Higher order concentration on Stiefel and Grassmann manifolds[*]

Friedrich Götze[†]   Holger Sambale[‡]

Abstract

We prove higher order concentration bounds for functions on Stiefel and Grassmann manifolds equipped with the uniform distribution. This partially extends previous work for functions on the unit sphere. Technically, our results are based on logarithmic Sobolev techniques for the uniform measures on the manifolds. Applications include Hanson–Wright type inequalities for Stiefel manifolds and concentration bounds for certain distance functions between subspaces of $\mathbb{R}^n$.

1 Introduction

In recent years, functions on the Stiefel manifold (i. e., the collection of all d-tuples of orthonormal vectors in $\mathbb{R}^n$) and on the closely related Grassmann manifold (the collection of all d-dimensional linear subspaces of $\mathbb{R}^n$) have attracted increasing attention from various fields of research. A central reason is that they admit a wealth of applications in various directions like data analysis, subspace estimation, computer vision or statistical learning. In addition, they have also been studied from more theoretical points of view, in particular in convex geometry but quite recently also in large deviations theory. For a selection of various results, see e. g. [TVC08, HL08, WLT11, ZZJH18, LLY20] (data analysis) and [LPT06, GKZ21, KP21] (including large deviation principles).

A frequent question which naturally arises in many applications is how to control the fluctuations of some (Stiefel or Grassmann) functional around a typical value like its expectation, i. e., one asks for suitable concentration of measure results. Often times, the functional is of Lipschitz-type, which in classical situations (e. g., functions of

independent sub-Gaussian random variables or in presence of a log-Sobolev inequality) leads to sub-Gaussian tail bounds. For a brief review especially adapted to Stiefel and Grassmann manifolds, cf. Section 1.2 below. However, this approach fails if the functionals under consideration are no longer Lipschitz. Still, one may hope for useful concentration bounds if they are Lipschitz of a "higher order", which typically means that their higher order derivatives are absolutely bounded. A classical example for order two are quadratic forms. In this situation, various questions of interest include exponential moment bounds, refined tail bounds with various levels of decay and centering around non-deterministic quantities which may correspond to a suitable decomposition.

In this note, we provide higher order concentration results for Stiefel and Grassmann manifolds equipped with the uniform distribution. Higher order concentration has been studied in various settings in the past two decades, e. g. in [La06] (polynomials in independent Gaussian variables) [AW15] (measures satisfying certain Sobolev-type inequalities) as well as [KV00, SS12, AL12, KL15, GSS21a, GSS21b]. In this paper, we especially continue the line of research begun in [BCG17, GSS19], where second and higher order concentration results for the unit sphere were established (noting that the unit sphere can be understood as a Stiefel manifold).

1.1 Grassmann and Stiefel manifolds

For any natural numbers $d \le n$, the Stiefel manifold $W_{n,d}$ is the set of all d-tupels of orthonormal vectors in $\mathbb{R}^n$. Clearly, $W_{n,d}$ may be written as

$$W_{n,d} = \{A \in \mathbb{R}^{n \times d} \colon A^T A = I_d\},$$

where $I_d \in \mathbb{R}^{d \times d}$ denotes the identity matrix. This is the representation we shall use throughout this note. $W_{n,d}$ is manifold of dimension $\dim(W_{n,d}) = nd - d(d+1)/2$. Obviously, $W_{n,1} = S^{n-1}$ (the unit sphere), $W_{n,n} = O(n)$ (the orthogonal group), and $W_{n,n-1}$ can be identified with the special orthogonal group $SO(n) = \{O \in O(n) \colon \det(O) = 1\}$.

We may equip $W_{n,d}$ with the subspace topology and distances inherited from $\mathbb{R}^{n \times d}$, including the scalar product on $\mathbb{R}^{n \times d}$ and the induced (Hilbert–Schmidt) norm

$$\langle A, B \rangle := \operatorname{tr}(A^T B), \qquad |A|^2 \equiv \|A\|_{\mathrm{HS}}^2 = \operatorname{tr}(A^T A) = \sum_{i,j} A_{ij}^2.$$

In particular, $W_{n,d}$ is a compact topological space. The product of the orthogonal groups $O(n) \times O(d)$ acts transitively on $W_{n,d}$ by the two-sided multiplication $O_n \times O_d \mapsto O_n A O_d$, turning $W_{n,d}$ into a homogeneous space. Hence, it may be equipped with a unique invariant (Haar) probability measure $\mu_{n,d}$ in the sense that if $A \sim \mu_{n,d}$ (i. e., A has distribution $\mu_{n,d}$), $O_n A O_d \sim \mu_{n,d}$ for any $O_n \in O(n)$, $O_d \in O(d)$. We call $\mu_{n,d}$ the uniform distribution on $W_{n,d}$. If $G = (G_{ij})_{i,j=1}^{d,n}$ is an $n \times d$ random matrix whose entries are i.i.d. standard normal, then $G(G^T G)^{-1/2} \sim \mu_{n,d}$, see [KPT20, Lemma 3.1].

There are several ways of introducing the closely related Grassmann manifold, i. e. the set of all d-dimensional subspaces of $\mathbb{R}^n$. For an overview, cf. e. g. [BZA20]. For instance, identifying a subspace with its basis (which is unique up to orthogonal transformations), we may understand the Grassmann manifold as the quotient $W_{n,d}/O(d)$, identifying any two $X, X' \in W_{n,d}$ such that $X' = XO$ for some $O \in O(d)$. However, the calculus of (higher order) derivatives on these equivalence classes turns out to be troublesome, and therefore, we choose a different approach.

For any $A \in W_{n,d}$, $\pi_{\mathrm{St}}(A) := AA^T \equiv P_A$ is a projection matrix of rank d (more precisely, the projection onto the subspace a basis of which is given by the columns of A), and for any $O_d \in O(d)$, we have $P_{AO_d} = P_A$. Therefore, identifying elements of the

Grassmannian with projection matrices, we may define

$$G_{n,d} := \{P \in \mathbb{R}_{\text{sym}}^{n \times n}\colon P^2 = P,\ \text{rank}(P) = d\},$$

where $\mathbb{R}_{\text{sym}}^{n \times n}$ denotes the space of the symmetric $n \times n$ matrices. $G_{n,d}$ is a manifold of dimension $\dim(G_{n,d}) = d(n-d)$. If $d = 1$, we get back the half-sphere (where θ and $-\theta$ are identified), which can also be regarded as the projective space $\mathbb{R}P^{n-1}$. Depending on the situation, we will either regard $G_{n,d}$ as a submanifold of $\mathbb{R}_{\text{sym}}^{n \times n}$ or $\mathbb{R}^{n \times n}$ (the latter is often more convenient when taking derivatives), equipping it with the inherited topology and distances similarly to the case of the Stiefel manifold (in particular, $G_{n,d}$ is compact).

Many of the properties of Stiefel manifolds can be extended to Grassmann manifolds. Especially, the group action $O(n) \ni O_n \mapsto P_{O_nA}$ turns $G_{n,d}$ into a symmetric (not just homogeneous) space. We denote the uniform distribution on $G_{n,d}$ by $\nu_{n,d}$. Clearly, $\nu_{n,d}$ is the pushforward of $\mu_{n,d}$ under the map π_{St}. If $G = (G_{ij})_{i,j=1}^{d,n}$ is an $n \times d$ random matrix whose entries are i.i.d. standard normal, it follows from the discussion above that $G(G^TG)^{-1}G^T \sim \nu_{n,d}$.

1.2 Concentration results

If $f\colon W_{n,d} \to \mathbb{R}$ is L-Lipschitz with mean $\mu_{n,d}(f)$ with respect to $\mu_{n,d}$, a standard concentration result (cf. e.g. [Le01, p. 27]) yields

$$\mu_{n,d}(f - \mu_{n,d}(f) \geq t) \leq \exp(-(n-1)t^2/(8L^2)). \tag{1.1}$$

By switching to the pushforward, the same result also holds for $(G_{n,d}, \nu_{n,d})$ with the constant 8 replaced by 16 (cf. Section 6 for details, in particular Lemma 6.1).

The aim of this note is to extend results of type (1.1) to higher orders, involving $\mathcal{C}^k$ functions and derivatives up to order k. Here, a $\mathcal{C}^k$ function on $W_{n,d}$ (likewise, $G_{n,d}$) may be understood as a function which admits an extension to some open neighborhood of the manifold which is $\mathcal{C}^k$-smooth. Let us first fix some notation. If $A = (a_{i_1 \dots i_k}) \in \mathbb{R}^{m^k}$ is any matrix or tensor of order k, we write

$$\|A\|_{\text{HS}} := \Big(\sum_{i_1, \dots, i_k} a_{i_1 \dots i_k}^2 \Big)^{1/2},$$

$$\|A\|_{\text{op}} := \sup \Big\{ \sum_{i_1, \dots, i_k} a_{i_1 \dots i_k} x_{i_1}^{(1)} \cdots x_{i_k}^{(k)}\colon x^{(j)} \in S^{m-1}\ \forall j \Big\}$$

for the respective Hilbert–Schmidt and operator type norms. For any function $g\colon W_{n,d} \to \mathbb{R}$ and any $p \geq 1$, we denote by $\|g\|_p$ the L^p norm of g with respect to $\mu_{n,d}$. If g is matrix-valued (e.g., some tensor of derivatives), we use the short-hand notation

$$\|g\|_{\text{HS},p} := \big\| \|g\|_{\text{HS}} \big\|_p = \Big(\int_{W_{n,d}} \|g\|_{\text{HS}}^p d\mu_{n,d} \Big)^{1/p}$$

and similarly $\|g\|_{\text{op},p}$.

In Section 3, we will introduce a notion of differentiability for functions on $W_{n,d}$. In particular, using the intrinsic (Stiefel) gradient $\nabla_W f \in \mathbb{R}^{n \times d}$ and assuming $f\colon W_{n,d} \to \mathbb{R}$ to be $\mathcal{C}^1$-smooth, we may reformulate (1.1) as

$$\mu_{n,d}(f - \mu_{n,d}(f) \geq t) \leq \exp\Big(-\frac{(n-1)t^2}{8\|\nabla_W f\|_{\text{HS},\infty}^2} \Big). \tag{1.2}$$

The same result holds for $\mathcal{C}^1$-smooth functions on $(G_{n,d}, \nu_{n,d})$, involving the Grassmann gradient $\nabla_G f \in \mathbb{R}_{\text{sym}}^{n \times n}$ and with 8 replaced by 16.

Our first theorem complements (1.2) by a second order concentration bound for functions on Stiefel manifolds of any order $d \leq n - 1$ in the spirit of [BCG17] (cf. also [SS21], where this approach to second order concentration has been generalized and adapted to a wealth of different situations). In addition to the intrinsic gradient, this also involves an intrinsic Hessian $f_W'' \in \mathbb{R}^{nd \times nd}$, a notion to be made precise in Section 3 as well.

Theorem 1.1. *Let* $f \colon W_{n,d} \to \mathbb{R}$ *be a* $\mathcal{C}^2$-*smooth function such that* $\mu_{n,d}(f) = 0$.

1. *Assuming* $\|\nabla_W f\|_{\mathrm{HS},2} \leq 2/\sqrt{n-2}$ *and* $\|f_W''\|_{\mathrm{op},\infty} \leq 1$, *we have*

$$\int_{W_{n,d}} \exp\Big(\frac{n-2}{32e}|f|\Big) d\mu_{n,d} \leq 2.$$

2. *For* $C = 16e^2/\log(2)$ *and any* $t \geq 0$,

$$\mu_{n,d}(|f - \mu_{n,d}(f)| \geq t) \leq 2\exp\Big(-\frac{n-2}{C}\min\Big(\frac{t^2}{\|\nabla_W f\|_{\mathrm{HS},2}^2}, \frac{t}{\|f_W''\|_{\mathrm{op},\infty}}\Big)\Big).$$

3. *If* $\mu_{n,d}(\nabla_W f) = 0$, *the bounds in parts 1 and 2 continue to hold with* $\|\nabla_W f\|_{\mathrm{HS},2}$ *replaced by* $\sqrt{8/(n-2-8d)}\|f_W''\|_{\mathrm{HS},2}$.

Note that in part 3, the integral $\mu_{n,d}(\nabla_W f)$ has to be understood componentwise. In fact, the condition in part 3 can also be modified to $\|f_W''\|_{\mathrm{HS},2} \leq b$ similarly as in [BCG17, Theorem 1.1] (in particular, arriving at b-dependent bounds). We skip the details. With only minor modifications, these results also hold for functions on Grassmann manifolds:

Theorem 1.2. *Let* $f \colon G_{n,d} \to \mathbb{R}$ *be a* $\mathcal{C}^2$-*smooth function such that* $\nu_{n,d}(f) = 0$.

1. *Assuming* $\|\nabla_G f\|_{\mathrm{HS},2} \leq \sqrt{8/(n-2)}$ *and* $\|f_G''\|_{\mathrm{op},\infty} \leq 1$, *we have*

$$\int_{G_{n,d}} \exp\Big(\frac{n-2}{64e}|f|\Big) d\mu_{n,d} \leq 2.$$

2. *For* $C = 32e^2/\log(2)$ *and any* $t \geq 0$,

$$\nu_{n,d}(|f - \mu_{n,d}(f)| \geq t) \leq 2\exp\Big(-\frac{n-2}{C}\min\Big(\frac{t^2}{\|\nabla_G f\|_{\mathrm{HS},2}^2}, \frac{t}{\|f_G''\|_{\mathrm{op},\infty}}\Big)\Big).$$

3. *If* $\nu_{n,d}(\nabla_G f) = 0$, *the bounds in parts 1 and 2 continue to hold with* $\|\nabla_G f\|_{\mathrm{HS},2}$ *replaced by* $\sqrt{16/(n-2-16d)}\|f_G''\|_{\mathrm{HS},2}$.

In general, once we have a function f which is defined and smooth in a neighborhood of $W_{n,d}$ (which may be achieved by choosing a suitable extension of f), we may also formulate concentration of measure results involving the usual (Euclidean) derivatives only. The same holds for functions defined in a neighborhood of $G_{n,d}$ in $\mathbb{R}_{\mathrm{sym}}^{n \times n}$ or $\mathbb{R}^{n \times n}$.

As shown in Sections 3 and 5, we always have $|\nabla_W f|, |\nabla_G f| \leq |\nabla f|$. In particular, results depending on Euclidean derivatives are somewhat less accurate than bounds involving intrinsic derivatives. On the other hand, the underlying calculus is typically much less involved, especially as the order of the derivatives increases, and in many situations, the usual derivatives may already be sufficient for meaningful concentration bounds. Here, we have the following result, which complements [BGS19].

Theorem 1.3. *Let* f *be a real-valued* $\mathcal{C}^k$-*smooth function defined in some neighborhood of* $W_{n,d}$ *such that* $\mu_{n,d}(f) = 0$.

