What is the gradient of a scalar function of a symmetric matrix?

Shriram Srinivasan · Nishant Panda

Abstract For a real valued function $\phi$ of a matrix argument, the gradient $\nabla \phi$ is calculated using a standard approach that follows from the definition of a Fréchet derivative for matrix functionals. In cases where the matrix argument is restricted to the space of symmetric matrices, the approach is easily modified to determine that the gradient ought to be $(\nabla \phi + \nabla \phi^T)/2$. However, perusal of research articles in the statistics and electrical engineering communities that deal with the topic of matrix calculus reveal a different approach that leads to a spurious result. In this approach, the gradient of $\phi$ is evaluated by explicitly taking into account the symmetry of the matrix, and this “symmetric gradient” $\nabla \phi_{sym}$ is reported to be related to the gradient $\nabla \phi$ which is computed by ignoring symmetry as $\nabla \phi_{sym} = \nabla \phi + \nabla \phi^T - \nabla \phi \circ I$, where $\circ$ denotes the elementwise Hadamard product of the two matrices and $I$ the identity matrix of the same size as $\nabla \phi$. The idea of the “symmetric gradient” has now appeared in several publications, as well as in textbooks and handbooks on matrix calculus which are often cited in this context. One of our important contributions has been to wade through the vague and confusing proofs of the result based on matrix calculus and cast the calculation of the “symmetric gradient” in a rigorous and concrete mathematical setting. After setting up the problem in a finite-dimensional inner-product space, we demonstrate rigorously that $\nabla \phi_{sym} = (\nabla \phi + \nabla \phi^T)/2$ is the correct relationship. Moreover, our derivation exposes that it is an incorrect lifting from the Euclidean space to the space of symmetric matrices, inconsistent with the underlying inner-product, that leads to the spurious result. We also discuss the implications of using the spurious gradient in different classes of problems, such as those where the gradient itself may be the quantity sought, or as part of an optimization algorithm such as gradient descent. We show that the spurious gradient has a relative error of 100% in the off-diagonal components, which makes it an egregious error if the gradient be a quantity of interest, but fortuitously, it proves to be an ascent direction, so that its use in gradient descent may not lead to major issues.

Keywords Matrix calculus · Symmetric matrix · Fréchet derivative · Gradient · Matrix functional

Mathematics Subject Classifications 15A60 · 15A63 · 26B12
1 Introduction

Matrix functionals defined over an inner-product space of square matrices are a common construct in applied mathematics. In most cases, the object of interest is not the matrix functional itself, but its derivative or gradient (if it be differentiable), and this notion is unambiguous. The Fréchet derivative, see for e.g. [26] and [5], being a linear functional readily yields the definition of the gradient via the Riesz Representation Theorem.

However, there is a sub-class of matrix functionals that frequently occurs in practice whose argument is a symmetric matrix. Such functionals and their gradients occur in the analysis and control of dynamical systems which are described by matrix differential equations [3], maximum likelihood estimation in statistics, econometrics and machine-learning [18], and in the theory of elasticity and continuum thermodynamics [9,30].

By working with the definition of the Fréchet derivative over the vector space of square matrices and specializing it to that of the symmetric matrices which is a proper subspace, the gradient of a matrix functional can be obtained as described in [13]. However, a different idea emerged through matrix calculus as practiced by the statistics and control systems community – that of a “symmetric gradient”. The root of this idea is the fact that while the space of square matrices in \( \mathbb{R}^{n \times n} \) has dimension \( n^2 \), the subspace of symmetric matrices has a dimension of \( n(n+1)/2 \). The second approach aims to take into account the symmetry of the matrix elements in some fashion to compute the “symmetric gradient” in \( \mathbb{R}^{n \times n} \). However, the gradients computed by the two different methods are not equal. Thus, for this sub-class of matrix functionals with a symmetric matrix argument, there appears to be two approaches to define the gradient whose results do not agree with each other. The question raised in the title of this article refers to this apparent dichotomy, which will be resolved before the conclusion.

A perusal of the literature reveals how the idea of the “symmetric gradient” came into being among the community of statisticians and electrical engineers that dominantly used matrix calculus. Early work in statistics in 1960s such as [6,39] does not make any mention of a need for special formulae to treat the case of a symmetric matrix, but does note that all the matrix elements must be “functionally independent”. The notion of “independence” of matrix elements was a recurring theme, and symmetric matrices, by dint of equality of the off-diagonal elements violated this condition. Gebhardt [7] in 1971 seems to have been the first to remark that the derivative formulae do not consider symmetry explicitly but he concluded that no adjustment was necessary in his case since the gradient obtained was already symmetric. Tracy and Singh [38] in 1975 echo the same sentiments as Gebhardt about the need for special formulae. By the end of the decade, the “symmetric gradient” makes its appearance in some form or the other in the work of Henderson [11] in 1979, a review by Nel [28], and a book by Rogers [33] in 1980. McCulloch [21] proves the expression for “symmetric gradient” that we quote here and notes that it applies to calculating derivatives with respect to variance-covariance matrices, and thus derives the information matrix for the multivariate normal distribution. By 1982, the “symmetric gradient” is included in the authoritative and influential textbook by Searle [35]. Today the idea is firmly entrenched as evidenced by the books [10,20,36] and the notes by Minka [25].

The idea of the “symmetric gradient” seems to have come up in the control systems community (as represented by publications in IEEE) as an offshoot of the extension of the Pontryagin Maximum principle for matrix of controls and states when Athans and Schweppe [3] remark that the formulae for gradient matrices are derived under the assumption of “functional independence” of matrix elements. Later, they warned [1,2] that special formulae were necessary for symmetric matrices. Geering [8] in 1976 exhibited an example calculation (gradient of the determinant of a symmetric \( 2 \times 2 \) matrix) to justify the definition of a “symmetric gradient”.

Brewer [4] in 1977 remarked that the formulae for gradient matrices in [1–3] can only be applied when the elements of the matrix are independent, which is not true for a symmetric matrix, and so proceeded to derive a general formula for the “symmetric gradient” (identical to McCulloch [21]) through the rules of matrix calculus for use in sensitivity analysis of optimal estimation systems. Following on from these works, it appears in other instances [31,40,43,45].

At present, the “symmetric gradient” formula is also recorded in [32], a handy reference for engineers and scientists working on inter-disciplinary topics using statistics and machine-learning, and the formula’s appearance in [27] shows that it is no longer restricted to a particular community of researchers.

Thus, the fact that these two well-established notions do not agree is a source of enormous confusion for researchers who straddle application areas, a point to which the authors can emphatically attest to. On the popular site Mathematics Stack Exchange, there are multiple questions (for example [12,14,16,22,23,37,44]) related to this theme, but their answers deepen and misguide rather than alleviate the existing confusion.

