MIT 18.S096 Matrix Calculus (2023)

Ian Poon

November 2024

hmmm...here is nice online matrix calculator tool.

Contents

1 Overview and Motivation

Theorem 1 (Differential Product Rule)
Let A, B be two matrices then we have the differential product rule for AB

$$d(AB) = (dA)B + A(dB)$$

By the differential of the matrix, we think of it as a small(unconstrained) change in the matrix A

Proof. we will cover this in later lectures

Example 2
By the product rule we have

1. $d(u^T v) = (du)^T v + u^T(dv) = v^T du + u^T dv$ since dot products commute
2. $d(uv^T) = (du)v^T + u(dv)^T$

We will cover more formal proofs for this kind of equations later

2 Derivatives as Linear Operators

Fact 3

Note the following notations

- derivative: $\boxed{f'}$

  note that $df = f(x+dx) - f(x) = f'(x)[dx]$

- gradient: $\boxed{\nabla f}$

  note that $df = \langle \nabla f, dx \rangle$

- difference: $\boxed{\delta x}$ and $\boxed{\delta f} = f(x+\delta x) - f(x)$

  These are small but not infinitesimal changes in the input x and output f(depending implicitly on both x and δx). This is *not* a linear operator, instead is a j ust an element of a vector space

- differential: $\boxed{df}$ and $\boxed{df} = f(x+dx) - f(x)$

  These are small and infinitesimal(we drop higher order terms) changes in the input x and output f. Again this is an element of a vector space not a linear operator.

- partial derivative: $\boxed{\dfrac{\partial f}{\partial x}}$, $f_x, \partial_x f$

  Note that $df = \frac{\partial f}{\partial x}[dx] + \frac{\partial f}{\partial y}[dy]$

Fact 4

We may write the **directional derivative** as

$$\frac{\partial}{\partial a} f(x+av)|_{a=0} = \lim_{\delta a \to 0} \frac{f(x+\delta av) - f(x)}{\delta a}$$

where we dropped higher terms in the limit of $\delta \to 0$ which gives

$$f(x + \underbrace{dav}_{dx}) - f(x) = f'[dx] = daf'(x)[v]$$

after factoring our da in the last step isnce $f'(x)$ is a linear operator and so

$$\boxed{\frac{\partial}{\partial a} f(x+av)|_{a=0} = f'(x)[v]}$$

The point is here is that it is perfectly reasonable to write $f'(x)[v]$ where v is not infinitesimal. So this term is not equal to df but instead simply a directional derivative

Fact 5
Now consider a scalar valued function f which takes in a column of vectors $x \in \mathbb{R}^n$ so we have

$$df = f(x+dx) - f(x) = f'(x)[dx] = \text{scalar}$$

because a scalar is produced it follows by the laws of matrix multiplication that df must be a row vector. We denote this row vector by

$$(\nabla f)^T$$

so that df is the dot product of dx with the gradient that is

$$df = \nabla f \cdot dx = \underbrace{(\nabla f)^T}_{f'(x)} dx \text{ where } dx = \begin{pmatrix} dx_1 \\ dx_2 \\ \vdots \\ dx_n \end{pmatrix}$$

The point here is that we will always define ∇f to have the same shape as x so that df is dot product of dx with the gradient

Example 6
Consider $f(x) = x^T Ax$ where $x \in \mathbb{R}^n$ and A is a square $n \times n$ matrix and thus $f(x) \in \mathbb{R}$. Compute $df, f'(x), \nabla f$

Solution. By definitions we have learnt previously we may write

$$\begin{aligned} df &= f(x+dx) - f(x) \\ &= (x+dx)^T A(x+dx) - x^T Ax \\ &= \cancel{x^T Ax} + dx^T Ax + x^T Adx \underbrace{\cancel{d^T Adx}}_{\text{higher order}} \cancel{-x^T Ax} \\ &= x^T (A + A^T) dx \Rightarrow \nabla f = (A + A^T)x \end{aligned}$$

where we can cancelled out the higher terms and $x^T Ax$ which cancels out in the summation. Finally in the last line we combined the two remaining terms noting that they are scalars so their transpose is the same.

