Indian J Pure Appl Math (2026) 57:1187–1210
https://doi.org/10.1007/s13226-024-00594-4

ORIGINAL RESEARCH

What is the gradient of a scalar function defined on a subspace of square matrices ?

Shriram Srinivasan · **Nishant Panda**

Received: 24 July 2023 / Accepted: 8 April 2024 / Published online: 24 April 2024

Abstract We illustrate a technique to calculate the gradient of scalar functions that are defined on *any* arbitrary matrix subspace. It generalizes our earlier work titled "What is the gradient of a scalar function of a symmetric matrix ?"(*Indian Journal of Pure and Applied Mathematics* (2022), https://doi.org/10.1007/s13226-022-00313-x), in which we considered the special case of the subspace of symmetric matrices. Extant methods to calculate the gradient in such cases have an inherent flaw which leads to spurious results that populate several publications, as well as respected textbooks and handbooks on matrix calculus. We examine these sources and results in a rigorous and concrete mathematical setting of a finite-dimensional inner-product space and discover the inherent flaw and also a remedy. We demonstrate two ways to calculate the derivative/gradient and second derivative for scalar functions of matrices defined over an arbitrary matrix subspace; the first method is by considering *any* (differentiable) extension to the space of square matrices and projection of its gradient onto the given subspace. The second method utilizes an ordered basis and computes each component of the gradient through evaluation of the directional derivative. All the ideas presented are illustrated by non-trivial examples, namely, considering the subspace of 3×3 circulant and Toeplitz matrices and presenting the results of gradient-descent with both the spurious and correct gradients. Moreover, our bibliography makes it clear that a rigorous approach to matrix calculus is not common in practice, and our presentation of matrix calculus in the language of inner-product spaces will be significant and meaningful for applied mathematicians, engineers and researchers working in inter-disciplinary fields to avoid the conceptual pitfalls that exist.

Keywords Matrix calculus, patterned matrix · Frechet Derivative · Gateaux derivative · Gradient · Hessian · Matrix functional

Mathematics Subject Classification 15A60 · 15A63 · 26B12

Communicated by Rahul Roy.

Updated April 9, 2024

S. Srinivasan (✉)
Applied Mathematics and Plasma Physics Group (T-5), Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
E-mail: shrirams@lanl.gov
URL: https://shrirams.gitlab.io

N. Panda
Information Sciences Group (CCS-3), Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
E-mail: nishpan@lanl.gov

1 Introduction

Matrix functionals, i.e., scalar functions of matrices, and their derivatives occur in disparate applications and fields such as statistics, econometrics [1], control systems [2], signal processing & communications [3], optimization [4], and mechanics [5,6]. In many diverse applications such as wireless communication [3,7], statistical learning [8] and mechanics [5,6], the matrix arguments possess symmetry, skew-symmetry, Toeplitz, or Hankel structure [9,10] which must be respected by the gradient. One salient feature of the matrix arguments in their appearance in various aforementioned applications is the fact that other than in mechanics, they do not generally describe physical quantities. For while a matrix of controls or correlations is certainly meaningful, they are not physical quantities like stress or strain. This observation is relevant because it is easy to regard a matrix as an array of scalars, especially when it does not represent a physical quantity, and think of functions of matrices as functions of matrix entries. Such an approach is antithetical to the modern viewpoint of calculus and linear algebra, where a study of derivatives of functions between normed linear spaces or inner-product spaces encompasses calculus of vector and matrix spaces and treats them in a unified manner. However, considering that influential books such as those by Halmos [11] and Dieudonnè [12] espousing the abstract viewpoint started to be adopted only in the 1960s, antithetical developments in the calculus of matrices were already becoming entrenched within research communities where the matrix as an entity did not represent any physical quantity.

The fundamental idea behind these developments appears to have been the conversion of matrices to vectors and use of the familiar rules of multivariable calculus for vectors. This succeeds when the domain is the vector space of square matrices, i.e., say $\mathbb{R}^{n\times n}$ where the dimension of the space is then n^2. In this case, the mapping from matrices in $\mathbb{R}^{n\times n}$ to vectors in $\mathbb{R}^{n^2\times 1}$ is isometric as well as isomorphic. However, if the domain is a matrix subspace of dimension $d < n^2$ (such as symmetric matrices, skew-symmetric matrices etc.), the approach to matrix calculus that seeks to convert matrices to vectors often leads to spurious results. The underlying reason (that we discover in the course of this article) is that in a d-dimensional subspace of $\mathbb{R}^{n\times n}$, while there are d basis elements, a matrix in the subspace is still represented by $n \times n$ entries. In contrast, for vectors in the d-dimensional subspace $\mathbb{R}^{d\times 1}$, there are exactly d basis vectors, and the representation has exactly d components. Thus, the mapping from a matrix to a vector while isomorphic is not necessarily isometric, and this is the underlying reason for the spurious results.

The development of matrix calculus along the aforementioned lines began in parallel in both statistics [13–17] and control systems [2,18,19]. One recurring theme in these works indicative of the underlying idea is that of "functionally independent" matrix elements, which is actually a surrogate for the dimension of the matrix subspace. Thus, in an $n \times n$ symmetric matrix, there are only $n(n+1)/2$ "functionally independent" elements due to the equality of the off-diagonal elements. The subspace of symmetric matrices is especially conspicuous in the literature, and Srinivasan and Panda [20] discuss in detail how spurious gradients appeared for this particular matrix subspace and became entrenched in published work [21–39]. The set of symmetric matrices is but one instance of what is termed in these works as "linear patterned matrices" because the matrix elements satisfy a linear constraint. Other linear patterned matrices, such as skew-symmetric and skew-hermitian matrices have also been considered but again with spurious results [3,40,41]. All these works share the same fundamental flaw that leads to incorrect results.

There are many works that have attempted to clarify the foundations of matrix calculus through the decades, but without success in our opinion. The works of Vetter [42,43] are some of the earliest, while at various times since, Pollock [44,45] and Magnus [46] have also made contributions in this regard. A perusal of these works will make clear that they suffer from the same malaise that has plagued this subject from the beginning - the intention to treat matrices as column vectors by utilizing various transformations such as "vec" operators and Kronecker products which obscure the true structure of the matrices as linear transformations. While some practitioners seem to understand the gains to be made if the underlying ideas are clarified with abstraction [47,48], the old way of doing things still returns when their exposition moves from the calculus of vectors to matrices. The exception is the article by Brinkhuis [49], which is in the same spirit of this current work, but is all too brief to be helpful. The text by Kollo and von Rossen [50] falls into a different category; it begins with the abstract definition of the Fréchet derivative but then lapses into the conventional practice of matrix calculus and deals with patterned matrices the same way and obtains similar spurious gradients. However, it then derives the correct result for derivative of scalar functions of symmetric matrices, albeit with tedious notation, but then declares (wrongly) - "*The results can not be directly applied to minimization or maximization problems, but the derivative is useful when obtaining moments for symmetric matrices, which will be illustrated later …*". This sums up the state of

matrix calculus in many ways, for there is much confusion and no clarity, as Fréchet, Gâteaux derivatives are thought to be different from "patterned matrix derivatives".

In this article, we consider the general case of calculating the gradient of a function defined on a lower-dimensional matrix subspace and unify the disparate ideas before laying out a mathematically correct procedure that is applicable for *any* matrix subspace. We illustrate the theory by specializing the results to a host of common matrix subspaces, such as symmetric, skew-symmetric, deviatoric, circulant matrices etc. However, the vision and purpose of this article is far wider. The rapid growth of machine-learning as a tool has given rise to a large research community with interest in the subject of matrix calculus [51] and the use of structured matrices within the layers of the neural network [52] who are being led astray by the conceptual pitfalls that exist in various aforementioned articles and textbooks on the subject. We intend this article to be a reference that facilitates a move away from the established practice. In order to enable that, in Section 2 we reformulate the basic ideas of calculus of matrix functions in the language of calculus on inner-product spaces . In particular, we focus on scalar functions of matrices, and outline different ways to obtain their first derivative and gradient in Sects. 3, 4, and the second derivative in Appendix B. We discuss how to use the methods described for derivatives in some common cases in Sect. 5 and conclude with Sect. 6.

2 Preliminaries for calculus on normed linear spaces

In the interest of brevity we state at the outset that all vector spaces considered in this article are assumed to be finite-dimensional and defined over the field of real numbers unless explicitly stated otherwise. However, the extension to the field of complex numbers is straightforward and remarks on the same will be made in context.

Before we get to derivatives of functions, we will state our setting required to define functions such that the idea of the derivative will make sense.

Definition 2.1 Given normed linear spaces $(\mathcal{V}, \|\cdot\|_{\mathcal{V}})$ and $(\mathcal{W}, \|\cdot\|_{\mathcal{W}})$, we shall omit reference to the underlying norm and denote the set of all linear transformations with domain in $\mathcal{V}$ and range in $\mathcal{W}$ by $\mathrm{Lin}(\mathcal{V}, \mathcal{W})$. A nonlinear function ϕ with domain a subset $\mathcal{D} \subset \mathcal{V}$ and range in the vector space $\mathcal{W}$ will be denoted by $\phi : \mathcal{V} \supset \mathcal{D} \longrightarrow \mathcal{W}$.

We now consider the definition of a Fréchet derivative on normed linear (Banach) spaces. This setting is more general than what we shall consider later but it allows us to make some remarks and clarify issues without loss of generality. Hilbert spaces (inner product spaces) are nested special cases of this general definition.

Consider a function $\phi : \mathcal{V} \supset \mathcal{D} \longrightarrow \mathcal{W}$ as in Definition 2.1. We say that ϕ is differentiable if its Fréchet derivative, defined below, exists [53].

Definition 2.2 (Fréchet derivative) Suppose $\mathcal{D} \subset \mathcal{V}$ is open and ϕ is as in Definition 2.1. The Fréchet derivative of ϕ is a mapping $d\phi : \mathcal{D} \longrightarrow \mathrm{Lin}(\mathcal{V}, \mathcal{W})$. In particular, the Fréchet derivative at $x \in \mathcal{D}$ is the unique linear transformation $d\phi(x) \in \mathrm{Lin}(\mathcal{V}, \mathcal{W})$ such that for any $\epsilon > 0$, there exists $\delta_\epsilon > 0$ so that for all $\|h\|_{\mathcal{V}} < \delta_\epsilon$, $x + h \in \mathcal{D}$ and

$$\|\phi(x+h) - \phi(x) - d\phi(x)(h)\|_{\mathcal{W}} < \epsilon \, \|h\|_{\mathcal{V}} \, .$$

In other words,

$$\lim_{\substack{\|h\|_{\mathcal{V}} \to 0 \\ x+h \in \mathcal{D}}} \frac{\|\phi(x+h) - \phi(x) - d\phi(x)(h)\|_{\mathcal{W}}}{\|h\|_{\mathcal{V}}} = 0.$$

A few remarks are in order to clarify certain implications of Definition 2.2.

Remark 2.3 (Requirement of open set) We require an open neighbourhood of $x \in \mathcal{D} \subset \mathcal{V}$ to be contained in $\mathcal{D}$ for $d\phi(x)$ to exist, which is why the derivative was defined assuming that $\mathcal{D}$ was open. If $\mathcal{D}$ is not open, then a derivative of ϕ cannot be defined for $x \in \mathcal{D}$, in the sense that we cannot define $d\phi : \mathcal{D} \longrightarrow \mathrm{Lin}(\mathcal{V}, \mathcal{W})$. However, if a subspace $\mathcal{U} \subset \mathcal{V}$ could be found such that $\mathcal{D}$ is open in $\mathcal{U}$, then $d\phi : \mathcal{D} \longrightarrow \mathrm{Lin}(\mathcal{U}, \mathcal{W})$ is still well-defined. In Definition 2.2, if no mention is made of an open subset of $\mathcal{V}$ to which x belongs, then the implicit assumption is that the domain $\mathcal{D} = \mathcal{V}$.

Remark 2.4 (Choice of norm) The Definition 2.2 does not depend on the choice of norm since in finite-dimensions, all norms are equivalent.

Remark 2.5 (Further properties) The Fréchet derivative is unique and obeys the product rule and chain rule just like differentiation in single variable calculus (see Cheney [53]).

The concept of the gradient appears only in the special case that $\mathcal{V}$ is a Hilbert (complete inner-product) space where the norm $\|\cdot\|_{\mathcal{V}}$ comes from an inner-product $\langle\cdot,\cdot\rangle_{\mathcal{V}}$ and the function ϕ is real or complex-valued, i.e., $\mathcal{W} = \mathbb{R}$ or $\mathcal{W} = \mathbb{C}$. We are interested in the case $\mathcal{W} = \mathbb{R}$ where the Riesz Representation theorem asserts the existence of the gradient in

Definition 2.6 (Gradient) Given Definition 2.1 for a scalar function ϕ on an inner-product space $\mathcal{V}$ with inner-product $\langle\cdot,\cdot\rangle_{\mathcal{V}}$ and $x \in \mathcal{V}$, the Fréchet derivative $d\phi(x) \in \mathrm{Lin}(\mathcal{V}, \mathbb{R})$ determines the gradient $\nabla\phi(x) \in \mathcal{V}$ uniquely through the identity

$$\langle \nabla\phi(x), h\rangle_{\mathcal{V}} = d\phi(x)(h) \text{ for every } h \in \mathcal{V}.$$

Remark 2.7 The derivative is clearly a more fundamental object than the gradient, since it is independent of the choice of norm and is an element of the dual space (being a linear functional), but the gradient belongs to the same vector space on which the scalar function is defined, and hence it has assumed great importance in applications.