Depending on the context, the disagreement between the two notions of gradient has implications that range from the serious to none. In the context of extremizing a matrix functional whose gradient can be written in a closed-form analytical expression, such as when calculating a maximum likelihood estimator for a Gaussian,
both approaches yield the same critical point. If the gradient be used in an optimization routine such as steepest
descent, one of the gradients is clearly not the steepest descent direction, and it is not clear what effect this has on
the iterates and their path to possible convergence. However, if the gradient itself be the quantity of interest, the
discrepancies are a serious issue. Such is the case in physics and mechanics since gradients of matrix functionals
are used to describe physical quantities like stress and strain in a body.

On a related note (see [24,41] for more details), we mention in passing that the idea to eliminate the redundant
degrees of freedom in a symmetric matrix is not a recent one; it dates back to Kelvin (1856) [15] and Voigt (1910)
[42] who proposed it in the context of the three-dimensional theory of elasticity to transform the algebra of fourth-
degree tensors into the algebra of second order tensors (represented as symmetric matrices).

The proof of the “symmetric gradient” that appears in various sources we cite here (for example [4,21]) uses
the heuristics and formal manipulations of matrix calculus as practiced by the statistics, economics and electrical
engineering communities, and we were unable to identify or pinpoint how their derivation went astray. However,
the example calculation of Geering [8] gave us an insight into the approach, using which we were able to put
those ideas on a rigorous mathematical footing and resolve the discrepancy.

We formulate the calculation of the “symmetric gradient” in its natural setting of finite-dimensional inner-
product spaces. If we wish to explicitly take into account the symmetry of the matrix elements, we view the
matrix functional as one defined on the vector space
$\mathbb{R}^{n \times n}$.
A careful evaluation led us to the comforting conclusion that the expression for the gradient is indeed unique; the flaw arises because the established “symmetric gradient” in $\mathbb{R}^{n \times n}$ is a misinterpretation of the gradient in $\mathbb{R}^{n(n+1)/2}$. In other words, the lifting from $\mathbb{R}^{n(n+1)/2}$ to $\mathbb{R}^{n \times n}$ is incorrect. When interpreted correctly, we are inexorably led to the same expression in both methods of evaluation described earlier.

Thus our main contributions are to interpret the established results from matrix calculus and formulate the
calculation of the “symmetric gradient” rigorously, demonstrate that there is no dichotomy, and explain the fatal
flaw that led to the wrong result. That finally brings us to the most important reason for writing this article,
which is that derivatives and gradients are fundamental ideas, and there should not be any ambiguity about their
definitions. Thus, we felt the urgent need to clarify the issues muddying the waters, and show that the “symmetric
gradient”, when calculated correctly, leads to the expected result.

The paper is organized as follows: after stating the problem, we begin with two illustrative examples in
Section 2 that allow us to see concretely what we later prove in the abstract. After that, Section 3 lays out all the
machinery of linear algebra that we shall need, ending with the proof of the main result. Finally, we discuss the
implications of the wrong result in different contexts in Section 4 and conclude with Section 5.

2 Problem formulation

To fix our notation, we introduce the following. We denote by $\mathbb{S}^{n \times n}$ the subspace of all symmetric matrices
in $\mathbb{R}^{n \times n}$. The space $\mathbb{R}^{n \times n}$ (and subsequently $\mathbb{S}^{n \times n}$) is an inner product space with the following 
*natural* inner product $\langle \cdot, \cdot \rangle_F$.

**Definition 2.1** For two matrices $A$, $B$ in $\mathbb{R}^{n \times n}$

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

defines an inner product and induces the Frobenius norm on $\mathbb{R}^{n \times n}$ via

$$\|A\|_F := \sqrt{\text{tr}(A^T A)}.$$  

**Corollary 2.2** We collect a few useful facts about the inner product defined above essential for this paper.

1. For $A$ symmetric, $B$ skew-symmetric in $\mathbb{R}^{n \times n}$, $\langle A, B \rangle_F = 0$.
2. If $\langle A, H \rangle_F = 0$ for every $H$ in $\mathbb{S}^{n \times n}$, then the symmetric part of $A$ given by $\text{sym}(A) := (A + A^T)/2$ is equal
to 0.
3. For $A$ in $\mathbb{R}^{n \times n}$ and $H$ in $\mathbb{S}^{n \times n}$, $\langle A, H \rangle_F = \langle \text{sym}(A), H \rangle_F$.

**Proof** See, for e.g. [34].

Consider a real valued function $\phi : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$. We say that $\phi$ is differentiable if its Fréchet derivative,
defined below, exists.
**Definition 2.3** The Fréchet derivative of $\phi$ at $A$ in $\mathbb{R}^{n \times n}$ is the unique linear transformation $D\phi(A)$ in $\text{Lin}(\mathbb{R}^{n \times n}, \mathbb{R})$ such that

$$\lim_{\|H\| \to 0} \frac{|\phi(A + H) - \phi(A) - D\phi(A)(H)|}{\|H\|} \to 0. \quad (2.1)$$

The Riesz Representation theorem then asserts the existence of the gradient $\nabla \phi(A)$ in $\mathbb{R}^{n \times n}$ such that

$$\langle \nabla \phi(A), H \rangle_F = D\phi(A)(H) \quad \text{for every } H \in \mathbb{R}^{n \times n}. \quad (2.2)$$

Note that if $A$ is a symmetric matrix, then by the Fréchet derivative defined above, the gradient $\nabla \phi(A)$ is not guaranteed to be symmetric. Also, observe that the dimension of $\mathbb{S}^{n \times n}$ is $m = n(n + 1)/2$, hence, it is natural to identify $\mathbb{S}^{n \times n}$ with $\mathbb{R}^m$. The reduced dimension along with the fact that Definition 2.3 doesn’t account for the symmetry structure. Harville [10] notes that the interior of $\mathbb{S}^{n \times n}$ is empty, and states that hence, Definition 2.3 is not applicable for symmetric matrices. The inference is incorrect, for while it is true that the interior of $\mathbb{S}^{n \times n}$ in $\mathbb{R}^{n \times n}$ is empty, the interior of $\mathbb{S}^{n \times n}$ in $\mathbb{S}^{n \times n}$ is non-empty, and this is the key to Definition 2.5. For completeness, we re-iterate the definition of Fréchet derivative of $\phi$ restricted to the subspace $\mathbb{S}^{n \times n}$.

**Claim 2.4** Let $\phi : \mathbb{R}^{n \times n} \to \mathbb{R}$, and let $\phi_{\text{sym}} : \mathbb{S}^{n \times n} \to \mathbb{R}$ be the real-valued function that is the restriction of $\phi$ to $\mathbb{S}^{n \times n}$, i.e., $\phi_{\text{sym}} := \phi|_{\mathbb{S}^{n \times n}}$. Let the gradient $\nabla \phi(A)$ be as defined in Definition 2.3. Then $\nabla \phi_{\text{sym}}(A)$ is the linear transformation in $\mathbb{S}^{n \times n}$ that is claimed to be the “symmetric gradient” of $\phi_{\text{sym}}$ and related to the gradient $\nabla \phi(A)$ as follows [10,21,32,36]:

$$\nabla \phi_{\text{sym}}(A) = \nabla \phi(A) + \nabla \phi^T(A) - \nabla \phi(A) \circ I. \quad (2.3)$$

where $\circ$ denotes the element-wise Hadamard product of $\nabla \phi(A)$ and the identity $I$.

