Matrix Calculus: Derivation and Simple Application

HU, Pili*

March 30, 2012†

Abstract

Matrix Calculus[3] is a very useful tool in many engineering problems. Basic rules of matrix calculus are nothing more than ordinary calculus rules covered in undergraduate courses. However, using matrix calculus, the derivation process is more compact. This document is adapted from the notes of a course the author recently attends. It builds matrix calculus from scratch. Only prerequisites are basic calculus notions and linear algebra operation. To get a quick executive guide, please refer to the cheat sheet in the end.

*hupili [at] ie [dot] cuhk [dot] edu [dot] hk

†Last compile:April 1, 2012

Contents

1 Introductory Example

We start with an one variable linear function:

$$f(x) = ax \tag{1}$$

To be coherent, we abuse the partial derivative notation:

$$\frac{\partial f}{\partial x} = a \tag{2}$$

Extending this function to be multivariate, we have:

$$f(x) = \sum_i a_i x_i = a^{\mathrm{T}} x \tag{3}$$

Where $a = [a_1, a_2, \ldots, a_n]^{\mathrm{T}}$ and $x = [x_1, x_2, \ldots, x_n]^{\mathrm{T}}$. We first compute partial derivatives directly:

$$\frac{\partial f}{\partial x_k} = \frac{\partial(\sum_i a_i x_i)}{\partial x_k} = a_k \tag{4}$$

for all $k = 1, 2, \ldots, n$. Then we organize n partial derivatives in the following way:

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = a \tag{5}$$

The first equality is by proper definition and the rest roots from ordinary calculus rules.

Eqn(5) is analogous to eqn(2), except the variable changes from a scalar to a vector. Thus we want to directly claim the result of eqn(5) without those intermediate steps solving for partial derivatives separately. Actually, we'll see soon that eqn(5) plays a core role in matrix calculus.

Following sections are organized as follows:

- Section(2) builds commonly used matrix calculus rules from ordinary calculus and linear algebra. Necessary and important properties of linear algebra is also proved along the way. This section is not organized afterhand. All results are proved when we need them.
- Section(3) shows some applications using matrix calculus.
- Section(4) concludes a cheat sheet of matrix calculus. Note that this cheat sheet may be different from others. Users need to figure out some basic definitions before applying the rules.

2 Derivation

2.1 Organization of Elements

From the introductary example, we already see that matrix calculus does not distinguish from ordinary calculus by fundamental rules. However, with better organization of elements and proving useful properties, we can simplify the derivation process in real problems.

The author would like to adopt the following definition:

Definition 1. *For a scalar valued function $f(x)$, the result $\frac{\partial f}{\partial x}$ has the same size with x. That is*

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{12}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \frac{\partial f}{\partial x_{21}} & \frac{\partial f}{\partial x_{22}} & \cdots & \frac{\partial f}{\partial x_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \frac{\partial f}{\partial x_{m2}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix} \tag{6}$$

In eqn(2), x is a 1-by-1 matrix and the result $\frac{\partial f}{\partial x} = a$ is also a 1-by-1 matrix. In eqn(5), x is a column vector(known as n-by-1 matrix) and the result $\frac{\partial f}{\partial x} = a$ has the same size.

Example 1. *By this definition, we have:*

$$\frac{\partial f}{\partial x^{\mathrm{T}}} = (\frac{\partial f}{\partial x})^{\mathrm{T}} = a^{\mathrm{T}} \tag{7}$$

Note that we only use the organization definition in this example. Later we'll show that with some matrix properties, this formula can be derived without using $\frac{\partial f}{\partial x}$ as a bridge.

2.2 Deal with Inner Product

Theorem 1. *If there's a multivariate scalar function $f(x) = a^{\mathrm{T}}x$, we have $\frac{\partial f}{\partial x} = a$.*

Proof. See introductary example. □

Since $a^{\mathrm{T}}x$ is scalar, we can write it equivalently as the trace of its own. Thus,

Proposition 2. *If there's a multivariate scalar function $f(x) = \mathrm{Tr}\left[a^{\mathrm{T}}x\right]$, we have $\frac{\partial f}{\partial x} = a$.*

$\mathrm{Tr}\left[\bullet\right]$ is the operator to sum up diagonal elements of a matrix. In the next section, we'll explore more properties of trace. As long as we can transform our target function into the form of theorem(1) or proposition(2), the result can be written out directly. Notice in proposition(2), a and x are both vectors. We'll show later as long as their sizes agree, it holds for matrix a and x.