1. Assuming $\|f^{(\ell)}\|_{\mathrm{op},2} \le (4/(n-2))^{(k-\ell)/2}$ for any $\ell = 1, \ldots, k-1$ and $\|f^{(k)}\|_{\mathrm{op},\infty} \le 1$, we have

$$\int_{W_{n,d}} \exp\Big(\frac{n-2}{32e}|f|^{2/k}\Big) d\mu_{n,d} \le 2.$$

2. For $C = 4e^2/\log(2)$ and any $t \ge 0$,

$$\mu_{n,d}(|f| \ge t) \le 2\exp\Big(-\frac{n-2}{Ck^2}\min\Big(\min_{\ell=1,\ldots,k-1}\frac{t^{2/\ell}}{\|f^{(\ell)}\|_{\mathrm{op},2}^{2/\ell}}, \frac{t^{2/k}}{\|f^{(k)}\|_{\mathrm{op},\infty}^{2/k}}\Big)\Big).$$

The Grassmann version of this theorem reads as follows:

Theorem 1.4. Let f be a real-valued C^k-smooth function defined in some neighborhood of $G_{n,d}$ such that $\nu_{n,d}(f) = 0$.

1. Assuming $\|f^{(\ell)}\|_{\mathrm{op},2} \le (8/(n-2))^{(k-\ell)/2}$ for any $\ell = 1, \ldots, k-1$ and $\|f^{(k)}\|_{\mathrm{op},\infty} \le 1$, we have

$$\int_{G_{n,d}} \exp\Big(\frac{n-2}{64e}|f|^{2/k}\Big) d\nu_{n,d} \le 2.$$

2. For $C = 8e^2/\log(2)$ and any $t \ge 0$,

$$\nu_{n,d}(|f| \ge t) \le 2\exp\Big(-\frac{n-2}{Ck^2}\min\Big(\min_{\ell=1,\ldots,k-1}\frac{t^{2/\ell}}{\|f^{(\ell)}\|_{\mathrm{op},2}^{2/\ell}}, \frac{t^{2/k}}{\|f^{(k)}\|_{\mathrm{op},\infty}^{2/k}}\Big)\Big).$$

Examples where these results are applied to concrete situations and are briefly compared to related bounds known from the literature are given in Section 7.

1.3 Overview

In view of the parametrizations of the manifolds via matrices, we first provide a brief review on some basic facts about derivatives in matrix arguments and related topics in Section 2. First and second order intrinsic derivatives for functions on Stiefel manifolds are introduced and discussed in Sections 3 and 4, including a number of inequalities relating first and second order. These results are adapted to Grassmann manifolds in Section 5. The proofs of our main results are given in Section 6. Applications to polynomial chaos (in particular of order 2), including concentration of the distance to a given subspace, are provided in Section 7.

2 Basic facts about matrix calculus

The aim of this section is to fix some notation and to collect some basic facts about matrix calculus mostly for the sake of reference. Most of the results are elementary and will be stated without proofs.

If $A = (A_{ij})$ is an $n \times m$ matrix (i.e., $A \in \mathbb{R}^{n \times m}$), we may vectorize it by setting

$$\mathrm{vec}(A) := (A_{11}, \ldots, A_{n1}, A_{12}, \ldots, A_{n2}, \ldots, A_{1m}, \ldots, A_{nm})^T,$$

i.e. $\mathrm{vec}(A)$ is the vector in $\mathbb{R}^{nm}$ with m n-blocks corresponding to the columns of A. Occasionally, we will also need the inverse operation of vec, which we denote by mat. For $B \in \mathbb{R}^{nd \times nd}$ and $U, V \in \mathbb{R}^{n \times d}$, we introduce the short-hand notation

$$BV := B\mathrm{vec}(V), \qquad \langle BV, U\rangle := \langle B\mathrm{vec}(V), \mathrm{vec}(U)\rangle. \tag{2.1}$$

There is an $nm \times nm$ permutation matrix $K_{n,m}$ such that for any $A \in \mathbb{R}^{n \times m}$,

$$\mathrm{vec}(A^T) = K_{n,m}\mathrm{vec}(A). \tag{2.2}$$

$K_{n,m}$ is uniquely defined by (2.2) and is called the (n, m) commutation matrix. Note that $K_{n,m}^T = K_{m,n}$.

Vectorization is closely related to the Kronecker product $A \otimes B$ for any $A \in \mathbb{R}^{n \times m}$ and any $B \in \mathbb{R}^{p \times q}$. The following lemma lists some of its elementary properties.

Lemma 2.1. *1. The Kronecker product is bilinear and associative.*

2. *We have $(A \otimes B)^T = A^T \otimes B^T$.*

3. *$A \otimes B$ is invertible whenever A and B are invertible, and in this case, $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.*

4. *For any $A \in \mathbb{R}^{n \times m}, B \in \mathbb{R}^{p \times q}, C \in \mathbb{R}^{m \times r}, D \in \mathbb{R}^{q \times s}$, we have $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$.*

5. *For any $A \in \mathbb{R}^{n \times m}, B \in \mathbb{R}^{m \times p}$, we have $\mathrm{vec}(AB) = (I_p \otimes A)\mathrm{vec}(B) = (B^T \otimes I_n)\mathrm{vec}(A)$.*

6. *For any $A \in \mathbb{R}^{n \times m}, B \in \mathbb{R}^{m \times p}, C \in \mathbb{R}^{p \times q}$, we have $\mathrm{vec}(ABC) = (C^T \otimes A)\mathrm{vec}(B) = (I_q \otimes AB)\mathrm{vec}(C) = (C^T B^T \otimes I_n)\mathrm{vec}(A)$.*

7. *For any $A \in \mathbb{R}^{n \times m}$ and any $B \in \mathbb{R}^{p \times q}$, we have $K_{p,n}(A \otimes B)K_{m,q} = B \otimes A$.*

If $f \colon \mathbb{R}^{n \times m} \to \mathbb{R}^{p \times q}$ is any differentiable (possibly matrix-valued) function of $X \in \mathbb{R}^{n \times m}$, we define

$$Df(X) := D\mathrm{vec}(f(X)) := \frac{df(X)}{dX} := \frac{d\mathrm{vec}(f(X))}{d\mathrm{vec}(X)},$$

which is a $pq \times mn$ matrix. Let us recall the usual differentiation rules for derivatives in matrix arguments together with the derivatives of several standard types of functions.

Lemma 2.2. *Let $X \in \mathbb{R}^{n \times m}$, $A \in \mathbb{R}^{p \times n}$ and $B \in \mathbb{R}^{m \times q}$.*

1. *Let $f(X) \in \mathbb{R}^{p \times q}$, $g(X) \in \mathbb{R}^{q \times k}$ be differentiable functions of X. Then, the* product rule *holds:*

$$D(f(X)g(X)) = (g(X)^T \otimes I_p)Df(X) + (I_k \otimes f(X))Dg(X).$$

2. *For differentiable functions f on $\mathbb{R}^{q \times k}$ and $g(X) \in \mathbb{R}^{q \times k}$, the usual* chain rule *holds:*

$$D[f(g(X))] = Df(g(X)) \cdot Dg(X).$$

3. *We have $\frac{dX^T}{dX} = K_{n,m}$.*

4. *We have $\frac{d(AXB)}{dX} = B^T \otimes A$.*

5. *We have $\frac{d\mathrm{tr}(AX)}{dX} = \mathrm{vec}(A^T)^T$.*

3 Derivatives on Stiefel manifolds

To formally introduce intrinsic derivatives of first and second order on Stiefel manifolds, we follow and extend the spherical case as discussed in [BCG17]. If f is any real-valued locally Lipschitz function on some metric space (M, d) (with no isolated points), we may always define the generalized modulus of the gradient of f in $x \in M$

$$|\nabla^* f(x)| = \limsup_{x' \to x} \frac{|f(x) - f(x')|}{d(x, x')}. \tag{3.1}$$

In particular, we may use (3.1) for functions $f(X)$ on the Stiefel manifold $W_{n,d}$ together with the Euclidean (Hilbert–Schmidt) metric $d(X, X') = |X - X'|$. Here we will always write $|\nabla^* f(X)| \equiv |\nabla_W f(X)|$, cf. the discussion after (3.2) below. Note we could also use

the (point-dependent) canonical instead of the Euclidean metric, which leads to different notions of differentiability, cf. e. g. [EAS98].

To introduce a notion of differentiability of a function f on $W_{n,d}$ which is consistent with (3.1), recall that the tangent space in $A \in W_{n,d}$ is given by

$$T_A := \{N \in \mathbb{R}^{n \times d} : A^T N + N^T A = 0\} = \{N \in \mathbb{R}^{n \times d} : A \circ N = 0\},$$

where for any $M, N \in \mathbb{R}^{n \times d}$ (or $M, N \in \mathbb{R}^{d \times d}$ as we shall also need later on), $M \circ N$ denotes the "symmetric product"

$$M \circ N := \frac{1}{2}(M^T N + N^T M),$$

which is a symmetric $d \times d$ matrix (for $d = 1$, this reduces to the Euclidean scalar product). Hence, T_A is the set of all the matrices $N \in \mathbb{R}^{n \times d}$ such that $A^T N$ is antisymmetric. Now, a function $f \colon W_{n,d} \to \mathbb{R}$ is differentiable at $A \in W_{n,d}$ if it admits a Taylor expansion

$$f(A') = f(A) + \langle M, A' - A \rangle + o(|A' - A|) \quad \text{as} \ \ A' \to A, \ \ A' \in W_{n,d} \qquad (3.2)$$

with some $M \in \mathbb{R}^{n \times d}$. The unique M_0 of smallest (Euclidean) length among all such M is called the intrinsic derivative or gradient of f at A and is denoted $\nabla_W f(A)$. The length of $\nabla_W f(A)$ agrees with (3.1).

Once $M \in \mathbb{R}^{n \times d}$ satisfies (3.2), any matrix $M - B$ satisfies (3.2) as well if

$$\langle B, A' - A \rangle = o(|A' - A|)$$

as $A' \to A$, $A, A' \in W_{n,d}$. The latter is equivalent to $\langle B, N \rangle = 0$ for all $N \in T_A$, i. e. $B \in T_A^\perp$. Therefore, the minimization problems translates into

$$\|M - B\|_{\mathrm{HS}} \to \min \qquad \text{over all } B \in T_A^\perp,$$

which is solved uniquely for the orthogonal projection M onto T_A.

Often, it is convenient to consider functions f which are defined and smooth in an open neighborhood of $W_{n,d}$, which may be achieved by choosing a suitable extension of the function under consideration. In this case, we may take the Euclidean derivative $Df(A) \in \mathbb{R}^{1 \times nd}$ of f in $A \in W_{n,d}$ (cf. the previous section) and set $\nabla f(A) := \mathrm{mat}(Df(A)^T) \in \mathbb{R}^{n \times d}$ (the usual gradient rewritten as a matrix). Clearly, $\nabla f(A)$ satisfies (3.2), and we may project it onto T_A. This way, we get back $\nabla_W f(A)$.

Proposition 3.1. *Let f be defined and $\mathcal{C}^1$-smooth in a neighborhood of $W_{n,d}$. Then, the intrinsic first derivative of f at $A \in W_{n,d}$ is given by the projection onto T_A*

$$\nabla_W f(A) = \nabla f(A) - A(A \circ \nabla f(A)). \qquad (3.3)$$

In particular, $|\nabla_W f(A)| \leq |\nabla f(A)|$ for any $A \in W_{n,d}$.

Defining the projection $\pi_A \colon \mathbb{R}^{n \times d} \to T_A$ by

$$\pi_A M := M - A(A \circ M), \qquad (3.4)$$

Proposition 3.1 states that $\nabla_W f(A) = \pi_A \nabla f(A)$. In particular, by the contractivity of orthogonal projections, this shows that indeed, $|\nabla_W f(A)| \leq |\nabla f(A)|$.

Next, we introduce second order (intrinsic) derivatives (Hessians). For any $\mathcal{C}^2$-smooth function $f \colon \mathbb{R}^{n \times d} \to \mathbb{R}$ at a given point $A \in W_{n,d}$, consider the Taylor expansion up to the quadratic term (using the notation (2.1))

$$f(A') = f(A) + \langle \nabla_W f(A), A' - A \rangle + \frac{1}{2}\langle B(A' - A), A' - A \rangle + o(|A' - A|^2) \qquad (3.5)$$

as $A' \to A$, $A' \in W_{n,d}$, where $B \in \mathbb{R}^{nd \times nd}$ is some matrix. The collection of all B's satisfying (3.5) represents an affine subspace of $\mathbb{R}^{nd \times nd}$. Therefore, among all of them, there exists a unique matrix of smallest Hilbert–Schmidt norm. It is called the (intrinsic) second derivative of f at the point X and will be denoted $f''_W(A)$.

If f is $\mathcal{C}^2$-smooth function on some open neighborhood of $W_{n,d}$, by the usual (Euclidean) Taylor expansion, (3.5) holds with

$$B = f''(A) - (A \circ \nabla f(A)) \otimes I_n, \tag{3.6}$$

where $f''(A)$ is the $nd \times nd$ matrix of the usual (Euclidean) second order derivatives of f at A. Indeed, first note that

$$\langle A(A \circ \nabla f(A)), A' - A \rangle + \frac{1}{2}\langle((A \circ \nabla f(A)) \otimes I_n)(A' - A), A' - A \rangle$$

$$= \frac{1}{2}\langle(A' + A)(A \circ \nabla f(A)), A' - A \rangle = \frac{1}{2}\mathrm{tr}((A \circ \nabla f(A))(A'^T + A^T)(A' - A))$$

$$= \frac{1}{2}\mathrm{tr}((A \circ \nabla f(A))(A'^T A - A^T A')) = 0,$$

where in the first identity we have used that by (2.1) and Lemma 2.1, part 5,

$$((A \circ \nabla f(A)) \otimes I_n)(A' - A) = \mathrm{vec}((A' - A)(A \circ \nabla f(A))),$$

the third step follows from $A, A' \in W_{n,d}$ and in the last one we use that the trace of a product of a symmetric and an antisymmetric matrix is 0. Therefore, plugging (3.3) and (3.6) into (3.5) leads to the usual Euclidean Taylor expansion (i. e. solely involving the gradient $\nabla f(W)$ and the Hessian $f''(W)$), as the additional terms in (3.3) and (3.6) cancel out.

Given $C \in \mathbb{R}^{nd \times nd}$, the matrix $B - C$ satisfies (3.5) if and only if

$$\langle C(A' - A), A' - A \rangle = o(|A' - A|^2)$$

for $A' \to A$, $A' \in W_{n,d}$. Similarly to the first order case, this is equivalent to $\langle CX, X \rangle = 0$ for all $X \in T_A$. This condition defines a linear subspace L of $\mathbb{R}^{nd \times nd}$, and the minimization problem translates into

$$\|B - C\|_{\mathrm{HS}} \to \min \qquad \text{over all } C \in L,$$

which is solved uniquely for the orthogonal projection of B onto the linear space $L^\perp$ of all matrices orthogonal to L. Since B is symmetric, we may restrict ourselves to symmetric matrices. As a result, we obtain the following description.