Theorem 3.8 in the next section will demonstrate that this claim is false. Before that, however, note that $\mathbb{S}^{n \times n}$ is a subspace of $\mathbb{R}^{n \times n}$ with the induced inner product in Definition 2.1. Thus, the derivative in Definition 2.3 is naturally defined for all scalar functions of symmetric matrices. The Fréchet derivative of $\phi$ when restricted to the subspace $\mathbb{S}^{n \times n}$ automatically accounts for the symmetry structure. Harville [10] notes that the interior of $\mathbb{S}^{n \times n}$ is empty, and states that hence, Definition 2.3 is not applicable for symmetric matrices. The inference is incorrect, for while it is true that the interior of $\mathbb{S}^{n \times n}$ in $\mathbb{R}^{n \times n}$ is empty, the interior of $\mathbb{S}^{n \times n}$ in $\mathbb{S}^{n \times n}$ is non-empty, and this is the key to Definition 2.5. For completeness, we re-iterate the definition of Fréchet derivative of $\phi$ restricted to the subspace $\mathbb{S}^{n \times n}$.

**Definition 2.5** The Fréchet derivative of the function $\phi_{\text{sym}} := \phi|_{\mathbb{S}^{n \times n}}$ at $A$ in $\mathbb{S}^{n \times n}$ is the unique linear transformation $D\phi_{\text{sym}}(A)$ in $\text{Lin}(\mathbb{S}^{n \times n}, \mathbb{R})$ such that

$$\lim_{\|H\| \to 0} \frac{|\phi_{\text{sym}}(A + H) - \phi_{\text{sym}}(A) - D\phi_{\text{sym}}(A)(H)|}{\|H\|} \to 0. \quad (2.4)$$

The Riesz Representation theorem then asserts the existence of the gradient $\nabla \phi_{\text{sym}}(A)$ in $\mathbb{S}^{n \times n}$ such that

$$\langle \nabla \phi_{\text{sym}}(A), H \rangle_F = D\phi_{\text{sym}}(A)(H) \quad \text{for every } H \in \mathbb{S}^{n \times n}. \quad (2.5)$$

There is a natural relationship between the gradient in the larger space $\mathbb{R}^{n \times n}$ and the restricted subspace $\mathbb{S}^{n \times n}$. The following corollary states this relationship.

**Corollary 2.6** If $\nabla \phi(A) \in \mathbb{R}^{n \times n}$ be the gradient of $\phi : \mathbb{R}^{n \times n} \to \mathbb{R}$, then $\nabla \phi_{\text{sym}}(A) = \text{sym}(\nabla \phi(A))$ is the gradient in $\mathbb{S}^{n \times n}$ of $\phi_{\text{sym}} := \phi|_{\mathbb{S}^{n \times n}}$.

**Proof** From Definition 2.3, we know that $D\phi(A)(H) = \langle \nabla \phi(A), H \rangle_F$ for every $H$ in $\mathbb{R}^{n \times n}$. If we restrict attention to $H$ in $\mathbb{S}^{n \times n}$, then

$$D\phi_{\text{sym}}(A)(H) = \langle \nabla \phi_{\text{sym}}(A), H \rangle_F = \langle \nabla \phi_{\text{sym}}(A), H \rangle_F. \quad \text{This is true for every } H \in \mathbb{S}^{n \times n}, \text{ so that by Corollary 2.2 and uniqueness of the gradient,}$$

$$\nabla \phi_{\text{sym}}(A) = \text{sym}(\nabla \phi(A))$$

is the gradient in $\mathbb{S}^{n \times n}$. \qed
2.1 An Illustrative Example 1

This example will illustrate the difference between the gradient on \( \mathbb{R}^{n \times n} \) and \( \mathbb{S}^{n \times n} \). Fix a non-symmetric matrix \( A \) in \( \mathbb{R}^{n \times n} \) and consider a linear functional, \( \phi : \mathbb{R}^{n \times n} \to \mathbb{R} \), given by \( \phi(X) = \text{tr}(A^T X) \) for every \( X \) in \( \mathbb{R}^{n \times n} \).

The gradient \( \nabla \phi(A) \) in \( \mathbb{R}^{n \times n} \) is equal to \( A \), as defined by the Fréchet derivative Definition 2.3. However, if \( \phi \) is restricted to \( \mathbb{S}^{n \times n} \), then observe that \( \nabla \phi_{\text{sym}}(A) = \text{sym}(A) = (A + A^T)/2 \) according to Corollary 2.6! This is different from what is predicted by Claim 2.4 as well as the online matrix derivative calculator [17].

Thus the definition of the gradient of a real-valued function defined on \( \mathbb{S}^{n \times n} \) in Corollary 2.6 is ensured to be symmetric. We will demonstrate that Claim 2.4 is incorrect. In fact, the correct symmetric gradient is the one given by the Fréchet derivative in Definition 2.5, Corollary 2.6, i.e. \( \text{sym}(\nabla \phi) \). To do this, we first illustrate through a simple example how \( \nabla \phi_{\text{sym}} \) as defined in Claim 2.4 gives an incorrect gradient.

2.2 An Illustrative Example 2

This short section is meant to highlight the inconsistencies that result from defining a symmetric gradient given by Claim 2.4.

We reconsider Geering’s example [8] and demonstrate the flaw in the argument that led to Claim 2.4. Define \( \phi : \mathbb{R}^{2 \times 2} \to \mathbb{R} \) given by \( \phi(A) = \text{det}(A) \) for any symmetric matrix \( A \) in \( \mathbb{S}^{2 \times 2} \). Let \( A = \begin{pmatrix} x & y \\ y & z \end{pmatrix} \).

The gradient, defined by the Fréchet derivative Definition 2.3 is

\[
\nabla \phi(A) = \det(A) A^{-T} = \begin{pmatrix} z & -y \\ -y & x \end{pmatrix}.
\]

If \( \phi \) is restricted to \( \mathbb{S}^{2 \times 2} \), then observe that \( \text{sym}(\nabla \phi) = \det(A) A^{-1} = \begin{pmatrix} z & -y \\ -y & x \end{pmatrix} \).

Geering identifies \( A \) through the triple \( [x, y, z]^T \) in \( \mathbb{R}^3 \), and consequently, we identify \( \phi(A) \) with \( \phi_s(x, y, z) = xz - y^2 \) as a functional on \( \mathbb{R}^3 \).

Then, \( \nabla \phi_s \) in \( \mathbb{R}^3 \) is given by \([z, -2y, x]^T\).