2.3 Properties of Trace

Definition 2. *Trace of square matrix is defined as:* $\mathrm{Tr}\left[A\right] = \sum_i A_{ii}$

Theorem 3. *Matrix trace has the following properties:*

- *(1)* $\mathrm{Tr}\left[A+B\right] = \mathrm{Tr}\left[A\right] + \mathrm{Tr}\left[B\right]$
- *(2)* $\mathrm{Tr}\left[cA\right] = c\mathrm{Tr}\left[A\right]$
- *(3)* $\mathrm{Tr}\left[AB\right] = \mathrm{Tr}\left[BA\right]$
- *(4)* $\mathrm{Tr}\left[A_1A_2\dots A_n\right] = \mathrm{Tr}\left[A_nA_1\dots A_{n-1}\right]$
- *(5)* $\mathrm{Tr}\left[A^{\mathrm{T}}B\right] = \sum_i\sum_j A_{ij}B_{ij}$
- *(6)* $\mathrm{Tr}\left[A\right] = \mathrm{Tr}\left[A^{\mathrm{T}}\right]$

where A, B *are matrices with proper sizes, and* c *is a scalar value.*

Proof. See wikipedia [5] for the proof. □

Here we explain the intuitions behind each property to make it easier to remenber. Property(1) and property(2) shows the linearity of trace. Property(3) means two matrices' multiplication inside a the trace operator is commutative. Note that the matrix multiplication without trace is not commutative and the commutative property inside the trace does not hold for more than 2 matrices. Property (4) is the proposition of property (3) by considering $A_1A_2\dots A_{n-1}$ as a whole. It is known as cyclic property, so that you can rotate the matrices inside a trace operator. Property (5) shows a way to express the sum of element by element product using matrix product and trace. Note that inner product of two vectors is also the sum of element by element product. Property (5) resembles the vector inner product by form$(A^{\mathrm{T}}B)$. The author regards property (5) as the extension of inner product to matrices(Generalized Inner Product).

2.4 Deal with Generalized Inner Product

Theorem 4. *If there's a multivariate scalar function* $f(x) = \mathrm{Tr}\left[A^{\mathrm{T}}x\right]$*, we have* $\frac{\partial f}{\partial x} = A$*. (*$A, x$ *can be matrices).*

Proof. Using property (5) of trace, we can write f as:

$$f(x) = \mathrm{Tr}\left[A^{\mathrm{T}}x\right] = \sum_{ij} A_{ij}x_{ij} \tag{8}$$

It's easy to show:

$$\frac{\partial f}{\partial x_{ij}} = \frac{\partial(\sum_{ij} A_{ij}x_{ij})}{\partial x_{ij}} = A_{ij} \tag{9}$$

Organize elements using definition(1), it is proved. □

With this theorem and properties of trace we revisit example(1).

Example 2. *For vector a, x and function $f(x) = a^{\mathrm{T}}x$*

$$\begin{aligned}
& \frac{\partial f}{\partial x^{\mathrm{T}}} && (10)\\
= \ & \frac{\partial (a^{\mathrm{T}}x)}{\partial x^{\mathrm{T}}} && (11)\\
\textit{(f is scalar)} \ = \ & \frac{\partial (\mathrm{Tr}\left[a^{\mathrm{T}}x\right])}{\partial x^{\mathrm{T}}} && (12)\\
\textit{(property(3))} \ = \ & \frac{\partial (\mathrm{Tr}\left[xa^{\mathrm{T}}\right])}{\partial x^{\mathrm{T}}} && (13)\\
\textit{(property(6))} \ = \ & \frac{\partial (\mathrm{Tr}\left[ax^{\mathrm{T}}\right])}{\partial x^{\mathrm{T}}} && (14)\\
\textit{(property of transpose)} \ = \ & \frac{\partial (\mathrm{Tr}\left[(a^{\mathrm{T}})^{\mathrm{T}}x^{\mathrm{T}}\right])}{\partial x^{\mathrm{T}}} && (15)\\
\textit{(theorem(4))} \ = \ & a^{\mathrm{T}} && (16)
\end{aligned}$$

The result is the same with example(1), where we used the basic definition.

