots\sin(\vartheta_{4n-2})\,d\vartheta_1\,d\vartheta_2\cdots d\vartheta_{4n-1}\\
& \sin^2\phi_1\sin\phi_2 d\rho d\phi_1 d\phi_2 d\phi_3.
\end{aligned}$$

Integrating out $\vartheta_1, \dots, \vartheta_{2n-1}, d\phi_1 d\phi_2 d\phi_3$ yields $C(1 - \rho^2)^{2n-1}\rho^3 d\rho$. And the constant C can be obtain by integral $1 = C\int_{-1}^{1}(1-\rho^2)^{2n-1}\rho^3 d\rho = C\int_0^1 x(1-x)^{2n-1}dx/2 = \frac{C}{4n(2n+1)}$.

So

$$f_{T_n}(t) = 4n(2n+1)(1-t^2)^{2n-1}t^3.$$

So

$$\begin{aligned}\mathbb{E}(|\langle U_n, V_n\rangle|^\alpha) &= \int_0^1 4n(2n+1)(1-t^2)^{2n-1}t^3t^\alpha dt\\ &= \int_0^1 2n(2n+1)(1-x)^{2n-1}x^{1+\alpha/2}dx\\ &= \frac{2n(2n+1)\Gamma(2n)\Gamma(2+\alpha/2)}{\Gamma(2n+2+\alpha/2)} = \frac{\Gamma(2n+2)\Gamma(2+\alpha/2)}{\Gamma(2n+2+\alpha/2)}.\end{aligned}$$

□

In fact, we have the following result:

Theorem 7.5.5. $K_\gamma^{\mathbb{H}} \geq \frac{64}{9\pi} - 1$, $K_{\gamma,d}^{\mathbb{H}} \geq \frac{32\Gamma^2(2d+1/2)}{9\pi d\Gamma^2(2d)} - 1$, $K_\Gamma^{\mathbb{H}} \geq 1.25709$.

Proof. We can use the same $B(f,f)$ defined by $K(x,y) = d\langle y, x\rangle$, $J(x) = -\rho$. Then by the same computation, we can get

$$\|B\|_{\gamma,d} \geq |1-\rho|, \|B\|_{\Gamma,d} \geq |1-\rho|$$

and

$$\|B\|_\theta = \sup\left\{\left|\Re d\int_{S^{d-1}\times S^{d-1}} \langle y,x\rangle\overline{f(x)}f(y)d\sigma(x)d\sigma(y) - \rho\right|, |f(x)| = 1\right\}.$$

We can further bound the first term by

$$\begin{aligned}&\Re\int_{S^{d-1}\times S^{d-1}} \langle y,x\rangle\overline{f(x)}f(y)d\sigma(x)d\sigma(y)\\ &= \|\int_{S^{d-1}} x\overline{f(x)}d\sigma(x)\|^2\\ &\leq \|\int_{S^{d-1}} |x_1|d\sigma(x)\|^2.\end{aligned}$$

The last inequality is sharp. In fact, we can assume $\int_{S^{d-1}} x\overline{f(x)}d\sigma(x) = (a, 0, \ldots, 0)$ in a suitable basis of $\mathbb{H}^d$. Then

$$\int_{S^{d-1}} x\overline{f(x)}d\sigma(x) = \int_{S^{d-1}} x_1\overline{f(x)}d\sigma(x) \leq \int_{S^{d-1}} |x_1|d\sigma(x).$$

So we have

$$\|B\|_\theta \leq \max\{|d| \int_{S^{d-1}} |x_1|d\sigma(x)|^2 - \rho|, \rho\}.$$

When $d \to \infty$, the distribution of $r = \sqrt{d}|x_1|$ tends to $8r^3 e^{-2r^2}dr$. So $d|\int_{S^{d-1}} |x_1|d\sigma(x)|^2$ converges to $\frac{9\pi}{32}$ and

$$K_\gamma^{\mathbb{H}} \geq \sup_{0<\rho<1} \frac{1-\rho}{\max\{\rho, |\frac{9\pi}{32} - \rho|\}} \geq \frac{64}{9\pi} - 1.$$

The lower bound matches the upper bound, so $K_\gamma^{\mathbb{H}} = \frac{64}{9\pi} - 1$. In fact, we can compute the constant for finite d.

$$d|\int_{S^{d-1}} |x_1|d\sigma(x)|^2 = \frac{9\pi d\Gamma^2(2d)}{16\Gamma^2(2d+1/2)},$$

So

$$K_{\gamma,d}^{\mathbb{H}} \geq \frac{32\Gamma^2(2d+1/2)}{9\pi d\Gamma^2(2d)} - 1.$$

For the third inequality, we have

$$\|B\|_\Theta = \sup\left\{ d\left| \left|\int_{S^{d-1}} x\overline{f(x)}d\sigma(x)\right|^2 - \rho\int_{S^{d-1}} |f(x)|^2 d\sigma(x)\right|, |f(x)| \leq 1 \right\}.$$