Most practitioners of matrix calculus will have greater familiarity with the idea of partial derivatives and might associate the gradient with the concept of directional derivative. We state the association precisely and explain the connection to Fréchet derivative in

Definition 2.8 (Gâteaux derivative) Suppose $\mathcal{D} \subset \mathcal{V}$ is open and as in Definition 2.1, $\phi : \mathcal{V} \supset \mathcal{D} \longrightarrow \mathcal{W}$. The Gâteaux derivative of ϕ at $x \in \mathcal{D}$ in the (non-zero) direction $h \in \mathcal{V}$ is the unique element $d\phi_x^h \in \mathcal{W}$ such that for any $\epsilon > 0$, there exists $\delta_\epsilon > 0$ so that for all $|\lambda| < \delta_\epsilon$, $x + \lambda h \in \mathcal{D}$ and

$$\left\| \phi(x + \lambda h) - \phi(x) - \lambda d\phi_x^h \right\|_{\mathcal{W}} < \epsilon\,|\lambda|\,.$$

In other words,

$$\lim_{\substack{\lambda\to 0 \\ x+\lambda h\in\mathcal{D}}} \frac{\phi(x+\lambda h) - \phi(x)}{\lambda} = d\phi_x^h.$$

Remark 2.9 (Relationship between Gâteaux and Fréchet derivatives at a point) The Gâteaux derivative is a generalized directional derivative and is a weaker notion than the Fréchet derivative, i.e., existence of a Fréchet derivative implies the existence of the Gâteaux derivative but not vice versa. Comparison with Definition 2.2 will show that when a Fréchet derivative exists, $d\phi_x^h = d\phi(x)(h)$. For scalar functions defined on an inner-product space, this is then related to the gradient through $\langle\nabla\phi, h\rangle_{\mathcal{V}} = d\phi_x^h$ emphasizing the directional nature of the Gâteaux derivative as the component of the gradient in a given direction.

Remark 2.10 (Connection to partial derivatives in elementary calculus) Suppose $\{v_1, \ldots, v_n\}$ is an ordered basis of $\mathcal{V}$, and for $x \in \mathcal{V}$,

$$x = \sum_{i=1}^{i=n} x_i v_i,\ x_i \in \mathbb{R}.$$

Then we may write, with slight abuse of notation,

$$\phi(x) = \phi(x_1, \ldots, x_n) \text{ and for } \lambda \in \mathbb{R},\ \phi(x + \lambda v_i) = \phi(x_1, \ldots, x_i + \lambda, \ldots, x_n).$$

The partial derivative $\dfrac{\partial\phi}{\partial x_i}$ in elementary calculus is the derivative with respect to x_i assuming other arguments x_j, $j \neq i$ of ϕ are unchanged, i.e.,

$$\begin{aligned}\frac{\partial\phi}{\partial x_i} &= \lim_{\lambda\to 0} \frac{\phi(x_1, \ldots, x_i + \lambda, \ldots, x_n) - \phi(x_1, \ldots, x_i, \ldots, x_n)}{\lambda} \\ &= \lim_{\lambda\to 0} \frac{\phi(x + \lambda v_i) - \phi(x)}{\lambda} = d\phi_x^{v_i} = \langle\nabla\phi, v_i\rangle_{\mathcal{V}}\,.\end{aligned}$$

Thus, if the Fréchet derivative is known to exist, then the components of the gradient with respect to a basis may be determined from the corresponding directional derivatives.

On inner-product spaces, we shall now define other concepts that will be found useful, starting with that of the tensor product in

Definition 2.11 (Tensor product) Given $a \in \mathcal{V}, b \in \mathcal{W}$ equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{V}}$, $\langle \cdot, \cdot \rangle_{\mathcal{W}}$, the tensor product $b \otimes a \in \text{Lin}(\mathcal{V}, \mathcal{W})$ such that

$$(b \otimes a)c = \langle a, c \rangle_{\mathcal{V}} b \ \in \mathcal{W} \ \forall c \in \mathcal{V}.$$

Remark 2.12 Note that if $\{v_1, \ldots, v_n\}$ be an ordered basis of $\mathcal{V}$ and $\{w_1, \ldots, w_m\}$ be an ordered basis of $\mathcal{W}$, then $\{w_i \otimes v_j \ \forall \ 1 \leq j \leq n, 1 \leq i \leq m\}$ is an ordered basis of $\text{Lin}(\mathcal{V}, \mathcal{W})$. In particular, $\{e_i \otimes e_j \ \forall \ 1 \leq j \leq n, 1 \leq i \leq n\}$ is an ordered basis of $\text{Lin}(\mathbb{R}^n, \mathbb{R}^n)$.

Definition 2.13 (Grammian) If $\{v_1, \ldots, v_d\}$ be an ordered basis of $\mathcal{V}$, the *Grammian* or *Gram matrix* of the basis is the matrix $\mathbf{G} \in \mathbb{R}^{d \times d}$ whose entries are $\mathbf{G}_{ij} = \left\langle v_i, v_j \right\rangle_{\mathcal{V}} \ \forall 1 \leq i, j \leq d$. Note that the Grammian is invertible and positive-definite.

Definition 2.14 (Coordinate Map) If $V = \{v_1, \ldots, v_d\}$ be an ordered basis of $\mathcal{V}$, and $x \in \mathcal{V}$ has the representation $\sum_{i=1}^{i=d} x_i v_i, \ x_i \in \mathbb{R}$, then $F \in \text{Lin}(\mathcal{V}, \mathbb{R}^{d \times 1})$, called the Coordinate Map associated with the basis V, is defined through

$$F(x) = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^{d \times 1}.$$

Remark 2.15 Strictly speaking, the notation for the coordinate map F ought to show its dependence on or association with the ordered basis V of $\mathcal{V}$, but in the interest of simplicity, we will continue to denote it by F since the context will make clear what basis and vector space is being referred to.

Proposition 2.16 *(Properties of the Coordinate Map) Let $V = \{v_1, \ldots, v_d\}$ be an ordered basis of $\mathcal{V}$ and let $F \in \text{Lin}(\mathcal{V}, \mathbb{R}^{d \times 1})$ be the Coordinate Map defined in Definition 2.14. The basis and inner-product assumed in $\mathbb{R}^{d \times 1}$ are the natural basis and the usual Euclidean dot product. Then the following are true:*

1. $F^{-1} \in \text{Lin}(\mathbb{R}^{d \times 1}, \mathcal{V})$ *exists.*
2. *For $x, y \in \mathcal{V}$, $\langle x, y \rangle_{\mathcal{V}} = \langle \mathbf{G}F(y), F(x) \rangle_{\mathbb{R}^{d \times 1}} = F(x)^T \mathbf{G} F(y)$.*
3. *For any $x \in \mathcal{V}$, $c \in \mathbb{R}^{d \times 1}$, $\left\langle F^T(c), x \right\rangle_{\mathcal{V}} = \langle c, F(x) \rangle_{\mathbb{R}^{d \times 1}}$.*
4. *For any $x \in \mathcal{V}$, $dF(x)(h) = F(h) \in \mathbb{R}^{d \times 1} \ \forall h \in \mathcal{V}$.*

2.1 Specialization to matrix calculus

In order to relate the abstract definitions stated earlier to the matrix spaces of interest, let $\mathcal{M} = \mathbb{R}^{n \times n}$ be the set of $n \times n$ *square* matrices while $\mathcal{M}^+$ denotes the subset with positive determinant. Let $\mathcal{S}$ denote the set of $n \times n$ *symmetric* matrices, and $\mathcal{S}^+$ denote the set of $n \times n$ *positive-definite (symmetric)* matrices. The set of real numbers will be denoted by $\mathbb{R}$ as is customary. Both $\mathcal{M}, \mathcal{S}$ are vector spaces of matrices and are equipped with the Frobenius inner-product defined in

Definition 2.17 (Frobenius inner product) For two matrices A, B in $\mathcal{M}$,

$$\langle A, B \rangle_{\mathcal{M}} := \text{tr}\,(A^T B)$$

defines an inner product and induces the Frobenius norm on $\mathcal{M}$ via

$$\|A\|_{\mathcal{M}} := \sqrt{\text{tr}\,(A^T A)}.$$

The inner product and norm are designated with the subscript $\mathcal{M}$ for clarity, but it can be dropped whenever the relevant vector space is obvious from context. Moreover, for any matrix subspace, the Frobenius inner-product and norm are the only ones considered in this article.

Remark 2.18 (Complex field) For a complex field, instead of the transpose A^T, the Hermitian (conjugate-transpose) A^* will appear in Definition 2.17.

Remark 2.19 The tensor product for elements of $\mathcal{M}$ can now be defined with respect to the inner-product $\langle \cdot, \cdot \rangle_{\mathcal{M}}$ exactly as outlined earlier in Definition 2.11, and Definition 2.13, Definition 2.14 and Proposition 2.16 also carry over.

Remark 2.20 (Basis of $\mathcal{M}$) Note that $\{\boldsymbol{e_i e_j}^T \ \forall\ 1 \leq i, j \leq n\}$ is an ordered basis of $\mathcal{M}$, where $\{\boldsymbol{e_1}, \boldsymbol{e_2}, \ldots, \boldsymbol{e_n} \in \mathbb{R}^{n\times 1}\}$ is the natural basis in $\mathbb{R}^{n\times 1}$. The $n \times n$ matrix $\boldsymbol{e_i e_j}^T$ has the ij^{th} element unity and zero elsewhere, analogous to the tensor-product basis of $\text{Lin}(\mathbb{R}^n, \mathbb{R}^n)$. In order to emphasize the abstract viewpoint, we shall, unlike the usual matrix calculus, discourage thinking about row and column stacking and hence denote this ordered basis by $\{\boldsymbol{M}^{(k)},\ 1 \leq k \leq n^2\}$. For concreteness, we specify an ordering by choosing $\boldsymbol{M}^{(k)} = \boldsymbol{e_i e_j}^T$, $k = (i-1)n + j\ \forall\ 1 \leq i, j \leq n$. Note that there is nothing special about this ordering other than convenience.

Definition 2.21 Given the normed linear spaces $(\mathcal{M}, \|\cdot\|_{\mathcal{M}})$ and $(\mathbb{R}, |\cdot|)$, a matrix functional ϕ with domain $\mathcal{D}$ is denoted as $\phi : \mathcal{M} \supset \mathcal{D} \longrightarrow \mathbb{R}$.

In order to emphasize the interplay between the domain of the functional and the normed linear space, let us consider the following

2.2 Illustrative example

Consider the mapping that takes matrix $A \mapsto \log\det(A)$, whose gradient is glibly reported to be A^{-T}. We could think of the domain of the function as being either $\mathcal{M}^+$ or $\mathcal{S}^+$ (neither is a vector space), and so proceed to write $\phi_1 : \mathcal{M} \supset \mathcal{M}^+ \longrightarrow \mathbb{R}$ or $\phi_2 : \mathcal{M} \supset \mathcal{S}^+ \longrightarrow \mathbb{R}$. In the former case, $\mathcal{M}^+$ is open in $\mathcal{M}$ and so $\nabla\phi_1(A) = A^{-T}$ makes sense, but in the latter case, $\mathcal{S}^+$ is open in the subspace $\mathcal{S}$ but not in $\mathcal{M}$ and hence in that setting, a derivative cannot be defined for ϕ_2. However, if we consider $\phi_3 : \mathcal{S} \supset \mathcal{S}^+ \longrightarrow \mathbb{R}$, then it is true that $\nabla\phi_3(A) = A^{-T} = A^{-1}$.

2.3 Derivatives on subspaces

In the previous example, the sets $\mathcal{M}^+, \mathcal{S}^+$ were not proper subspaces of the parent (ambient) vector space. Given the definition of the Fréchet derivative in Definition 2.2 and a matrix functional with range in $\mathbb{R}$ (Definition 2.21), our primary interest in this article lies in the case when the set $\mathcal{D}$ is a proper subspace of the ambient vector space $\mathcal{M}$. On one hand, a lower-dimensional subspace cannot be open in a vector space of higher-dimension that contains it. Thus $\mathcal{D}$ is not open in $\mathcal{M}$ and a derivative cannot be defined as asserted. However, consider

Definition 2.22 For a given function ϕ on a proper subspace $\mathcal{D} \subset \mathcal{M}$, ϕ^{amb} is an ambient function or extension if $\phi^{\text{amb}} : \mathcal{M} \longrightarrow \mathbb{R}$ is well-defined and $\phi^{\text{amb}}(X) = \phi(X)\ \forall X \in \mathcal{D}$. Often, ϕ continues to be well-defined for $X \in \mathcal{M} \setminus \mathcal{D}$, in which case we can define $\phi^{\text{amb}}(X) = \phi(X)\ \forall X \in \mathcal{M}$.

For ϕ^{amb} defined as above and differentiable on $\mathcal{M}$, $d\phi^{\text{amb}} : \mathcal{M} \longrightarrow \text{Lin}(\mathcal{M}, \mathbb{R})$ and $\nabla\phi^{\text{amb}} : \mathcal{M} \longrightarrow \mathcal{M}$ are still well-defined, and so too are $d\phi : \mathcal{D} \longrightarrow \text{Lin}(\mathcal{D}, \mathbb{R})$ and $\nabla\phi : \mathcal{D} \longrightarrow \mathcal{D}$.