Geering identifies \( \nabla \phi_s \) with a matrix \( \begin{pmatrix} z & -2y \\ -2y & x \end{pmatrix} \) in \( \mathbb{S}^{2 \times 2} \) by retracing the earlier identification, which can be seen to agree with Claim 2.4 that \( \nabla \phi_{\text{sym}}(A) = \begin{pmatrix} z & -2y \\ -2y & x \end{pmatrix} \).

However, this identification is inconsistent because the gradients \( \nabla \phi_s(x, y, z) \) in \( \mathbb{R}^3 \) and \( \nabla \phi_{\text{sym}}(A) \) in \( \mathbb{S}^{2 \times 2} \) are not independent; rather, for every perturbation \( H = \begin{pmatrix} h_1 & h_2 \\ h_2 & h_3 \end{pmatrix} \) identified by \( h_s = [h_1, h_2, h_3]^T \), the inner product

\[
\langle \nabla \phi_s(x, y, z), h_s \rangle_{\mathbb{R}^3} = \langle \nabla \phi_{\text{sym}}(A), H \rangle_{\mathbb{S}^{2 \times 2}}.
\]

This relationship (2.6) (derived rigorously in a later section) expresses the simple idea that follows from the chain and product rule for derivatives applied to different representations of the same function, in this case identified as either \( \phi_{\text{sym}}(A) \) or \( \phi_s(x, y, z) \).

The crux of the issue is that Geering’s identification that agrees with Claim 2.4 violates the relationship, but the correct identification \( \begin{pmatrix} z & -y \\ -y & x \end{pmatrix} \) satisfies it. Thus, we have shown using Geering’s example, that Claim 2.4 cannot hold and that \( \nabla \phi_{\text{sym}} = \text{sym}(\nabla \phi) \) is the correct gradient.

In the subsequent sections, we shall prove (2.6) in greater generality and rigour for the restriction of any differentiable function \( \phi : \mathbb{R}^{n \times n} \to \mathbb{R} \) to the subspace \( \mathbb{S}^{n \times n} \), and also show how the same incorrect identification leads to the spurious Claim 2.4.
3 Gradient of real-valued functions of symmetric matrices

Matrices in $\mathbb{R}^{n \times n}$ can be naturally identified with vectors in $\mathbb{R}^{n^2}$. Thus a real valued function defined on $\mathbb{R}^{n \times n}$ can be naturally identified with a real valued function defined on $\mathbb{R}^{n^2}$. Moreover, the inner product on $\mathbb{R}^{n \times n}$ defined in Definition 2.1 is naturally identified with the Euclidean inner product on $\mathbb{R}^{n^2}$. This identification is useful when the goal is to find derivatives of scalar functions in $\mathbb{R}^{n \times n}$. The scheme then is to identify the scalar function on $\mathbb{R}^{n \times n}$ with a scalar function on $\mathbb{R}^{n^2}$, compute its gradient and use the identification to go back to construct the gradient in $\mathbb{R}^{n \times n}$. In case of symmetric matrices, the equation in Claim 2.4 is claimed to be the identification of the gradient in $\mathbb{S}^{n \times n}$ after computations in $\mathbb{R}^m$, since symmetric matrices are identified with $\mathbb{R}^m$ where $m = n(n + 1)/2$. In this section, we show that the claim is false. We first begin by formalizing these natural identifications we discussed in this paragraph.

Definition 3.1 The function $\text{vec}: \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n^2}$ given by

$$\text{vec}(A) := \begin{bmatrix} A_{11} \\ \vdots \\ A_{n1} \\ A_{12} \\ \vdots \\ A_{nn} \end{bmatrix}$$

identifies a matrix $A$ in $\mathbb{R}^{n \times n}$ with a vector $\text{vec}(A)$ in $\mathbb{R}^{n^2}$.

This operation can be inverted in obvious fashion, i.e., given the vector, one can reshape to form the matrix through the mat operator defined below.

Definition 3.2 $\text{mat}: \mathbb{R}^{n^2} \rightarrow \mathbb{R}^{n \times n}$ is the function given by

$$\text{mat}(\text{vec}(A)) = A, \text{ for every } A \in \mathbb{R}^{n^2}.$$ 

The subspace $\mathbb{S}^{n \times n}$ of $\mathbb{R}^{n \times n}$ is the subspace of all symmetric matrices and the object of investigation in this paper. Since this subspace has a dimension $m = n(n + 1)/2$, a symmetric matrix is naturally identified with a vector in $\mathbb{R}^m$. This identification is given by the elimination operation $P$ defined below.

Definition 3.3 Let $\mathcal{V}$ be the range of $\text{vec}$ restricted to $\mathbb{S}^{n \times n}$, i.e., $\mathcal{V} = \text{vec}(\mathbb{S}^{n \times n})$. The elimination operator $P$ is the function $P: \mathcal{V} \rightarrow \mathbb{R}^m$ that eliminates the redundant entries of a vector $v$ in $\mathcal{V}$ by extracting the entries corresponding to the lower-triangular part of the parent matrix. An explicit expression is as follows:

$$P = \sum_{1 \leq i, j \leq n} u_{ij} \text{vec}(E_{ij})^T,$$

where $E_{ij} \in \mathbb{R}^{n \times n}$ is zero everywhere except for unity in the $i, j$ position, and $u_{ij} \in \mathbb{R}^m$ is a unit vector whose component $i + (j - 1)n - j(j - 1)/2$ is unity and rest zero. The interested reader may consult [19] for more details about the elimination operator.

The operator $P$ lets us identify symmetric matrices in $\mathbb{S}^{n \times n}$ with vectors in $\mathbb{R}^m$ via the vech operator defined below.

Definition 3.4 The operator $\text{vech}$ is the function $\text{vech}: \mathbb{S}^{n \times n} \rightarrow \mathbb{R}^m$ given by

$$\text{vech}(A) = P \text{vec}(A), \text{ for every symmetric matrix } A \in \mathbb{S}^{n \times n}.$$ 

Definition 3.5 The duplication operator $D: \mathbb{R}^m \rightarrow \mathcal{V}$ given by

$$\text{vec}(A) = D \text{vech}(A) \text{ for every } A \in \mathbb{S}^{n \times n}$$

acts as the inverse of the elimination operator $P$. 
Lemma 3.6 Any a in $\mathbb{R}^n$ can be lifted to a symmetric matrix in $\mathbb{S}^{n \times n}$.

Proof $\text{mat}(D(a))$ lifts $a$ in $\mathbb{R}^n$ to a symmetric matrix in $\mathbb{S}^{n \times n}$. \hfill \square

We record some properties of the duplication operator $D$ that will be useful in proving our main theorem Theorem 3.8 later.