The above example actually demonstrates the usual way of handling a matrix derivative problem.

2.5 Define Matrix Differential

Although we want matrix derivative at most time, it turns out matrix differential is easier to operate due to the form invariance property of differential. Matrix differential inherit this property as a natural consequence of the following definition.

Definition 3. *Define matrix differential:*

$$\mathrm{d}A = \begin{bmatrix} \mathrm{d}A_{11} & \mathrm{d}A_{12} & \dots & \mathrm{d}A_{1n} \\ \mathrm{d}A_{21} & \mathrm{d}A_{22} & \dots & \mathrm{d}A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{d}A_{m1} & \mathrm{d}A_{m2} & \dots & \mathrm{d}A_{mn} \end{bmatrix} \tag{17}$$

Theorem 5. *Differential operator is distributive through trace operator:* $\mathrm{d}\mathrm{Tr}\left[A\right] = \mathrm{Tr}\left[\mathrm{d}A\right]$

Proof.

$$\text{LHS} = \mathrm{d}(\sum_i A_{ii}) = \sum_i \mathrm{d}A_{ii} \tag{18}$$

$$\text{RHS} = \text{Tr}\begin{bmatrix} \mathrm{d}A_{11} & \mathrm{d}A_{12} & \dots & \mathrm{d}A_{1n} \\ \mathrm{d}A_{21} & \mathrm{d}A_{22} & \dots & \mathrm{d}A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{d}A_{m1} & \mathrm{d}A_{m2} & \dots & \mathrm{d}A_{mn} \end{bmatrix} \tag{19}$$

$$= \sum_i \mathrm{d}A_{ii} = \text{LHS} \tag{20}$$

□

Now that matrix differential is well defined, we want to relate it back to matrix derivative. The scalar version differential and derivative can be related as follows:

$$\mathrm{d}f = \frac{\partial f}{\partial x}\mathrm{d}x \tag{21}$$

So far, we're dealing with scalar function f and matrix variable x. $\frac{\partial f}{\partial x}$ and $\mathrm{d}x$ are both matrix according to definition. In order to make the quantities in eqn(21) equal, we must figure out a way to make the RHS a scalar. It's not surprising that trace is what we want.

Theorem 6.

$$\mathrm{d}f = \text{Tr}\left[(\frac{\partial f}{\partial x})^{\mathrm{T}}\mathrm{d}x\right] \tag{22}$$

for scalar function f and arbitrarily sized x.

Proof.

$$\text{LHS} = \mathrm{d}f \tag{23}$$

$$(\text{definition of scalar differential}) = \sum_{ij} \frac{\partial f}{\partial x_{ij}}\mathrm{d}x_{ij} \tag{24}$$

$$\text{RHS} = \text{Tr}\left[(\frac{\partial f}{\partial x})^{\mathrm{T}}\mathrm{d}x\right] \tag{25}$$

$$(\text{trace property (5)}) = \sum_{ij}(\frac{\partial f}{\partial x})_{ij}(\mathrm{d}x)_{ij} \tag{26}$$

$$(\text{definition(3)}) = \sum_{ij}(\frac{\partial f}{\partial x})_{ij}\mathrm{d}x_{ij} \tag{27}$$

$$(\text{definition(1)}) = \sum_{ij} \frac{\partial f}{\partial x_{ij}}\mathrm{d}x_{ij} \tag{28}$$

$$= \text{LHS} \tag{29}$$

□

Theorem(6) is the bridge between matrix derivative and matrix differential. We'll see in later applications that matrix differential is more convenient to manipulate. After certain manipulation we can get the form of theorem(6). Then we can directly write out matrix derivative using this theorem.