A natural question that follows is whether one can obtain the gradient of ϕ defined on $\mathcal{D}$ directly from the definition or from consideration of the extension function defined on $\mathcal{M}$. This question has led researchers astray to needlessly complicated methods and spurious results. As an example, Harville [32] justifies the complicated methods of matrix calculus by noting that the interior of $\mathcal{S}$ is empty in $\mathcal{M}$, and that hence, Definition 2.2 is not applicable.

If interest is restricted to the action of the derivative or gradient on elements of the subspace $\mathcal{D}$, i.e., $d\phi(X)(H)$ or $\langle\nabla\phi(X), H\rangle_{\mathcal{D}}$ for $X, H \in \mathcal{D}$, then it is sufficient to evaluate the action of the derivative $d\phi^{\text{amb}}(X)$ or gradient $\nabla\phi^{\text{amb}}(X)$ (with the arguments $X, H \in \mathcal{D}$ evaluated as elements of $\mathcal{M}$, a fact that will become apparent in the subsequent proof of Theorem 4.5. However, sometimes the object of interest is the derivative or gradient at $X \in \mathcal{D}$ viewed as an element of $\text{Lin}(\mathcal{D}, \mathbb{R})$ or $\mathcal{D}$ respectively.

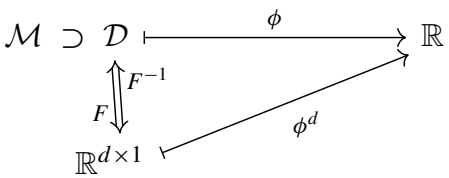
Thus, we have arrived at the central question addressed in this article which is the following - how do we conceptualize and compute the Fréchet derivative and gradient for a scalar function defined on a domain $\mathcal{D}$ which is a *lower-dimensional subspace* of $\mathcal{M}$? At this juncture, we note that for scalar functions of matrices, these lower-dimensional subspaces represent the space of matrices with particular structure or pattern. Specific cases that we shall consider include symmetric, skew-symmetric, upper/lower triangular, diagonal, deviatoric and circulant matrices. The set of orthogonal matrices is exluded from consideration since it does not form a subspace.

3 A critical view of matrix calculus development and practice

In this section, we attempt to explain the approach towards gradients and derivatives prevalent in matrix calculus and also detail how to get correct results following those methods. The viewpoint taken by the majority of practitioners of matrix calculus is the idea of viewing a matrix through its constituent entries, regarding them as degrees of freedom and linking them to the dimension of the matrix subspace. Thus, if $\dim(\mathcal{D}) = d$, a matrix in the subspace is supposed to have d "independent" entries whereas $\dim(\mathcal{M}) = n^2$ so that a matrix in the ambient space has n^2 degrees of freedom. This is indeed correct, though the reasoning is backwards. The dimension of the subspace determines the number of *basis vectors*, and the matrix entries are expansion coefficients of the linear transformation expressed in that basis. This interpretation of a matrix is different from the abstract viewpoint expressed through the definition of the Fréchet derivative, which was that of a basis-independent linear transformation. The following quote by Dwyer [14] confirms this - "*In one sense this theory of matrix derivatives is not necessary since each matrix derivative is simply a collection of scalar derivatives. In this sense the general theory of matrices is not necessary either since the matrix results are simply collections of scalar results*". The basis-dependent or component view is less abstract and somewhat limiting. One obvious (and favourable) interpretation of these remarks would be the conclusion that it is enough to compute the Gâteaux derivatives which would agree with the sentiment expressed in the quote that"*each matrix derivative is simply a collection of scalar derivatives*".

If conventional matrix calculus had followed this procedure with fidelity, there would have been no spurious results, as we shall see later in Theorem 4.8. However, the problems stem from the fact that, in the desire to stay in the familiar realm of multivariable calculus rather than matrix calculus, the scalar derivatives are associated with a function $\phi^d : \mathbb{R}^{d\times 1} \longrightarrow \mathbb{R}$ and the natural basis $\{e_1, e_2, \ldots, e_d \in \mathbb{R}^{d\times 1}\}$ (instead of the original function and matrix basis of $\mathcal{D}$), thus obtaining a *vector* gradient of ϕ^d lying in $\mathbb{R}^{d\times 1}$. Fortunately, the action of this vector gradient on elements of $\mathbb{R}^{d\times 1}$ is identical to that of the correct matrix gradient on elements of $\mathcal{D}$. The problem comes when inverting (or reshaping) the transformation to obtain the gradient in $\mathcal{D}$. The inversion recorded in published articles, books and cookbooks results in a spurious gradient that does not preserve the action. In the literature, we can see the spurious gradients recorded for the subspaces of symmetric, skew-symmetric, hermitian, and skew-hermitian matrices. While we could not find any published results for gradients with the other subspaces considered here (such as deviatoric, circulants etc.), it is certain that application of the same procedure will lead to spurious results in those cases as well. However, it is possible to obtain the correct gradient despite this long-winded procedure. This, and the observations earlier made in this paragraph will now be demonstrated rigorously.

By means of the coordinate mapping F (see Definition 2.14), the following commutative diagram will now make things clear:



(3.1)

Theorem 3.1 *(The circuitous method to calculate gradient on a subspace) Given a function $\phi : \mathcal{D} \longrightarrow \mathbb{R}$ where $\mathcal{D} \subset \mathcal{M}$ is a proper subspace of dimension d, an ordered basis $\{\boldsymbol{M}^{(i)} \in \mathcal{M} | i = 1, 2, \ldots, d\}$ of $\mathcal{D}$, and its Grammian $\mathbf{G}$:*

(a) There exists a function $\phi^d : \mathbb{R}^{d\times 1} \longrightarrow \mathbb{R}$, $\phi^d = \phi \circ F^{-1}$, with F being the coordinate map (Definition 2.14) associated with the given basis.
(b) $\langle \nabla\phi(X), H\rangle_{\mathcal{D}} = \left\langle \nabla\phi^d(x), h\right\rangle_{\mathbb{R}^{d\times 1}}$ $\forall H \in \mathcal{D}$. Here $x = F(X), h = F(H)$.
(c) $\nabla\phi(X) = F^T \nabla\phi^d(x)$, $F(\nabla\phi(X)) = \mathbf{G}^{-1}\nabla\phi^d(x)$, where $x = F(X)$.

Proof Any $X \in \mathcal{D}$ may be expressed as $X = \sum_{i=1}^{i=d} x_i \boldsymbol{M}^{(i)}$ with $x_i \in \mathbb{R}$. The commutative diagram shows how the linear map F extracts the components of the argument $X \in \mathcal{D}$ of ϕ as an element of $\mathbb{R}^{d\times 1}$, and allows us to instead think about a scalar-valued function ϕ^d of vectors in $\mathbb{R}^{d\times 1}$ that maps to the same real number $\phi(X)$. From (3.1), we obtain $\phi^d = \phi \circ F^{-1}$ which proves (a).

Again, starting from (3.1), we get $\phi = \phi^d \circ F$. Applying the chain rule for Fréchet derivatives yields

$$d\phi(X) = d\phi^d(F(X)) \circ dF(X).$$

Note that in this statement, $d\phi : \mathcal{D} \longrightarrow \text{Lin}(\mathcal{D}, \mathbb{R})$ but $d\phi^d : \mathbb{R}^{d\times 1} \longrightarrow \text{Lin}(\mathbb{R}^{d\times 1}, \mathbb{R})$. Now considering the action on an element $H \in \mathcal{D}$, denoting $x = F(X), h = F(H)$ and introducing the gradients with respect to respective inner-products, we get

$$\begin{aligned}
&\langle \nabla\phi(X), H\rangle_{\mathcal{D}} = d\phi(X)(H) = d\phi^d(F(X)) \circ dF(X)(H) = d\phi^d(F(X))(F(H)),\\
&\implies \langle \nabla\phi(X), H\rangle_{\mathcal{D}} = \left\langle \nabla\phi^d(x), h\right\rangle_{\mathbb{R}^{d\times 1}}.
\end{aligned}$$

This equation (b) proves our earlier claim that even after the transformation to a vector gradient, the action of the gradient is preserved. The fundamental flaw in current practice of matrix calculus occurs at this juncture, arguing that since $\nabla\phi^d(x) \in \mathbb{R}^{d\times 1}$, one can use the inverse map to get $F^{-1}\nabla\phi^d(x)$ as the required gradient in $\mathcal{D}$. Indeed, published (spurious) results for various matrix subspaces [3] can be recovered by this flawed approach that uses the inverse map. However, the error in this approach is that the gradient action is not preserved. The correct way to recover the gradient ensures that (b) still holds during the transformation:

$$\left\langle \nabla\phi^d(x), h\right\rangle_{\mathbb{R}^{d\times 1}} = \left\langle \nabla\phi^d(x), FF^{-1}h\right\rangle_{\mathbb{R}^{d\times 1}} = \left\langle F^T\nabla\phi^d(x), H\right\rangle_{\mathcal{D}} = \langle \nabla\phi(X), H\rangle_{\mathcal{D}},$$

from which we obtain

$$\nabla\phi(X) = F^T\nabla\phi^d(x).$$

In computations, to get the coefficients in the basis representation of $\nabla\phi(X)$, we argue as follows using Proposition 2.16:

$$\left\langle \nabla\phi^d(x), h\right\rangle_{\mathbb{R}^{d\times 1}} = \langle \nabla\phi(X), H\rangle_{\mathcal{D}} = \langle \mathbf{G}F(\nabla\phi(X)), h\rangle_{\mathbb{R}^{d\times 1}},$$

so that

$$F(\nabla\phi(X)) = \mathbf{G}^{-1}\nabla\phi^d(x).$$

□

Remark 3.2 The proof now allows us to tell at a glance why the result for diagonal matrices (see Table 1) was correct in Hjørungnes [3] while that of symmetric, hermitian, skew-symmetric, skew-hermitian were all spurious.

The (incorrect) expressions

$$\nabla\phi(X) = F^{-1}\nabla\phi^d(x), \quad F(\nabla\phi(X)) = FF^{-1}\nabla\phi^d(x) = \nabla\phi^d(x)$$

used in these published works coincide with the correct expressions if and only if $F^T = F^{-1}$ or equivalently, $\mathbf{G}$ is the identity matrix, i.e., when the transformation F is orthogonal or the basis in $\mathcal{D}$ is orthonormal. Thus, this proof validates our claim that the procedure in Hjørungnes [3] will lead to spurious results for other matrix subspaces where the natural basis is not orthonormal.

Corollary 3.3 *The following assertions follow naturally:*

$$\begin{aligned}
&\text{If } x^* \text{ solves } \arg\min_{x\in\mathbb{R}^{d\times 1}} \ \phi^d(x),\ F^{-1}(x^*) \text{ solves } \arg\min_{X\in\mathcal{D}} \ \phi(X).\\
&\text{For } x \in \mathbb{R}^{d\times 1}, X = F^{-1}(x) \in \mathcal{D},\ \nabla\phi^d(x) = 0 \iff \nabla\phi(X) = 0.
\end{aligned}$$

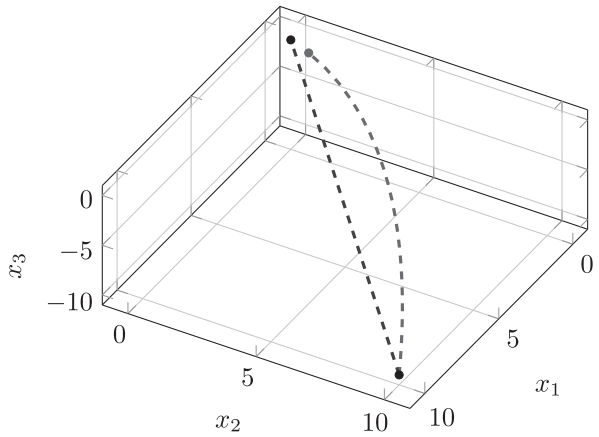
Proof The first assertion is a consequence of the identity $\phi = \phi^d \circ F$, while the second follows from the invertibility of F that also implies invertibility of F^T. □

Remark 3.4 Thus, the spurious gradient $F^{-1}\nabla\phi^d$ still preserves zeros of $\nabla\phi$.

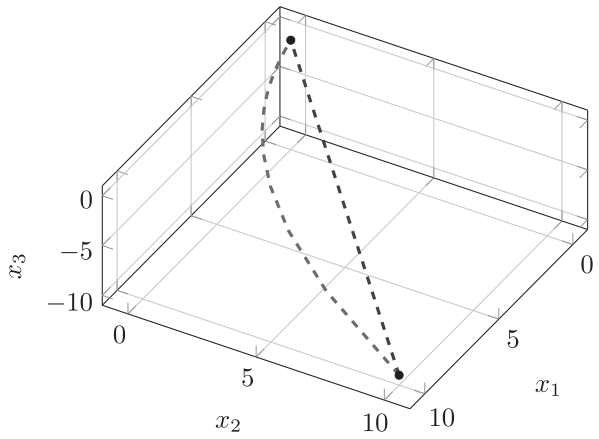
The component of the spurious gradient along the true gradient is

$$\begin{aligned}
\left\langle F^{-1}\nabla\phi^d(x), F^T\nabla\phi^d(x)\right\rangle_{\mathcal{D}} &= \left\langle FF^{-1}\nabla\phi^d(x), \nabla\phi^d(x))\right\rangle_{\mathbb{R}^{d\times 1}}\\
&= \left\langle \nabla\phi^d(x), \nabla\phi^d(x)\right\rangle_{\mathbb{R}^{d\times 1}} \geq 0,
\end{aligned}$$

implying that the spurious gradient is still *an* ascent direction. Figure 1 compares the use of both gradients in a gradient-descent algorithm to find the minimum.