If $\rho\int_{S^{d-1}} |f(x)|^2 d\sigma(x) > \left|\int_{S^{d-1}} x\overline{f(x)}d\sigma(x)\right|^2$, then

$$\|B\|_\Theta \leq \rho\int_{S^{d-1}} |f(x)|^2 d\sigma(x) \leq \rho.$$

In the other case,

$$\|B\|_\Theta \leq d\left|\int_{S^{d-1}} x\overline{f(x)}d\sigma(x)\right|^2 - \rho\int_{S^{d-1}} |f(x)|^2 d\sigma(x).$$

By choose a suitable basis, we can replace the first term by $\left|\int_{S^{d-1}} x_1\overline{f(x)}d\sigma(x)\right|^2$. By replacing $\overline{f(x)}$ with $f(x)\bar{x}_1/|x_1|$, we can then consider instead

$$\sup\{d\left|\int_{S^{d-1}} |x_1||f(x)|d\sigma(x)\right|^2 - \rho\int_{S^{d-1}} |f(x)|^2 d\sigma(x)\}.$$

Then we can proceed as before and get

$$\begin{aligned}\|B\|_\Theta \leq &\left|\int_\lambda^\infty 8r^4e^{-2r^2}dr + \frac{1}{\lambda}\int_0^\lambda 8r^5e^{-2r^2}dr\right|^2\\ &-\rho(\int_\lambda^\infty 8r^3e^{-2r^2}dr + \frac{1}{\lambda^2}\int_0^\lambda 8r^5e^{-2r^2}dr)\\ :=& G(\lambda,\rho).\end{aligned}$$

We can take derivative with respect to λ and find that

$$\frac{\partial}{\partial\lambda}G(\lambda,\rho) = h(\lambda)(\rho/2 - \theta(\lambda)),$$

where $h(\lambda) := (1-(2\lambda^4+2\lambda^2+1)e^{-2\lambda^2})/\lambda^3 \geq 0$ and $\theta(\lambda) := \frac{1}{4}(2-e^{-2\lambda^2}(\lambda^2+2)+3\lambda\int_\lambda^\infty e^{-2r^2}dr)$ as before.

Furthermore, $\theta(\lambda)$ is an increaing function with $\theta(0)=0$ and $\lim_{\lambda\to\infty}\theta(\lambda)=1/2$. For fixed $\rho\geq 1$, $\frac{\partial}{\partial\lambda}G(\lambda,\rho)$ is nonnegative, so $G(\lambda,\rho)\leq\lim_{\lambda\to\infty}G(\lambda,\rho)\leq 0$. For fixed $\rho<1$, $h(\lambda)(\rho-\theta(\lambda))$ changes sign once as λ increasing from 0 to ∞, so $G(\lambda,\rho)\leq F(\theta^{-1}(\rho/2),\rho)$.

$$\sup_{0<\rho<1}\frac{1-\rho}{\|B\|_\Theta} \geq \sup_\rho\frac{1-\rho}{\max\{\rho, G(\theta^{-1}(\rho/2),\rho)\}} = \sup_\lambda\frac{1-2\theta(\lambda)}{\max\{2\theta(\lambda), G(\lambda, 2\theta(\lambda))\}} \approx 1.25709,$$

where $\lambda \approx 0.473831, \rho \approx 0.443047$. $\square$

REFERENCES

[1] Karim M Abadir and Paolo Paruolo. Two mixed normal densities from cointegration analysis. *Econometrica*, 65(3):671–680, 1997.

[2] Hiraku Abe and Tomoo Matsumura. Schur polynomials and weighted Grassmannians. *J. Algebraic Combin.*, 42(3):875–892, 2015.

[3] P.-A. Absil, R. Mahony, and R. Sepulchre. Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. *Acta Appl. Math.*, 80(2):199–220, 2004.

[4] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, Princeton, NJ, 2008.

[5] P.-A. Absil and Jérôme Malick. Projection-like retractions on matrix manifolds. *SIAM J. Optim.*, 22(1):135–158, 2012.

[6] Roy L. Adler, Jean-Pierre Dedieu, Joseph Y. Margulies, Marco Martens, and Mike Shub. Newton's method on Riemannian manifolds and a geometric model for the human spine. *IMA J. Numer. Anal.*, 22(3):359–390, 2002.

[7] Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Conference on Learning Theory*, pages 28–40, 2010.

[8] Alekh Agarwal, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Alexander Rakhlin. Stochastic convex optimization with bandit feedback. In *Advances in Neural Information Processing Systems*, pages 1035–1043, 2011.

[9] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide*. SIAM, Philadelphia, PA, third edition, 1999.

[10] Patrice Assouad. Espaces p-lisses et q-convexes, inégalités de Burkholder. In *Séminaire Maurey-Schwartz 1974–1975: Espaces L^p, applications radonifiantes et géométrie des espaces de Banach, Exp. No. XV*. 1975.