2.6 Matrix Differential Properties

Theorem 7. *We claim the following properties of matrix differential:*

- $\mathrm{d}(cA) = c\mathrm{d}A$
- $\mathrm{d}(A+B) = \mathrm{d}A + \mathrm{d}B$
- $\mathrm{d}(AB) = \mathrm{d}AB + A\mathrm{d}B$

Proof. They're all natural consequences given the definition(3). We only show the 3rd one in this document. Note that the equivalence holds if LHS and RHS are equivalent element by element. We consider the (ij)-th element.

$$\begin{aligned}
\mathrm{LHS}_{ij} &= \mathrm{d}(\sum_k A_{ik}B_{kj}) && (30)\\
&= \sum_k (\mathrm{d}A_{ik}B_{kj} + A_{ik}\mathrm{d}B_{kj}) && (31)\\
\mathrm{RHS}_{ij} &= (\mathrm{d}AB)_{ij} + (A\mathrm{d}B)_{ij} && (32)\\
&= \sum_k \mathrm{d}A_{ik}B_{kj} + \sum_k A_{ik}\mathrm{d}B_{kj} && (33)\\
&= \mathrm{LHS}_{ij} && (34)
\end{aligned}$$

□

Example 3. *Given the function $f(x) = x^{\mathrm{T}}Ax$, where A is square and x is a column vector, we can compute:*

$$\begin{aligned}
\mathrm{d}f &= \mathrm{dTr}\left[x^{\mathrm{T}}Ax\right] && (35)\\
&= \mathrm{Tr}\left[\mathrm{d}(x^{\mathrm{T}}Ax)\right] && (36)\\
&= \mathrm{Tr}\left[\mathrm{d}(x^{\mathrm{T}})Ax + x^{\mathrm{T}}\mathrm{d}(Ax)\right] && (37)\\
&= \mathrm{Tr}\left[\mathrm{d}(x^{\mathrm{T}})Ax + x^{\mathrm{T}}\mathrm{d}Ax + x^{\mathrm{T}}A\mathrm{d}x\right] && (38)\\
\textit{(A is constant)} &= \mathrm{Tr}\left[\mathrm{d}x^{\mathrm{T}}Ax + x^{\mathrm{T}}A\mathrm{d}x\right] && (39)\\
&= \mathrm{Tr}\left[\mathrm{d}x^{\mathrm{T}}Ax\right] + \mathrm{Tr}\left[x^{\mathrm{T}}A\mathrm{d}x\right] && (40)\\
&= \mathrm{Tr}\left[x^{\mathrm{T}}A^{\mathrm{T}}\mathrm{d}x\right] + \mathrm{Tr}\left[x^{\mathrm{T}}A\mathrm{d}x\right] && (41)\\
&= \mathrm{Tr}\left[x^{\mathrm{T}}A^{\mathrm{T}}\mathrm{d}x + x^{\mathrm{T}}A\mathrm{d}x\right] && (42)\\
&= \mathrm{Tr}\left[(x^{\mathrm{T}}A^{\mathrm{T}} + x^{\mathrm{T}}A)\mathrm{d}x\right] && (43)
\end{aligned}$$

Using theorem(6), we obtain the derivative:

$$\frac{\partial f}{\partial x} = (x^{\mathrm{T}}A^{\mathrm{T}} + x^{\mathrm{T}}A)^{\mathrm{T}} = Ax + A^{\mathrm{T}}x \tag{44}$$

When A is symmetric, it simplifies to:

$$\frac{\partial f}{\partial x} = 2Ax \tag{45}$$

Let $A = I$, we have:

$$\frac{\partial (x^{\mathrm{T}}x)}{\partial x} = 2x \tag{46}$$

Example 4. *For a non-singular square matrix X, we have $XX^{-1} = I$. Take matrix differentials at both sides:*

$$0 = \mathrm{d}I = \mathrm{d}(XX^{-1}) = \mathrm{d}XX^{-1} + X\mathrm{d}(X^{-1}) \tag{47}$$

Rearrange terms:

$$\mathrm{d}(X^{-1}) = -X^{-1}\mathrm{d}XX^{-1} \tag{48}$$

2.7 Schema of Hanlding Scalar Function

The above example already demonstrates the general schema. Here we conclude the process:

1. $\mathrm{d}f = \mathrm{dTr}\left[f\right] = \mathrm{Tr}\left[\mathrm{d}f\right]$
2. Apply trace properties(see theorem(3)) and matrix differential properties(see theorem(7)) to get the following form:

$$\mathrm{d}f = \mathrm{Tr}\left[A^{\mathrm{T}}x\right] \tag{49}$$

3. Apply theorem(6) to get:

$$\frac{\partial f}{\partial x} = A \tag{50}$$

To this point, you can handle many problems. In this schema, matrix differential and trace play crucial roles. Later we'll deduce some widely used formula to facilitate potential applications. As you will see, although we rely on matrix differential in the schema, the deduction of certain formula may be more easily done using matrix derivatives.