Lemma 3.7 Let $D$ be the duplication operator defined in Definition 3.5. The following are true:

1. Null $D = \{0\}$.
2. $D^T \vec{v}(A) = \text{vech}(A) + \text{vech}(A^T) - \text{vech}(A \circ I) \forall A \in \mathbb{S}^{n \times n}$.
3. $D^T D$ in $\mathbb{S}^{m \times m}$ is a positive-definite, symmetric matrix.
4. $(D^T D)^{-1}$ exists.

Proof See [19] \hfill \square

Consider a real valued function $\phi : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ and its restriction to $\mathbb{S}^{n \times n}$ $\phi_{\text{sym}} := \phi|_{\mathbb{S}^{n \times n}}$. Then $\phi_{\text{sym}}$ can be identified with a scalar function $\phi_s : \mathbb{R}^n \rightarrow \mathbb{R}$. Moreover, there is a relationship between the gradients calculated from the different representations of the function. The next theorem formalizes this concept and demonstrates two fundamental ideas - (1) the notion of Fréchet derivative naturally carries over to the subspace of symmetric matrices, hence there is no need to identify an equivalent representation of the functional in a lower dimensional space and, (2) if such an equivalent representation is constructed, a careful analysis leads to the correct gradient defined by the Fréchet derivative, not Claim 2.4.

Theorem 3.8 Consider a real-valued function $\phi : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ whose restriction to $\mathbb{S}^{n \times n}$ is the function $\phi_{\text{sym}} := \phi|_{\mathbb{S}^{n \times n}}$. As before $\nabla \phi(A)$ and $\nabla \phi_{\text{sym}}(A) = \text{sym}(\nabla \phi)$ denote the gradients of $\phi$, $\phi_{\text{sym}}$ at $A \in \mathbb{S}^{n \times n}$.

1. $\phi_{\text{sym}}$ can be identified with a scalar function $\phi_s : \mathbb{R}^n \rightarrow \mathbb{R}$ given by $\phi_s = \phi_{\text{sym}} \circ \text{mat} \circ D$ with $m = n(n + 1)/2$.
2. If $\nabla \phi_s(\text{vech}(A))$ in $\mathbb{R}^n$ is the gradient of $\phi_s$, then the symmetric matrix in $\mathbb{S}^{n \times n}$ given by $\text{mat}(D(D^T D)^{-1} \nabla \phi_s(\text{vech}(A)))$ represents the correct “symmetric gradient” of $\phi$ in the sense that

$$\left\{ \text{mat}(D(D^T D)^{-1} \nabla \phi_s(\text{vech}(A))), H \right\}_F = \langle \nabla \phi_s(\text{vech}(A)), \text{vech}(H) \rangle_{\mathbb{R}^m} = \langle \text{sym}(\nabla \phi(A)), H \rangle_F,$$

for all $H$ in $\mathbb{S}^{n \times n}$. Thus, $\text{mat}(D(D^T D)^{-1} \nabla \phi_s(\text{vech}(A))) = \text{sym}(\nabla \phi(A))$.

Before proving Theorem 3.8 we establish a few useful Lemmas that are interesting in their own right. Remark 3.11 will illustrate a plausible argument for Claim 2.4.

Lemma 3.9 Let $A, B$ be two symmetric matrices in $\mathbb{S}^{n \times n}$. Then we have the following equivalence

$$\langle A, B \rangle_F = \langle \text{vec}(A), \text{vec}(B) \rangle_{\mathbb{R}^n} = \langle D^T D \text{vech}(A), \text{vech}(B) \rangle_{\mathbb{R}^m} = \langle \text{vech}(A), D^T D \text{vech}(B) \rangle_{\mathbb{R}^m},$$

where $\langle \cdot, \cdot \rangle_{\mathbb{R}^2}$, $\langle \cdot, \cdot \rangle_{\mathbb{R}^n}$ are the usual Euclidean inner products.

Proof We start with the inner-product

$$\langle A, B \rangle_F = \text{tr}(A^T B) = \langle \text{vec}(A), \text{vec}(B) \rangle_{\mathbb{R}^2} \quad \text{(by Definition 2.1)}$$

$$= \langle D \text{vech}(A), D \text{vech}(B) \rangle_{\mathbb{R}^2} \quad \text{(from Equation (3.2))}$$

$$= \langle D^T D \text{vech}(A), \text{vech}(B) \rangle_{\mathbb{R}^m} \quad \text{(by definition of the transpose operator).} \hfill \square$$

The observation that $\langle A, B \rangle_F \neq \langle \text{vech}(A), \text{vech}(B) \rangle$ is a crucial one and lies at the heart of the discrepancy alluded to in the title of this article. Instead, if we want to refactor the inner product of two elements in $\mathbb{R}^n$ into one in $\mathbb{R}^n$ one has the following

Lemma 3.10 For any $a, b$ in $\mathbb{R}^n$,

$$\langle a, b \rangle_{\mathbb{R}^n} = \left\{ \text{mat}(D(D^T D)^{-1} a), \text{mat}(Db) \right\}_F.$$

\( \odot \) Springer
Proof: Given that \(a, b\) in \(\mathbb{R}^m\),

\[
(a, b)_{\mathbb{R}^m} = \left((D^T D)(D^T D)^{-1}a, b\right)_{\mathbb{R}^m} \quad \text{(by Lemma 3.7)}
\]

\[
= \left(D(D^T D)^{-1}a, Db\right)_{\mathbb{R}^{n^2}}
\]

\[
= \left(\text{mat}(D(D^T D)^{-1}a), \text{mat}(Db)\right)_{F} \quad \text{(by Lemma 3.9)}.
\]

\[\square\]

We are now ready to prove, in full generality for \(\phi_{\text{sym}} := \phi\big|_{\mathbb{S}^{n\times n}}\) what we demonstrated through the example earlier – that Claim 2.4 is false, and that the “symmetric gradient” should be \(\text{sym}(\nabla \phi)\), identical to that computed using Corollary 2.6.

Proof of Theorem 3.8

\[
\begin{array}{ccc}
\mathbb{S}^{n\times n} & \xrightarrow{\phi_{\text{sym}} := \phi\big|_{\mathbb{S}^{n\times n}}} & \mathbb{R} \\
\downarrow \text{mat} & & \uparrow \phi_s \\
\mathbb{V} \subset \mathbb{R}^{n^2} & \xleftarrow{P} & \mathbb{R}^m \\
\downarrow \text{vec} & & \uparrow D \\
\end{array}
\]

\(\text{(3.3)}\)

1. The operators defined above establish the commutative diagram given in Equation (3.3). These yield the following relation

\[
\phi_s = \phi_{\text{sym}} \circ \text{mat} \circ D, \tag{3.4}
\]

where \(\circ\) represents the usual composition of functions and \(m = n(n + 1)/2\).