Remark 7. *Compare this if we used the noob way of doing $x^T Ax = \sum_{ij} x_i A_{i,j} x_j$ to find $\frac{\partial f}{\partial x_k}$. So much more trouble ugh*

Example 8
Consider the function $f(x) = Ax$ where A is a constant $m \times n$ matrix. Then applying the distributive law for matrix-vector products we have

$$df = f(x+dx) - f(x) = A(x+dx) - Ax = Adx = f'(x)dx$$

Therefore $f'(x) = A$.Again see how quick this was

Let us consider some derivative rules

- **Sum rule**: given $f(x) = g(x) + h(x)$ we get that

$$df = dg + dh \quad \Rightarrow \quad f'(x)dx = g'(x)dx + h'(x)dx$$

- **product rule** suppose $f(x) = g(x)h(x)$ then

$$\begin{aligned} df &= f(x+dx) - f(x) \\ &= g(x+dx)h(x+dx) - g(x)h(x) \\ &= (g(x) + g'(x)dx)(h(x) + h'(x)dx) - g(x)h(x) \\ &= gh + dgh + gdh + \underbrace{\cancel{dgdh}}_{\text{higher order}} - gh \\ &= dgh + gdh \end{aligned}$$

where the second line follows since $g(x+dx) - g(x) = dg$ so simple rearrangement yields our result $\square$

Now we revisit the 2 examples we encountered but this time applying our sum and product rules directly

Example 9

Let us revisit this same example. Consider the function $f(x) = Ax$ where A is a constant $m \times n$ matrix. Then applying the distributive law for matrix-vector products we have

$$df = d(Ax) = \underbrace{\cancel{dAx}} = 0 + Adx = Adx \Rightarrow f'(x) = A$$

Notice this time how we applied product rule directly.Notice we have $dA = 0$ since A is a constant with respect to x

Example 10

Let $f(x) = x^T Ax$(a mapping from $\mathbb{R}^n \to \mathbb{R}$) then

$$df = dx^T(Ax) + x^T d(Ax) = \underbrace{dx^T Ax}_{x^T A^T dx} + x^T Adx = x^T(A + A^T)dx = (\nabla f)^T dx$$

where we could group terms since $dx^T Ax$ is a scalar so tranpose same(we have done this before)! In particular these show that

$$(x^T(A + A^T))^T = (A + A^T)x = \nabla f$$

as we had derived before and this also simplifies to $2Ax$ if A is symmetric.

Now let's derive another familiar calculus rule that is relevant in the context of matrix calculus. No surprise

- **Chain rule**: let $f(x) = g(h(x))$ then

$$\begin{aligned} df &= f(x+dx) - f(x) = g(h(x+dx)) - g(h(x)) \\ &= g'(h(x))(dh) = g'(h(x))[h(x+dx) - h(x)] \\ &= g'(h(x))[h'(x)[dx]] \\ &= g'(h(x))h'(x)[dx] \end{aligned}$$

where $g'(h(x))h'(x)$ is a composition of g' and h' as matrices. So we see that $f'(x) = g'(h(x))h'(x)$ is a product composition of jacobians $g'h'$

To make sense of this "compositional product" consider the following examples

Example 11

Let $x \in \mathbb{R}^n, h(x) \in \mathbb{R}^p$ and $g(h(x)) \in \mathbb{R}^m$ then $f(x) = g(h(x))$ is a mapping from $\mathbb{R}^n \to \mathbb{R}^m$. The chainn rule then states that $f'(x) = g'(h(x))h'(x)$. Note that the order of multiplication matters.Simple dimensional analysis will show that g' is an $m \times p$ matrix while h' is a $p \times n$ matrix.