2.8 Determinant

For a background of determinant, please refer to [6]. We first quote some definitions and properties without proof:

Theorem 8. *Let A be a square matrix:*

- *The minor M_{ij} is obtained by remove i-th row and j-th column of A and then take determinant of the resulting (n-1) by (n-1) matrix.*
- *The ij-th cofactor is defined as $C_{ij} = (-1)^{i+j} M_{ij}$.*
- *If we expand determinant with respect to the i-th row, $\det(A) = \sum_j A_{ij} C_{ij}$.*
- *The adjugate of A is defined as $\text{adj}(A)_{ij} = (-1)^{i+j} M_{ji} = C_{ji}$. So $\text{adj}(A) = C^{\mathrm{T}}$*
- *For non-singular matrix A, we have: $A^{-1} = \frac{\text{adj}(A)}{\det(A)} = \frac{C^{\mathrm{T}}}{\det(A)}$*

Now we're ready to show the derivative of determinant. Note that determinant is just a scalar function, so all techniques discussed above is applicable. We first write the derivative element by element. Expanding determinant on the i-th row, we have:

$$\frac{\partial \det(A)}{\partial A_{ij}} = \frac{\partial (\sum_j A_{ij} C_{ij})}{\partial A_{ij}} = C_{ij} \tag{51}$$

First equality is from determinant definition and second equality is by the observation that only coefficient of A_{ij} is left. Grouping all elements using definition(1), we have:

$$\frac{\partial \det(A)}{\partial A} = C = \text{adj}(A)^{\mathrm{T}} \tag{52}$$

If A is non-singular, we have:

$$\frac{\partial \det(A)}{\partial A} = (\det(A) A^{-1})^{\mathrm{T}} = \det(A)(A^{-1})^{\mathrm{T}} \tag{53}$$

Next, we use theorem(6) to give the differential relationship:

$$\begin{aligned} \mathrm{d} \det(A) &= \mathrm{Tr}\left[(\frac{\partial \det(A)}{\partial A})^{\mathrm{T}} \mathrm{d}A \right] && (54) \\ &= \mathrm{Tr}\left[(\det(A)(A^{-1})^{\mathrm{T}})^{\mathrm{T}} \mathrm{d}A \right] && (55) \\ &= \mathrm{Tr}\left[\det(A) A^{-1} \mathrm{d}A \right] && (56) \end{aligned}$$

In many practical problem, the log determinant is more widely used:

$$\frac{\partial \ln \det(A)}{\partial A} = \frac{1}{\det(A)} \frac{\partial \det(A)}{\partial A} = (A^{-1})^{\mathrm{T}} \tag{57}$$

The first equality comes from chain rule of ordinary calculus($\ln \det(A)$ and $\det(A)$ are both scalars). Similarly, we derive for differential:

$$\mathrm{d} \ln \det(A) = \mathrm{Tr}\left[A^{-1} \mathrm{d}A \right] \tag{58}$$

3 Application

3.1 The 2nd Induced Norm of Matrix

The induced norm of matrix is defined as [4]:

$$||A||_p = \max_x \frac{||Ax||_p}{||x||_p} \tag{59}$$

where $||\bullet||_p$ denotes the p-norm of vectors. Now we solve for $p = 2$. (By default, $||\bullet||$ means $||\bullet||_2$)

The problem can be restated as:

$$||A||^2 = \max_x \frac{||Ax||^2}{||x||^2} \tag{60}$$

since all quantities involved are non-negative. Then we consider a scaling of vector $x' = tx$, thus:

$$||A||^2 = \max_{x'} \frac{||Ax'||^2}{||x'||^2} = \max_x \frac{||tAx||^2}{||tx||^2} = \max_x \frac{t^2||Ax||^2}{t^2||x||^2} = \max_x \frac{||Ax||^2}{||x||^2} \tag{61}$$