2. From Equation (3.3),

\[
\phi_{\text{sym}} = \phi_s \circ P \circ \text{vec}. \tag{3.5}
\]

Thus, when the derivatives are evaluated at \(A \in \mathbb{S}^{n\times n}\), the chain-rule for Fréchet derivatives yields

\[
D\phi_{\text{sym}}(A) = D\phi_s(P \circ \text{vec}(A)) \circ DP(\text{vec}(A)) \circ D\text{vec}(A). \tag{3.6}
\]

Noting that \(P\) and \(\text{vec}\) are linear operators and considering the action on an arbitrary symmetric matrix \(H\), the equation above yields the following relationship between the Fréchet derivatives at \(A\):

\[
D\phi_{\text{sym}}(A)(H) = D\phi_s(\text{vech}(A))(\text{vech}(H)) \text{ for every } H \in \mathbb{S}^{n\times n}. \tag{3.7}
\]

With the usual inner products defined earlier, the Riesz Representation Theorem gives us the following relationship between the gradients,

\[
\langle \nabla \phi_{\text{sym}}(A), H \rangle_F = \langle \nabla \phi_s(\text{vech}(A)), \text{vech}(H) \rangle_{\mathbb{R}^m} \text{ for every } H \in \mathbb{S}^{n\times n}. \tag{3.8}
\]

Thus far, we have shown the argument of each gradient or derivative explicitly. However, from here on for ease of readability we drop the arguments and refer to the gradients as \(\nabla \phi_{\text{sym}}\) and \(\nabla \phi_s\).

Recall that from Corollary 2.6 and Subequation 2.1, we already know that the gradient \(\nabla \phi_{\text{sym}}\) is given by \(\text{sym}(\nabla \phi)\). Since \(\nabla \phi_s\) is a vector in \(\mathbb{R}^m\), it may appear that it can be lifted to the space of symmetric matrices to yield a matrix as in Lemma 3.6. However such a lifting fails to satisfy the inner product relationship given by (3.8) and we will demonstrate subsequently that it will instead yield the incorrect gradient (see Remark 3.11).
By Lemma 3.10 and the fact that \( \mat(D \vech(H)) = \mat(\vec(H)) = H \), we find that the correct lifting is given by \( \mat(D(D^T D)^{-1} \nabla \phi_s) \) in \( \mathbb{R}^{n \times n} \) such that,
\[
\langle \nabla \phi_s, \vech(H) \rangle_{\mathbb{R}^n} = \left( \mat(D(D^T D)^{-1} \nabla \phi_s), H \right)_F \text{ for every } H \in \mathbb{R}^{n \times n}.
\] (3.9)

To show that it is indeed the correct expression for the “symmetric gradient” we need to show that \( \mat(D(D^T D)^{-1} \nabla \phi_s) = \text{sym}(\nabla \phi) \), but this follows immediately from Equation (3.8).

This completes our proof of Theorem 3.8. \( \square \)

**Remark 3.11** If we ignore Lemma 3.10 and the constraint Equation (3.8), and instead naively use Lemma 3.6, we obtain \( \mat(D \nabla \phi_n) = \nabla^{\text{claim}} \phi_{\text{sym}} \) as illustrated in the example in Section 2.2 and stated in Claim 2.4.

In order to demonstrate the result, first note that, from Equation 3.8, Lemma 3.9 and the properties of the duplication operator \( D \) stated in Lemma 3.7, we can relate \( \nabla \phi_n \) to \( \nabla \phi_{\text{sym}} \) in the following way:
\[
\nabla \phi_s = D^T \vec(\nabla \phi_{\text{sym}}) = \vec(\nabla \phi_{\text{sym}}) + \vec(\nabla \phi_{\text{sym}}^T) - \vec(\nabla \phi_{\text{sym}} \circ I).
\] (3.10)

This is equivalent to
\[
\nabla \phi_s = \nabla(\vec(\phi + \nabla \phi^T - \nabla \phi \circ I))
\]
so that \( \mat(D \nabla \phi_n) = \mat(D \vec(\nabla \phi + \nabla \phi^T - \nabla \phi \circ I)) \) which simplifies to
\[
\mat(D \nabla \phi_n) = \mat(\vec(\nabla \phi + \nabla \phi^T - \nabla \phi \circ I)) = \nabla \phi + \nabla \phi^T - \nabla \phi \circ I = \nabla^{\text{claim}} \phi_{\text{sym}}.
\]

Thus, the same fundamental flaw discovered in Section 2.2 underpins the “proof” of the spurious Claim 2.4, and the crux of the theorem was the recognition that while the lifting of an element in \( \mathbb{R}^m \) to \( \mathbb{R}^{n \times n} \) follows Lemma 3.6, the gradient in \( \mathbb{R}^m \) must satisfy Equation (3.8) and instead must be lifted by Lemma 3.10.

**Remark 3.12** The possible claim that one is free to choose any arbitrary lifting rule to transform the gradient from \( \mathbb{R}^m \) to \( \mathbb{R}^{n \times n} \) is mathematically unsound. What is true is that one is free to choose an inner-product on \( \mathbb{R}^{n \times n} \), but once the choice is made, everything else is imposed upon us as a logical consequence - the gradient from the Riesz Representation Theorem, the lifting rules from Lemma 3.6, Lemma 3.9 and Lemma 3.10.

**Remark 3.13** While our analysis has assumed that the field in question is \( \mathbb{R} \), the same arguments will be valid for matrix functionals defined over the complex field \( \mathbb{C} \) with an appropriate modification of the definition of the inner-product in Definition 2.1.

**Remark 3.14** A larger theme of this article is that the Fréchet derivative over linear manifolds (subspaces) of \( \mathbb{R}^{n \times n} \) can be obtained from the Fréchet derivative over \( \mathbb{R}^{n \times n} \) by an appropriate projection/restriction to the relevant linear manifold as shown in Corollary 2.6. In this paper, the linear manifold was the set of symmetric matrices designated as \( \mathbb{S}^{n \times n} \). One can adapt the same ideas expressed here to obtain the derivative over the subspace of skew-symmetric, diagonal, upper-triangular, or lower triangular matrices. However, note that this remark does not apply to the set of orthogonal matrices since it is not a linear manifold.

### 4 Implication and significance

The implication of Theorem 3.8 depends on the context of the application and the quantity of interest. The gradient itself may be the quantity sought, or it could be necessary as part of an optimization algorithm such as gradient descent.

If the quantity of interest is the gradient of a scalar function defined over \( \mathbb{S}^{n \times n} \), then the correct gradient is the one given by Theorem 3.8, irrespective of the method used to evaluate it. The error in using Claim 2.4 is quantified by Lemma 4.1. In simple terms, the diagonal entries are exact, but the off-diagonal terms are twice the correct value so if calculating stresses from a strain energy function using Claim 2.4, the shear stresses have a relative error of 100%!