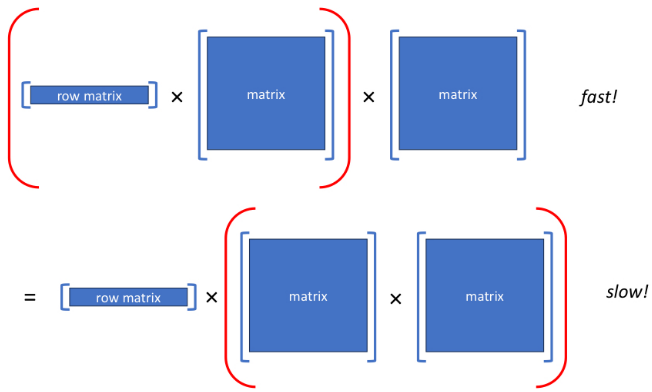

Figure 1: Importance of getting the order of matrix multiplication right: see that the 1st one is $O(n^2)$ but the latter is $O(n^3)$

Let see some exaples of matrix valued functions

Example 12

Let $f(A) = A^3$ where A is a square matrix. Compute df

Solution. here we apply the chain rule one step at a time

$$df = dAA^3 + AdAA + A^2dA = f'(A)[dA]$$

Remark 13. *Notice that this is not equal to* $3A^2$*(unless* dA *and* A *commute)*

Example 14

Let $f(A) = A^{-1}$ where A is a square invertible matrix. Compute $df = d(A^{-1})$

Solution. Here we use a slight trick. Notice that $AA^{-1} = I$. Thus we can compute the differential using the product rule knowing that $dI = 0$ so

$$d(AA^{-1}) = dAA^{-1} + Ad(A^{-1}) = d(I) = 0 \Rightarrow d(A^{-1}) = -A^{-1}dAA^{-1}$$

3 Jacobians of Matrix Functions

Definition 15

The **vectorization** $\mathrm{vec}A \in \mathbb{R}^{mn}$ of any $m \times n$ matrix $A \in \mathbb{R}^{m\times n}$ is defined by simply stacking the columns of A from left to right into a column vector $\mathrm{vec}A$. That is if we denote the n columns of A by m-component vectors $\vec{a}_1, \vec{a}_2, \ldots \in \mathbb{R}^m$ then

$$\mathrm{vec}A = \mathrm{vec}\underbrace{(\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n)}_{A\in\mathbb{R}^{m\times n}} = \begin{pmatrix} \vec{a}_1 \\ \vec{a}_2 \\ \vdots \\ \vec{a}_n \end{pmatrix} \in \mathbb{R}^{mn}$$

is an mn component column vector containing all the entries of A

Example 16

For a 2×2 matrix

$$A = \begin{pmatrix} p & r \\ q & s \end{pmatrix}$$

the matrix square function is

$$f(A) = A^2 = \begin{pmatrix} p & r \\ q & s \end{pmatrix}\begin{pmatrix} p & r \\ q & s \end{pmatrix} = \begin{pmatrix} p^2+qr & pr+rs \\ pq+qs & qr+s^2 \end{pmatrix}$$

Now we want to write f in a "vectorized" form like so

$$\tilde{f}(\begin{pmatrix} p \\ q \\ r \\ s \end{pmatrix}) = \begin{pmatrix} p^2+qr \\ pq+qs \\ pr+rs \\ qr+s^2 \end{pmatrix}$$

To do we may use the operation "vec" like so

$$\mathrm{vec}A = \mathrm{vec}\begin{pmatrix} p & r \\ q & s \end{pmatrix} = \begin{pmatrix} p \\ q \\ r \\ s \end{pmatrix}$$

Recall that vec stacks columns now. Now consider

$$\mathrm{vec}f(A) = \mathrm{vec}\begin{pmatrix} p^2+qr & pr+rs \\ pq+qs & qr+s^2 \end{pmatrix} = \begin{pmatrix} p^2+qr \\ pq+qs \\ pr+rs \\ qr+s^2 \end{pmatrix}$$

Therefore we observe the following relations

$$\tilde{f}(\mathrm{vec}A) = \mathrm{vec}f(A) = \mathrm{vec}(A^2)$$

We basically vectorized the input and output as tasked.