(a) Toeplitz matrix $X = \begin{pmatrix} x_1 & x_2 \\ x_3 & x_1 \end{pmatrix}$



(b) Hankel matrix $X = \begin{pmatrix} x_1 & x_2 \\ x_2 & x_3 \end{pmatrix}$

Fig. 1 Iterates of gradient descent with the true (straight dashed line) and spurious (curved dashed line) gradient for the function $\phi(X) = \text{tr}\,(X^T X)$ defined over the subspace of 2×2 Toeplitz and Hankel matrices with the starting point $x_1 = x_2 = -x_3 = 10$. Compare the convergence and trajectory obtained with the true gradient and the spurious gradient towards the unique minimum $x_1 = x_2 = x_3 = 0$

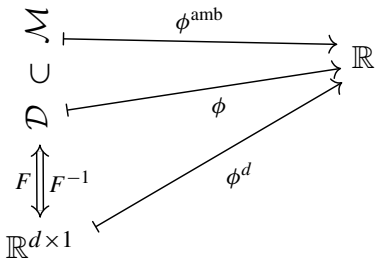
Table 1 For function $\phi : \mathcal{M} \longrightarrow \mathbb{R}$, $\nabla\phi(X)$ determines the gradient at $X \in \mathcal{D}$ for subspace $\mathcal{D}$. The table shows the gradient of ϕ restricted to $\mathcal{D}$ in terms of $\nabla\phi(X)$. The argument X is suppresed for economy of notation. The entries in the second column appear in Hjørungnes [3] in Table 6.2 on Page-185. In keeping with the conclusion of Theorem 3.1, the only subspace for which the reported gradient is correct is that of the subspace of diagonal matrices since the matrix basis considered is orthonormal

Subspace	Reported Gradient	Correct Gradient
Diagonal	$\text{diag}(\nabla\phi)$	$\text{diag}(\nabla\phi)$
Symmetric	$\nabla\phi + \nabla\phi^T - \text{diag}(\nabla\phi)$	$(\nabla\phi + \nabla\phi^T)/2$
Skew-Symmetric	$\nabla\phi - \nabla\phi^T$	$(\nabla\phi - \nabla\phi^T)/2$
Hermitian	$\nabla\phi + \nabla\phi^*$	$(\nabla\phi + \nabla\phi^*)/2$
Skew-Hermitian	$\nabla\phi - \nabla\phi^*$	$(\nabla\phi - \nabla\phi^*)/2$

4 Matrix calculus on subspaces done right

In the previous section, we illustrated the steps to get the correct gradient on matrix subspaces if one were following the rationale of existing work on matrix calculus through (3.1) and Theorem 3.1. However, it is contrary to the approach we espouse, which is the logical development of matrix calculus on subspaces starting from the definition of the Fréchet derivative. Recall that the motivating question was whether one can obtain the gradient in a matrix subspace $\mathcal{D}$ knowing the gradient in the ambient space $\mathcal{M}$ in the context of Definition 2.22. Theorem 4.5 in this section will affirm this possibility, and we shall illustrate two methods to obtain the gradient for various matrix subspaces (patterned matrices), with one based on Theorem 4.5, the other using Gâteaux derivatives in Theorem 4.8.

The *unification* of these ideas is embodied in the commutative diagram (4.1), and it also summarizes the content of this section. It shows that to compute the gradient on a subspace, one could obtain it in one of three equivalent ways: from the gradient of *any* ambient function (Theorem 4.5), directly from the function defined on the subspace (Theorem 4.8), or via Theorem 3.1. However, before we prove these assertions, we define some terms and state some results that will be be useful in subsequent discussion.



(4.1)

Definition 4.1 Given a proper subspace $\mathcal{D}$ of $\mathcal{M}$, let $\mathbb{P} \in \text{Lin}(\mathcal{M}, \mathcal{M})$ be the projection operator onto $\mathcal{D}$ such that $\text{Ran}\,(\mathbb{P}) = \mathcal{D}$ and $\text{Null}\,(\mathbb{P}) = \mathcal{D}^{\perp}$.

The construction of the projection operator proceeds through the following

Lemma 4.2 *(Projection onto subspace) Let $\{M^{(i)} \in \mathcal{M} | i = 1, 2, \ldots, d\}$ be any ordered basis of the proper matrix subspace $\mathcal{D}$ of $\mathcal{M}$ and* $\mathbf{G}$ *be the corresponding Grammian as in Definition 2.13. Then the action of the projection operator $\mathbb{P}$ on $X \in \mathcal{M}$ is defined through*

$$\mathbb{P}(X) = \sum_{i=1}^{i=d} \alpha_i M^{(i)},$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{d\times 1}$ satisfies the linear system $\mathbf{G}\boldsymbol{\alpha} = \mathbf{b}$ with $\mathbf{b} \in \mathbb{R}^{d\times 1}$ and $\mathbf{b}_i = \left\langle X, M^{(i)} \right\rangle_{\mathcal{M}}$.

Proof The statement is nothing but the projection theorem (see for example, Cheney [53]) restated for matrix spaces and may be equivalently stated as

$$\mathbb{P}(X) = \underset{Y\in\mathcal{D}}{\arg\min} \|Y - X\|_{\mathcal{M}}.$$

The projection operator ensures that any $X \in \mathcal{M}$ may be written uniquely as $X = X_p + X_0$ with $X_p = \mathbb{P}(X) \in \mathcal{D}$, $X_0 \in \mathcal{D}^{\perp}$ through the direct sum decomposition $\mathcal{M} = \text{Null}\,(\mathbb{P}) \oplus \text{Ran}\,(\mathbb{P})$. The representation for $\mathbb{P}(X)$ is a consequence of the property that $\mathbb{P}(X) - X \in \mathcal{D}^{\perp}$. □

Remark 4.3 (Orthonormal basis) If the basis $\{M^i \in \mathcal{M} | i = 1, 2, \ldots, d\}$ in Lemma 4.2 be orthonormal, then the Grammian simplifies to the identity matrix and $\mathbb{P}(X) = \sum_{i=1}^{i=d} \left\langle X, M^{(i)} \right\rangle_{\mathcal{M}} M^{(i)}$.

Thus far, we have assumed the existence of an ambient function or extension in Definition 2.22, but the following lemma shows that one can construct a multitude of functions satisfying Definition 2.22.

Lemma 4.4 *(Existence of ambient function) If $\phi : \mathcal{D} \longrightarrow \mathbb{R}$ be differentiable, then there exists differentiable $\phi^{\text{amb}} : \mathcal{M} \longrightarrow \mathbb{R}$, such that $\phi^{\text{amb}}(X) = \phi(X)\ \forall X \in \mathcal{D}$.*

Proof From the proof of Lemma 4.2, we know that $\mathcal{M} = \mathcal{D} \oplus \mathcal{D}^{\perp}$. Thus, for *any* differentiable function $\phi^{\perp} : \mathcal{D}^{\perp} \longrightarrow \mathbb{R}$ with $\phi^{\perp}(0) = 0$, we can define,

$$\text{for } X \in \mathcal{M},\ \phi^{\text{amb}}(X) = \phi(X_p) + \phi^{\perp}(X_0)$$

satisfying the required condition with the unique decomposition

$$X = X_p + X_0,\ X_p \in \mathcal{D},\ X_0 \in \mathcal{D}^{\perp}.$$

Clearly, the choice of ϕ^{amb} is non-unique.

The differentiability of ϕ^{amb} is verified using Definition 2.2 with $H \in \mathcal{M}$,

$$H = H_p + H_0,\ H_p \in \mathcal{D},\ H_0 \in \mathcal{D}^{\perp},$$

and the linear operator $d\phi^{\text{amb}}(X)$ defined through

$$d\phi^{\text{amb}}(X)(H) = d\phi(X_p)(H_p) + d\phi^{\perp}(X_0)(H_0).$$

As $\|H\|^2 = \left\|H_p + H_0\right\|^2 = \left\|H_p\right\|^2 + \|H_0\|^2$, we have

$$\left\|H_p\right\| \leq \|H\|, \|H_0\| \leq \|H\|.$$

Note that

$$\|H\| \to 0 \implies \left\|H_p\right\| \to 0,\ \|H_0\| \to 0.$$

Now consider

$$\begin{aligned} 0 &\leq \left|\phi^{\text{amb}}(X+H) - \phi^{\text{amb}}(X) - d\phi^{\text{amb}}(X)(H)\right| \ / \ \|H\| \\ &\leq \left|\phi(X_p + H_p) - \phi(X_p) - d\phi(X_p)(H_p)\right| \ / \ \left\|H_p\right\| \\ &\quad + \left|\phi^{\perp}(X_0 + H_0) - \phi^{\perp}(X_0) - d\phi^{\perp}(X_0)(H_0)\right| \ / \ \|H_0\|, \end{aligned}$$

so that

$$\lim_{\substack{\|H\| \to 0 \\ X+H \in \mathcal{M}}} \left|\phi^{\text{amb}}(X+H) - \phi^{\text{amb}}(X) - d\phi^{\text{amb}}(X)(H)\right| \ / \ \|H\| = 0,$$

thus showing that ϕ^{amb} is differentiable on $\mathcal{M}$. □

The first major result of this section provides a direct relationship for the gradient in $\mathcal{D}$ knowing the gradient in the ambient space $\mathcal{M}$ in the context of Definition 2.22. We state and prove this result in

Theorem 4.5 *(Gradient projection) If ϕ^{amb}, ϕ defined as in Definition 2.22 are Fréchet differentiable at $X \in \mathcal{D}$, then*

$$\nabla\phi(X) = \mathbb{P} \circ \nabla\phi^{\text{amb}}(X),$$

where $\mathbb{P}$ is the projection operator defined in Definition 4.1. Thus, $\nabla\phi(X)$ is independent of the choice of ϕ^{amb}.

Proof From Definition 2.2 and Definition 2.6 we see that in order to obtain a gradient at $X \in \mathcal{D}$ that is in $\mathcal{D}$, we need to start with the function ϕ rather than ϕ^{amb}. Doing so yields the defining equation

$$\langle \nabla\phi(X), H \rangle_{\mathcal{D}} = d\phi(X)(H)\ \forall H \in \mathcal{D}.$$

Additionally, if we start with the function ϕ^{amb}, we have the equation

$$\left\langle \nabla\phi^{\text{amb}}(X), H \right\rangle_{\mathcal{M}} = d\phi^{\text{amb}}(X)(H)\ \forall H \in \mathcal{M}.$$

If we restrict H to $H \in \mathcal{D}$, and observe that $\phi^{\text{amb}}(X) = \phi(X)$ on $\mathcal{D}$, the uniqueness of the derivative in Definition 2.2 gives

$$\langle \nabla\phi(X), H\rangle_{\mathcal{D}} = d\phi(X)(H) = d\phi^{\text{amb}}(X)(H) = \left\langle \nabla\phi^{\text{amb}}(X), H\right\rangle_{\mathcal{M}} \quad \forall H \in \mathcal{D}.$$

This statement is proof of the fact (stated earlier) that though $\nabla\phi(X) \in \mathcal{D}$ and $\nabla\phi^{\text{amb}}(X) \in \mathcal{M}$, they have the same action on an element $H \in \mathcal{D}$. However, in order to obtain $\nabla\phi(X)$ from $\nabla\phi^{\text{amb}}(X)$, we appeal to Lemma 4.2 which yields the unique decomposition $\nabla\phi^{\text{amb}}(X) = \mathbb{P}\left(\nabla\phi^{\text{amb}}(X)\right) + \nabla\phi^{\text{amb}}(X)^{\perp}$, where $\mathbb{P}\left(\nabla\phi^{\text{amb}}(X)\right) \in \mathcal{D}$, and $\nabla\phi^{\text{amb}}(X)^{\perp} \in \mathcal{D}^{\perp}$. Thereafter, simplification after invoking orthogonality yields

$$\nabla\phi(X) = \mathbb{P}\left(\nabla\phi^{\text{amb}}(X)\right) = \mathbb{P} \circ \nabla\phi^{\text{amb}}(X).$$

The invariance of $\nabla\phi(X)$ with respect to the choice of ϕ^{amb} is apparent since different functions can have the the same image under a projection. □

Remark 4.6 The utility of Theorem 4.5 is due to the fact that evaluating $\nabla\phi(X)$ requires us to consider only those $H \in \mathcal{D}$ which is easily accomplished only when $\mathcal{D}$ has a convenient structure. This also becomes challenging when $\phi(X)$ does not have a closed-form expression but is specified as a computational routine. In such cases, it is far easier to compute $\nabla\phi^{\text{amb}}(X)$ numerically (since $H \in \mathcal{M}$) or use the gradient $\nabla\phi^{\text{amb}}(X)$ if it be already known and solve for $\nabla\phi(X)$ as the image of the projection onto $\mathcal{D}$.

Remark 4.7 Note that we have been very deliberate and careful in using the coordinate-independent identity of elements that belong to the vector spaces and subspaces and never the matrix representation of the elements in an ordered basis. For example, as objects of the vector space, symmetric matrices are special cases of n^2 matrices, but their representation in an ordered basis has only $n(n+1)/2$ elements. Thus, the relationship between the gradients that we have derived is an abstract one, and we do not try to relate their matrix representations.