**Lemma 4.1** For \( \nabla \phi_{\text{sym}}, \nabla^{\text{claim}} \phi_{\text{sym}} \) defined by Corollary 2.6 and Claim 2.4 respectively,
\[
\nabla^{\text{claim}} \phi_{\text{sym}} - \nabla \phi_{\text{sym}} = \nabla \phi_{\text{sym}} - \nabla \phi_{\text{sym}} \circ I.
\]
Fig. 1 Gradient descent applied to the function $\text{tr}(X^2)$ and $\text{tr}(X^4)$. If $X \in S^{2 \times 2}$, and $\lambda_{\min}, \lambda_{\max}$ are the eigenvalues of $X$, $\text{tr}(X^2) = \lambda_{\min}^2 + \lambda_{\max}^2$, $\text{tr}(X^4) = \lambda_{\min}^4 + \lambda_{\max}^4$. The minimum in both cases occurs at $X = 0$. The iterates of gradient descent are superposed on the contours of $\phi$, showing how the correct gradient takes the iterates on the shortest path from the initial guess to the optimal point. Tracking the residual $|\phi - \phi_{\min}|$ also confirms that using the true gradient results in faster decay to the optimal solution.

Proof

\[
\nabla^{\text{claim}} \phi_{\text{sym}} = \nabla \phi + \nabla \phi^T - \nabla \phi \circ I
\]

\[
= 2(\nabla \phi + \nabla \phi^T)/2 - \left(\frac{\nabla \phi + \nabla \phi^T}{2}\right) \circ I
\]

\[
= 2\nabla \phi_{\text{sym}} - \nabla \phi_{\text{sym}} \circ I
\]

If we consider a simple unconstrained problem of minimization of a scalar matrix function $\phi(X)$, one can state it in two equivalent ways using the commutative diagram (3.5):

\[
\arg\min_{X \in S^{n \times n}} \phi(X). \tag{4.1a}
\]

\[
\arg\min_{x \in \mathbb{R}^m} \phi_s(x). \tag{4.1b}
\]

If the gradient is expressible analytically, then it is clear from Lemma 4.1 that the critical points found by both the true and spurious gradient are identical. However, the use of the gradient in the context of gradient descent for Eq. (4.1) leads to an interesting situation, for gradient descent may be provided with either the true gradient or the spurious gradient due to Claim 2.4. If Eq. (4.1b) is used, there is no ambiguity about the gradient, so supposing the optimal argument is found to be $x^*$, then $\phi_{\min} = \phi_s(x^*) = \phi(X^*)$ where $X^* = \text{mat}(Dx^*)$ has been lifted to $S^{n \times n}$ by Lemma 3.6.
However, if the formulation (4.1a) is used, it is interesting to consider the effect of the spurious gradient \( \nabla_{\text{claim}} \phi_{\text{sym}} \) relative to the correct gradient \( \nabla \phi_{\text{sym}} \) in a gradient descent algorithm [29]. We do so for the functions \( \text{tr} \left( X^2 \right) \) and \( \text{tr} \left( X^4 \right) \) where \( X \in \mathbb{S}^{2 \times 2} \). Restricting ourselves to the simplest non-trivial setting of \( \mathbb{S}^{2 \times 2} \) allows us to parametrize the iterates in terms of their eigenvalues and visualize the trajectory, for we already know the optimal value is attained at \( 0 \in \mathbb{S}^{2 \times 2} \) that has both eigenvalues zero. The results are shown in (1). In both cases, we see that the iterates using the true gradient \( \nabla \phi_{\text{sym}} \) move along a straight line (the shortest path) in the plane of eigenvalues towards the optimum; however the iterates using the spurious gradient \( \nabla_{\text{claim}} \phi_{\text{sym}} \) also converge to the optimum. Tracking the residual \( |\phi - \phi_{\text{min}}| \) also confirms that using the true gradient results in faster decay to the optimal solution. The convergence to an optimal solution observed in these two special cases despite using the spurious gradient reflects a fortuitous property that though \( \nabla_{\text{claim}} \phi_{\text{sym}} \) is not the true gradient, it is still an ascent direction as proved in Lemma 4.2.

**Lemma 4.2** For \( \nabla \phi_{\text{sym}}, \nabla_{\text{claim}} \phi_{\text{sym}} \) defined by Corollary 2.6 and Claim 2.4 respectively,

\[
\frac{\langle \nabla_{\text{claim}} \phi_{\text{sym}}, \nabla \phi_{\text{sym}} \rangle_F}{\| \nabla \phi_{\text{sym}} \|_F \| \nabla_{\text{claim}} \phi_{\text{sym}} \|_F} = \frac{2 - r^2}{\sqrt{4 - 3r^2}} \geq 0, \quad \text{where} \quad r = \frac{\| \nabla \phi_{\text{sym}} \circ I \|_F^2}{\| \nabla \phi_{\text{sym}} \|_F^2}.
\]

**Proof** Consider the inner products

\[
\begin{align*}
\langle \nabla_{\text{claim}} \phi_{\text{sym}}, \nabla \phi_{\text{sym}} \rangle_F &= 2 \| \nabla \phi_{\text{sym}} \|_F^2 - \| \nabla \phi_{\text{sym}} \circ I \|_F^2, \\
\langle \nabla_{\text{claim}} \phi_{\text{sym}}, \nabla_{\text{claim}} \phi_{\text{sym}} \rangle_F &= \| \nabla_{\text{claim}} \phi_{\text{sym}} \|_F^2 = 4 \| \nabla \phi_{\text{sym}} \|_F^2 - 3 \| \nabla \phi_{\text{sym}} \circ I \|_F^2.
\end{align*}
\]

The inner product of \( \nabla_{\text{claim}} \phi_{\text{sym}}, \nabla \phi_{\text{sym}} \) being non-negative shows that \( \nabla_{\text{claim}} \phi_{\text{sym}} \), though not the gradient, is always an ascent direction. It also allows us to see that \( \nabla_{\text{claim}} \phi_{\text{sym}} \) is aligned with \( \nabla \phi_{\text{sym}} \) when either the diagonal components or the off-diagonal components of \( \nabla \phi_{\text{sym}} \) vanish, i.e., when \( r = 0 \) or \( r = 1 \).

## 5 Conclusions

In this article, we investigated the two different notions of a gradient that exist for a real-valued function when the argument is a symmetric matrix. The first notion is the mathematical definition of a Fréchet derivative on the space of symmetric matrices. The other definition aims to eliminate the redundant degrees of freedom present in a symmetric matrix and perform the gradient calculation in the space of reduced-dimension and finally map the result back into the space of symmetric matrices. We showed, both through an example and rigorously through a theorem, that the problem in the second approach lies in the final step as the gradient in the reduced-dimension space is mapped into a symmetric matrix. Moreover, the approach does not recognize that Definition 2.5, restricted to \( \mathbb{S}^{n \times n} \), already accounts for the symmetry in the matrix argument; thus there is no theoretical need to identify an equivalent representation of the functional in a lower-dimensional space of dimension \( m = n(n+1)/2 \). However, we demonstrated that if such an equivalent representation is constructed, then a consistent approach does lead to the correct gradient. We also discussed the implications of using the spurious gradient in different classes of problems, such as those where the gradient itself may be the quantity sought, or as part of an optimization algorithm such as gradient descent. We showed that the spurious gradient has a relative error of 100% in the off-diagonal components, but fortuitously, it proves to be an ascent direction, so that its use in gradient descent may not lead to major issues. Since derivatives and gradients are fundamental ideas, we feel there should be no ambiguity about their definitions and hence there is an urgent need to clarify the issues muddying the waters. We thus lay to rest all the confusion, and unambiguously answer the question posed in the title of this article.