[11] Christine Bachoc, Renaud Coulangeon, and Gabriele Nebe. Designs in Grassmannian spaces and lattices. *J. Algebraic Combin.*, 16(1):5–19, 2002.

[12] R Raj Bahadur. A note on quantiles in large samples. *The Annals of Mathematical Statistics*, 37(3):577–580, 1966.

[13] Gah-Yi Ban and Cynthia Rudin. The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67(1):90–108, 2019.

[14] Alexander Barg and Dmitry Yu. Nogin. Bounds on packings of spheres in the Grassmann manifold. *IEEE Trans. Inform. Theory*, 48(9):2450–2454, 2002.

[15] Jonathan Barzilai and Jonathan M. Borwein. Two-point step size gradient methods. *IMA J. Numer. Anal.*, 8(1):141–148, 1988.

[16] Amir Beck and Luba Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM J. Optim.*, 23(4):2037–2060, 2013.

[17] Rajendra Bhatia. Linear algebra to quantum cohomology: the story of Alfred Horn's inequalities. *Amer. Math. Monthly*, 108(4):289–318, 2001.

[18] Rajendra Bhatia and John Holbrook. Noncommutative geometric means. *Math. Intelligencer*, 28(1):32–39, 2006.

[19] Rajendra Bhatia and Fuad Kittaneh. On the singular values of a product of operators. *SIAM J. Matrix Anal. Appl.*, 11(2):272–277, 1990.

[20] Rajendra Bhatia and Fuad Kittaneh. Notes on matrix arithmetic-geometric mean inequalities. *Linear Algebra Appl.*, 308(1-3):203–211, 2000.

[21] Rajendra Bhatia and Fuad Kittaneh. The matrix arithmetic-geometric mean inequality revisited. *Linear Algebra Appl.*, 428(8-9):2177–2191, 2008.

[22] Rabi Bhattacharya and Vic Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds. *The Annals of Statistics*, 31(1):1–29, 2003.

[23] Julius R Blum. Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics*, 25(4):737–744, 1954.

[24] Léon Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. In *Proceedings of the symposium on learning and data science, Paris*, 2009.

[25] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Physica-Verlag/Springer, Heidelberg, 2010.

[26] Jop Briët, Fernando Mário de Oliveira Filho, and Frank Vallentin. The positive semidefinite Grothendieck problem with rank constraint. In *Automata, languages and programming. Part I*, volume 6198 of *Lecture Notes in Comput. Sci.*, pages 31–42. Springer, Berlin, 2010.

[27] Jop Briët, Fernando Mário de Oliveira Filho, and Frank Vallentin. Grothendieck inequalities for semidefinite programs with rank constraint. *Theory Comput.*, 10:77–105, 2014.

[28] Mark Broadie, Deniz Cicek, and Assaf Zeevi. General bounds and finite-time improvement for the kiefer-wolfowitz stochastic approximation algorithm. *Operations Research*, 59(5):1211–1224, 2011.

[29] Roger W. Brockett. *Finite dimensional linear systems*, volume 74 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2015. Reprint of the 1970 original.

[30] Darshan Bryner. Endpoint geodesics on the stiefel manifold embedded in euclidean space. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1139–1159, 2017.

[31] A. R. Calderbank, R. H. Hardin, E. M. Rains, P. W. Shor, and N. J. A. Sloane. A group-theoretic framework for the construction of packings in Grassmannian spaces. *J. Algebraic Combin.*, 9(2):129–140, 1999.

[32] Joao R. Cardoso and F. Silva Leite. Exponentials of skew-symmetric matrices and logarithms of orthogonal matrices. *J. Comput. Appl. Math.*, 233(11):2867–2875, 2010.

[33] Raymond J Carroll. On almost sure expansions for m-estimates. *The Annals of Statistics*, 6(2):314–318, 1978.

[34] Elena Celledoni and Arieh Iserles. Methods for the approximation of the matrix exponential in a Lie-algebraic setting. *IMA J. Numer. Anal.*, 21(2):463–488, 2001.

[35] Shih-Kang Chao and Guang Cheng. A generalization of regularized dual averaging and its dynamics. *arXiv preprint arXiv:1909.10072*, 2019.

[36] Jeff Cheeger and David G. Ebin. *Comparison theorems in Riemannian geometry*. AMS Chelsea Publishing, Providence, RI, 2008.

[37] Elynn Y Chen, Rui Song, and Michael I Jordan. Reinforcement learning with heterogeneous data: estimation and inference. *arXiv preprint arXiv:2202.00088*, 2022.

[38] Han-Fu Chen, Tyrone E Duncan, and Bozenna Pasik-Duncan. A kiefer-wolfowitz algorithm with randomized differences. *IEEE Transactions on Automatic Control*, 44(3):442–453, 1999.

[39] Haoyu Chen, Wenbin Lu, and Rui Song. Statistical inference for online decision making: In a contextual bandit setting. *Journal of the American Statistical Association*, 116(533):240–255, 2021.