We are now in a position to state the second major result we alluded to at the start of this section, namely that when it exists, the gradient can be computed with respect to an ordered basis through the corresponding directional or Gâteaux derivatives as follows:

Theorem 4.8 *(Gradient from Gâteaux derivatives) For an ordered basis of $\mathcal{D}$, $\{M^{(i)} \in \mathcal{M} | i = 1, 2, \ldots, d\}$, any $X \in \mathcal{D}$ has the unique representation $X = \sum_{i=1}^{i=d} x_i M^{(i)}$ with $x_i \in \mathbb{R}$. For $\phi : \mathcal{D} \longrightarrow \mathbb{R}$, $\phi(X) = \phi\left(\sum_{i=1}^{i=d} x_i M^{(i)}\right)$, and we may write*

$$\nabla\phi(X) = \sum_{i=1}^{i=d}\sum_{j=1}^{j=d}\left(\mathbf{G}^{-1}\right)_{ij}\frac{\partial\phi}{\partial x_j}M^{(i)} \in \mathcal{D},$$

where $\mathbf{G}$ *is the Grammian in $\mathcal{D}$ as in Definition 2.13. If the basis be* orthonormal*, then this expression simplifies to*

$$\nabla\phi(X) = \sum_{i=1}^{i=d}\frac{\partial\phi}{\partial x_i}M^{(i)} \in \mathcal{D}.$$

Proof If

$$\nabla\phi(X) = \sum_{j=1}^{i=d}\beta_j M^{(j)},$$

one can solve for $\boldsymbol{\beta} \in \mathbb{R}^{d\times 1}$ from Remark 2.9 and Lemma 4.2, since

$$\frac{\partial\phi}{\partial x_i} = \left\langle \nabla\phi(X), M^{(i)}\right\rangle_{\mathcal{D}}.$$

to obtain

$$\nabla\phi(X) = \sum_{j=1}^{i=d}\beta_j M^{(j)} = \sum_{i=1}^{i=d}\sum_{j=1}^{j=d}\left(\mathbf{G}^{-1}\right)_{ij}\frac{\partial\phi}{\partial x_j}M^{(i)}.$$

For an orthonormal basis $\mathbf{G}_{ij} = \left(\mathbf{G}^{-1}\right)_{ij} = \delta_{ij}$, simplifying the expression further. □

Remark 4.9 Theorem 4.8 and Theorem 4.5 both aim to construct the gradient on the matrix subspace, and are equivalent. Theorem 4.5 provides a basis-independent statement, but the equivalence becomes clear when we consider that Lemma 4.2 is used in practice in conjunction with an ordered basis of the subspace, which will lead to the same expression as Theorem 4.8.

Remark 4.10 From the two representations of the gradient, Theorem 4.5 and Theorem 4.8, it would appear that the latter is more convenient since we do not need to find the projection and instead all that is needed is the set of partial (directional) derivatives of ϕ with respect to x_j. This inference is correct if the (directional) partial derivatives of ϕ with respect to x_j are to be calculated by a computational approximation. However, there can be cases where $\nabla\phi^{\text{amb}}(X)$ is known as a closed form analytical expression in which case it is easier to use Theorem 4.5 since the projector can be obtained without reference to ϕ. In contrast, Theorem 4.8 tacitly assumes that an explicit closed form expression for $\phi(X)$ is known in terms of the components x_j. This becomes clear if we consider the function $\det(\cdot)$ in the context of $\mathbb{R}^{10\times 10}$ and some subspace of it. In order to use Theorem 4.8, one needs to first evaluate the 10×10 determinant in terms of the matrix elements whereas since one has a closed form expression for the gradient in $\mathbb{R}^{10\times 10}$ it is easy to project it onto the subspace as in Theorem 4.5.

We have thus demonstrated the assertions made in (4.1) and shown three equivalent methods for determining the gradient of a scalar function defined on a matrix subspace.

4.1 Derivatives on some example subspaces

We now apply the results of Theorem 4.5 by choosing the matrix subspace $\mathcal{D}$ to be some well-known patterned matrix. In most cases, the projection is self-explanatory and it is easy to guess what $\mathcal{D}^{\perp}$ should be so we provide details only where necessary, instead quoting the final expression for $\nabla\phi(X) = \mathbb{P}\left(\nabla\phi^{\text{amb}}(X)\right)$.

4.1.1 Symmetric matrices

A matrix $X \in \mathcal{M}$ is symmetric if $X = X^T$. Let $\text{sym}(X) := (X + X^T)/2$. In this case, $\dim(\mathcal{D}) = n(n+1)/2$ and we have

$$\nabla\phi(X) = \text{sym}(\nabla\phi^{\text{amb}}(X)).$$

4.1.2 Skew-symmetric matrices

A matrix $X \in \mathcal{M}$ is skew-symmetric (anti-symmetric) if $X = -X^T$. Let $\text{skw}(X) := (X - X^T)/2$. In this case, $\dim(\mathcal{D}) = n(n-1)/2$ and we have

$$\nabla\phi(X) = \text{skw}(\nabla\phi^{\text{amb}}(X)).$$

4.1.3 Diagonal matrices

A matrix $X \in \mathcal{M}$ is diagonal if its entries $X_{ij} = 0\ \forall i \neq j$. We define $\text{diag}(X)$ to be the matrix whose main diagonal has the same entries as X but is zero elsewhere. Here, $\dim(\mathcal{D}) = n$ and

$$\nabla\phi(X) = \text{diag}(\nabla\phi^{\text{amb}}(X)).$$

4.1.4 Upper/lower triangular matrices

A matrix $X \in \mathcal{M}$ is upper-triangular if its entries $X_{ij} = 0\ \forall i > j$. We define $\text{uppertriangular}(X)$ to be the matrix whose upper-triangular elements ($i \leq j$) are the same as X but zero elsewhere. Here, $\dim(\mathcal{D}) = n(n+1)/2$ and

$$\nabla\phi(X) = \text{uppertriangular}(\nabla\phi^{\text{amb}}(X)).$$

We do not repeat these ideas for lower-triangular matrices since they are transposes of upper-triangular matrices.

4.1.5 Deviatoric matrices

A matrix $X \in \mathcal{M}$ is deviatoric if tr $(X) = 0$. Let $\text{dev}(X) := X - \frac{\text{tr}\,(X)}{n} I$. In this case, $\dim(\mathcal{D}) = n^2 - 1$ and we have

$$\nabla\phi(X) = \text{dev}(\nabla\phi^{\text{amb}}(X)).$$

4.1.6 Circulant matrices

A circulant matrix [54,55] $X \in \mathcal{M}$ is a (special kind of Toeplitz) matrix where each column is a cyclic shift of the previous column. Suppose the first column is the sequence of entries $x := [x_0, x_1, \ldots, x_{n-1}]^T$. The next column will be $[x_{n-1}, x_0, x_1, \ldots, x_{n-2}]^T$, etc. Then we write $X = \text{circ}(x)$. Note that we expect the dimension of the subspace to be n but the projection in this case is not obvious and depends on some properties of the circulant matrices that we now state.

1. For a circulant matrix X, $X^*X = XX^*$, so that X is a *normal* matrix. The spectral theorem guarantees the existence of a unitary matrix U whose columns are the eigen basis. Furthermore, U is known to be the *normalized* DFT matrix, i.e, a *normalized* Vandermonde matrix comprising the roots of unity and this diagonalizes *every* circulant matrix as $U^*XU = \Lambda$ where Λ is diagonal.
2. Thus, with the basis U, every circulant matrix is diagonal, and we confirm from this that $\dim(\mathcal{D}) = n$.
3. Let $e_2 = (0, 1, 0, , \ldots, 0) \in \mathbb{R}^n$. Define $W = \text{circ}(e_2)$, $W^0 = I$. Then, $X = \text{circ}(x) = \sum_{i=0}^{i=n-1} x_i W^i$. Note that $\{I, W, \ldots, W^{n-1}\}$ is an ordered basis of $\mathcal{D}$.

The projection can now be expressed in two equivalent ways. In the orthogonal (but not orthonormal) basis listed above, one can write, following Lemma 4.2,

$$\nabla\phi(X) = \sum_{i=0}^{i=n-1} \frac{1}{n} \left\langle \nabla\phi^{\text{amb}}(X), W^i \right\rangle_{\mathcal{M}} W^i.$$

Otherwise, using the spectral decomposition and unitary invariance of the Frobenius norm, we can transform the search space to diagonal matrices as follows:

$$\underset{Y\in\mathcal{D}}{\arg\min} \|Y - X\|_{\mathcal{M}} = \underset{\Lambda \text{ diagonal}}{\arg\min} \left\|\Lambda - U^*XU\right\|_{\mathcal{M}}.$$

The projection in this case is already discussed above and so we can write

$$\nabla\phi(X) = U\text{diag}(U^*\nabla\phi^{\text{amb}}(X)U)U^*.$$

As an example, both expressions yield

$$\nabla\phi(X) = \begin{pmatrix} 5 & 5 & 5 \\ 5 & 5 & 5 \\ 5 & 5 & 5 \end{pmatrix} \text{ when } \nabla\phi^{\text{amb}}(X) = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}.$$

4.1.7 Toeplitz, Hankel matrices and others

The significance of Lemma 4.2 and Theorem 4.5 is that they offer a straightforward way to compute the gradient on *any* matrix subspace as long as an ordered basis for the subspace is known. Thus, other matrix subspaces consisting of the Toeplitz matrices, Hankel matrices (both have dimension $2n - 1$) etc. can also be considered.

4.1.8 Nonlinear patterned matrices

The matrix subspaces considered here are instances of a "linear pattern" [56], i.e., the entries of the matrix satisfy a linear constraint. There are also "nonlinear patterned matrices" [57], but on closer examination one finds that the theory is simply a restatement of the chain rule for Fréchet derivatives used in conjunction with our statement proved earlier in Theorem 4.5 – that the gradient of any ambient function acting on a subspace has the same action as the gradient of the given function restricted to the subspace.

4.2 An example with circulant matrices

We shall demonstrate the three different ways to calculate the gradient on an example. Consider

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}, \text{ and the circulant matrix } X = \begin{pmatrix} x & y & z \\ z & x & y \\ y & z & x \end{pmatrix} = xW_1 + yW_2 + zW_3.$$

The space of circulant matrices $\mathcal{D}$ in $\mathbb{R}^{3\times 3}$ is three-dimensional and the three basis vectors W_1, W_2, W_3 and the Grammian $\mathbf{G}$ are

$$W_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \; W_2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \; W_3 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \mathbf{G} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

We assume $\phi : X \in \mathcal{D} \mapsto \text{tr}\,((AX)^3) \in \mathbb{R}$, so that

$$\phi(X) = \phi^d(x, y, z) = 36x^3 + 144xyz + 18y^3 + 18z^3.$$

Following Definition 2.2, for $\phi^{\text{amb}} : Y \in \mathbb{R}^{3\times 3} \mapsto \text{tr}\,((AY)^3) \in \mathbb{R}$,

$$\nabla\phi^{\text{amb}}(X) = 3A^T(AX)^T(AX)^T = \begin{pmatrix} 3x^2 + 15yz & 18xz + 18y^2 & 36xy + 18z^2 \\ 18xy + 18z^2 & 24x^2 + 48yz & 90xz + 18y^2 \\ 36xz + 18y^2 & 90xy + 18z^2 & 81x^2 + 81yz \end{pmatrix}.$$

The same expression for $\nabla\phi^{\text{amb}}(X)$ follows from using Theorem 4.8, when we note that $Y = \sum_{i=1}^{i=9} y_i M^{(i)}$ as per Remark 2.20 and

$$\nabla\phi^{\text{amb}}(X) = \sum_{i=1}^{i=9} \frac{\partial\phi^{\text{amb}}(Y)}{\partial y_i}\Big|_{Y=X} M^{(i)}.$$

The gradient $\nabla\phi^{\text{amb}}(X)$ is clearly not a circulant, and we now project it using Lemma 4.2 to get

$$\nabla\phi(X) = \begin{pmatrix} 36x^2 + 48yz & 48xy + 18z^2 & 48xz + 18y^2 \\ 48xz + 18y^2 & 36x^2 + 48yz & 48xy + 18z^2 \\ 48xy + 18z^2 & 48xz + 18y^2 & 36x^2 + 48yz \end{pmatrix}.$$

Now coming to the second method using Gâteaux derivatives,

$$\frac{\partial\phi}{\partial x} = 108x^2 + 144yz, \; \frac{\partial\phi}{\partial y} = 144xz + 54y^2, \; \frac{\partial\phi}{\partial z} = 144xy + 54z^2,$$

and using Theorem 4.8, we again get

$$\nabla\phi(X) = (36x^2 + 48yz)W_1 + (48xy + 18z^2)W_2 + (48xz + 18y^2)W_3.$$

Starting with $\nabla\phi^d(x, y, z) = (108x^2+144yz)e_1+(144xz+54y^2)e_2+(144xy+54z^2)e_3$ and using Theorem 3.1 also yields the same result as above. We can also confirm that for $H = h_1W_1 + h_2W_2 + h_3W_3 \in \mathcal{D}$, $F(H) = h = h_1e_1 + h_2e_2 + h_3e_3$, we have

$$\left\langle \nabla\phi^{\text{amb}}(X), H \right\rangle_{\mathcal{M}} = \langle \nabla\phi(X), H \rangle_{\mathcal{D}} = \left\langle \nabla\phi^d(x, y, z), h \right\rangle_{\mathbb{R}^{3\times 1}}.$$

4.3 An example with Toeplitz matrices

Consider the same matrix function ϕ as in Subsection 4.2 but the argument this time is a Toeplitz matrix

$$X = \begin{pmatrix} x & y & z \\ v & x & y \\ u & v & x \end{pmatrix} = uW_1 + vW_2 + xW_3 + yW_4 + zW_5.$$

The space of Toeplitz matrices $\mathcal{D}$ in $\mathbb{R}^{3\times 3}$ is five-dimensional and the five basis vectors $W_1, \ldots, W_5$ and the Grammian $\mathbf{G}$ are

$$W_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \; W_2 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \; W_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$W_4 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \; W_5 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \mathbf{G} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Note that in this case,

$$H = h_1W_1 + h_2W_2 + h_3W_3 + h_4W_4 + h_5W_5 \in \mathcal{D},$$
$$F(H) = h = h_1e_1 + h_2e_2 + h_3e_3 + h_4e_4 + h_5e_5.$$

We also have

$$\phi(X) = \phi^d(u, v, x, y, z) = 36uxz + 108vxy + 18uy^2 + 18v^2z + 36x^3.$$

As circulants are special cases of Toeplitz matrices, it is logical that one recovers the expression derived earlier for $\phi(X) = \phi^d(x, y, z)$ on substituting $u = y$, $v = z$.