Proposition 3.2. *The intrinsic second derivative of f at each $A \in W_{n,d}$ is the symmetric matrix which is given by the orthogonal projection*

$$f''_W(A) = P_{L^\perp}B, \qquad B = f''(A) - (A \circ \nabla f(A)) \otimes I_n,$$

to the orthogonal complement of the linear subspace $L = L_A$ of all symmetric matrices C in $\mathbb{R}^{nd \times nd}$ such that $\langle CX, X \rangle = 0$ for all $X \in T_A$. Equivalently,

$$f''_W(A) = P_A B P_A,$$

where $P_A = AA^T$ and $P_A B P_A$ is the $nd \times nd$ matrix which is defined by the relation

$$(P_A B P_A)\mathrm{vec}(V) = \mathrm{vec}(P_A \mathrm{mat}[B\mathrm{vec}(P_A V)]) \tag{3.7}$$

for any $V \in \mathbb{R}^{n \times d}$.

Note that unless $d = 1$, the notation $P_A B P_A$ is not to be understood in the sense of matrix multiplication but as defined by the relation (3.7). In particular, Proposition 3.2 implies that for all $A \in W_{n,d}$ and $V \in \mathbb{R}^{n \times d}$, $f_W''(A)V \in T_A$, $f_W''(A)A = 0$, and hence $\langle f_W''(A)V, A \rangle = 0$ for all V. Furthermore, we have the contraction property

$$\|f_W''(A)\|_{\mathrm{HS}} \leq \|f''(A) - (A \circ \nabla f(A)) \otimes I_n\|_{\mathrm{HS}},$$

which also holds for the operator norm.

In the notation of (2.1), it is not hard to show that for any $V \in \mathbb{R}^{n \times d}$

$$f''(X)V = \nabla \langle \nabla f(X), V \rangle$$

for the usual Euclidean derivatives. The analogue for intrinsic derivatives reads as follows:

Proposition 3.3. *Given a $\mathcal{C}^2$-smooth function f on $W_{n,d}$, for all $A \in W_{n,d}$ and $V \in \mathbb{R}^{n \times d}$, we have*

$$f_W''(A)V = \nabla_W \langle \nabla_W f(A), V \rangle + P_A(\nabla_W f(A)(A \circ V)).$$

Here, the left-hand side has to be read as $f''(W)\mathrm{vec}(V)$ and the right-hand side has to be understood as vectorized.

In particular, if $V \in T_A$, then $P_A V = V$ and $A \circ V = 0$, so that we obtain

$$f_W''(A)V = \nabla_W \langle \nabla_W f(A), V \rangle.$$

Proof. Consider the function $\psi_V(A) := \langle \nabla_W f(A), V \rangle$. In view of Proposition 3.1, any smooth extension of f to a neighborhood of $W_{n,d}$ also yields a smooth extension of ψ_V which is given by

$$\psi_V(X) = \langle \nabla f(X), V \rangle - \langle X(X \circ \nabla f(X)), V \rangle$$
$$= \langle \nabla f(X), V \rangle - \mathrm{tr}(V^T X(X \circ \nabla f(X))).$$

Let us calculate $\nabla \psi_V(X)$. First, we have

$$\mathrm{vec}(\nabla \langle \nabla f(X), V \rangle) = f''(X)\mathrm{vec}(V).$$

Next, we consider

$$\frac{d}{dX}[V^T X(X \circ \nabla f(X))] = ((X \circ \nabla f(X)) \otimes I_d)\frac{d(V^T X)}{dX}$$
$$+ (I_d \otimes (V^T X))\frac{d(X \circ \nabla f(X))}{dX},$$

where we have applied Lemma 2.2, part 1. By Lemma 2.2, part 4,

$$\frac{d(V^T X)}{dX} = I_d \otimes V^T,$$

so that

$$((X \circ \nabla f(X)) \otimes I_d)\frac{d(V^T X)}{dX} = (X \circ \nabla f(X)) \otimes V^T.$$

Moreover, using Lemma 2.2, parts 1&3, we have

$$\frac{d(X^T \nabla f(X))}{dX} = (\nabla f(X)^T \otimes I_d)K_{n,d} + (I_d \otimes X^T)f''(X)$$

as well as

$$\frac{d(\nabla f(X)^T X)}{dX} = (X^T \otimes I_d)K_{n,d}f''(X) + I_d \otimes \nabla f(X)^T,$$

and hence, using Lemma 2.1, part 7, it follows that

$$\frac{d(X \circ \nabla f(X))}{dX} = \frac{1}{2}(I_{d^2} + K_{d,d})((I_d \otimes \nabla f(X)^T) + (I_d \otimes X^T)f''(X)).$$

Putting everything together, it follows by Lemma 2.2, parts 2&5, and (switching from $d/(dX)$ to $\mathrm{vec}(\nabla)$, hence transposing) Lemma 2.1, part 2, that

$$\mathrm{vec}(\nabla \psi_V(X)) = f''(X)\mathrm{vec}(V) - ((X \circ \nabla f(X)) \otimes V)\mathrm{vec}(I_d)$$
$$- \frac{1}{2}((I_d \otimes \nabla f(X)) + f''(X)(I_d \otimes X))(I_{d^2} + K_{d,d})(I_d \otimes (X^T V))\mathrm{vec}(I_d).$$

Using Lemma 2.1, part 6, is possible to simplify this impression further. First, note that

$$((X \circ \nabla f(X)) \otimes V)\mathrm{vec}(I_d) = \mathrm{vec}(V(X \circ \nabla f(X))).$$

Moreover, we have

$$(I_d \otimes (X^T V))\mathrm{vec}(I_d) = \mathrm{vec}(X^T V)$$

and by (2.2), it follows that

$$K_{d,d}\mathrm{vec}(X^T V) = \mathrm{vec}(V^T X),$$

so that

$$\frac{1}{2}(I_{d^2} + K_{d,d})(I_d \otimes (X^T V))\mathrm{vec}(I_d) = \mathrm{vec}(X \circ V).$$

To continue,

$$(I_d \otimes \nabla f(X))\mathrm{vec}(X \circ V) = \mathrm{vec}(\nabla f(X)(X \circ V))$$

as well as

$$(I_d \otimes X)\mathrm{vec}(X \circ V) = \mathrm{vec}(X(X \circ V)).$$

Putting everything together, we thus obtain

$$\mathrm{vec}(\nabla \psi_V(X)) = f''(X)\mathrm{vec}(V) - \mathrm{vec}(V(X \circ \nabla f(X)))$$
$$- \mathrm{vec}(\nabla f(X)(X \circ V)) - f''(X)\mathrm{vec}(X(X \circ V)).$$

Restricting this from $X \in \mathbb{R}^{n \times d}$ to $A \in W_{n,d}$ and recalling the projection P_A, we therefore have

$$\mathrm{vec}(\nabla \psi_V(A)) = f''(A)\mathrm{vec}(P_A V) - \mathrm{vec}(V(A \circ \nabla f(A)) + \nabla f(A)(A \circ V)).$$

In terms of intrinsic derivatives (and now using the short-hand notation introduced in Proposition 3.2), we obtain

$$\nabla_W \psi_V(A) = P_A f''(A) P_A V - P_A[V(A \circ \nabla f(A))] - P_A[\nabla f(A)(A \circ V)].$$

Recall the matrix $B = f''(A) - (A \circ \nabla f(A)) \otimes I_n$ from (3.6). By Proposition 2.1, part 5,

$$\tilde{B}\mathrm{vec}(V) = ((A \circ \nabla f(A)) \otimes I_n)\mathrm{vec}(V) = \mathrm{vec}(V(A \circ \nabla f(A))),$$

so that if we write $B = f''(A) - \tilde{B}$, we obtain

$$\nabla_W \psi_V(A) = P_A f''(A) P_A V - P_A \tilde{B} V - P_A[\nabla f(A)(A \circ V)]$$
$$= P_A f''_W(A) P_A V - P_A \tilde{B}(V - P_A V) - P_A[\nabla f(A)(A \circ V)]$$

in view of Proposition 3.2. As above, we note that

$$\tilde{B}V = (V - P_A V)(A \circ \nabla f(A)),$$

so that altogether, we arrive at

$$f_W''(A)V = \nabla_W \langle \nabla_W f(A), V \rangle + P_A[(V - P_A V)(A \circ \nabla f(A))]$$
$$+ P_A[\nabla f(A)(A \circ V)].$$

To finish the proof, it remains to note that

$$P_A[(V - P_A V)(A \circ \nabla f(A))] + P_A[\nabla f(A)(A \circ V)] - P_A[\nabla_W f(A)(A \circ V)]$$
$$= P_A[A(A \circ V)(A \circ \nabla f(A))] + P_A[A(A \circ \nabla f(A))(A \circ V)]$$
$$= P_A[A((A \circ V) \circ (A \circ \nabla f(A)))] = 0,$$

where the last step follows by an easy calculation using $A^T A = I_d$. $\qquad\square$

4 Second order modulus of gradient

Recall that the generalized second order modulus of the gradient on the Stiefel manifold $W_{n,d}$ is defined by

$$|\nabla^{(2)} f(A)| = |\nabla_W |\nabla_W f(A)||$$
$$= \limsup_{A' \to A} \frac{||\nabla_W f(A)| - |\nabla_W f(A')||}{|A - A'|}.$$

Typically, explicitly calculating $|\nabla^{(2)} f(A)|$ is not easy, however, and therefore, motivated by the Euclidean or spherical calculus, we may hope for an estimate of the form $|\nabla^{(2)} f(A)| \le \|f_W''(A)\|_{\mathrm{op}}$. Indeed, we have the following result.

Proposition 4.1. *For any $\mathcal{C}^2$-smooth function f on $W_{n,d}$, $|\nabla_W f|$ has finite Lipschitz semi-norm, and for any $A \in W_{n,d}$,*

$$|\nabla^{(2)} f(A)| = |\nabla_W f(A)|^{-1} |f_W''(A) \nabla f_W(A)|,$$

where the right-hand side has to be understood as $\|f_W''(A)\|_{\mathrm{op}}$ if $|\nabla_W f(A)| = 0$. In particular, we always have

$$|\nabla^{(2)} f(A)| \le \|f_W''(A)\|_{\mathrm{op}}.$$

Proof. Let us first prove that the function $|\nabla_W f(A)|$ has finite Lipschitz semi-norm. Since the first two intrinsic derivatives of f are continuous and therefore bounded on the compact manifold $W_{n,d}$, we obtain from Proposition 3.3 that

$$|\nabla_W \langle \nabla_W f(A), V \rangle| \le C$$

for any $V \in \mathbb{R}^{n \times d}$ such that $|V| \equiv \|V\|_{\mathrm{HS}} = 1$, where C is some constant independent of A and V. Hence, the function $A \mapsto \langle \nabla_W f(A), V \rangle$ has Lipschitz semi-norm bounded by C, i. e.

$$|\langle \nabla_W f(A'), V \rangle - \langle \nabla_W f(A), V \rangle| \le C d(A, A')$$

for all $A, A' \in W_{n,d}$. Therefore, taking the supremum over all V and applying the triangle inequality yields

$$||\nabla_W f(A')| - |\nabla_W f(A)|| \le |\nabla_W f(A') - \nabla_W f(A)| \le C d(A', A),$$

which had to be proven.

To show the identity for the second order modulus of the gradient, fix $A \in W_{n,d}$. By the definition of the intrinsic gradient and Proposition 3.3, we have

$$\langle \nabla_W f(A'), V \rangle = \langle \nabla_W f(A), V \rangle + \langle \tilde{V}, A' - A \rangle + o(|A' - A|),$$

where
$$\tilde{V} = f_W''(A)V - P_A(\nabla_W f(A)(A \circ V)).$$

Moreover, a closer analysis (using the integral form of the Taylor formula and the compactness of $W_{n,d}$, which implies that every continuous function is already uniformly continuous) yields that the remainder term in the Taylor expansion can be bounded independently of $V \in \mathbb{R}^{n \times d}$ such that $|V| = 1$, i.e.

$$\sup_{|V|=1} |\langle \nabla_W f(A'), V \rangle - \langle \nabla_W f(A), V \rangle - \langle \tilde{V}, A' - A \rangle| \le \varepsilon(|A - A'|),$$

where $\varepsilon(t)$ is some function which satisfies $\varepsilon(t) \to 0$ as $t \to 0$.

The next step is to rewrite the Taylor formula as

$$\langle \nabla_W f(A'), V \rangle = \langle \nabla_W f(A) + L, V \rangle + o(|A' - A|) \tag{4.1}$$

for some L to be determined later. To this end, we first choose a suitable $\tilde{L}$ such that $\langle \tilde{V}, A' - A \rangle = \langle \tilde{L}, V \rangle$. Obviously,

$$\langle f_W''(A)V, A' - A \rangle = \langle f_W''(A)(A' - A), V \rangle.$$

Moreover,

$$\begin{aligned}
&P_A(\nabla_W f(A)(A \circ V)) \\
&= \nabla_W f(A)(A \circ V) - A(A \circ (\nabla_W f(A)(A \circ V))) \\
&= \frac{1}{2}(\nabla_W f(A)A^T V + \nabla_W f(A)V^T A) - \frac{1}{4}(AA^T \nabla_W f(A)A^T V \\
&\quad + AA^T \nabla_W f(A)V^T A + AA^T V \nabla_W f(A)^T A + AV^T A \nabla_W f(A)^T A).
\end{aligned}$$

Recall that $\langle U, V \rangle = \text{tr}(U^T V)$ and that the trace is invariant under cyclic permutations. Therefore, we may easily verify the general rules

$$\langle UVW, X \rangle = \langle U^T X W^T, V \rangle, \qquad \langle UV^T W, X \rangle = \langle WX^T U, V \rangle, \tag{4.2}$$

where $V, X \in \mathbb{R}^{n \times d}$ and $U \in \mathbb{R}^{n \times n}$, $W \in \mathbb{R}^{d \times d}$ in the first identity or $U, W \in \mathbb{R}^{n \times d}$ in the second one. We now apply these identities to all six terms appearing in $P_A(\nabla_W f(A)(A \circ V))$ with V as above and $X = A' - A$. For instance, for the first term this yields

$$\langle \nabla_W f(A)A^T V, A' - A \rangle = \langle A \nabla_W f(A)^T (A' - A), V \rangle.$$

Proceeding similarly, we arrive at

$$\langle \nabla_W f(A)(A \circ V), A' - A \rangle = \langle A(\nabla_W f(A) \circ (A' - A)), V \rangle$$

as well as

$$\langle A(A \circ (\nabla_W f(A)(A \circ V))), A' - A \rangle = \langle A((A^T \nabla_W f(A)) \circ (A \circ (A' - A))), V \rangle.$$

Altogether, we obtain

$$\tilde{L} = f_W''(A)(A' - A) - A[\nabla_W f(A) \circ (A' - A) - (A^T \nabla_W f(A)) \circ (A \circ (A' - A))].$$

Now we define

$$L := f_W''(A)(A' - A) - A(\nabla_W f(A) \circ (A' - A))$$

To see that (4.1) holds, it remains to show that $\tilde{L} - L = o(|A' - A|)$ as $A' \to A$, $A, A' \in W_{n,d}$. To this end, note that by an easy calculation,

$$A \circ (A' - A) = \frac{1}{2}(A' - A)^T (A - A') = o(|A' - A|).$$

From here, the claim immediately follows by compactness arguments.