This shows the invariance under scaling. Now we can restrict our attention to those x with $||x|| = 1$, and reach the following formulation:

$$\text{Maximize} \qquad f(x) = ||Ax||^2 \tag{62}$$

$$s.t. \qquad ||x||^2 = 1 \tag{63}$$

The standard way to handle this constrained optimization is using Lagrange relaxation:

$$L(x) = f(x) - \lambda(||x||^2 - 1) \tag{64}$$

Then we apply the general schema of handling scalar function on $L(x)$. First take differential:

$$\begin{aligned}
\mathrm{d}L(x) &= \mathrm{dTr}\left[L(x)\right] && (65)\\
&= \mathrm{Tr}\left[\mathrm{d}(L(x))\right] && (66)\\
&= \mathrm{Tr}\left[\mathrm{d}(x^{\mathrm{T}}A^{\mathrm{T}}Ax - \lambda(x^{\mathrm{T}}x - 1))\right] && (67)\\
&= \mathrm{Tr}\left[2x^{\mathrm{T}}A^{\mathrm{T}}A\mathrm{d}x - \lambda(2x^{\mathrm{T}}\mathrm{d}x)\right] && (68)\\
&= \mathrm{Tr}\left[(2x^{\mathrm{T}}A^{\mathrm{T}}A - 2\lambda x^{\mathrm{T}})\mathrm{d}x\right] && (69)
\end{aligned}$$

Next write out derivative:

$$\frac{\partial L}{\partial x} = 2A^{\mathrm{T}}Ax - 2\lambda x \tag{70}$$

Let $\frac{\partial L}{\partial x} = 0$, we have:

$$(A^{\mathrm{T}}A)x = \lambda x \tag{71}$$

That means x is the eigen vector of $(A^{\mathrm{T}}A)$ (normalized to $||x|| = 1$), and λ is corresponding eigen value. We plug this result back to objective function:

$$f(x) = x^{\mathrm{T}}(A^{\mathrm{T}}Ax) = x^{\mathrm{T}}(\lambda x) = \lambda \tag{72}$$

which means, to maximize $f(x)$, we should pick the maximum eigen value:

$$||A||^2 = \max f(x) = \lambda_{\max}(A^{\mathrm{T}}A) \tag{73}$$

That is:

$$||A|| = \sqrt{\lambda_{\max}(A^{\mathrm{T}}A)} = \sigma_{\max}(A) \tag{74}$$

where $\sigma_{\max}$ denotes the maximum singular value. If A is real symmetric, $\sigma_{\max}(A) = \lambda_{\max}(A)$.

Now we consider a real symmetric A and check whether:

$$\lambda^2_{\max}(A) = \max_x \frac{||Ax||^2}{||x||^2} = \max_x \frac{x^{\mathrm{T}}A^{\mathrm{T}}Ax}{x^{\mathrm{T}}x} \tag{75}$$

Proof. Since $A^{\mathrm{T}}A$ is real symmetric, it has an orthnormal basis formed by n eigen vectors, $v_1, v_2, \ldots, v_n$, with eigen values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. We can write $x = \sum_i c_i v_i$, where $c_i =< x, v_i >$. Then,

$$\begin{aligned}
& \frac{x^{\mathrm{T}}A^{\mathrm{T}}Ax}{x^{\mathrm{T}}x} && (76)\\
(v_k \text{ is an orthnormal set}) \quad &= \frac{\sum_i \lambda_i c_i^2}{\sum_i c_i^2} && (77)\\
&\leq \frac{\sum_i \lambda_1 c_i^2}{\sum_i c_i^2} && (78)\\
&= \lambda_1 && (79)
\end{aligned}$$

Now we have proved an upper bound for $||A||^2$. We show this bound is achievable by assigning $x = v_1$. □

4 Cheat Sheet

Acknowledgements

Thanks prof. XU, Lei's tutorial on matrix calculus. Besides, the author also benefit a lot from other online materials.

References

[1] Pili Hu. Matrix calculus. GitHub, https://github.com/hupili/tutorial/tree/master/matrix-calculus, 3 2012. HU, Pili's tutorial collection.

[2] Pili Hu. Tutorial collection. GitHub, https://github.com/hupili/tutorial, 3 2012. HU, Pili's tutorial collection.

[3] Matrix Calculus, Wikipedia, http://en.wikipedia.org/wiki/Matrix_calculus

[4] Matrix Norm, Wikipedia, http://en.wikipedia.org/wiki/Matrix_norm

[5] Matrix Trace, Wikipedia, http://en.wikipedia.org/wiki/Trace_%28linear_algebra%29

[6] Matrix Determinant, Wikipedia, http://en.wikipedia.org/wiki/Determinant

Appendix