Clearly, ϕ^{amb} and ϕ^d are same as earlier so

$$\nabla\phi^{\text{amb}}(X) = \begin{pmatrix} 9uz + 6vy + 3x^2 & 18uy + 18vx & 36ux + 18v^2 \\ 18vz + 18xy & 24x^2 + 48vy & 90vx + 18uy \\ 36xz + 18y^2 & 90xy + 18vz & 27uz + 54vy + 81x^2 \end{pmatrix},$$

while

$$\nabla\phi^d(u, v, x, y, z) = (36xz + 18y^2)e_1 + (36vz + 108xy)e_2 + (36uz + 108vy + 108x^2)e_3 + (36uy + 108vx)e_4 + (36ux + 18v^2)e_5.$$

All methods to evaluate $\nabla\phi(X)$ lead to the expression

$$\nabla\phi(X) = \begin{pmatrix} 12uz + 36vy + 36x^2 & 18uy + 54vx & 36ux + 18v^2 \\ 18vz + 54xy & 12uz + 36vy + 36x^2 & 18uy + 54vx \\ 36xz + 18y^2 & 18vz + 54xy & 12uz + 36vy + 36x^2 \end{pmatrix}.$$

5 Use and significance in applications

The discussion thus far was at a conceptual level, and understandably so, since we were demonstrating the rigorous way to think about derivatives of functions defined on subspaces of matrices. The same techniques can be applied to obtain the second derivative of scalar matrix functions, and this has been recorded in Appendix A and Appendix B. What follows is a brief explanation of how to use the theory in various cases that one is likely to encounter in applications. In our view, they can be grouped into two categories, namely, cases where the function has a closed-form analytical representation, and those where the function is specified as a computational "black-box" routine.

5.1 Function has closed-form analytical representation

The examples illustrated in this article, and indeed almost all examples in books fall under this category. Starting with the closed-form representation for the function $\phi : \mathcal{D} \to \mathbb{R}$, it is ideal if one can write ϕ as a sum, product, or composition of other functions whose Fréchet derivative is known and use the product/chain rules [53,58] to obtain the expression for the derivative. If this fails, then one could try to evaluate the function in terms of the matrix components. Note that this may be difficult or even infeasible sometimes when dealing with determinants or inverses of large matrices. However, if successful, one can compute the derivative using the directional (Gâteaux) derivative along each basis vector following Theorem 4.8. If this approach also fails, then one has no choice but to use a numerical approximation method that we describe next.

5.2 Function specified as a computational routine

Suppose we are provided with a routine that computes $\phi(x_1, \ldots, x_d) \in \mathbb{R}$. In the absence of any other information, one may assume that the domain is $\mathbb{R}^{d\times 1}$ and proceed to compute $\left(\frac{\partial \phi}{\partial x_1}, \ldots, \frac{\partial \phi}{\partial x_d}\right)$ by a finite-difference approximation around a given point $(x_1, \ldots, x_d)$. These are the components of the gradient in the natural basis. If the requirement is the action of the derivative or gradient on another vector $h = (h_1, \ldots, h_d) \in \mathbb{R}^{d\times 1}$, i.e., $\langle \nabla\phi((x_1, \ldots, x_d), h\rangle_{\mathbb{R}^{d\times 1}}$, then the approximated gradient is sufficient. If further information is available regarding the d-dimensional domain $\mathcal{D}$, in particular if an ordered basis of $\mathcal{D}$ is known, then one can use Theorem 4.8 to associate the components (approximate partial derivatives) with the basis and determine the gradient. Note that this can be done for the second derivative too (see Theorem B.11) to get the bilinear form.

6 Conclusion

We showed three ways to calculate the derivative/gradient and second derivative for scalar functions of matrices defined over a matrix subspace. The first method is by definition of the Fréchet derivative, the second method utilizes an ordered basis and computes each component of the Fréchet derivative through a Gâteaux (directional) derivative. The third method follows the reasoning behind existing results in matrix calculus and casts the derivative calculation in terms of a scalar function of a vector. In this process, we also uncover the reason behind the spurious results obtained for various matrix subspaces, and propose an alternative to correct the fundamental flaw. All the methods were illustrated by an example calculation. Most importantly, we have reformulated the basic ideas of calculus of matrix functions in the language of calculus on inner-product spaces so that engineers and researchers who are working in inter-disciplinary fields may find it useful and avoid the conceptual pitfalls that exist in various articles and textbooks on the subject.

Acknowledgements SS thanks Los Alamos National Laboratory (US) for funding through LDRD #20230480ECR and #20220006ER while NP was funded by LDRD #20230254ER. The opinions expressed in this paper are those of the authors and do not necessarily reflect that of the sponsors.

Author Contributions SS conceived the idea and both authors (SS, NP) formulated the proofs and edited the manuscript.

Funding SS thanks Los Alamos National Laboratory (US) for funding through LDRD #20230480ECR and #20220006ER while NP was funded by LDRD #20230254ER

Availability of Data and Materials No datasets were generated or analyzed in the current study.

Declarations

Competing Interest The authors declare they have no competing interests.

Appendix A Extension to functions that are not scalar-valued

Thus far, we worked exclusively with functions of matrices whose range was real numbers and elucidated three methods for computing gradients on a matrix subspace, i.e, using the Frechet, Gateaux derivatives as well as

using the commutative diagram. It is natural to ask if analogous results exist for functions whose range is a vector space of dimension greater than one, especially since the second derivative of scalar functions, which are of interest in the context of optimization, fall under this category. Of course, the concept of gradient does not enter the picture, but one can derive similar results with derivatives instead.
To start with, consider

Definition A.1 Denote $g^{\text{amb}} : \mathcal{M} \longrightarrow \mathcal{W}$, and $g : \mathcal{D} \longrightarrow \mathcal{W}$ by $g(X) = g^{\text{amb}}(X)\ \forall X \in \mathcal{D}$ for a proper subspace $\mathcal{D} \subset \mathcal{M}$.

We would like to know if the derivative $dg(X) \in \text{Lin}(\mathcal{D}, \mathcal{W})$ can be obtained from $dg^{\text{amb}}(X) \in \text{Lin}(\mathcal{M}, \mathcal{W})$ in a manner similar to that of Theorem 4.5. The following lemma answers this.

Lemma A.2 *If g^{amb}, g defined as in Definition A.1 are Fréchet differentiable at $X \in \mathcal{D}$, then*

$$dg(X) = dg^{\text{amb}}(X) \circ \mathbb{P},$$

where $\mathbb{P}$ is the projection operator defined in Definition 4.1.

Proof Consider Definition 2.2 for the function g^{amb} for $X \in \mathcal{D}$. Observe that, $\mathbb{P}^2(H) = \mathbb{P} \circ \mathbb{P}(H) \in \mathcal{D}\ \forall H \in \mathcal{M}$, hence

$$\begin{aligned} & g^{\text{amb}}(X + \mathbb{P}(H)) - g^{\text{amb}}(X) - dg^{\text{amb}}(X)(\mathbb{P}(H)) \\ = \ & g(X + \mathbb{P}(H)) - g(X) - dg^{\text{amb}}(X) \circ \mathbb{P}(\mathbb{P}(H)). \end{aligned}$$

On comparison with the definition and appealing to the uniqueness of the Fréchet derivative, we have

$$dg(X) = dg^{\text{amb}}(X) \circ \mathbb{P}.$$

□

Let us also record how to compute the Fréchet derivative componentwise using the Gâteaux derivative as in Definition 2.8. This is a direct application of Theorem 4.8.

Lemma A.3 *If $\{D^{(i)} \in \mathcal{D} | i = 1, 2, \ldots, d\}$ be an ordered basis of $\mathcal{D}$, $\{W^{(i)} \in \mathcal{W} | i = 1, 2, \ldots, p\}$ be an ordered basis of $\mathcal{W}$, and $g : \mathcal{D} \longrightarrow \mathcal{W}$ have the representation*

$$g(X) = \sum_{i=1}^{i=p} g_i \left(\sum_{j=1}^{i=d} x_i D^{(i)} \right) W^{(i)},$$

then we may write

$$dg(X) = \sum_{i=1}^{i=p} \sum_{j=1}^{j=d} \sum_{k=1}^{k=d} \frac{\partial g_i}{\partial x_k} \left(\mathbf{G}^{-1} \right)_{kj} W^{(i)} \otimes D^{(j)},$$

where $\mathbf{G}$ is the Grammian in $\mathcal{D}$ as in Definition 2.13. If we construct a matrix $dg^{\text{mat}} \in \mathbb{R}^{p \times d}$ as $dg^{\text{mat}}_{ij} = \dfrac{\partial g_i}{\partial x_j}$, then we can rephrase the statement above in terms of the matrix of $dg(X)$ with respect to the two bases in $\mathcal{D}, \mathcal{W}$ as

$$\text{mat}(dg(X)) = dg^{\text{mat}} \mathbf{G}^{-1}.$$

If the basis in $\mathcal{D}$ be orthonormal*, then these expressions simplify to*

$$dg(X) = \sum_{i=1}^{i=p} \sum_{j=1}^{j=d} \frac{\partial g_i}{\partial x_j} W^{(i)} \otimes D^{(j)} \text{ and } \text{mat}(dg(X)) = dg^{\text{mat}} \text{ respectively.}$$

Proof From Definition 2.11, and the remarks following Definition 2.17, it is clear that $\{W^{(i)} \otimes D^{(j)} \ \forall\ 1 \leq i \leq p, 1 \leq j \leq d\}$ is an ordered basis of $\mathrm{Lin}(\mathcal{D}, \mathcal{W})$ containing $dg(X)$, hence

$$dg(X) = \sum_{i=1}^{i=p} \sum_{j=1}^{j=d} \alpha_{ij} W^{(i)} \otimes D^{(j)},\ \alpha_{ij} \in \mathbb{R}.$$

The rest of the proof proceeds by developing an expression for α_{ij} as follows:

$$\left\langle dg(X)(D^{(k)}), W^{(l)} \right\rangle_{\mathcal{W}} = \left\langle dg_X^{D^{(k)}}, W^{(l)} \right\rangle_{\mathcal{W}} = \left\langle \sum_{r=1}^{r=p} \frac{\partial g_r}{\partial x_k} W^{(r)}, W^{(l)} \right\rangle_{\mathcal{W}},$$

so that

$$\left\langle dg(X)(D^{(k)}), W^{(l)} \right\rangle_{\mathcal{W}} = \sum_{i=1}^{i=p} \sum_{j=1}^{j=d} \alpha_{ij} \left\langle W^{(i)}, W^{(l)} \right\rangle_{\mathcal{W}} \left\langle D^{(j)}, D^{(k)} \right\rangle_{\mathcal{D}}.$$

Introducing the Grammians of both bases to expand the inner-products and then simplifying, we get

$$\alpha_{ij} = \sum_{k=1}^{k=d} \frac{\partial g_i}{\partial x_k} \left(\mathbf{G}^{-1} \right)_{kj}.$$

□

Remark A.4 It is not obvious why the Grammian of the basis in $\mathcal{W}$ has no role to play in the representation of the derivative until one recognizes that the derivative of g is essentially the derivative of each of its components, which are all scalar functions. The correspondence with Theorem 4.8 then becomes clear.

The third method to calculate the derivative is through a commutative diagram like (3.1), with the only change being to replace $\mathbb{R}$ with $\mathcal{W}$ and write g, g^d as the functions instead of ϕ, ϕ^d. With these modifications, the statement of the chain rule

$$dg(X)(H) = dg^d(F(X))(F(H)) = dg^d(x)(h) \tag{A.1}$$

still holds with $dg : \mathcal{D} \longrightarrow \mathrm{Lin}(\mathcal{D}, \mathcal{W}), dg^d : \mathbb{R}^{d\times 1} \longrightarrow \mathrm{Lin}(\mathbb{R}^{d\times 1}, \mathcal{W})$ but of course in this instance we cannot simplify this relation further by transforming to gradients. However, we can relate the matrix representations of $dg(X)$ and $dg^d(x)$ relative to their respective ordered bases as follows:

$$\mathrm{mat}(dg(X)) = \mathrm{mat}(dg^d(x)). \tag{A.2}$$

Appendix B The second derivative of a scalar matrix functional

For a twice-differentiable scalar function of a matrix, the critical points may be obtained by finding the zeroes of the derivative or gradient. However, to classify the nature of the extremum, it is important to obtain the second derivative. We shall obtain the second derivatives from three different approaches, but first we record

Proposition B.1 *Given a vector space* $\mathcal{D}$, $\mathrm{Lin}(\mathcal{D}, \mathrm{Lin}(\mathcal{D}, \mathbb{R}))$ *and* $\mathrm{Lin}(\mathcal{D} \times \mathcal{D}, \mathbb{R})$ *are isometrically isomorphic (see Amann and Escher [59]).*

Proposition B.1 allows us to speak of the space of bilinear forms $\mathrm{Lin}(\mathcal{D} \times \mathcal{D}, \mathbb{R})$ instead of the convoluted $\mathrm{Lin}(\mathcal{D}, \mathrm{Lin}(\mathcal{D}, \mathbb{R}))$ and use the properties of the former to interpret the latter.