Now we take an absolute value on both sides of the Taylor formula (4.1) and take the supremum over all V such that $\mathrm{vec}(V) \in S^{nd-1}$. This leads to

$$|\nabla_W f(A')| = |\nabla_W f(A) + L| + o(|A' - A|).$$

Next, we write

$$|\nabla_W f(A) + L|^2 = |\nabla_W f(A)|^2 + 2\langle \nabla_W f(A), L \rangle + |L|^2.$$

Noting that

$$\langle \nabla_W f(A), A(A \circ \nabla f(A)) \rangle = \mathrm{tr}(\nabla_W f(A)^T A(A \circ \nabla f(A))) = 0$$

since $\nabla_W f(A)^T A$ is antisymmetric (as $\nabla_W f(A) \in T_A$) and $A \circ \nabla f(A)$ is symmetric, we obtain

$$\langle \nabla_W f(A), L \rangle = \langle \nabla_W f(A), f_W''(A)(A' - A) \rangle = \langle U, A' - A \rangle,$$

where

$$U := f_W''(A)\nabla_W f(A) \equiv f_W''(A)\mathrm{vec}(\nabla_W f(A)).$$

Since $|L|^2 = O(|A' - A|^2)$, we obtain

$$|\nabla_W f(A) + L|^2 = |\nabla_W f(A)|^2 + 2\langle U, A' - A \rangle + o(|A' - A|).$$

If $|\nabla_W f(A)| > 0$, it therefore follows that

$$|\nabla_W f(A) + L| = |\nabla_W f(A)| + |\nabla_W f(A)|^{-1}\langle U, A' - A \rangle + o(|A' - A|).$$

Hence,

$$|\nabla_W f(A')| - |\nabla_W f(A)| = |\nabla_W f(A)|^{-1}\langle U, A' - A \rangle + o(|A' - A|)$$

and thus

$$\limsup_{A' \to A} \frac{\left| |\nabla_W f(A')| - |\nabla_W f(A)| \right|}{|A' - A|} = |\nabla_W f(A)|^{-1} \limsup_{A' \to A} \frac{\left| \langle U, A' - A \rangle \right|}{|A' - A|}$$
$$= |\nabla_W f(A)|^{-1}|\nabla_W \psi_U(A)|,$$

where $\psi_U(A) := \langle U, A \rangle$. As noted after Proposition 3.2, $U \in T_A$, so that $\nabla_W \psi_U(A) = U$. Thus, we arrive at

$$|\nabla_W^{(2)}(A)| = |\nabla_W f(A)|^{-1}|f_W''(A)\nabla_W f(A)|$$

if $|\nabla_W f(A)| > 0$.

It remains to consider the case where $|\nabla_W f(A)| = 0$. Here, $L = f_W''(A)(A' - A)$, and the Taylor formula reads

$$|\nabla_W f(A')| = |L| + o(|A' - A|).$$

It follows that

$$|\nabla_W^{(2)} f(A)| = \limsup_{A' \to A} \frac{|\nabla_W f(A')|}{|A' - A|} = \limsup_{A' \to A} \frac{|f_W''(A)(A' - A)|}{|A' - A|}$$
$$= \limsup_{V \to 0, V \in T_A^\perp} \frac{|f_W''(A)V|}{|V|} = \|f_W''(A)\|_{\mathrm{op}},$$

which finishes the proof. $\qquad\square$

5 Derivatives on Grassmann manifolds

Let us adapt the results of the previous two sections to functions on Grassmann manifolds. Much of what follows relies on similar arguments as in the Stiefel case, and for this reason we will often only sketch the arguments. To introduce a notion of differentiability on $G_{n,d}$, first recall that the tangent space in $P \in G_{n,d}$ is given by

$$T_P := \{S \in \mathbb{R}_{\mathrm{sym}}^{n \times n} \colon S = SP + PS\} \equiv \{S \in \mathbb{R}_{\mathrm{sym}}^{n \times n} \colon S = [[S, P], P]\},$$

where for any $M, N \in \mathbb{R}^{n \times n}$, $[M, N] = MN - NM$ denotes the matrix commutator. A function $f \colon G_{n,d} \to \mathbb{R}$ is differentiable at $P \in G_{n,d}$ if it admits a Taylor expansion

$$f(P') = f(P) + \langle M, P' - P \rangle + o\big(|P' - P|\big) \quad \text{as } P' \to P, \ P' \in G_{n,d} \tag{5.1}$$

with some $M \in \mathbb{R}_{\mathrm{sym}}^{n \times n}$. Among all such M, there exists a unique M_0 of smallest (Euclidean) length, called the intrinsic (first) derivative or gradient of f at P and denoted $\nabla_G f(P)$. The length of $\nabla_G f(P)$ agrees with the generalized modulus of the gradient (3.1) applied to the Grassmann manifold. In passing, note that unlike in case of the Stiefel manifold, for Grassmann manifolds the Euclidean metric and the canonical metric lead to the same notion of differentiablity. As in case of the Stiefel manifold, the minimization problem (5.1) translates into

$$\|M - B\|_{\mathrm{HS}} \to \min \quad \text{over all } B \in T_P^{\perp},$$

which is solved uniquely for the orthogonal projection M onto T_P.

If we consider functions which are defined and smooth in an open neighborhood of $G_{n,d}$ in the ambient space, e.g. $\mathbb{R}_{\mathrm{sym}}^{n \times n}$, we may take the Euclidean gradient $\nabla f(P) = \mathrm{mat}(Df(P)^T)$ (cf. Section 2) and project it onto T_P, which gives back $\nabla_G f(P)$. As the projection $\pi_P \colon \mathbb{R}_{\mathrm{sym}}^{n \times n} \to T_P$ is given by

$$\pi_P M := [P, [P, M]] = PM + MP - 2PMP, \tag{5.2}$$

we immediately arrive at the following analogue of Proposition 3.1.

Proposition 5.1. *Let f be defined and $\mathcal{C}^1$-smooth in some open neighborhood of $G_{n,d}$ in $\mathbb{R}_{\mathrm{sym}}^{n \times n}$. Then, the intrinsic first derivative of f at $P \in G_{n,d}$ is given by the projection onto T_P*

$$\nabla_G f(P) = \pi_P \nabla f(P) = [P, [P, \nabla f(P)]].$$

In particular, $|\nabla_G f(P)| \leq |\nabla f(P)|$ for any $P \in G_{n,d}$.

In fact, sometimes yet a further embedding might be convenient, so that we regard $G_{n,d}$ as a submanifold of $\mathbb{R}^{n \times n}$ and take the Euclidean derivatives of some extension of f to an open neighborhood in $\mathbb{R}^{n \times n}$. However, in this case, we may project $\nabla f(P)$ onto the tangent space of $\mathbb{R}_{\mathrm{sym}}^{n \times n}$ (which equals $\mathbb{R}_{\mathrm{sym}}^{n \times n}$) by applying the projection $\pi_{\mathrm{sym}} \colon \mathbb{R}^{n \times n} \to \mathbb{R}_{\mathrm{sym}}^{n \times n}$ given by $\pi_{\mathrm{sym}}(M) := (M + M^T)/2$. Then, we may proceed as in Proposition 5.1 for $\pi_{\mathrm{sym}}(\nabla f(P))$.

For any $\mathcal{C}^2$-smooth function f on $G_{n,d}$ at a given point $P \in W_{n,d}$, the intrinsic second order derivative is the matrix $B \in \mathbb{R}^{n^2 \times n^2}$ of smallest Hilbert–Schmidt norm which satisfies

$$f(P') = f(P) + \langle \nabla_G f(P), P' - P \rangle + \frac{1}{2} \langle B(P' - P), P' - P \rangle + o\big(|P' - P|^2\big) \tag{5.3}$$

as $P' \to P$, $P' \in G_{n,d}$. It will be denoted $f_G''(X)$. Considering functions f which are defined and smooth in some open neighborhood of $G_{n,d}$ in $\mathbb{R}_{\mathrm{sym}}^{n \times n}$, it is possible to express the intrinsic second order derivative in terms of the Euclidean derivatives of f. Instead of providing the details, we refer to [HHT07, Theorem 2.4] where corresponding calculations have been done. Adapting them to our framework yields the following analogue of Proposition 3.2.

Proposition 5.2. *The intrinsic second derivative of f at each $P \in G_{n,d}$ is given by the relation*

$$f_G''(P)V = \pi_P f''(P)\pi_P V - [P, [\nabla f(P), \pi_P V]]$$

for any $V \in \mathbb{R}_{\mathrm{sym}}^{n \times n}$. Here, the expression on the right hand side has to be understood as

$$\mathrm{vec}(\pi_P \mathrm{mat}[f''(P)\mathrm{vec}(\pi_P V)]) - \mathrm{vec}([P, [\nabla f(P), \pi_P V]]).$$

Note that $[P, [\nabla f(P), \pi_P V]] \in T_P$. In particular, for all $P \in G_{n,d}$ and $V \in \mathbb{R}_{\mathrm{sym}}^{n \times n}$, $f_G''(P)V \in T_P$, $f_G''(P)P = 0$, and hence $\langle f_G''(P)V, P \rangle = 0$ for all $V \in \mathbb{R}_{\mathrm{sym}}^{n \times n}$. Furthermore, we have the contraction property

$$\|f_G''(P)\|_{\mathrm{HS}} \le \|f''(P) - [P, [\nabla f(P), \pi_P V]]\|_{\mathrm{HS}},$$

which also holds for the operator norm. Similarly to the first order case, sometimes it is necessary to extend f to a smooth function on some neighborhood in $\mathbb{R}^{n \times n}$. In this case, we may take the usual Euclidean Hessian $f''(P)$, from which we get back the Hessian on $\mathbb{R}_{\mathrm{sym}}^{n \times n}$ by considering $\pi_{\mathrm{sym}} f''(P)\pi_{\mathrm{sym}}$.

Moreover, we have the following analogue of Proposition 3.3.

Proposition 5.3. *Given a $\mathcal{C}^2$-smooth function f on $G_{n,d}$, for all $P \in G_{n,d}$ and $V \in \mathbb{R}_{\mathrm{sym}}^{n \times n}$, we have*

$$f_G''(P)V = \nabla_G \psi_V(P) - [P, [\nabla_G f(P), V]].$$

Here, the left-hand side has to be read as $f_G''(P)\mathrm{vec}(V)$ and the right-hand side has to be understood as vectorized.

In particular, if $V \in T_P$ it follows from [HHT07, Lemma 2.2] that

$$f_G''(P)V = \nabla_G \langle \nabla_G f(P), V \rangle.$$

Finally, we also have an analogue of Proposition 4.1.

Proposition 5.4. *For any $\mathcal{C}^2$-smooth function f on $G_{n,d}$, $|\nabla_G f|$ has finite Lipschitz seminorm, and for any $P \in G_{n,d}$,*

$$|\nabla_G^{(2)} f(P)| = |\nabla_G f(P)|^{-1}|f_G''(P)\nabla f_G(P)|,$$

where the right-hand side has to be understood as $\|f_G''(P)\|_{\mathrm{op}}$ if $|\nabla_G f(P)| = 0$. In particular, we always have

$$|\nabla_G^{(2)} f(P)| \le \|f_G''(P)\|_{\mathrm{op}}.$$

As the proofs of Proposition 5.3 and Proposition 5.4 are mostly an adaption of the arguments known from the Stiefel case, we defer them to the appendix.

6 Proofs

To prepare the proofs of our main results, we briefly revisit the representations of the Stiefel and Grassmann manifolds we use in this paper. In particular, we shall discuss the Lipschitz properties of the map $\pi_{\mathrm{St}} \colon W_{n,d} \to G_{n,d}$, $A \mapsto AA^T \equiv P_A$. Clearly, π_{St} is $2\sqrt{d}$-Lipschitz as

$$\|AA^T - A'A'^T\|_{\mathrm{HS}} \le \|A(A^T - A'^T)\|_{\mathrm{HS}} + \|(A - A')A'^T\|_{\mathrm{HS}} \le 2\sqrt{d}\|A - A'\|_{\mathrm{HS}}$$

since $\|A\|_{\mathrm{HS}}, \|A'\|_{\mathrm{HS}} = \sqrt{d}$. However, this Lipschitz constant is too weak for our purposes. To avoid the dependency on d, more subtle arguments are needed.

Recall that if $M_1, M_2 \subset \mathbb{R}^n$ are two subspaces of dimension d, the principal angles θ_j, $j = 1, \ldots, d$, between M_1 and M_2 are recursively defined by

$$\cos \theta_j = \max_{u \in M_1} \max_{v \in M_2} |\langle u, v \rangle| = \langle u_j, v_j \rangle$$

subject to the constraints

$$\langle u_i, u \rangle = 0, \quad \langle v_i, v \rangle = 0, \quad i = 1, 2, \ldots, j-1.$$

The vectors $\{u_1, \ldots, u_d\}$, $\{v_1, \ldots, v_d\}$ are called the principal vectors of the pair (M_1, M_2). The concept of principal angles goes back to Jordan. Here we mainly follow the survey article [Ga08].

Note that the principal angles are uniquely defined and satisfy $0 \le \theta_1 \le \ldots \le \theta_d \le \pi/2$, while the principal vectors are not unique. However, they form orthonormal d-frames and can thus be interpreted as elements $U, V \in W_{n,d}$. Moreover, we have $\langle u_i, v_j \rangle = \delta_{ij} \cos \theta_i$. By [Af57], the respective orthogonal projections $P_U = \pi_{\mathrm{St}}(U)$ and $P_V = \pi_{\mathrm{St}}(V)$ satisfy

$$P_U P_V u_j = (\cos^2 \theta_j) u_j, \quad P_V P_U v_j = (\cos^2 \theta_j) v_j, \quad j = 1, \ldots, d,$$

i. e., the non-zero eigenvalues of the matrices $P_U P_V$ and $P_V P_U$ are $\cos^2 \theta_j$, $j = 1, \ldots, d$. This relation extends to all $A, B \in W_{n,d}$ such that $A = UO$ and $B = VO'$ for some $O, O' \in O(d)$ (since $P_A = P_U$ and $P_B = P_V$). Note also that by definition of the principal angles, for any $A, A' \in W_{n,d}$

$$\max_{O \in O(d)} \langle A, A'O \rangle = \langle A, A'' \rangle = \mathrm{tr}(A^T A'') = \sum_{j=1}^{d} \cos \theta_j,$$

where θ_j are the principal angles between the two subspaces induced by A and A' and A'' is the Stiefel matrix maximizing the term on the left-hand side.