Definition 17

If A is an $m \times n$ matrix with entries $a_{i,j}$ and B is a $p \times q$ matrix then their **Kronecker product** $A \otimes B$ is defined by

$$A = \begin{pmatrix} a_{1,1} & \dots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \dots & a_{m,n} \end{pmatrix} \Rightarrow \underbrace{A}_{m\times n} \otimes \underbrace{B}_{p\times q} = \underbrace{\begin{pmatrix} a_{1,1}B & \dots & a_{1,n}B \\ \vdots & \ddots & \vdots \\ a_{m,1}B & \dots & a_{m,n}B \end{pmatrix}}_{mp\times nq}$$

so that $A \otimes B$ is an $mp \times nq$ matrix formed by "pasting in" in a copy of B mutliplying every element of A

Example 18

Order of multiplication matters

$$A \otimes B = \begin{pmatrix} pB & rB \\ qB & sB \end{pmatrix} = \begin{pmatrix} p\mathbf{a} & p\mathbf{c} & ra & rc \\ p\mathbf{b} & p\mathbf{d} & rb & rd \\ qa & qc & sa & sc \\ qb & qd & sb & sd \end{pmatrix} \neq B \otimes A = \begin{pmatrix} aA & cA \\ bA & dA \end{pmatrix} = \begin{pmatrix} \mathbf{a}p & ar & \mathbf{c}p & cr \\ aq & as & cq & cs \\ \mathbf{b}p & br & \mathbf{d}p & dr \\ bq & bs & dq & ds \end{pmatrix}$$

Problem 19

From the definition of the Kronecker product derive the following identities

1. $(A \otimes B)^T = A^T \otimes B^T$
2. $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$
3. $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ (follows from property 2)
4. $A \otimes B$ is orthgonal if A and B are orthogonal (from properties 1 and 3)
5. $\det(A \otimes B) = \det(A)^m \det(B)^n$ where $A \in \mathbb{R}^{n,n}$ and $B \in \mathbb{R}^{m,m}$
6. $\operatorname{tr}(A \otimes B) = (\operatorname{tr} A)(\operatorname{tr} B)$
7. Given eigenvector/values $Au = \lambda u$ and $Bv = \mu v$ of A and B then $\lambda\mu$ is an eigenvalue of $A \otimes B$ with eigenvector $u \otimes v$

Proof. Most of them can be proven by dimensional analysis.

1. consider $(m \times n \otimes p \times q)^T = (mp \times nq)^T = (nq \times mp) = (n \times m) \otimes (q \times m) = A^T \otimes B^T$
2. consider $(A \otimes B)(C \otimes D) = (mp \times nq)(nq \times mp) = (mp \times mp)$ where $A = (m \times n)$, $C = (n \times m)$,$B = (p \times q)$ and $D = (q \times p)$.So the only other way to get that is $AC \otimes BD$ as desired
3. (3) follows from (2). Consider $(A \otimes B)(A^{-1} \otimes B^{-1}) = (AA^{-1} \otimes BB^{-1}) = I \otimes I = (nn \otimes mm) = (nm \times nm)$ - identity matrix. Then we also know that

$$(A^{-1} \otimes B^{-1}) = (A \otimes B)^{-1}(I \otimes I) = (A \otimes B)^{-1}$$

as desired

4. (4) is obvious when you consider (1) and (3) and that orthogonal means inverse is the same as tranpose.
5. Consider $\det(A \otimes B) = \det(A \otimes I)(I \otimes B) = \det(nn \otimes mm)(nn \otimes mm) = \det(nm \times nm)$ but we also know by properties of determinant that

$$\det(A \otimes I)(I \otimes B) = \det(A \otimes I)\det(I \otimes B) = \det(A)^m \det(B)^n$$

because we multiplying block diagonals of A and B
6. Consider $\operatorname{tr}(A \otimes B) = \sum_k \sum_j A_{kk} B_{jj} = \operatorname{tr} A \operatorname{tr} B$ as desired.(think about how their matrix product looks like to reason this out)
7. Finally recalling that $\det(A \otimes B)$ is the product of eigenvalues while $\operatorname{tr}(A \otimes B)$ is sum of eigenvalues. Their identities very clearly reflect that

$$(\lambda_i \nu_j), \quad i = 1, \ldots, n \quad j = 1, \ldots, m$$

since the new sum reflects $(\sum \lambda_i)(\sum \lambda_j) = \sum_{ij} \lambda_i \mu_j$. So does their product $\prod_i \lambda_i \prod \mu_j$.