**Acknowledgements** SS thanks Los Alamos National Laboratory (US) for funding through #20170508DR while NP was funded by #20060599ECR & ASC-Directive. The authors also thank Leello Tadesse Dadi for suggesting the illustration of the trajectory of the iterates in terms of their eigenvalues. The opinions expressed in this paper are those of the authors and do not necessarily reflect that of the sponsors.
No datasets were generated or analyzed in the current study.

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

### Competing Interests
The authors declare that they have no competing interests.

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

### References

1. Michael Athans. Matrix minimum principle. Technical Report Report ESL-R-317, Massachusets Institute of Technology Lincoln Laboratory, Aug 1967. [Online; accessed 4. Mar. 2020].
2. Michael Athans. The matrix minimum principle. *Information and Control*, 11(5):592–606, Nov 1967.
3. 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. [Online; accessed 1. Nov. 2019].
4. J. Brewer. The gradient with respect to a symmetric matrix. *IEEE Transactions on Automatic Control*, 22(2):265–267, Apr 1977.
5. Ward Cheney. *Analysis for Applied Mathematics*, volume 208. Springer Science & Business Media, 2013.
6. Paul S. Dwyer. Some Applications of Matrix Derivatives in Multivariate Analysis. *Journal of the American Statistical Association*, 62(318):607–625, Jun 1967.
7. Friedrich Gebhardt. Maximum likelihood solution to factor analysis when some factors are completely specified. *Psychometrika*, 36(2):155–163, Jun 1971.
8. H. Geering. On calculating gradient matrices. *IEEE Transactions on Automatic Control*, 21(4):615–616, Aug 1976.
9. Morton E. Gurtin, Eliot Fried, and Lallit Anand. *Mechanics and Thermodynamics of Continua*. Cambridge University Applied Mathematics Research eXpress, 2010.
10. David A. Harville. *Matrix Algebra from a Statistician’s Perspective*. Springer-Verlag, New York, 1997.
11. 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.
12. TopDog (https://math.stackexchange.com/users/569224/topdog). Matrix gradient of det(x). Mathematics Stack Exchange. URL: https://math.stackexchange.com/q/2816512 (version: 2018-06-12).
13. Mikhail Itskov. Tensor Algebra and Tensor Analysis for Engineers. Springer, Cham, 2019.
14. jds (https://math.stackexchange.com/users/31891/jds). Derivative with respect to symmetric matrix. Mathematics Stack Exchange. URL: https://math.stackexchange.com/q/3005374 (version: 2018-11-19).
15. William Thomson Lord Kelvin. Elements of a mathematical theory of elasticity. *Philosophical Transactions of the Royal Society of London*, 146:481–498, 1856.
16. kMaster (https://math.stackexchange.com/users/58360/kmaster). Derivative of the inverse of a symmetric matrix. Mathematics Stack Exchange. URL: https://math.stackexchange.com/q/982386 (version: 2014-10-20).
17. Sören Laue, Matthias Mitterreiter, and Joachim Giesen. Computing higher order derivatives of matrix and tensor expressions. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.
18. J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley, New York, 1988.
19. Jan R. Magnus and H. Neudecker. The elimination matrix: Some lemmas and applications. *SIAM Journal on Algebraic Discrete Methods*, 1(4):422–449, 1980.
20. A. M. Mathai. *Jacobians of Matrix Transformations and Functions of Matrix Argument*. World Scientific, Singapore, 1997.
21. Charles E. McCulloch. Symmetric Matrix Derivatives with Applications. *Journal of the American Statistical Association*, Mar 1980.
22. me10240 (https://math.stackexchange.com/users/66158/me10240). Can derivative formulae in matrix cookbook be interpreted as frechet derivatives? Mathematics Stack Exchange. URL: https://math.stackexchange.com/q/2508276 (version: 2017-11-06).
23. me10240 (https://math.stackexchange.com/users/66158/me10240). Making sense of matrix derivative formula for determinant of symmetric matrix as a frechet derivative? Mathematics Stack Exchange. URL: https://math.stackexchange.com/q/2436680 (version: 2017-09-21).
24. Morteza Mehrabadi and Stephen C. Cowin. Eigentensors of linear anisotropic elastic materials. *Quarterly Journal of Mechanics and Applied Mathematics*, 43:15 – 41, 1990.
25. Thomas Minka. Old and New Matrix Algebra Useful for Statistics, 2001. [Online; accessed 1. Nov. 2019].
26. James Munkres. *Analysis on Manifolds*. Addison Wesley, New York, 1991.
27. Iain Murray. Differentiation of the Cholesky decomposition. ArXiv e-prints, Feb 2016.
28. D. G. Nel. On matrix differentiation in statistics. *South African Statistical Journal*, 14(2):137–193, Jan 1980.
29. Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, 2006.
30. Ray W. Ogden. *Non-linear Elastic Deformations*. Dover Publications, New York, 1997.
31. A-M. Parring. About the concept of the matrix derivative. *Linear Algebra and its Applications*, 176:223–235, Nov 1992.
32. Kaare Brandt Petersen and Michael Syskind Pedersen. The Matrix Cookbook, 11 2012. Version 20121115.
33. G.S. Rogers. *Matrix Derivatives*. Marcel Dekker, New York, 1980.
34. Steven Roman. *Advanced Linear Algebra*, volume 3. Springer.
35. S.R. Searle. *Matrix Algebra for Statistics*. John Wiley, New York, 1982.
36. George Arthur Frederick Seber. *A Matrix Handbook for Statisticians*. John Wiley & Sons, New Jersey, 2008.
37. tomka (https://math.stackexchange.com/users/118706/tomka). What is the derivative of the determinant of a symmetric positive definite matrix? Mathematics Stack Exchange. URL: https://math.stackexchange.com/q/1981210 (version: 2017-04-13).
38. 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.
39. 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.
40. 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.
41. Paolo Vannucci. *Anisotropic Elasticity*. Springer, Singapore, 2018.
42. W. Voigt. *Lehrbuch der Kristallphysik*. B.G. Teubner, Leipzig, 1910.
43. P. Walsh. On symmetric matrices and the matrix minimum principle. *IEEE Transactions on Automatic Control*, 22(6):995–996, Dec 1977.
44. wueb (https://math.stackexchange.com/users/238307/wueb). Understanding notation of derivatives of a matrix. Mathematics Stack Exchange. URL: https://math.stackexchange.com/q/2131708 (version: 2017-04-13).
45. 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.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.