Lemma 6.1. *The map* $\pi_{\mathrm{St}} \colon W_{n,d} \to G_{n,d}$, $A \mapsto AA^T = P_A$ *is* $\sqrt{2}$-*Lipschitz.*

Proof. Using principle angles and the notation introduced above, we have

$$\|P_A - P_{A'}\|_{\mathrm{HS}}^2 = 2d - 2\mathrm{tr}(P_A P_{A'}) = 2d - 2\sum_{j=1}^{d} \cos^2 \theta_j = 2\sum_{j=1}^{d}(1 - \cos^2 \theta_j)$$

$$\le 2\sum_{j=1}^{d} 2(1 - \cos \theta_j) = 2(2d - 2\mathrm{tr}(A^T A''))$$

$$\le 2(2d - 2\mathrm{tr}(A^T A')) = 2\|A - A'\|_{\mathrm{HS}}^2,$$

where in the first inequality we used that $(1 - x^2)/(1 - x) \le 2$ for any $x \in [0, 1]$. $\qquad\square$

One may wonder whether the map P_A might even be 1-Lipschitz. However, simple examples are sufficient to show that this cannot be true (e. g., consider $d = 1$ and the vectors $A = (1, 0, \ldots, 0)^T$ and $A' = (1/\sqrt{n}, \ldots, 1/\sqrt{n})^T$).

The core of our arguments is a logarithmic Sobolev inequality for Stiefel and Grassmann manifolds. Even if we are not aware of a source where log-Sobolev inequalities for Stiefel and Grassmann manifolds are rigorously formulated, they may easily be derived by a simple projection argument. We emphasize that these inequalities are designed for the representations of $W_{n,d}$ and $G_{n,d}$ we use all over this paper.

Proposition 6.2. *1. For any* $d < n$, $W_{n,d}$ *satisfies a logarithmic Sobolev inequality with constant* $4/(n-2)$, *i. e. for any* $f \colon W_{n,d} \to \mathbb{R}$ *sufficiently smooth,*

$$\mathrm{Ent}_{\mu_{n,d}}(f^2) \le \frac{8}{n-2} \int_{W_{n,d}} |\nabla_W f|^2 d\mu_{n,d}.$$

2. *For any $d < n$, $G_{n,d}$ satisfies a logarithmic Sobolev inequality with constant $8/(n-2)$, i.e. for any $f : G_{n,d} \to \mathbb{R}$ sufficiently smooth,*

$$\mathrm{Ent}_{\nu_{n,d}}(f^2) \leq \frac{16}{n-2} \int_{G_{n,d}} |\nabla_G f|^2 d\nu_{n,d}.$$

Let us briefly mention that for $d = 1$ (i.e. the sphere), the optimal Sobolev constant is known to be $1/(n-1)$ as shown in [MW82]. In particular, even if Proposition 6.2 is not fully accurate, the behavior of the Sobolev constant for large values of n agrees with the optimal result in case of the sphere. Moreover, note that if $d = n$, $W_{n,n} = O(n)$ has two connected components, which in particular implies that a log-Sobolev inequality cannot hold.

Proof of Proposition 6.2. First recall that if $d < n$, we have

$$W_{n,d} \cong SO(n)/SO(n-d).$$

Indeed, identifying any matrix in $SO(n)$ with its first $n - 1$ columns $e_1, \ldots, e_{n-1}$, this follows readily using the projection map $\varphi \colon SO(n) \to W_{n,d}$ which is given by $\varphi(e_1, \ldots, e_{n-1}) := (e_1, \ldots, e_d)$. By [Me19, Theorem 5.16], the special orthogonal group $SO(n)$ equipped with the uniform probability measure and Hilbert–Schmidt metric satisfies a log-Sobolev inequality with constant $4/(n-2)$. Noting that the map φ is 1-Lipschitz, part 1 of the proposition therefore follows immediately.

To see part 2, it remains to note that $(G_{n,d}, \nu_{n,d})$ is the the pushforward of $(W_{n,d}, \mu_{n,d})$ under π_{St}, which is $\sqrt{2}$-Lipschitz according to Lemma 6.1 (so that the Sobolev constant is doubled). $\qquad\square$

In the sequel, we will give the proofs of the concentration bound for function on Stiefel manifolds, i.e., Theorem 1.1 and Theorem 1.3. As already shown in [AS94], logarithmic Sobolev inequalities imply certain L^p norm inequalities. In the situation under consideration, for any locally Lipschitz function $g \colon W_{n,d} \to \mathbb{R}$ and for any $p \geq 2$,

$$\|g\|_p^2 \leq \|g\|_2^2 + 4(n-2)^{-1}(p-2)\|\nabla_W g\|_p^2. \tag{6.1}$$

Using Proposition 3.1, we moreover have

$$\|g\|_p^2 \leq \|g\|_2^2 + 4(n-2)^{-1}(p-2)\|\nabla g\|_p^2. \tag{6.2}$$

Let us also recall that since $\mu_{n,d}$ satisfies a log-Sobolev inequality, it also satisfies a Poincaré inequality with the same constant, i.e.

$$\mathrm{Var}_{\mu_{n,d}}(f) \equiv \int_{W_{n,d}} (f - \mu_{n,d}(f))^2 d\mu_{n,d} \leq \frac{4}{n-2} \int_{W_{n,d}} |\nabla_W f|^2 d\mu_{n,d} \tag{6.3}$$

for all $f \colon W_{n,d} \to \mathbb{R}$ sufficiently smooth.

We now first prove Theorem 1.3. Note that its proof follow the lines of the proofs established in [BGS19, GSS21b].

Proof of Theorem 1.3. Applying (6.2) to $g = \|f^{(\ell)}\|_{\mathrm{op}}$ and recalling that for Euclidean derivatives, $|\nabla \|f^{(\ell)}\|_{\mathrm{op}}| \leq \|f^{(\ell+1)}\|_{\mathrm{op}}$ (cf. [BGS19, Lemma 4.1]), we obtain that

$$\|f^{(\ell)}\|_{\mathrm{op},p}^2 \leq \|f^{(\ell)}\|_{\mathrm{op},2}^2 + 4(n-2)^{-1}(p-2)\|f^{(\ell+1)}\|_{\mathrm{op},p}^2$$

for any $\ell = 1, \ldots, k-1$, which yields

$$\|f\|_p^2 \leq \|f\|_2^2 + \sum_{\ell=1}^{k-1} \Big(\frac{4(p-2)}{n-2}\Big)^\ell \|f^{(\ell)}\|_{\mathrm{op},2}^2 + \Big(\frac{4(p-2)}{n-2}\Big)^k \|f^{(k)}\|_{\mathrm{op},\infty}^2$$

$$\leq \frac{4}{n-2}\|f^{(1)}\|_2^2 + \sum_{\ell=1}^{k-1} \Big(\frac{4(p-2)}{n-2}\Big)^\ell \|f^{(\ell)}\|_{\mathrm{op},2}^2 + \Big(\frac{4(p-2)}{n-2}\Big)^k \|f^{(k)}\|_{\mathrm{op},\infty}^2$$

$$\leq \sum_{\ell=1}^{k-1} \Big(\frac{4(p-1)}{n-2}\Big)^\ell \|f^{(\ell)}\|_{\mathrm{op},2}^2 + \Big(\frac{4(p-1)}{n-2}\Big)^k \|f^{(k)}\|_{\mathrm{op},\infty}^2 \qquad (6.4)$$

for any $p \geq 2$, where the second step follows by Poincaré inequality.

To see part 1 of the theorem, plugging in the assumptions we arrive at

$$\|f\|_p^2 \leq \Big(\frac{4}{n-2}\Big)^k \sum_{\ell=1}^k p^\ell \leq \frac{1}{1-p^{-1}} \Big(\frac{4p}{n-2}\Big)^k \leq \Big(\frac{8p}{n-2}\Big)^k$$

and therefore $\|f\|_p \leq (8p/(n-2))^{k/2}$ for any $p \geq 2$. If $p \leq 2$, we may estimate $\|f\|_p \leq \|f\|_2 \leq (16/(n-2))^{k/2}$. In particular, we obtain that for any $m \geq 1$,

$$\||f|^{2/k}\|_m = \|f\|_{2m/k}^{2/k} \leq \gamma m$$

with $\gamma = 16/(n-2)$. In particular, by an easy calculation (cf. [BGS19, Eq. (2.17)]), this yields that $\int_{W_{n,d}} \exp(c'|f|)d\mu_{n,d} \leq 2$ for $c' = 1/(2\gamma e)$ as stated in part 1 of the theorem.

To show part 2, taking roots in (6.4) yields

$$\|f\|_p \leq \sum_{\ell=1}^{k-1} \Big(\frac{4(p-1)}{n-2}\Big)^{\ell/2} \|f^{(\ell)}\|_{\mathrm{op},2} + \Big(\frac{4(p-1)}{n-2}\Big)^{k/2} \|f^{(k)}\|_{\mathrm{op},\infty}.$$

Now, [GSS21b, Proposition 4] yields

$$\mu_{n,d}(|f - \mu_{n,d}(f)| \geq t) \leq 2\exp\Big(-\frac{1}{C}\min\Big(\min_{\ell=1,\ldots,k-1}\frac{t^{2/\ell}}{\|f^{(\ell)}\|_{\mathrm{op},2}^{2/\ell}}, \frac{t^{2/k}}{\|f^{(k)}\|_{\mathrm{op},\infty}^{2/k}}\Big)\Big),$$

where we may choose

$$C = \frac{4e^2 k^2}{\log(2)(n-2)},$$

which completes the proof. $\qquad\square$

For the proof of Theorem 1.1, we moreover need the following lemma which relates the L^2 norms of the derivatives of first and second order under a centering assumption.

Lemma 6.3. Let $f : W_{n,d} \to \mathbb{R}$ be a $\mathcal{C}^2$-smooth function which has centered first order intrinsic derivatives, i.e. $\mu_{n,d}(\nabla_W f) = 0$. Then,

$$\int_{W_{n,d}} |\nabla_W f|^2 d\mu_{n,d} \leq \frac{8}{n-2-8d} \int_{W_{n,d}} \|f_W''\|_{\mathrm{HS}}^2 d\mu_{n,d}.$$

Proof. Write

$$|\nabla_W f(A)|^2 = \sum_{i,j} \langle \nabla_W f(A), e_{ij}\rangle^2,$$

where $i \leq n$, $j \leq d$ and e_{ij} is the "indicator matrix" whose (i,j)-th entry equals 1 while all other entries are zero. By assumption, the functions $A \mapsto \langle \nabla_W f(A), e_{ij}\rangle$ are centered. Therefore, applying the Poincaré inequality (6.3) yields

$$\frac{n-2}{4} \int_{W_{n,d}} \langle \nabla_W f(A), e_{ij}\rangle^2 d\mu_{n,d} \leq \int_{W_{n,d}} |\nabla_W(\nabla_W f(A), e_{ij}\rangle)|^2 d\mu_{n,d}.$$

By Proposition 3.3, we have

$$\nabla_W(\nabla_W f(A), e_{ij}\rangle) = f''_W(A)e_{ij} - P_A(\nabla_W f(A)(A \circ e_{ij})),$$

so that we may estimate

$$\int_{W_{n,d}} |\nabla_W(\nabla_W f(A), e_{ij}\rangle)|^2 d\mu_{n,d}$$
$$\leq 2 \int_{W_{n,d}} |f''_W(A)e_{ij}|^2 d\mu_{n,d} + 2 \int_{W_{n,d}} |P_A(\nabla_W f(A)(A \circ e_{ij}))|^2 d\mu_{n,d}.$$

By contractivity and using $|A_{ij}|^2 \leq 1$,

$$|P_A(\nabla_W f(A)(A \circ e_{ij}))|^2 \leq |(\nabla_W f(A)(A \circ e_{ij})|^2 \leq d|\nabla_W f(A)|^2.$$

Plugging in, summing up and noting that

$$\sum_{i,j} |f''_W(A)e_{ij}|^2 = \|f''_W(A)\|^2_{\mathrm{HS}}$$

(recall that $i \leq n$, $j \leq d$ and $f''_W(A)$ is an $nd \times nd$ matrix) completes the proof. $\qquad\square$

Proof of Theorem 1.1. Obviously, we may argue as in the proof of Theorem 1.3 with (6.2) replaced by (6.1) and [BGS19, Lemma 4.1] replaced by Proposition 4.1 (thus involving intrinsic derivatives) to obtain

$$\|f\|_p^2 \leq \frac{4(p-1)}{n-2} \|\nabla_W f\|_2^2 + \Big(\frac{4(p-1)}{n-2}\Big)^2 \|f''_W\|^2_{\mathrm{op},\infty},$$

which is the "intrinsic analogue" of (6.4) for $k = 2$. Therefore, to see parts 1 and 2, we may just copy the proof of Theorem 1.3. Moreover, part 3 follows by applying Lemma 6.3. $\qquad\square$

Finally, to address Theorem 1.2 and Theorem 1.4, note that for any locally Lipschitz function $g\colon G_{n,d} \to \mathbb{R}$ and for any $p \geq 2$,

$$\|g\|_p^2 \leq \|g\|_2^2 + 8(n-2)^{-1}(p-2)\|\nabla_G g\|_p^2,$$

which is an analogue of (6.1). From here, we may essentially copy the proofs of Theorem 1.1 and Theorem 1.3 with the obvious adaptions (which, in particular, lead to slightly different constants).

7 Concentration results for polynomial chaos

Arguably the most classical object in the study of higher order concentration is polynomial chaos. In the context of Stiefel manifolds, by a polynomial chaos of order k we refer to functionals of the form

$$f_k(A) := \sum_{i_1,j_1=1}^{n,d} \cdots \sum_{i_k,j_k=1}^{n,d} c_{i_1 j_1,\ldots,i_k j_k} A_{i_1 j_1} \cdots A_{i_k j_k},$$

where $c_{i_1 j_1,\ldots,i_k j_k}$ are real-valued coefficients. For $k = 1, 2$, these are linear or quadratic forms, respectively. Functionals of this type are for instance related to angles and distances between such random subspaces like $\det(A^T C)$, where $A \in W_{n,d}$ and $C \in \mathbb{R}^{n \times d}$ is a matrix of coefficients which might be random with, say, independent entries. (In this case, $\det(A^T C)$ is the scalar product of the exterior d-forms induced by A and C

representing subspaces of dimension d.) In this section, we will mainly be interested in the case of $k = 2$ but also provide some basic comments on other values of k.