□

Proposition 20

Given (compatibly sized) matrices A, B, C we have

$$(A \otimes B)\operatorname{vec}(C) = \operatorname{vec}(BCA^T)$$

We can thus view $A \otimes B$ as a vectorized equivalent of the linear operation $C \mapsto BCA^T$

Proof. First consider the case where either A or B is an identity matrix I(of the appropriate size). To start with suppose that $A = I$ so that $BCA^T = BC$. Now let $\vec{c}_1, \vec{c}_2, \ldots$ denote the columns of C then recall that BC simpy multiplies B on the left with each of the columns of C that is

$$BC = B\,(\vec{c}_1 \;\; \vec{c}_2 \;\; \ldots) = (B\vec{c}_1 \;\; B\vec{c}_2 \;\; \ldots) \Rightarrow \operatorname{vec}(BC) = \begin{pmatrix} B\vec{c}_1 \\ B\vec{c}_2 \\ \vdots \end{pmatrix}$$

To get $\operatorname{vec}(BC)$vector as something multpyling vec$C$ we can can guess that we have

$$\operatorname{vec}(BC) = \begin{pmatrix} B\vec{c}_1 \\ B\vec{c}_2 \\ \vdots \end{pmatrix} = \underbrace{\begin{pmatrix} B & & \\ & B & \\ & & \ddots \end{pmatrix}}_{I \otimes B} \underbrace{\begin{pmatrix} \vec{c}_1 \\ \vec{c}_2 \\ \vdots \end{pmatrix}}_{\text{vec } C}$$

so immediately we have verified that

$$(I \otimes B)\operatorname{vec} C = \operatorname{vec}(BC)$$

Now we need to somehow add in the A^T term. To that end we simplify to the case where $B = I$ in which case

$BCA^T = CA^T$. To vectorize this again we need to look at the columns. COnsider that

$$\text{vec}(CA^T) = \begin{pmatrix} \sum_j a_{1j}\vec{c}_j \\ \sum_j a_{2j}\vec{c}_j \\ \vdots \end{pmatrix} = \underbrace{\begin{pmatrix} a_{11}I & a_{11}I & \dots \\ a_{11}I & a_{11}I & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}}_{A\otimes I} \underbrace{\begin{pmatrix} \vec{c}_1 \\ \vec{c}_2 \\ \vdots \end{pmatrix}}_{\text{vec } C}$$

and so similarly we have derived

$$(A \otimes I)\text{vec}C = \text{vec}(CA^T)$$

Our proof essentially relied on the concept that changes by A and B are independent of each other. So we isolate the effects and slightly perturb to figure out what goes on.Explicitly

$$(I \otimes I)\text{vec C}$$

when I is replaced with B it appears on $\text{vec}(BC)$ on the left while if the other I is replaced it appears as a transpose on the right of C. Therefore we have the proposition when both Is are replaced simultaneously as desired.

Remark 21. *This is quite a common and effective method used in proofs. Great application!*

Example 22

Let us use 20 to calculate the vectorized jacobian of $f(A) = A^2$. Consider that

$$df = f'(A)[dA] \quad \Rightarrow \quad \text{vec}(df) = \text{vec}(f'(A)[dA]) = \tilde{f}'(\text{vec}(A))[\text{vec}(dA)]$$

now consider

$$\begin{aligned} \text{vec}(df) &= \text{vec}(AdA + dAA) = \text{vec}(AdA) + \text{vec}(dAA) \\ &= (I \otimes A)\text{vec}(dA) + (A^T \otimes I)\text{vec}(dA) \\ &= (I \otimes A + A^T \otimes I)[\text{vec}(dA)] \end{aligned}$$

on comparison with above we immediately see that

$$(I \otimes A + A^T \otimes I) = \tilde{f}'(\text{vec}(A))$$

and so because our example is only 2×2 we can explicitly calculate to obtain

$$\underbrace{\begin{pmatrix} 1 & \\ & 1 \end{pmatrix}}_{I} \otimes \underbrace{\begin{pmatrix} p & r \\ q & s \end{pmatrix}}_{A} \otimes \underbrace{\begin{pmatrix} p & q \\ r & 1s \end{pmatrix}}_{A^T} + \underbrace{\begin{pmatrix} 1 & \\ & 1 \end{pmatrix}}_{I} = \begin{pmatrix} 2p & r & q & 0 \\ q & p+s & 0 & q \\ r & 0 & p+s & r \\ 0 & r & q & 2s \end{pmatrix} = \tilde{f}'$$