[40] Haoyu Chen, Wenbin Lu, and Rui Song. Statistical inference for online decision making via stochastic gradient descent. *Journal of the American Statistical Association*, 116(534):708–719, 2021.

[41] Haoyu Chen, Wenbin Lu, and Rui Song. Statistical inference for online decision making via stochastic gradient descent. *Journal of the American Statistical Association*, 116(534):708–719, 2021.

[42] Hung Chen. Lower rate of convergence for locating a maximum of a function. *The Annals of Statistics*, pages 1330–1334, 1988.

[43] Xi Chen, Jason D Lee, Xin T Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 2020.

[44] Xi Chen, Jason D Lee, Xin T Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 2020.

[45] Xi Chen, Zachary Owen, Clark Pixton, and David Simchi-Levi. A statistical learning approach to personalization in revenue management. *Management Science*, 68(3):1923–1937, 2022.

[46] Yasuko Chikuse. *Statistics on special manifolds*, volume 174 of *Lecture Notes in Statistics*. Springer-Verlag, New York, NY, 2003.

[47] E. S. Coakley, V. Rokhlin, and M. Tygert. A fast randomized algorithm for orthogonal projection. *SIAM J. Sci. Comput.*, 33(2):849–868, 2011.

[48] Andrew R Conn, Katya Scheinberg, and Luis N Vicente. *Introduction to derivative-free optimization*. Society for Industrial and Applied Mathematics, 2009.

[49] John H. Conway, Ronald H. Hardin, and Neil J. A. Sloane. Packing lines, planes, etc.: packings in Grassmannian spaces. *Experiment. Math.*, 5(2):139–159, 1996.

[50] J. X. da Cruz Neto, L. L. de Lima, and P. R. Oliveira. Geodesic algorithms in Riemannian geometry. *Balkan J. Geom. Appl.*, 3(2):89–100, 1998.

[51] Y. H. Dai and Y. Yuan. A nonlinear conjugate gradient method with a strong global convergence property. *SIAM J. Optim.*, 10(1):177–182, 1999.

[52] A. M. Davie. Lower bound for k_g. *unpublished note*, 1984.

[53] Christopher M De Sa. Random reshuffling is not always better. *Advances in Neural Information Processing Systems*, 33:5957–5967, 2020.

[54] N. Del Buono, L. Lopez, and R. Peluso. Computation of the exponential of large sparse skew-symmetric matrices. *SIAM J. Sci. Comput.*, 27(1):278–293, 2005.

[55] James W. Demmel. *Applied numerical linear algebra*. SIAM, Philadelphia, PA, 1997.

[56] Yash Deshpande, Lester Mackey, Vasilis Syrgkanis, and Matt Taddy. Accurate inference for adaptive linear models. In *International Conference on Machine Learning*, pages 1194–1203. PMLR, 2018.

[57] Jürgen Dippon. Accelerated randomized stochastic optimization. *The Annals of Statistics*, 31(4):1260–1281, 2003.

[58] Manfredo Perdigao do Carmo. *Riemannian geometry*. Mathematics: Theory & Applications. Birkhäuser, Boston, MA, 1992.

[59] David W. Dreisigmeyer. Direct search methods on reductive homogeneous spaces. *J. Optim. Theory Appl.*, 176(3):585–604, 2018.

[60] John C. Duchi. Commentary on "Towards a noncommutative arithmetic-geometric mean inequality" by B. Recht and C. Ré. *JMLR: Workshop and Conference Proceedings*, 23:11.25–11.27, 2012.

[61] John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

[62] John C Duchi and Feng Ruan. Asymptotic optimality in stochastic optimization. *The Annals of Statistics*, 49(1):21–48, 2021.

[63] Marie Duflo. *Random iterative models, volume 34 of Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1997.

[64] Rick Durrett. *Probability: theory and examples*. Cambridge University Press, Cambridge, 2019. Fifth edition.

[65] Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, 1999.

[66] Martin Ehler and Manuel Gräf. Reproducing kernels for the irreducible components of polynomial spaces on unions of Grassmannians. *Constr. Approx.*, 49(1):29–58, 2019.

[67] Vaclav Fabian. Stochastic approximation of minima with improved asymptotic speed. *The Annals of Mathematical Statistics*, pages 191–200, 1967.

[68] Vaclav Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4):1327–1332, 1968.

[69] Väclav Fabian. Stochastic approximation methods for constrained and unconstrained systems. *SIAM Review*, 22(3):382–384, 1980.

[70] Cong Fang, Zhouchen Lin, and Tong Zhang. Sharp analysis for nonconvex sgd escaping from saddle points. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1192–1234, Phoenix, USA, 2019.

[71] Yixin Fang, Jinfeng Xu, and Lei Yang. On scalable inference with stochastic gradient descent. *arXiv preprint arXiv:1707.00192*, 2017.

[72] Yixin Fang, Jinfeng Xu, and Lei Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *Journal of Machine Learning Research*, 2018.

[73] Miriam Farber and Alexander Postnikov. Arrangements of equal minors in the positive Grassmannian. *Adv. Math.*, 300:788–834, 2016.