In particular, information about the (mixed) moments of the Stiefel matrix entries A_{ij} of order up to k is needed. To this end, note that the A_{ij} are identically distributed uncorrelated random variables with mean zero and variance $1/n$. Indeed, using the invariance under orthogonal transformations we clearly have $A_{ij} \overset{d}{=} A_{k\ell}$, where $\overset{d}{=}$ means equality in distribution, and also, multiplying suitable columns or rows by -1,

$$\int_{W_{n,d}} A_{ij} A_{k\ell} d\mu_{n,d} = - \int_{W_{n,d}} A_{ij} A_{k\ell} d\mu_{n,d} = 0$$

for any $(i, j) \neq (k, \ell)$. Moreover, the A_{ij} are centered since $A \overset{d}{=} -A$, and we have

$$d = \sum_{k,\ell=1}^{n,d} A_{k\ell}^2 = \int_{W_{n,d}} \sum_{k,\ell=1}^{n,d} A_{k\ell}^2 d\mu_{n,d} = nd \int_{W_{n,d}} A_{ij}^2 d\mu_{n,d}$$

for any (i, j). Some of these relations may also deduced from the fact that the marginal distribution of each column of A is the uniform distribution on the sphere.

These results also imply that the entries of a Grassmannian (projection) matrix P satisfy

$$\int_{G_{n,d}} P_{ij} d\nu_{n,d} = \begin{cases} d/n, & i = j \\ 0, & i \neq j \end{cases}.$$

To see this, it suffices to note that writing $P = P_A = AA^T$, we have $P_{ij} = \sum_{k=1}^{d} A_{ik} A_{jk}$.

7.1 First order results

To start, let us briefly study linear forms on $W_{n,d}$, which we may rewrite as

$$f_1(A) = \langle V, A \rangle$$

for some matrix of coefficients $V \in \mathbb{R}^{n \times d}$. Since $\nabla_W f_1(A) = P_A(V)$ and the entries of A are centered, (1.2) immediately yields

$$\mu_{n,d}(|f_1| \geq t) \leq 2 \exp\left(- \frac{(n-1)t^2}{8\|P_A(V)\|_{\mathrm{HS},\infty}^2} \right). \tag{7.1}$$

One may also choose the vector of coefficients V at random according to some distribution with independent entries or dependent entries with higher order uncorrelatedness conditions. In this case, results like (7.1) hold conditionally given V, for instance. Functionals of this type may be regarded as natural generalizations of weighted sums $\langle X, \theta \rangle = \theta_1 X_1 + \cdots + \theta_n X_n$ with uniformly distributed weights $\theta = (\theta_1, \ldots, \theta_n) \in S^{n-1}$ and some random vector X with higher order uncorrelated entries (cf. [BCG18, BCG20]), replacing the sphere by the Stiefel manifold.

7.2 Second order results

The central object of this section are quadratic forms on $W_{n,d}$. Recall that a classical (second order) concentration result for quadratic forms in independent subgaussian random variables is the famous Hanson–Wright inequality, cf. [HW71]. It states that for a random vector $X = (X_1, \ldots, X_n)$ with independent centered components with variance 1,

$$\mathbb{P}(|X^T M X - \mathrm{tr}(M)| \geq t) \leq 2 \exp\left(- \frac{1}{C} \min\left(\frac{t^2}{\|M\|_{\mathrm{HS}}^2}, \frac{t}{\|M\|_{\mathrm{op}}} \right) \right)$$

for any $t \geq 0$, where the constant $C > 0$ depends on the subgaussian norm of the coordinates X_i.

We derive analogues of the Hanson–Wright inequality for Stiefel manifolds, i.e., we consider

$$f_2(A) = \mathrm{vec}(A)^T M \mathrm{vec}(A) \equiv A^T M A \tag{7.2}$$

for some matrix $M \in \mathbb{R}^{nd \times nd}$ which, for simplicity, we assume to be symmetric. Note that the informal notation $A^T M A$ extends (2.1). For instance, to go back to the example sketched at the beginning of this section, we may consider $\det(A^T C)$, where for simplicity we assume $C \in \mathbb{R}^{n \times d}$ to be fixed, in the case of $d = 2$. Here,

$$\det(A^T C) = \Big(\sum_{i=1}^{n} a_{i1} c_{i1} \Big) \Big(\sum_{j=1}^{n} a_{i2} c_{i2} \Big) - \Big(\sum_{i=1}^{n} a_{i1} c_{i2} \Big) \Big(\sum_{j=1}^{n} a_{i2} c_{i1} \Big) = A^T M A$$

for the matrix M with entries $M_{ij,kl} = (c_{ij} c_{kl} - c_{il} c_{kj})/2$. More advanced examples will be discussed in a further subsection.

In the next theorem, we present three different Hanson–Wright type inequalities for Stiefel manifolds. Write

$$U := \Big(\sum_{k,l} M_{ij,kl} A_{kl} \Big)_{i,j}, \qquad B := M - (A \circ U) \otimes I_n$$

for $i, k \leq n$, $j, l \leq d$, noting that $\nabla_W f_2(A) = 2 P_A U$ and $f_{2W}''(A) = 2 P_A B P_A$.

Theorem 7.1 (Hanson–Wright type inequalities for Stiefel manifolds). *1. For any $t \geq 0$, we have*

$$\mu_{n,d}(|f_2 - \mathrm{tr}(M)/n| \geq t) \leq 2 \exp\Big(-\frac{1}{C} \min\Big(\frac{(n-2)^2 t^2}{\|M\|_{\mathrm{HS}}^2}, \frac{(n-2)t}{\|M\|_{\mathrm{op}}} \Big) \Big),$$

where we may choose $C = 128 e^2 / \log(2)$.

2. For any $t \geq 0$, we have

$$\mu_{n,d}(|f_2 - \mathrm{tr}(M)/n| \geq t) \leq 2 \exp\Big(-\frac{1}{C} \min\Big(\frac{(n-2)t^2}{\|P_A U\|_{\mathrm{HS},2}^2}, \frac{(n-2)t}{\|P_A B P_A\|_{\mathrm{op},\infty}} \Big) \Big),$$

where we may choose $C = 32 e^2 / \log(2)$.

3. For any $t \geq 0$, we have

$$\mu_{n,d}(|f_2 - \mathrm{tr}(M)/n| \geq t) \leq 2 \exp\Big(-\frac{1}{C} \min\Big(\frac{(n-2-8d)^2 t^2}{\|P_A B P_A\|_{\mathrm{HS},2}^2}, \frac{(n-2)t}{\|P_A B P_A\|_{\mathrm{op},\infty}} \Big) \Big),$$

where we may choose $C = 256 e^2 / \log(2)$.

For the notation used in parts 2 and 3, cf. (3.7). The three different analogues of the classical Hanson–Wright inequality presented in Theorem 7.1 reflect a certain flexibility which can turn out to be useful in applications (in particular, it would be hard to speak of a single canonical Stiefel variant of the Hanson–Wright inequality). Certainly, part 1 is the most immediate analogue, involving quantities which can easily be calculated. However, unlike parts 2 and 3, it does not make use of intrinsic derivatives and can therefore be less accurate.

For instance, in the trivial case $M = I_{nd}$ (thus, $f_2(A) - \mathrm{tr}(M)/n \equiv 0$) part 1 does not yield an optimal result, while parts 2 and 3 do. On the other hand, if $M_{ij,kl} = 1$ for every i, j, k, l, parts 1, 2 and 3 are not substantially different. As compared to part 2, part 3 is nearer to the classical Hanson–Wright inequality, but this is the cost of an additional d-dependence in the subgaussian term, which can be avoided in part 2.

For explicit calculations of the L^2 norms appearing in parts 2 and 3, note that we may construct a uniformly distributed Stiefel element via $A = O_n I_{n,d} O_d$, where O_n and O_d are distributed according to the Haar measure on $SO(n)$ and on $SO(d)$, respectively, and $I_{n,d}$ is the element in $W_{n,d}$ whose columns are the first d unit vectors. Then $P_A = AA^T = O_n I_{n,d} I_{n,d}^T O_n^{-1} = (\langle p_d \theta_j, p_d \theta_k \rangle)_{j,k=1,\dots,n}$ (where $O_n = (\theta_1, \dots, \theta_n)$ and p_d is the projection to the first d coordinates). Thus rewriting $\|P_A U\|_{\mathrm{HS},2}^2$ resp. $\|P_A B P_A\|_{\mathrm{HS},2}^2$ as linear combinations of products of components of orthogonal matrices, we may now average them over $O(n)$ and $O(d)$ by applying formulas from the Weingarten calculus. The only non-vanishing products in this average are those whose indices are equal subject to a scheme of two pair partitions which will determine a so-called Weingarten function which is a rational function of n of order $O(n^{-2})$ or smaller. For explicit formulas covering our cases see in particular [CM09], table VII. For an introduction to the underlying theory we refer to [CMN22] and for more details to e. g. [CS06, BCS11].

Proof of Theorem 7.1. To see part 1, we extend f_2 to $X \in \mathbb{R}^{n \times d}$ by the canonical choice $f_2(X) = \mathrm{vec}(X)^T M \mathrm{vec}(X)$ and apply Theorem 1.3, part 2. Clearly, $f''(A) = 2M$, and in the subgaussian term, we use that by Poincaré inequality,

$$\|\nabla f\|_{\mathrm{HS},2} \leq \frac{4}{n-2} \|f''\|_{\mathrm{HS},2}$$

whenever ∇f has centered components (indeed, the proof is a simplification of the proof of Lemma 6.3). To see parts 2 and 3, it remains to apply Theorem 1.1, parts 2 and 3, respectively, using that $\nabla_W f(A) = 2P_A U$ and $f''_W(A) = 2P_A B P_A$. $\qquad\square$

7.3 Higher orders

Let us briefly comment on possible results for polynomial chaos of order $k \geq 3$. In general, similarly as in Theorem 7.1, part 1, one may derive concentration bounds for f_k by applying Theorem 1.3 (for order k). Here, for $k = 3$, by similar arguments as for orders 1 and 2 (including symmetry) we may use that $\int_{W_{n,d}} A_{ij} A_{kl} A_{pq} d\mu_{n,d} = 0$ for any choice of the indices. However, already for $k = 4$ the calculation of the mixed moments (e. g. with each index appearing twice) can turn out to be quite involved.

Using the L^p norm inequalities from Section 6, it is also possible to derive results in the style of [AW15] (not only for polynomial chaos but also for general $\mathcal{C}^k$ functions) which yield highly elaborate and accurate tail bounds in terms of a large family of tensor-type norms. We will not pursue this direction in this note.

Note though that all these results will make use of Euclidean derivatives only. In fact, the calculus of intrinsic derivatives gets increasingly involved as the order increases. For functions on the sphere, in [BGS19] higher order concentration results depending on spherical partial derivatives have been derived, which can be thought of as a sort of intermediate object between Euclidean and intrinsic derivatives.

7.4 Applications

Typical applications of concentration inequalities for Stiefel and Grassmann functionals include bounds for distances between subspaces. If $E, F \subset \mathbb{R}^n$ are two subspaces of dimension d, canonical choices for the distance between them are $\|P_E - P_F\|_{\mathrm{HS}}$ or $\|P_E - P_F\|_{\mathrm{op}}$. In the sequel, we will always choose one subspace at random, while the other one is fixed.

As a start, we consider the following simple example, in which the function $G_{n,d} \ni P \mapsto |P - P_F| \equiv \|P - P_F\|_{\mathrm{HS}}$ is studied.

Proposition 7.2. *Let $F \subset \mathbb{R}^n$ be any fixed d-dimensional subspace of $\mathbb{R}^n$, and denote by P_F the corresponding projection matrix. Then, for any $t \geq 0$,*

$$\nu_{n,d}(||P - P_F|^2 - 2d(1 - d/n)| \geq t) \leq 2\exp(-(n-1)t^2/(64d)).$$

Proof. Note that $|P - P_F|^2 = 2(d - \langle P, P_F \rangle)$, which has expectation $2d(1 - d/n)$ since

$$\int_{G_{n,d}} \langle P, P_F \rangle dP = d/n \, \mathrm{tr}(P_F) = d^2/n.$$

Furthermore, the function $\mathbb{R}^{n \times n} \ni X \mapsto 2(d - \langle X, P_F \rangle)$ has derivative $-2P_F$, which has Euclidean (Hilbert–Schmidt) norm $\sqrt{d}$. The result now follows immediately by combining (1.2) (in the Grassmann version) and Proposition 5.1. $\qquad\square$

It is possible to remove the square and consider fluctuations of $|P - P_F|$ by adapting some of the arguments used in the proof of Lemma 7.4 below. We skip the details, however, especially since we are rather interested in applying higher order concentration results in what follows.

To this end, let us consider another distance between two subspaces E and F of dimension d which has been studied in [WWF06] and subsequent papers. Here, we chose any orthonormal basis of E, say, $e_1, \ldots, e_d$, and set

$$\mathrm{dist}(E, F) := \Big(\sum_{j=1}^d \mathrm{dist}(e_j, F)^2\Big)^{1/2},$$

where for any $x \in \mathbb{R}^n$, $\mathrm{dist}(x, F) = \min_{y \in F} |x - y|$ is the usual point-to-subspace distance induced by the Euclidean norm. It is not hard to see that $\mathrm{dist}(E, F)$ does not depend on the choice of the orthonormal basis of E, cf. [WWF06, Theorem 1].

As above, let us fix F and consider the distance as a function of the subspace E. In view of the definition via orthonormal bases (and the invariance under change of the basis), we may then regard dist as a function on the Stiefel manifold $W_{n,d}$. That is, writing $A_{\bullet j}$ for the columns of $A \in W_{n,d}$, we shall study the fluctuations of

$$\mathrm{dist}(A, F) := \Big(\sum_{j=1}^d \mathrm{dist}(A_{\bullet j}, F)^2\Big)^{1/2}$$

under $\mu_{n,d}$. In this situation, the function fits into a framework which allows to derive concentration bounds via Hanson–Wright inequalities similarly as in [RV13, Corollary 3.1], where distances of random vectors to a fixed subspace have been stuied. Under the uniform distribution on the Stiefel manifold, we obtain the following result.

Proposition 7.3. *Let $F \subset \mathbb{R}^n$ be any fixed d-dimensional subspace of $\mathbb{R}^n$. Then, for any $t \geq 0$,*

$$\mu_{n,d}(|\mathrm{dist}(A, F) - d/\sqrt{n}| \geq t) \leq 2\exp(-(n-2)t^2/C)$$

with $C = 384e^2/\log(2)$.