Proposition B.2 *For a function* $\phi : \mathcal{D} \longrightarrow \mathbb{R}$, *at* $X \in \mathcal{D}$, *the second derivative* $d^2\phi(X) \in \mathrm{Lin}(\mathcal{D}, \mathrm{Lin}(\mathcal{D}, \mathbb{R}))$ *is a* symmetric *bilinear form, i.e.,* $d^2\phi(X)(H, K) = d^2\phi(X)(K, H)\ \forall H, K \in \mathcal{D}$. *(see Amann and Escher [59])*

Remark B.3 The isometric isomorphism allows us to write $d^2\phi(X)(H, K)$ instead of $d^2\phi(X)(H)(K)$.

We are now ready to state the results for the second derivative.

Corollary B.4 *For a function $\phi^{\text{amb}} : \mathcal{M} \longrightarrow \mathbb{R}$, and $\phi : \mathcal{D} \longrightarrow \mathcal{R}$, we have the relationships*

$$d\nabla\phi(X) = \mathbb{P} \circ d\nabla\phi^{\text{amb}}(X) \circ \mathbb{P}, \text{ and } d^2\phi(X) = d^2\phi^{\text{amb}}(X) \circ \mathbb{P} \circ \mathbb{P},$$

where $d^2\phi(X) \in \text{Lin}(\mathcal{D}, \text{Lin}(\mathcal{D}, \mathbb{R}))$ but $d\nabla\phi(X) \in \text{Lin}(\mathcal{D}, \mathcal{D})$.

Proof The result follows from Lemma A.2 and Theorem 4.5. □

Remark B.5 Very often, for a function $\phi : \mathcal{M} \longrightarrow \mathbb{R}$, we find the second derivative at X denoted by $\nabla\nabla\phi(X)$ or $\nabla^2\phi(X)$. This notation is inappropriate and misleading because strictly speaking, such quantities are not defined; one cannot determine the gradient of a function $\nabla\phi$ that is not scalar-valued. Moreover, we see from Corollary B.4 that $d\nabla\phi(X)$ is not a bilinear form but an element of $\text{Lin}(\mathcal{D}, \mathcal{D})$. In order to find extrema of smooth functions, we are interested in the second derivative due to its appearance in the Taylor series representation of $\phi(X)$, and thus only $d^2\phi(X)$ makes sense in this context, as $d^2\phi(X)(H, H) \in \mathbb{R}$. Thus, even if the term is denoted by $\nabla^2\phi(X)$ or $d\nabla\phi(X)$, this is what is actually meant.

Now we provide a representation of the second derivative in terms of components. The proof is omitted since it is a straightforward consequence of Theorem 4.8 and Lemma A.3.

Lemma A.5 *(Second derivative in terms of components) If $\{D^{(i)} \in \mathcal{D} | i = 1, 2, \ldots, d\}$ be an ordered basis of $\mathcal{D}$, then $\phi : \mathcal{D} \longrightarrow \mathbb{R}$ has the representation $\phi(X) = \phi\left(\sum_{i=1}^{i=d} x_i D^{(i)}\right)$, and we may write*

$$d^2\phi(X) = \sum_{i=1}^{i=d}\sum_{j=1}^{j=d}\sum_{k=1}^{k=d}\sum_{l=1}^{l=d}\left(\mathbf{G}^{-1}\right)_{il}\frac{\partial^2\phi}{\partial x_l \partial x_k}\left(\mathbf{G}^{-1}\right)_{kj} D^{(i)} \otimes D^{(j)},$$

where $\mathbf{G}$ is the Grammian of the basis. If we construct a matrix (the Hessian)

$$\mathbf{H}(\phi(X)) \in \mathbb{R}^{d\times d}, \ \mathbf{H}(\phi(X))_{lk} = \frac{\partial^2\phi(X)}{\partial x_l \partial x_k},$$

then we may write

$$\text{mat}(d^2\phi(X)) = \mathbf{G}^{-1}\mathbf{H}(\phi(X))\mathbf{G}^{-1}.$$

If the basis be orthonormal, then the representations simplify to

$$d^2\phi(X) = \sum_{i=1}^{i=d}\sum_{j=1}^{j=d}\frac{\partial^2\phi}{\partial x_i \partial x_j} D^i \otimes D^j \text{ and } \text{mat}(d^2\phi(X)) = \mathbf{H}(\phi(X)) \text{ respectively.}$$

Remark B.6 This result demonstrates how the Hessian arises in the context of the second derivative, and how to interpret terse but imprecise statements that suggest that the second derivative of a scalar function is the Hessian matrix.

Remark B.7 If we expand the bilinear form in terms of components, say $H = \sum_{i=1}^{i=d} h_i D^i$, $K = \sum_{j=1}^{j=d} k_j D^j$,

$$d^2\phi(X)(H, K) = (\mathbf{G}\ F(H))^T\ \text{mat}(d^2\phi(X))\ (\mathbf{G}\ F(K)) = F(H)^T\mathbf{H}(\phi(X))F(K).$$

Before we get to the calculation of $d^2\phi$ from the commutative diagram (3.1), we state

Proposition B.8 *(Generalized product rule, Amann and Escher [59]) If $\mathcal{U}, \mathcal{V}, \mathcal{W}$ are Banach spaces, and*

$$\pi : \text{Lin}(\mathcal{V} \times \mathcal{W}, \mathbb{R}),\ f : \mathcal{U} \longrightarrow \mathcal{V},\ g : \mathcal{U} \longrightarrow \mathcal{W}, \text{ then}$$
$$d\pi(f, g) = \pi(df, g) + \pi(f, dg).$$

Lemma B.9 *(Second derivative from the commutative diagram)*

$$\text{For } \phi, \phi^d, F \text{ as in (3.1)},\ d^2\phi(X)(H, K) = d^2\phi^d(x)(h, k),$$

where $X, H, K \in \mathcal{D}$, $x = F(X)$, $h = F(H)$, $k = F(K) \in \mathbb{R}^{d\times 1}$.

Proof Recall that the chain rule applied to $\phi = \phi^d \circ F$ yields

$$d\phi(X) = d\phi^d(F(X)) \circ dF(X) \; \forall X \in \mathcal{D}.$$

Now before we apply the chain rule again, we make the point that the chain rule is stated for composition of functions, not their values, thus, one cannot (and should not) invoke it on the expression for $d\phi(X)$ but instead express $d\phi$ as a composition first. This can be done by invoking Proposition B.8 after writing

$$d\phi = \pi((d\phi^d \circ F), dF)$$

with $\pi \in \text{Lin}(\text{Lin}(\mathcal{D}, \text{R}) \times \text{Lin}(\mathcal{D}, \mathbb{R}^{d\times 1}), \mathbb{R})$. Thus, we have simply rewritten the composition of two linear transformations (which is bilinear) in terms of π to make it easy to see the way forward. Invoking the chain rule now yields

$$d^2\phi = \pi((d^2\phi^d \circ F) \circ dF, dF) + \pi((d\phi^d \circ F), d^2F).$$

The second term on the right hand side vanishes since F is a constant linear map, but in the first term the symbol $\circ$ has to be interpreted in context in each appearance. For example, we know that $d^2\phi^d(F(X)) \in \text{Lin}(\mathbb{R}^{d\times 1}, \text{Lin}(\mathbb{R}^{d\times 1}, \mathbb{R}))$ and for $H, K \in \mathcal{D}$,

$$d^2\phi^d(F(X)) \circ dF(H) \in \text{Lin}(\mathbb{R}^{d\times 1}, \mathbb{R}) \text{ while } (d^2\phi^d(F(X)) \circ dF(H)) \circ dF(K) \in \mathbb{R}.$$

Thus, with this understanding, we may write

$$d^2\phi(X) = \left(d^2\phi^d(F(X)) \circ F\right) \circ F.$$

Considering the action on $H, K \in \mathcal{D}$, we get

$$d^2\phi(X)(H, K) = d^2\phi^d(x)(h, k).$$

□

Remark B.10 In terms of matrices, the relationship may be restated as

$$F(H)^T \mathbf{H}(\phi(X))F(K) = h^T \mathbf{H}(\phi^d(x))k.$$

However, note that $F(X) = x \; \forall X \in \mathcal{D}$, and hence this equation reduces to

$$\mathbf{H}(\phi(X)) = \mathbf{H}(\phi^d(x)).$$

This is hardly surprising, considering both matrices are constructed from the same scalar function of the matrix components.

Theorem B.11 *(Invariant nature of quadratic form) For $\phi^{\text{amb}} : \mathcal{M} \longrightarrow \mathbb{R}$, ϕ, ϕ^d, F as in* (3.1)*, the following are equivalent:*

(a) $d^2\phi(X)$ *is positive-definite (positive-semidefinite) (indefinite)*
(b) $\mathbf{H}(\phi(X))$ *is positive-definite (positive-semidefinite) (indefinite)*
(c) $\mathbf{H}(\phi^d(x))$ *is positive-definite (positive-semidefinite) (indefinite)*
(d) $\mathbf{H}(\phi^{\text{amb}}(X))$ *is positive-definite (positive-semidefinite) (indefinite) on* $\mathcal{D}$

Proof Following the remarks in Lemma A.5 and Lemma B.9, the quadratic forms for given $X, H, K \in \mathcal{D}$ satisfy

$$d^2\phi(X)(H, K) = F(H)^T \mathbf{H}(\phi(X))F(K) = h^T \mathbf{H}(\phi^d(x))k,$$

and have the same value in (a), (b), (c). But we also have

$$d^2\phi(X)(H, K) = d^2\phi^{\text{amb}}(X)(\mathbb{P}(H), \mathbb{P}(K)).$$

Recall that $\dim(\mathcal{M}) = n^2$, hence we have $\mathbf{H}(\phi^{\text{amb}}(X)) \in \mathbb{R}^{n^2\times n^2}$, $F(H), F(K) \in \mathbb{R}^{n^2\times 1}$ since the coordinate map is for the orthonormal natural basis in $\mathcal{M}$ Remark 2.20, and

$$d^2\phi^{\text{amb}}(X)(\mathbb{P}(H), \mathbb{P}(K)) = d^2\phi^{\text{amb}}(X)(H, K) = F(H)^T \mathbf{H}(\phi^{\text{amb}}(X))F(K).$$

Thus, they are all positive-definite (positive-semidefinite) (indefinite). □

Remark B.12 The significance of this result is that to determine the nature of the extremum existing at a critical point, one can test the quadratic form obtained from the second derivative using any of the methods outlined above.

Appendix B.1 Illustrative examples continued

We continue with the example considered in Subsection 4.2 to now calculate the second derivative. By direct calculation, we can show that for $H, K \in \mathcal{D}$,

$$d^2\phi(X)(H, K) = 3\,\mathrm{tr}\,(AK^TAX^TAH^T + AX^TAK^TAH^T).$$

On the other hand, from the Hessian matrix

$$\mathbf{H}(\phi(X)) = \mathbf{H}(\phi^d(x)) = \begin{pmatrix} 216x & 144z & 144y \\ 144z & 108y & 144x \\ 144y & 144x & 108z \end{pmatrix},$$

while

$$\mathbf{H}(\phi^{\mathrm{amb}}(X)) = \begin{pmatrix} 6x & 6z & 9y & 6y & 0 & 0 & 9z & 0 & 0 \\ 6z & 0 & 0 & 18x & 12z & 18y & 18y & 0 & 0 \\ 9y & 0 & 0 & 18z & 0 & 0 & 36x & 18z & 27y \\ 6y & 18x & 18z & 0 & 12y & 0 & 0 & 18z & 0 \\ 0 & 12z & 0 & 12y & 48x & 36z & 0 & 36y & 0 \\ 0 & 18y & 0 & 0 & 36z & 0 & 18y & 90x & 54z \\ 9z & 18y & 36x & 0 & 0 & 18y & 0 & 0 & 27z \\ 0 & 0 & 18z & 18z & 36y & 90x & 0 & 0 & 54y \\ 0 & 0 & 27y & 0 & 0 & 54z & 27z & 54y & 162x \end{pmatrix}.$$

Note that $\mathbf{H}(\phi^{\mathrm{amb}}(X))_{lk} = \dfrac{\partial^2\phi^{\mathrm{amb}}(Y)}{\partial y_l \partial y_k}\Big|_{Y=X}$. By direct calculation, we find that all bilinear forms have the same expression

$$36(k_1(6h_1x + 4h_2z + 4h_3y) + k_2(4h_1z + 3h_2y + 4h_3x) + k_3(4h_1y + 4h_2x + 3h_3z)).$$