[74] Douglas R Farenick and Barbara AF Pidkowich. The spectral theorem in quaternions. *Linear algebra and its applications*, 371:75–102, 2003.

[75] Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394, 2005.

[76] R. Fletcher and C. M. Reeves. Function minimization by conjugate gradients. *Comput. J.*, 7:149–154, 1964.

[77] Shmuel Friedland and Lek-Heng Lim. Symmetric grothendieck inequality. *arXiv preprint arXiv:2003.07345*, 2020.

[78] Shmuel Friedland, Lek-Heng Lim, and Jinjie Zhang. An elementary and unified proof of grothendieck's inequality. *arXiv preprint arXiv:1711.10595*, 2017.

[79] William Fulton and Joe Harris. *Representation theory*, volume 129 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1991. A first course, Readings in Mathematics.

[80] Pavel Galashin and Pavlo Pylyavskyy. Ising model and the positive orthogonal Grassmannian. *Duke Math. J.*, 169(10):1877–1942, 2020.

[81] Karin Gatermann and Pablo A. Parrilo. Symmetry groups, semidefinite programs, and sums of squares. *J. Pure Appl. Algebra*, 192(1-3):95–128, 2004.

[82] Evan S. Gawlik, Yuji Nakatsukasa, and Brian D. Sutton. A backward stable algorithm for computing the CS decomposition via the polar decomposition. *SIAM J. Matrix Anal. Appl.*, 39(3):1448–1469, 2018.

[83] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[84] D. G. Giovanis and M. D. Shields. Data-driven surrogates for high dimensional models using Gaussian process regression on the Grassmann manifold. *Comput. Methods Appl. Mech. Engrg.*, 370:113269, 2020.

[85] Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. Assoc. Comput. Mach.*, 42(6):1115–1145, 1995.

[86] Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013.

[87] Alexander Grothendieck. Résumé de la théorie métrique des produits tensoriels topologiques. *Bol. Soc. Mat. São Paulo*, 8:1–79, 1953.

[88] M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo. Why random reshuffling beats stochastic gradient descent. *Math. Program.*, 2019.

[89] Uffe Haagerup. The Grothendieck inequality for bilinear forms on C^*-algebras. *Adv. in Math.*, 56(2):93–116, 1985.

[90] Uffe Haagerup. A new upper bound for the complex Grothendieck constant. *Israel J. Math.*, 60(2):199–224, 1987.

[91] Peter Hall and Christopher C Heyde. *Martingale limit theory and its application.* Academic press, 1980.

[92] Peter Hall and Ilya Molchanov. Sequential methods for design-adaptive estimation of discontinuities in regression curves and surfaces. *The Annals of Statistics*, 31(3):921–941, 2003.

[93] John Hammersley. *Monte carlo methods.* Springer Science & Business Media, 2013.

[94] Botao Hao, Yasin Abbasi Yadkori, Zheng Wen, and Guang Cheng. Bootstrapping upper confidence bound. *Advances in Neural Information Processing Systems*, 32, 2019.

[95] Mehrtash Harandi, Richard Hartley, Chunhua Shen, Brian Lovell, and Conrad Sanderson. Extrinsic methods for coding and dictionary learning on Grassmann manifolds. *Int. J. Comput. Vis.*, 114(2-3):113–136, 2015.

[96] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities.* Cambridge Mathematical Library. Cambridge University Press, Cambridge, 1988. Reprint of the 1952 edition.

[97] Xuming He and Qi-Man Shao. A general bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *The Annals of Statistics*, 24(6):2608–2630, 1996.

[98] Ying He, Michael C Fu, and Steven I Marcus. Convergence of simultaneous perturbation stochastic approximation for nondifferentiable optimization. *IEEE Transactions on Automatic Control*, 48(8):1459–1463, 2003.

[99] Sigurdur Helgason. *Differential geometry, Lie groups, and symmetric spaces*, volume 34 of *Graduate Studies in Mathematics.* AMS, Providence, RI, 2001.

[100] Uwe Helmke, Knut Hüper, and Jochen Trumpf. Newton's method on Graßmann manifolds. *preprint*, arXiv:0709.2205, 2007.

[101] J. William Helton. "Positive" noncommutative polynomials are sums of squares. *Ann. of Math. (2)*, 156(2):675–694, 2002.

[102] J. William Helton, Igor Klep, and Scott McCullough. The convex Positivstellensatz in a free algebra. *Adv. Math.*, 231(1):516–534, 2012.

[103] Magnus R. Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *J. Research Nat. Bur. Standards*, 49:409–436 (1953), 1952.

[104] Nicholas J. Higham. *Accuracy and stability of numerical algorithms.* SIAM, Philadelphia, PA, second edition, 2002.

[105] Nicholas J Higham. The scaling and squaring method for the matrix exponential revisited. *SIAM Journal on Matrix Analysis and Applications*, 26(4):1179–1193, 2005.

[106] Nicholas J. Higham. *Functions of matrices.* SIAM, Philadelphia, PA, 2008.