To prove Proposition 7.3, we need the following lemma, which is itself of independent interest since it generalizes [RV13, Theorem 2.1] (cf. also [S23, Proposition 2.2]). In essence, given any matrix $M \in \mathbb{R}^{nd \times nd}$ it controls the fluctuations of the Euclidean norm of MA around the Hilbert–Schmidt norm of M, rescaled by $n^{-1/2}$ according to the standard deviation of the entries of $A \sim \mu_{n,d}$.

Lemma 7.4. *For any fixed $M \in \mathbb{R}^{nd \times nd}$, we have*

$$\mu_{n,d}(|\|MA| - \|M\|_{\mathrm{HS}}/\sqrt{n}| \geq t) \leq 2\exp\Big(-\frac{(n-2)t^2}{C\|M\|_{\mathrm{op}}^2}\Big)$$

with $C = 384e^2/\log(2)$.

Proof. Let $Q = M^T M$, so that $|MA|^2 = A^T Q A$ (recall (2.1) and (7.2)) and $\operatorname{tr}(Q) = \|M\|_{\mathrm{HS}}^2$. Assuming that $\|M\|_{\mathrm{HS}} = \sqrt{n}$, Theorem 7.1, part 1, yields

$$\mu_{n,d}(||MA|^2 - 1| \geq t) \leq 2 \exp\left(-\frac{1}{C} \min\left(\frac{(n-2)^2 t^2}{n\|M\|_{\mathrm{op}}^2}, \frac{(n-2)t}{\|M\|_{\mathrm{op}}^2} \right) \right)$$

$$\leq 2 \exp\left(-\frac{n-2}{3C\|M\|_{\mathrm{op}}^2} \min(t^2, t) \right),$$

where the first step follows from $\|Q\|_{\mathrm{HS}}^2 \leq \|M\|_{\mathrm{op}}^2 \|M\|_{\mathrm{HS}}^2 = n\|M\|_{\mathrm{op}}^2$ and $\|Q\|_{\mathrm{op}} \leq \|M\|_{\mathrm{op}}^2$, and $C = 128e^2/\log(2)$.

Now, as in [RV13], we use the inequality $|z - 1| \leq \min(|z^2 - 1|, |z^2 - 1|^{1/2})$, giving for any $t \geq 0$

$$\mu_{n,d}\left(||MA| - 1| \geq t\right) \leq \mu_{n,d}(||MA|^2 - 1| \geq \max(t, t^2)).$$

Combining this with the first step and using that $\min(\max(t, t^2)^2, \max(t, t^2)) = t^2$ yields

$$\mu_{n,d}(||MA| - 1| \geq t) \leq 2 \exp\left(-\frac{(n-2)t^2}{3C\|M\|_{\mathrm{op}}^2} \right),$$

i.e. the claim for for $\|M\|_{\mathrm{HS}} = \sqrt{n}$. The general case now follows by considering $\tilde{M} := \sqrt{n} M \|M\|_{\mathrm{HS}}^{-1}$, noting that

$$\mu_{n,d}(||MA| - \|M\|_{\mathrm{HS}}/\sqrt{n}| \geq t) = \mu_{n,d}(||\tilde{M}A| - 1| \geq \sqrt{n}t\|M\|_{\mathrm{HS}}^{-1}). \qquad \square$$

Proof of Proposition 7.3. Let P_F be the orthogonal projection onto F, and denote by $A_{\bullet j}$ the columns of A. Then, $\operatorname{dist}(A_{\bullet j}, F) = |P_F A_{\bullet j}|$. In particular, if $M := I_d \otimes P_F$ is the $nd \times nd$ block diagonal matrix whose diagonal consists of d copies of the matrix P_F, $\operatorname{dist}(A, F) = |MA|$, so that we may apply Lemma 7.4. Here we use that $\|M\|_{\mathrm{HS}}^2 = d\|P_F\|_{\mathrm{HS}}^2 = d^2$ and moreover that since d eigenvalues of P_F are 1 and all other eigenvalues are 0, we have $\|M\|_{\mathrm{op}} = \|P_F\|_{\mathrm{op}} = 1$. $\qquad \square$

In fact, it is not even necessary to consider subspaces F of the same dimension d. If $F \subset \mathbb{R}^n$ is a subspace of any dimension $m < n$, the definitions of $\operatorname{dist}(E, F)$ and $\operatorname{dist}(A, F)$ continue to hold (even if $\operatorname{dist}(E, F)$ is no longer symmetric in E and F in this case), and with minimal adaptions in the proof of Proposition 7.3, one can show that

$$\mu_{n,d}(|\operatorname{dist}(A, F) - \sqrt{md/n}| \geq t) \leq 2 \exp(-(n-2)t^2/C)$$

with $C = 384e^2/\log(2)$.

A Proofs of the results from Section 5

Proof of Proposition 5.3. Considering the function

$$\psi_V(P) := \langle \nabla_G f(P), V \rangle = \langle [P, [P, \nabla f(P)]], V \rangle,$$

let us calculate $\nabla \psi_V(P)$. Choosing some extension of f onto an open neighborhood of $G_{n,d}$ in $\mathbb{R}^{n \times n}$, the function

$$\psi_V(X) := \langle [X, [X, \pi_{\mathrm{sym}}(\nabla f(X))]], V \rangle$$
$$= \operatorname{tr}(V^T(X\pi_{\mathrm{sym}}(\nabla f(X)) + \pi_{\mathrm{sym}}(\nabla f(X))X - 2P\pi_{\mathrm{sym}}(\nabla f(X))X))$$

is a smooth extension of $\psi_V(P)$. Note that π_{sym} has derivative

$$\frac{d\pi_{\mathrm{sym}}(Y)}{dY} = \frac{1}{2}\frac{d(Y + Y^T)}{dY} = \frac{1}{2}(I_{n^2} + K_{n,n}).$$

As a consequence, by chain rule,

$$\frac{d\pi_{\mathrm{sym}}(\nabla f(X))}{dX} = \frac{1}{2}(I_{n^2} + K_{n,n})f''(X).$$

Let us now calculate

$$\frac{d}{dX}[V^T(X\pi_{\mathrm{sym}}(\nabla f(X)) + \pi_{\mathrm{sym}}(\nabla f(X))X - 2X\pi_{\mathrm{sym}}(\nabla f(X))X)].$$

First, by Leibniz rule,

$$\frac{d}{dX}[V^T X\pi_{\mathrm{sym}}(\nabla f(X))]$$
$$= (\pi_{\mathrm{sym}}(\nabla f(X)) \otimes I_n)(I_n \otimes V^T) + \frac{1}{2}(I_n \otimes (V^T X))(I_{n^2} + K_{n,n})f''(X)$$
$$= \pi_{\mathrm{sym}}(\nabla f(X)) \otimes V^T + \frac{1}{2}(I_n \otimes (V^T X))(I_{n^2} + K_{n,n})f''(X).$$

Next, note that similarly,

$$\frac{d(\pi_{\mathrm{sym}}(\nabla f(X))X)}{dX} = \frac{1}{2}(X^T \otimes I_n)(I_{n^2} + K_{n,n})f''(X) + I_n \otimes \pi_{\mathrm{sym}}(\nabla f(X)).$$

In particular, we obtain that

$$\frac{d}{dX}[V^T \pi_{\mathrm{sym}}(\nabla f(X))X]$$
$$= (I_n \otimes V^T)\Big(\frac{1}{2}(X^T \otimes I_n)(I_{n^2} + K_{n,n})f''(X) + I_n \otimes \pi_{\mathrm{sym}}(\nabla f(X))\Big)$$
$$= \frac{1}{2}(X^T \otimes V^T)(I_{n^2} + K_{n,n})f''(X) + I_n \otimes (V^T \pi_{\mathrm{sym}}(\nabla f(X))).$$

Finally,

$$\frac{d}{dX}[V^T X\pi_{\mathrm{sym}}(\nabla f(X))X]$$
$$= ((X^T \pi_{\mathrm{sym}}[\nabla f(X)]) \otimes I_n)\frac{d(V^T X)}{dX} + (I_n \otimes (V^T X))\frac{d}{dX}[\pi_{\mathrm{sym}}(\nabla f(X))X]$$
$$= ((X^T \pi_{\mathrm{sym}}[\nabla f(X)]) \otimes I_n)(I_n \otimes V^T)$$
$$\quad + (I_n \otimes (V^T X))\Big(\frac{1}{2}(X^T \otimes I_n)(I_{n^2} + K_{n,n})f''(X) + I_n \otimes \pi_{\mathrm{sym}}(\nabla f(X))\Big)$$
$$= (X^T \pi_{\mathrm{sym}}[\nabla f(X)]) \otimes V^T + \frac{1}{2}(X^T \otimes (V^T X))(I_{n^2} + K_{n,n})f''(X)$$
$$\quad + I_n \otimes (V^T X\pi_{\mathrm{sym}}(\nabla f(X))).$$

Consequently, by chain rule,

$$\frac{d}{dX}\mathrm{tr}(V^T(X\pi_{\mathrm{sym}}(\nabla f(X)) + \pi_{\mathrm{sym}}(\nabla f(X))X - 2X\pi_{\mathrm{sym}}(\nabla f(X))X))$$
$$= \mathrm{vec}(I_n)^T\Big(\pi_{\mathrm{sym}}(\nabla f(X)) \otimes V^T + \frac{1}{2}(I_n \otimes (V^T X))(I_{n^2} + K_{n,n})f''(X)$$
$$\quad + \frac{1}{2}(X^T \otimes V^T)(I_{n^2} + K_{n,n})f''(X) + I_n \otimes (V^T \pi_{\mathrm{sym}}(\nabla f(X)))$$
$$\quad - 2(X^T \pi_{\mathrm{sym}}[\nabla f(X)]) \otimes V^T - (X^T \otimes (V^T X))(I_{n^2} + K_{n,n})f''(X)$$
$$\quad - 2I_n \otimes (V^T X\pi_{\mathrm{sym}}(\nabla f(X)))\Big).$$

Switching from $d/(dX)$ to $\mathrm{vec}(\nabla)$ (and thus, transposing), we obtain

$$\mathrm{vec}(\nabla\psi_V(X)) = \Big(\pi_{\mathrm{sym}}(\nabla f(X)) \otimes V + \frac{1}{2}f''(X)(I_{n^2} + K_{n,n})(I_n \otimes (X^TV))$$
$$+ \frac{1}{2}f''(X)(I_{n^2} + K_{n,n})(X \otimes V) + I_n \otimes (\pi_{\mathrm{sym}}(\nabla f(X))V)$$
$$- 2(\pi_{\mathrm{sym}}(\nabla f(X))X) \otimes V - f''(X)(I_{n^2} + K_{n,n})(X \otimes (X^TV))$$
$$- 2I_n \otimes (\pi_{\mathrm{sym}}(\nabla f(X))X^TV)\Big)\mathrm{vec}(I_n).$$

Let us simplify this further. We have

$$(\pi_{\mathrm{sym}}(\nabla f(X)) \otimes V)\mathrm{vec}(I_n) = \mathrm{vec}[V\pi_{\mathrm{sym}}(\nabla f(X))],$$
$$(I_n \otimes (\pi_{\mathrm{sym}}(\nabla f(X))V))\mathrm{vec}(I_n) = \mathrm{vec}[(\pi_{\mathrm{sym}}(\nabla f(X))V)],$$
$$((\pi_{\mathrm{sym}}(\nabla f(X))X) \otimes V)\mathrm{vec}(I_n) = \mathrm{vec}[VX^T\pi_{\mathrm{sym}}(\nabla f(X))],$$
$$(I_n \otimes (\pi_{\mathrm{sym}}(\nabla f(X))X^TV))\mathrm{vec}(I_n) = \mathrm{vec}[\pi_{\mathrm{sym}}(\nabla f(X))X^TV].$$

Recalling that $\pi_{\mathrm{sym}} = (I_{n^2} + K_{n,n})/2$, we may further rewrite

$$\frac{1}{2}f''(X)(I_{n^2} + K_{n,n})(I_n \otimes (X^TV))\mathrm{vec}(I_n) = f''(X)\pi_{\mathrm{sym}}\mathrm{vec}(X^TV),$$
$$\frac{1}{2}f''(X)(I_{n^2} + K_{n,n})(X \otimes V)\mathrm{vec}(I_n) = f''(X)\pi_{\mathrm{sym}}\mathrm{vec}(VX^T),$$
$$f''(X)(I_{n^2} + K_{n,n})(X \otimes (X^TV))\mathrm{vec}(I_n) = f''(X)\pi_{\mathrm{sym}}\mathrm{vec}(X^TVX^T)$$

Summing up and restricting to $P \in G_{n,d}$ (which is symmetric, in particular),

$$\mathrm{vec}(\nabla\psi_V(P)) = f''(P)\pi_{\mathrm{sym}}\mathrm{vec}([P, [P, V]]) + 2\mathrm{vec}[\pi_{\mathrm{sym}}(V\pi_{\mathrm{sym}}(\nabla f(P)))]$$
$$- 4\mathrm{vec}[\pi_{\mathrm{sym}}(VP\pi_{\mathrm{sym}}(\nabla f(P)))].$$

Next we calculate the intrinsic derivatives. First, we need to project onto $\mathbb{R}^{n\times n}_{\mathrm{sym}}$ by applying π_{sym}. Using the notation from Proposition 5.2, this yields

$$\pi_{\mathrm{sym}}\nabla\psi_V(P) = \pi_{\mathrm{sym}}f''(P)\pi_{\mathrm{sym}}\mathrm{vec}([P, [P, V]]) + 2\mathrm{vec}[\pi_{\mathrm{sym}}(V\pi_{\mathrm{sym}}(\nabla f(P)))]$$
$$- 4\mathrm{vec}[\pi_{\mathrm{sym}}(VP\pi_{\mathrm{sym}}(\nabla f(P)))].$$

Now we project onto T_P by applying π_P. This yields

$$\nabla_G\psi_V(P) = \pi_P\pi_{\mathrm{sym}}f''(P)\pi_{\mathrm{sym}}\pi_P(V) + 2\pi_P\mathrm{vec}[\pi_{\mathrm{sym}}(V\pi_{\mathrm{sym}}(\nabla f(P)))]$$
$$- 4\pi_P\mathrm{vec}[\pi_{\mathrm{sym}}(VP\pi_{\mathrm{sym}}(\nabla f(P)))].$$

By Proposition 5.2, it follows that

$$f''_G(P)V = \nabla_G\psi_V(P) - 2\pi_P\mathrm{vec}[\pi_{\mathrm{sym}}(V\pi_{\mathrm{sym}}(\nabla f(P)))]$$
$$+ 4\pi_P\mathrm{vec}[\pi_{\mathrm{sym}}(VP\pi_{\mathrm{sym}}(\nabla f(P)))] - [P, [\pi_{\mathrm{sym}}(\nabla f(P)), \pi_PV]].$$

Now, a direct calculation yields

$$- 2\pi_P\mathrm{vec}[\pi_{\mathrm{sym}}(V\pi_{\mathrm{sym}}(\nabla f(P)))] + 4\pi_P\mathrm{vec}[\pi_{\mathrm{sym}}(VP\pi_{\mathrm{sym}}(\nabla f(P)))]$$
$$- [P, [\pi_{\mathrm{sym}}(\nabla f(P)), \pi_PV]]$$
$$= -[P, [\pi_P\nabla f(P), V]] = -[P, [\nabla_G f(P), V]],$$

which finishes the proof. $\qquad\square$

Proof of Proposition 5.4. Similarly as in the proof of Proposition 4.1, it follows from Proposition 5.3 applied to any $V \in \mathbb{R}^{n \times n}_{\mathrm{sym}}$ such that $|V| \equiv \|V\|_{\mathrm{HS}} = 1$ together with compactness arguments that $|\nabla_G f|$ has finite Lipschitz semi-norm.