Continuing with Subsection 4.3, we find

$$\mathbf{H}(\phi(X)) = \mathbf{H}(\phi^d(x)) = \begin{pmatrix} 0 & 0 & 36z & 36y & 36x \\ 0 & 36z & 108y & 108x & 36v \\ 36z & 108y & 216x & 108v & 36u \\ 36y & 108x & 108v & 36u & 0 \\ 36x & 36v & 36u & 0 & 0 \end{pmatrix},$$

while

$$\mathbf{H}(\phi^{\mathrm{amb}}(X)) = \begin{pmatrix} 6x & 6v & 9u & 6y & 0 & 0 & 9z & 0 & 0 \\ 6v & 0 & 0 & 18x & 12v & 18u & 18y & 0 & 0 \\ 9u & 0 & 0 & 18v & 0 & 0 & 36x & 18v & 27u \\ 6y & 18x & 18v & 0 & 12y & 0 & 0 & 18z & 0 \\ 0 & 12v & 0 & 12y & 48x & 36v & 0 & 36y & 0 \\ 0 & 18y & 0 & 0 & 36v & 0 & 18y & 90x & 54v \\ 9z & 18y & 36x & 0 & 0 & 18y & 0 & 0 & 27z \\ 0 & 0 & 18v & 18z & 36y & 90x & 0 & 0 & 54y \\ 0 & 0 & 27u & 0 & 0 & 54v & 27z & 54y & 162x \end{pmatrix}.$$

It can be verified that all the three associated bilinear forms have the same expression

$$\begin{aligned} &36k_1(h_3z + h_4y + h_5x) + 36k_2(h_2z + 3h_3y + 3h_4x + h_5v) \\ &+ 36k_3(h_1z + 3h_2y + 6h_3x + 3h_4v + h_5u) \\ &+ 36k_4(h_1y + 3h_2x + 3h_3v + h_4u) + 36k_5(h_1x + h_2v + h_3u). \end{aligned}$$

References

1. J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics.* John Wiley, New York, 1988.
2. Michael Athans and Fred C. Schweppe. Gradient matrices and matrix calculations. Technical Report Technical Note 1965-53, Massachusets Institute of Technology Lincoln Laboratory, Nov 1965. URL https://apps.dtic.mil/docs/citations/AD0624426. [Online; accessed 1. Nov. 2019].
3. Are Hjørungnes. Generalized complex-valued matrix derivatives. In *Complex-Valued Matrix Derivatives: With Applications in Signal Processing and Communications*, pages 133–200. Cambridge University Press, Cambridge, England, UK, February 2011. https://doi.org/10.1017/CBO9780511921490.008.
4. John Dattorro. *Convex Optimization & Euclidean Distance Geometry*. Meboo Publishing, Palo Alto, California, 2005.
5. Ray W. Ogden. *Non-linear Elastic Deformations*. Dover Publications, New York, 1997.
6. Morton E. Gurtin, Eliot Fried, and Lallit Anand. *Mechanics and Thermodynamics of Continua*. Cambridge University Applied Mathematics Research eXpress, 2010.
7. Aravindh Krishnamoorthy and Robert Schober. Downlink MIMO-RSMA with successive null-space precoding. *IEEE Transactions on Wireless Communications*, 21(11):9170–9185, May 2022. https://doi.org/10.1109/TWC.2022.3173463.
8. Gaëtan Louvet, Jakob Raymaekers, Germain Van Bever, and Ines Wilms. The influence function of graphical Lasso estimators. *ArXiv e-prints*, September 2022. https://doi.org/10.48550/arXiv.2209.07374.
9. Bart L. R. De Moor. *Structured Total Least Squares for Hankel Matrices*, pages 243–258. Springer, Boston, MA, Boston, MA, USA, 1997. https://doi.org/10.1007/978-1-4615-6281-8_13.
10. J. P. Burg, D. G. Luenberger, and D. L. Wenger. Estimation of structured covariance matrices. *Proceedings of the IEEE*, 70(9):963–974, September 1982. https://doi.org/10.1109/PROC.1982.12427.
11. Paul R. Halmos. *Finite-Dimensional Vector Spaces*. D. Van Nostrand Company, Inc., Princeton, NJ, second edition, 1958.
12. John Dieudonnè. *Foundations of Modern Analysis*. Academic Press, New York, 1960.
13. P.S. Dwyer and M.S. Macphail. Symbolic matrix derivatives. *Annals of Mathematical Statistics*, 19(4):517–534, December 1948. URL https://www.jstor.org/stable/2236019.
14. Paul S. Dwyer. Some applications of matrix derivatives in multivariate analysis. *Journal of the American Statistical Association*, 62(318):607–625, June 1967. https://doi.org/10.1080/01621459.1967.10482934.
15. Derrick S. Tracy and Paul S. Dwyer. Multivariate Maxima and Minima with Matrix Derivatives. *Journal of the American Statistical Association*, 64(328):1576–1594, Dec 1969. ISSN 0162-1459. https://doi.org/10.2307/2286090.
16. Friedrich Gebhardt. Maximum likelihood solution to factor analysis when some factors are completely specified. *Psychometrika*, 36(2):155–163, Jun 1971. ISSN 1860-0980. https://doi.org/10.1007/BF02291395.
17. D. S. Tracy and R. P. Singh. Some Applications of Matrix Differentiation in the General Analysis of Covariance Structures. Sankhyā: The Indian Journal of Statistics, Series A (1961-2002), 37(2):269–280, Apr 1975. ISSN 0581-572X. https://doi.org/10.2307/25049980.
18. Michael Athans. The matrix minimum principle. *Information and Control*, 11(5):592–606, Nov 1967a. https://doi.org/10.1016/S0019-9958(67)90803-0.
19. Michael Athans. Matrix minimum principle. Technical Report Report ESL-R-317, Massachusets Institute of Technology Lincoln Laboratory, Aug 1967b. [Online; accessed 4. Mar. 2020].
20. Shriram Srinivasan and Nishant Panda. What is the gradient of a scalar function of a symmetric matrix ? *Indian Journal of Pure & Applied Mathematics*, August 2022. https://doi.org/10.1007/s13226-022-00313-x.
21. H. Geering. On calculating gradient matrices. *IEEE Transactions on Automatic Control*, 21(4):615–616, Aug 1976. https://doi.org/10.1109/TAC.1976.1101267.
22. J. Brewer. The gradient with respect to a symmetric matrix. *IEEE Transactions on Automatic Control*, 22(2):265–267, Apr 1977. https://doi.org/10.1109/TAC.1977.1101459.
23. P. Walsh. On symmetric matrices and the matrix minimum principle. *IEEE Transactions on Automatic Control*, 22(6):995–996, Dec 1977. https://doi.org/10.1109/TAC.1977.1101664.
24. H. V. Henderson and S. R. Searle. Vec and vech operators for matrices, with some uses in jacobians and multivariate statistics. *Canadian Journal of Statistics*, 7:65–81, 1979.
25. D. G. Nel. On matrix differentiation in statistics. *South African Statistical Journal*, 14(2):137–193, Jan 1980.
26. Charles E. McCulloch. Symmetric Matrix Derivatives with Applications. *Journal of the American Statistical Association*, Mar 1980.
27. G.S. Rogers. *Matrix Derivatives*. Marcel Dekker, New York, 1980.
28. A. Graham. *Kronecker Products and Matrix Calculus with Applications*. Ellis Horwood Limited, 1981.
29. A. E. Yanchevsky and V. J. Hirvonen. Optimization of feedback systems with constrained information flow. *International Journal of Systems Science*, 12(12):1459–1468, 1981. https://doi.org/10.1080/00207728108963830.
30. S.R. Searle. *Matrix Algebra for Statistics*. John Wiley, New York, 1982.
31. A.-M. Parring. About the concept of the matrix derivative. *Linear Algebra and its Applications*, 176:223–235, Nov 1992. https://doi.org/10.1016/0024-3795(92)90220-5.
32. David A. Harville. *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag, New York, 1997.
33. A. M. Mathai. *Jacobians of Matrix Transformations and Functions of Matrix Argument*. World Scientific, Singapore, 1997.
34. Thomas Minka. Old and New Matrix Algebra Useful for Statistics, 2001. URL https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.115.5808. [Online; accessed 1. Nov. 2019].
35. D. H. van Hessem and O. H. Bosgra. A full solution to the constrained stochastic closed-loop mpc problem via state and innovations feedback and its receding horizon implementation. In *42nd IEEE International Conference on Decision and Control (IEEE Cat. No.03CH37475)*, volume 1, pages 929–934 Vol.1, 2003.
36. George Arthur Frederick Seber. *A Matrix Handbook for Statisticians*. John Wiley & Sons, New Jersey, 2008.

37. P. P. Vaidyanathan, See-May Phoong, and Yuan-Pei Lin. Matrix differentiation. In *Signal Processing and Optimization for Transceiver Systems*, pages 660–693. Cambridge University Press, Cambridge, England, UK, March 2010. https://doi.org/10.1017/CBO9781139042741.021.
38. Kaare Brandt Petersen and Michael Syskind Pedersen. The Matrix Cookbook, 11 2012. https://www2.imm.dtu.dk/pubdb/pubs/3274-full.html
39. Iain Murray. Differentiation of the Cholesky decomposition. *ArXiv e-prints*, Feb 2016. arxiv:1602.07527.
40. Are Hjørungnes and Daniel P. Palomar. Patterned complex-valued matrix derivatives. In *2008 5th IEEE Sensor Array and Multichannel Signal Processing Workshop*, pages 293–297. IEEE, July 2008a. https://doi.org/10.1109/SAM.2008.4606875.
41. Are Hjørungnes and Daniel P. Palomar. Finding patterned complex-valued matrix derivatives by using manifolds. In *2008 First International Symposium on Applied Sciences on Biomedical and Communication Technologies*, pages 1–5. IEEE, October 2008b. https://doi.org/10.1109/ISABEL.2008.4712619.
42. W. Vetter. Derivative operations on matrices. *IEEE Transactions on Automatic Control*, 15(2):241–244, April 1970. https://doi.org/10.1109/TAC.1970.1099409.
43. William J. Vetter. Matrix calculus operations and taylor expansions. *SIAM Review*, 15(2):352–369, August 1973. https://doi.org/10.1137/1015034.
44. D. S. G. Pollock. Tensor products and matrix differential calculus. *Linear Algebra and its Applications*, 67:169–193, June 1985. https://doi.org/10.1016/0024-3795(85)90194-6.
45. D. S. G. Pollock. On kronecker products, tensor products and matrix differential calculus. *International Journal of Computer Mathematics*, 90(11):2462–2476, November 2013. https://doi.org/10.1080/00207160.2013.783696.
46. Jan R. Magnus. On the concept of matrix derivative. *Journal of Multivariate Analysis*, 101(9):2200–2206, October 2010. https://doi.org/10.1016/j.jmva.2010.05.005.
47. Richard Golden. *Statistical Machine Learning:A Unified Framework*. Taylor & Francis, Andover, England, UK, July 2020. https://doi.org/10.1201/9781351051507.
48. Shuangzhe Liu, Víctor Leiva, Dan Zhuang, Tiefeng Ma, and Jorge I. Figueroa-Zúñiga. Matrix differential calculus with applications in the multivariate linear model and its diagnostics. *Journal of Multivariate Analysis*, 188:104849, March 2022. https://doi.org/10.1016/j.jmva.2021.104849.
49. Jan Brinkhuis. On the use of coordinate-free matrix calculus. *Journal of Multivariate Analysis*, 133:377–381, January 2015. https://doi.org/10.1016/j.jmva.2014.09.019.
50. Tõnu Kollo and Dietrich von Rossen. *Advanced Multivariate Statistics with Matrices*. Springer, Dordrecht, The Netherlands, 2005.
51. Alan Edelman and Steven G. Johnson. Matrix calculus for machine learning and beyond. Massachusetts Institute of Technology: MIT OpenCouseWare, 2022. URL https://ocw.mit.edu/courses/18-s096-matrix-calculus-for-machine-learning-and-beyond-january-iap-2022/. License: Creative Commons BY-NC-SA.
52. Matthias Kissel and Klaus Diepold. Structured matrices and their application in neural networks: A survey. *New Generation Computing*, 41(3):697–722, 2023. https://doi.org/10.1007/s00354-023-00226-1.
53. Ward Cheney. *Analysis for Applied Mathematics*, volume 208. Springer Science & Business Media, 2013.
54. Robert M. Gray. Toeplitz and circulant matrices: a review. *Communications and Information Theory*, 2(3):155–239, August 2005. https://doi.org/10.1561/0100000006.
55. Irwin Kra and Santiago R. Simanca. On circulant matrices. *American Mathematical Monthly*, 3(59):368–377, March 2012. https://doi.org/10.1090/noti804.
56. Douglas P. Wiens. On some pattern-reduction matrices which appear in statistics. *Linear Algebra and its Applications*, 67:233–258, June 1985. https://doi.org/10.1016/0024-3795(85)90199-5.
57. D. S. Tracy and K. G. Jinadasa. Patterned matrix derivatives. *Canadian Journal of Statistics*, 16(4):411–418, December 1988. https://doi.org/10.2307/3314938.
58. Vladimir A. Zorich. *Differential Calculus from a More General Point of View*, pages 41–108. Springer, Berlin, Germany, February 2016. https://doi.org/10.1007/978-3-662-48993-2_2.
59. Herbert Amann and Joachim Escher. *Multilinear maps*, pages 173–180. Birkhäuser, Basel, Switzerland, 2006. https://doi.org/10.1007/3-7643-7402-0_13.