[107] James E. Humphreys. *Linear algebraic groups.* Graduate Texts in Mathematics, No. 21. Springer-Verlag, New York-Heidelberg, 1975.

[108] Dale Husemoller. *Fibre bundles*, volume 20 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, third edition, 1994.

[109] Arie Israel, Felix Krahmer, and Rachel Ward. An arithmetic-geometric mean inequality for products of three matrices. *Linear Algebra Appl.*, 488:1–12, 2016.

[110] Kevin G Jamieson, Robert Nowak, and Ben Recht. Query complexity of derivative-free optimization. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2012.

[111] Bo Jiang and Yu-Hong Dai. A framework of constraint preserving update schemes for optimization on Stiefel manifold. *Math. Program.*, 153(2, Ser. A):535–575, 2015.

[112] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1724–1732, International Convention Centre, Sydney, Australia, 2017.

[113] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

[114] Siddharth Joshi and Stephen Boyd. Sensor selection via convex optimization. *IEEE Transactions on Signal Processing*, 57(2):451–462, 2008.

[115] Jürgen Jost. *Riemannian geometry and geometric analysis.* Universitext. Springer, Cham, seventh edition, 2017.

[116] V. Jurdjevic, I. Markina, and F. Silva Leite. Extremal curves on Stiefel and Grassmann manifolds. *J. Geom. Anal.*, 108(4):289–318, 2019.

[117] H. Karcher. Riemannian center of mass and mollifier smoothing. *Comm. Pure Appl. Math.*, 30(5):509–541, 1977.

[118] H. Karcher. Riemannian center of mass and so called Karcher mean. *preprint*, arXiv:1407.2087, 2014.

[119] Steven N. Karp. Sign variation, the Grassmannian, and total positivity. *J. Combin. Theory Ser. A*, 145:308–339, 2017.

[120] Maximilian Kasy and Anja Sautmann. Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1):113–132, 2021.

[121] Tosio Kato. *Perturbation theory for linear operators.* Classics in Mathematics. Springer-Verlag, Berlin, 1995.

[122] Koulik Khamaru, Yash Deshpande, Lester Mackey, and Martin J Wainwright. Near-optimal inference in adaptive linear regression. *arXiv preprint arXiv:2107.02266*, 2021.

[123] Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

[124] Nicholas M Kiefer, Timothy J Vogelsang, and Helle Bunzel. Simple robust testing of regression hypotheses. *Econometrica*, 68(3):695–714, 2000.

[125] Edward S Kim, Roy S Herbst, Ignacio I Wistuba, J Jack Lee, George R Blumenschein, Anne Tsao, David J Stewart, Marshall E Hicks, Jeremy Erasmus, Sanjay Gupta, et al. The BATTLE trial: Personalizing therapy for lung cancerthe BATTLE trial: Personalizing therapy for lung cancer. *Cancer Discovery*, 1(1):44–53, 2011.

[126] Shoshichi Kobayashi and Katsumi Nomizu. *Foundations of differential geometry. Vol. II.* Wiley Classics Library. John Wiley and Sons, New York, NY, 1996.

[127] Roger Koenker. *Quantile Regression.* Cambridge University Press, 2005.

[128] Jacek Koronacki. Random-seeking methods for the stochastic unconstrained optimization. *International Journal of Control*, 21(3):517–527, 1975.

[129] Jean-Louis Krivine. Constantes de Grothendieck et fonctions de type positif sur les sphères. *Adv. in Math.*, 31(1):16–30, 1979.

[130] Zehua Lai and Lek-Heng Lim. Recht-ré noncommutative arithmetic-geometric mean conjecture is false. In *International Conference on Machine Learning*, 2020.

[131] Zehua Lai, Lek-Heng Lim, and Ke Ye. Simpler grassmannian optimization, 2020.

[132] John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in Neural Information Processing Systems*, 20, 2007.

[133] J. B. Lasserre. A new Farkas lemma for positive semidefinite matrices. *IEEE Trans. Automat. Control*, 40(6):1131–1133, 1995.

[134] Jean Bernard Lasserre. *An introduction to polynomial and semi-algebraic optimization*, volume 52. Cambridge University Press, 2015.

[135] Ian Le and Chris Fraser. Tropicalization of positive Grassmannians. *Selecta Math. (N.S.)*, 25(5):Paper No. 75, 55 pp., 2019.

[136] Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Fast and robust online inference with stochastic gradient descent via random scaling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[137] Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Fast inference for quantile regression with millions of observations. *arXiv preprint arXiv:2209.14502*, 2022.

[138] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recoMendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[139] Tengyuan Liang and Weijie Su. Statistical inference for the population landscape via moment-adjusted stochastic gradients. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2):431–456, 2019.

[140] Joram Lindenstrauss and Aleksander Pełczyński. Absolutely summing operators in L_p-spaces and their applications. *Studia Math.*, 29:275–326, 1968.