To show the identity for second order modulus of gradient, fix $P \in G_{n,d}$ and note that by the definition of the intrinsic gradient and Proposition 5.3, we have

$$\langle \nabla_G f(P'), V \rangle = \langle \nabla_G f(P), V \rangle + \langle \tilde{V}, P' - P \rangle + o(|P' - P|),$$

where

$$\tilde{V} = f''_G(P)V + [P, [\nabla_G f(P), V]].$$

Moreover, using the integral form of the Taylor formula and the compactness of $G_{n,d}$, which implies that every continuous function is already uniformly continuous, we see that the remainder term in the Taylor expansion can be bounded independently of $V \in \mathbb{R}^{n \times n}_{\mathrm{sym}}$ such that $|V| = 1$, i. e.

$$\sup_{|V|=1} |\langle \nabla_G f(P'), V \rangle - \langle \nabla_G f(P), V \rangle - \langle \tilde{V}, P' - P \rangle| \leq \varepsilon(|P' - P|),$$

where $\varepsilon(t)$ is some function which satisfies $\varepsilon(t) \to 0$ as $t \to 0$.

We now rewrite the Taylor formula as

$$\langle \nabla_G f(P'), V \rangle = \langle \nabla_G f(P) + L, V \rangle + o(|P' - P|) \tag{A.1}$$

by choosing a suitable L such that $\langle L, V \rangle = \langle \tilde{V}, P' - P \rangle$. Obviously,

$$\langle f''_G(P)V, P' - P \rangle = \langle f''_G(P)(P' - P), V \rangle.$$

Moreover, write

$$[P, [\nabla_G f(P), V]] = P \nabla_G f(P) V - P V \nabla_G f(P) - \nabla_G f(P) V P + V \nabla_G f(P) P.$$

Adapting the identities (4.2) to $U, W \in \mathbb{R}^{n \times n}$ and $V, X (= P' - P) \in \mathbb{R}^{n \times n}_{\mathrm{sym}}$, we obtain

$$\langle U^T X W^T, V \rangle = \langle UVW, X \rangle = \langle WXU, V \rangle,$$

where we have used the symmetry of V and X. If we apply the first identity to $P \nabla_G f(P) V$ and $P V \nabla_G f(P)$ and the second one to $\nabla_G f(P) V P$ and $V \nabla_G f(P) P$, we obtain

$$\langle [P, [\nabla_G f(P), V]], P' - P \rangle = 2 \langle [\nabla_G f(P), P(P' - P)], V \rangle,$$

so that altogether, we may also set

$$L := f''_G(P)(P' - P) + 2[\nabla_G f(P), P(P' - P)].$$

Now we take an absolute value on both sides of the Taylor formula (A.1) and take the supremum over all symmetric V such that $\mathrm{vec}(V) \in S^{nd-1}$. This leads to

$$|\nabla_G f(P')| = |\nabla_G f(P) + L| + o(|P' - P|).$$

Next, we write

$$|\nabla_G f(P) + L|^2 = |\nabla_G f(P)|^2 + 2\langle \nabla_G f(P), L \rangle + |L|^2.$$

Note that as a trace of the product of a symmetric and an anti-symmetric matrix, $\langle \nabla_G f(P), [\nabla_G f(P), P(P' - P)] \rangle = 0$. Therefore,

$$\langle \nabla_G f(P), L \rangle = \langle \nabla_G f(P), f''_G(P)(P' - P) \rangle = \langle U, P' - P \rangle$$

with $U := f_G''(P)\nabla_G f(P)$. Since $|L|^2 = O(|P' - P|^2)$, we obtain

$$|\nabla_G f(P) + L|^2 = |\nabla_G f(P)|^2 + 2\langle U, P' - P\rangle + o(|P' - P|).$$

If $|\nabla_G f(P)| > 0$, it therefore follows that

$$|\nabla_G f(P) + L| = |\nabla_G f(P)| + |\nabla_G f(P)|^{-1}\langle U, P' - P\rangle + o(|P' - P|).$$

Hence,

$$|\nabla_G f(P')| - |\nabla_G f(P)| = |\nabla_G f(P)|^{-1}\langle U, P' - P\rangle + o(|P' - P|)$$

and thus

$$\limsup_{P' \to P} \frac{\big||\nabla_G f(P')| - |\nabla_G f(P)|\big|}{|P' - P|} = |\nabla_G f(P)|^{-1}\limsup_{P' \to P} \frac{|\langle U, P' - P\rangle|}{|P' - P|}$$
$$= |\nabla_G f(X)|^{-1}|\nabla_G \psi_U(P)|,$$

where $\psi_U(P) := \langle U, P\rangle$. As noted after Proposition 5.2, it holds that $U \in T_P$, so that $\nabla_G \psi_U(P) = U$. Thus, we arrive at

$$|\nabla_G^{(2)} f(P)| = |\nabla_G f(P)|^{-1}|f_G''(P)\nabla_G f(P)|$$

if $|\nabla_W f(P)| > 0$.

It remains to consider the case where $|\nabla_G f(P)| = 0$. Here, $L = f_G''(P)(P' - P)$, and the Taylor formula reads

$$|\nabla_G f(P')| = |L| + o(|P' - P|).$$

It follows that

$$|\nabla_G^{(2)} f(P)| = \limsup_{P' \to P} \frac{|\nabla_G f(P')|}{|P' - P|} = \limsup_{P' \to P} \frac{|f_G''(P)(P' - P)|}{|P' - P|}$$
$$= \limsup_{V \to 0, V \in T_P^\perp} \frac{|f_G''(P)V|}{|V|} = \|f_G''(P)\|_{\mathrm{op}},$$

which finishes the proof. $\qquad\square$

References

[AL12] Adamczak, R., and Latała, R.; *Tail and moment estimates for chaoses generated by symmetric random variables with logarithmically concave tails*, Ann. Inst. Henri Poincaré Probab. Stat. **48**(4) (2012), 1103–1136. MR3052405

[AW15] Adamczak, R., and Wolff, P.: *Concentration inequalities for non-Lipschitz functions with bounded derivatives of higher order*, Probab. Theory Relat. Fields **162**(3-4) (2015), 531–586. MR3383337

[Af57] Afriat, S. N.: *Orthogonal and oblique projectors and the characteristics of pairs of vector spaces*, Math. Proc. Camb. Philos. Soc. **53** (1957), 800–816. MR0094880

[AS94] Aida, S., and Stroock, D.: *Moment estimates derived from Poincaré and logarithmic Sobolev inequalities*, Math. Res. Lett. **1**(1) (1994), 75–86. MR1258492

[BCS11] Banica, T., Collins, B., and Schlenker, J. M.: *On polynomial integrals over the orthogonal group*, J. Comb. Theory Ser. A. **118**(3) (2001), 778–795. MR2745424

[BZA20] Bendokat, T., Zimmermann, R., and Absil, P.-A.: *A Grassmann Manifold Handbook: Basic Geometry and Computational Aspects*, arXiv:2011.13699.

[BCG17] Bobkov, S. G., Chistyakov, G. P., and Götze, F.: *Second-order concentration on the sphere*, Commun. Contemp. Math. **19**(5) (2017), Article 1650058, 20 pp. MR3670792

[BCG18] Bobkov, S. G., Chistyakov, G. P., and Götze, F.: *Berry–Esseen bounds for typical weighted sums*, Electron. J. Probab. **23** (2018), no. 23, 22 pp. MR3858920

[BCG20] Bobkov, S. G., Chistyakov, G. P., and Götze, F.: *Normal approximation for weighted sums under a second-order correlation condition*, Ann. Probab. **48**(3) (2020), 1202–1219. MR4112712

[BGS19] Bobkov, S. G., Götze, F., and Sambale, H.: *Higher order concentration of measure*, Commun. Contemp. Math. **21**(3) (2019), Article 1850043, 36 pp. MR3947067

[CM09] Collins, B., and Matsumoto, S.: *On some properties of orthogonal Weingarten functions*, J. Math. Phys. **50**(11) (2009), Article 113516, 14 pp. MR2567222

[CMN22] Collins, B., Matsumoto, S., and Novak, J.: *The Weingarten Calculus*, Notices Amer. Math. Soc. **69**(5) (2022), 734–745. MR4415894

[CS06] Collins, B., and Śniady, P.: *Integration with respect to the Haar measure on unitary, orthogonal and symplectic group*, Commun. Math. Phys. **264**(3) (2006), 773–795. MR2217291

[EAS98] Edelman, A., Arias, T. A., and Smith, S. T.: *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl. **20**(2) (1998), 303–353. MR1646856

[Ga08] Galántai, A.: *Subspaces, angles and pairs of orthogonal projections*, Linear Multilinear Algebra **56**(3) (2008), 227–260. MR2384652

[GKZ21] Götze, F., Kabluchko, Z., Zaporozhets, D: *Grassmann angles and absorption probabilities of Gaussian convex hulls*, Zap. Nauchn. Sem. POMI **501** (2021), 126–148. MR4328185

[GS20] Götze, F., and Sambale, H.: *Second order concentration via logarithmic Sobolev inequalities*, Bernoulli **26**(1) (2020), 93–126. MR4036029

[GSS19] Götze, F., Sambale, H., and Sinulis, A.: *Higher order concentration for functions of weakly dependent random variables*, Electron. J. Probab. **24** (2019), no. 85, 19 pp. MR4003138

[GSS21b] Götze, F., Sambale, H., and Sinulis, A.: *Concentration Inequalities for Bounded Functionals via Log-Sobolev-Type Inequalities*, J. Theor. Probab. **34**(3) (2021), 1623–1652. MR4289897

[GSS21a] Götze, F., Sambale, H., and Sinulis, A.: *Concentration inequalities for polynomials in α-sub-exponential random variables*, Electron. J. Probab. **26** (2021), no. 48, 22 pp. MR4247973

[HL08] Hamm, J., and Lee, D. D.: *Grassmann discriminant analysis: a unifying view on subspace-based learning*, In: Proceedings of the 25th international conference on Machine learning, 2008, 376–383.

[HW71] Hanson, D. L., and Wright, F. T.: *A bound on tail probabilities for quadratic forms in independent random variables*, Ann. Math. Stat. **42** (1971), 1079–1083. MR0279864

[HHT07] Helmke, U., Hüper, K., and Trumpf, J.: *Newton's method on Graßmann manifolds*, arXiv:0709.2205.

[KP21] Kabluchko, Z., and Prochno, J.: *Large deviations for random matrices in the orthogonal group and Stiefel manifold with applications to random projections of product distributions*, arXiv:2110.12977.

[KPT20] Kabluchko, Z., Prochno, J., and Thäle, C.: *A new look at random projections of the cube and general product measures*, Bernoulli **27**(3) (2021), 2117–2138. MR4260512

[KV00] Kim, J. H., and Vu, V. H.: *Concentration of multivariate polynomials and its applications*, Combinatorica **20**(3) (2000), 417–434. MR1774845

[KL15] Kolesko, K., and Latała, R.: *Moment estimates for chaoses generated by symmetric random variables with logarithmically convex tails*, Stat. Probab. Letters **107** (2015), 210–214. MR3412778

[LLY20] Lai, Z., Lim, L. H., and Ye, K.: *Simpler Grassmannian optimization*, arXiv:2009.13502.

[La06] Latała, R.: *Estimates of moments and tails of Gaussian chaoses*, Ann. Probab. **34**(6) (2006), 2315–2331. MR2294983

[Le01] Ledoux, M.: *The concentration of measure phenomenon*, American Mathematical Society, 2001. MR1849347

[LPT06] Litvak, A. E., Pajor, A., and Tomczak-Jaegermann, N.: *Diameters of sections and coverings of convex bodies*, J. Funct. Anal. **231**(2) (2006), 438–457. MR2195339

[Me19] Meckes, E.: *The Random Matrix Theory of the Classical Compact Groups*, Cambridge University Press, 2019. MR3971582

[MW82] Mueller, C. E., and Weissler, F. B.: *Hypercontractivity for the heat semigroup for ultraspherical polynomials and on the n-sphere*, J. Funct. Anal. **48** (1992), 252–283. MR0674060

[RV13] Rudelson, M., and Vershynin, R.: *Hanson–Wright inequality and sub-Gaussian concentration*, Electron. Commun. Probab. **18** (2013), no. 82, 9 pp. MR3125258

[S23] Sambale, H.: *Some notes on concentration for α-subexponential random variables*, In: High Dimensional Probability IX, Eds. R. Adamczak, N. Gozlan, K. Lounici and M. Madiman, Birkhäuser, Springer, Progress in Probability **80** (2023), 167–192.

[SS21] Sambale, H., and Sinulis, A.: *Modified log-Sobolev inequalities and two-level concentration*, ALEA, Lat. Am. J. Probab. Math. Stat. **18** (2021), 855–885. MR4243518

[SS12] Schudy, W., and Sviridenko, M.: *Concentration and moment inequalities for polynomials of independent random variables*, In: Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, 2012, 437–446. MR3205227

[TVC08] P. Turaga, A. Veeraraghavan, and R. Chellappa: *Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision*, 2008 IEEE Conference on Computer Vision and Pattern Recognition. (2008), 1–8.

[WLT11] Wang, X., Li, Z., and Tao, D.: *Subspaces indexing model on Grassmann manifold for image search*, IEEE Trans. Image Process. **20**(9) (2011), 2627–2635. MR2866258

[WWF06] Wang, L., Wang, X., and Feng, J.: *Subspace Distance Analysis with Application to Adaptive Bayesian Algorithm for Face Recognition*, Pat. Rec. **39**(3) (2006), 456–464.

[ZZJH18] Zhang, J., Zhu, G., Heath Jr., R. W., and Huang, K.: *Grassmannian Learning: Embedding Geometry Awareness in Shallow and Deep Learning*, arXiv:1808.02229.