[141] Nathan Mankovich, Emily J King, Chris Peterson, and Michael Kirby. The flag median and flagirls. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10339–10347, 2022.

[142] Pertti Mattila. *Geometry of sets and measures in Euclidean spaces*, volume 44 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1995.

[143] Ahmed Medra and Timothy N. Davidson. Incremental Grassmannian feedback schemes for multi-user MIMO systems. *IEEE Trans. Signal Process.*, 63(5):1130–1143, 2015.

[144] Bamdev Mishra, Hiroyuki Kasai, Pratik Jawanpuria, and Atul Saroop. A Riemannian gossip approach to subspace learning on Grassmann manifold. *Mach. Learn.*, 108(10):1783–1803, 2019.

[145] Bamdev Mishra and Rodolphe Sepulchre. Riemannian preconditioning. *SIAM J. Optim.*, 26(1):635–660, 2016.

[146] Abdelkader Mokkadem and Mariane Pelletier. A companion for the kiefer–wolfowitz–blum stochastic approximation algorithm. *The Annals of Statistics*, 35(4):1749–1772, 2007.

[147] Cleve Moler and Charles Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Rev.*, 45(1):3–49, 2003.

[148] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24, pages 451–459, 2011.

[149] Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. SGD without replacement: Sharper rates for general smooth convex functions. In *International Conference on Machine Learning*, pages 4703–4711, 2019.

[150] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2008.

[151] A. Nemirovski, C. Roos, and T. Terlaky. On maximization of quadratic form over intersection of ellipsoids with common center. *Math. Program.*, 86(3, Ser. A):463–473, 1999.

[152] Yu. Nesterov. Semidefinite relaxation and nonconvex quadratic optimization. *Optim. Methods Softw.*, 9(1-3):141–160, 1998.

[153] Yu. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM J. Optim.*, 22(2):341–362, 2012.

[154] Yurii Nesterov. Lexicographic differentiation of nonsmooth functions. *Mathematical programming*, 104(2):669–700, 2005.

[155] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[156] Du Nguyen. Closed-form geodesics and optimization for riemannian logarithms of stiefel and flag manifolds. *Journal of Optimization Theory and Applications*, 194(1):142–166, 2022.

[157] Liviu I. Nicolaescu. *Lectures on the geometry of manifolds.* World Scientific, Hackensack, NJ, second edition, 2007.

[158] Jorge Nocedal and Stephen J. Wright. *Numerical optimization.* Springer Series in Operations Research and Financial Engineering. Springer, New York, NY, second edition, 2006.

[159] J. E. Pascoe. Positivstellensätze for noncommutative rational expressions. *Proc. Amer. Math. Soc.*, 146(3):933–937, 2018.

[160] E. Polak and G. Ribière. Note sur la convergence de méthodes de directions conjuguées. *Rev. Française Informat. Recherche Opérationnelle*, 3(16):35–43, 1969.

[161] B. T. Polyak and A. B. Tsybakov. Optimal order of accuracy of search algorithms in stochastic optimization. *Problemy Peredachi Informatsii*, pages 126–133, 1990.

[162] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[163] Sheng Qiang and Mohsen Bayati. Dynamic pricing with demand covariates. *arXiv preprint arXiv:1604.07463*, 2016.

[164] Pratik Ramprasad, Yuantong Li, Zhuoran Yang, Zhaoran Wang, Will Wei Sun, and Guang Cheng. Online bootstrap inference for policy evaluation in reinforcement learning. *Journal of the American Statistical Association*, pages 1–14, 2022.

[165] Benjamin Recht and Christopher Ré. Toward a noncommutative arithmetic-geometric mean inequality: Conjectures, case-studies, and consequences. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 11.1–11.24, Edinburgh, Scotland, 2012.

[166] Ronald E. Rietz. A proof of the Grothendieck inequality. *Israel J. Math.*, 19:271–276, 1974.

[167] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

[168] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[169] R Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471, 2002.

[170] David Ruppert. Almost Sure Approximations to the Robbins-Monro and Kiefer-Wolfowitz Processes with Dependent Noise. *The Annals of Probability*, 10(1):178 – 187, 1982.

[171] David Ruppert. A newton-raphson version of the multivariate robbins-monro procedure. *The Annals of Statistics*, 13(1):236–245, 1985.

[172] David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

[173] Alain Sarlette and Rodolphe Sepulchre. Consensus optimization on manifolds. *SIAM J. Control Optim.*, 48(1):56–76, 2009.

[174] Ohad Shamir. Without-replacement sampling for stochastic gradient methods. In *Advances in neural information processing systems*, pages 46–54, 2016.

[175] Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(1):1703–1713, 2017.

[176] Qi-Man Shao and Zhuo-Song Zhang. Berry–esseen bounds for multivariate nonlinear statistics with applications to M-estimators and stochastic gradient descent algorithms. *Bernoulli*, 28(3):1548–1576, 2022.

[177] Chengchun Shi, Rui Song, Wenbin Lu, and Runze Li. Statistical inference for high-dimensional models via recursive online-score estimation. *Journal of the American Statistical Association*, 116(535):1307–1318, 2021.

[178] Chengchun Shi, Shengxing Zhang, Wenbin Lu, and Rui Song. Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 2021.

[179] Chengchun Shi, Jin Zhu, Shen Ye, Shikai Luo, Hongtu Zhu, and Rui Song. Off-policy confidence interval estimation with confounded markov decision process. *Journal of the American Statistical Association*, pages 1–12, 2022.

[180] Michael Shub. Some remarks on dynamical systems and numerical analysis. In *Dynamical systems and partial differential equations (Caracas, 1984)*, pages 69–91. Univ. Simon Bolivar, Caracas, 1986.

[181] G. Sonnevend, J. Stoer, and G. Zhao. On the complexity of following the central path of linear programs by linear extrapolation. II. volume 52, pages 527–553. 1991.

[182] James C Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.

[183] James C Spall. Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Transactions on Automatic Control*, 45(10):1839–1853, 2000.

[184] G. W. Stewart. Computing the CS decomposition of a partitioned orthonormal matrix. *Numer. Math.*, 40(3):297–306, 1982.

[185] G. W. Stewart. *Matrix algorithms I: Basic decompositions.* SIAM, Philadelphia, PA, 1998.

[186] Gilbert Strang. On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.*, 5:506–517, 1968.

[187] Thomas Strohmer and Roman Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *J. Fourier Anal. Appl.*, 15(2):262–278, 2009.

[188] Jos F Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization methods and software*, 11(1-4):625–653, 1999.

[189] Weijie J Su and Yuancheng Zhu. Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent. *arXiv preprint arXiv:1802.04876*, 2018.

[190] Yue Sun, Nicolas Flammarion, and Maryam Fazel. Escaping from saddle points on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, pages 7276–7286, 2019.

[191] Panos Toulis and Edoardo M Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017.

[192] Lloyd N. Trefethen and David Bau, III. *Numerical linear algebra.* SIAM, Philadelphia, PA, 1997.

[193] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[194] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[195] Martin J Wainwright and Michael Irwin Jordan. *Graphical models, exponential families, and variational inference.* Now Publishers Inc, 2008.

[196] Yining Wang, Simon Du, Sivaraman Balakrishnan, and Aarti Singh. Stochastic zeroth-order optimization in high dimensions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 1356–1365. PMLR, 2018.

[197] David S. Watkins. *Fundamentals of matrix computations.* Pure and Applied Mathematics. John Wiley and Sons, Hoboken, NJ, third edition, 2010.

[198] Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Math. Program.*, 142(1-2, Ser. A):397–434, 2013.

[199] J. H. Wilkinson. *The algebraic eigenvalue problem.* Monographs on Numerical Analysis. Oxford University Press, New York, NY, 1988.

[200] Yung-Chow Wong. Differential geometry of Grassmann manifolds. *Proc. Nat. Acad. Sci. U.S.A.*, 57:589–594, 1967.

[201] Michael Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979.

[202] Stephen J. Wright. Coordinate descent algorithms. *Math. Program.*, 151(1, Ser. B):3–34, 2015.

[203] Ke Ye and Lek-Heng Lim. Schubert varieties and distances between subspaces of different dimensions. *SIAM J. Matrix Anal. Appl.*, 37(3):1176–1197, 2016.

[204] Ke Ye, K. S. Wong, and L. Lim. Optimization on flag manifolds. *arXiv: Optimization and Control*, 2019.

[205] Kelly Zhang, Lucas Janson, and Susan Murphy. Statistical inference with M-estimators on adaptively collected data. *Advances in Neural Information Processing Systems*, 34:7460–7471, 2021.

[206] Kelly W Zhang, Lucas Janson, and Susan A Murphy. Statistical inference after adaptive sampling in non-markovian environments. *arXiv preprint arXiv:2202.07098*, 2022.

[207] Teng Zhang. A note on the matrix arithmetic-geometric mean inequality. *Electron. J. Linear Algebra*, 34:283–287, 2018.

[208] Gongyun Zhao. Representing the space of linear programs as the Grassmann manifold. *Math. Program.*, 121(2, Ser. A):353–386, 2010.

[209] Lizhong Zheng and David N. C. Tse. Communication on the Grassmann manifold: a geometric approach to the noncoherent multiple-antenna channel. *IEEE Trans. Inform. Theory*, 48(2):359–383, 2002.

[210] Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, page preprint, 2021.

[211] Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, pages 1–12, 2021.

[212] Wolfgang Ziller. Examples of Riemannian manifolds with non-negative sectional curvature. In *Surveys in differential geometry. Vol. XI*, volume 11 of *Surv. Differ. Geom.*, pages 63–102. International Press, Somerville, MA, 2007.

[213] Ralf Zimmermann and Knut Hüper. Computing the riemannian logarithm on the stiefel manifold: Metrics, methods, and performance. *SIAM Journal on Matrix Analysis and Applications*, 43(2):953–980, 2022.