Example 23

For the matrix cube function A^3 where A is an $m \times m$ square matrix compute the $m^2 \times m^2$ jacobian of the vectorized function $\text{vec}(A^3)$. Using the same way as above we obtain

$$(A^3)'[dA] = dAA^2 + AdAA + A^2dA$$

and so

$$\text{vec}(dAA^2 + AdAA + A^2dA) = ((A^2)^T \otimes I + A^T \otimes A + I \otimes A^2)\text{vec}(dX)$$

4 Finite Difference Approximations

5 Derivatives in General vector spaces

Definition 24

The **Frobenius inner product** of two $m \times n$ matrices A and B is

$$\langle A, B \rangle_F = \sum_{ij} A_{ij}B_{ij} = \text{vec}(A)^T\text{vec}(B) = \text{tr}(A^TB)$$

The above is basically pointwise multiplication.Given this inner product we also have the corresponding **Frobenius norm**

$$||A||_F = \sqrt{\langle A, A \rangle_F} = \sqrt{\text{tr}(A^TA)} = ||\text{vec}A|| = \sqrt{\sum_{i,j} |A_{i,j}|^2}$$

For the rest of the notes we will assume this to be our default matrix inner product and hence drop the F subscript

Example 25

Consider the function

$$f(A) = ||A||_F = \sqrt{\text{tr}(A^TA)}$$

What is df?

Solution. Firstly by familiar scalar differentiation chain(2) and power rules we have that

$$df = \frac{1}{2\sqrt{\text{tr}(A^TA)}}d(\text{tr}\, A^TA)$$

Then note that by learity of trace we have

$$d(\text{tr}\, B) = \text{tr}(B + dB) - \text{tr}(B) = \text{tr}(B) + \text{tr}(dB) - \text{tr}(B) = \text{tr}(dB)$$

that is we may interchange the differential with the trace function. Hence

$$
\begin{aligned}
df &= \frac{1}{2\,||A||_F}\operatorname{tr}(d(A^T A)) \\
&= \frac{1}{2\,||A||_F}\operatorname{tr}(dA^T A + A^T dA) \\
&= \frac{1}{||A||_F}(\operatorname{tr}(dA^T A) + \operatorname{tr}(A^T dA)) \\
&= \frac{1}{||A||_F}\operatorname{tr}(A^T dA) = \langle \frac{A}{||A||_F}, dA\rangle
\end{aligned}
$$

where in the penultimate step we used the fact that $\operatorname{tr} B = \operatorname{tr} B^T$.Hence recall since $\nabla f \cdot dA = df$ we immediately conclude

$$\nabla f = \nabla\,||A||_F = \frac{A}{||A||_F}$$

Example 26

Fix some constant $x \in \mathbb{R}^m, y \in \mathbb{R}^n$ and consider the function $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ given by

$$f(A) = x^T Ay$$

What is ∇f?

Solution. We have that

$$
\begin{aligned}
df &= x^T dAy \\
&= \operatorname{tr}(x^T dAy) \quad \text{(since trace of real number is itself)} \\
&= \operatorname{tr}(yx^T dA) \quad (\text{since } \operatorname{tr} B = \operatorname{tr} B^T) \\
&= \langle xy^T, dA\rangle
\end{aligned}
$$

and immediately we similarly see that

$$xy^T = \nabla f$$

Fact 27

More generally for any scalar valued function $f(A)$ from the definition of the Frobenius inner product it follows that

$$df = f(A + dA) - f(A) = \langle \nabla f, dA\rangle = \sum_{i,j}(\nabla f)_{i,j} dA_{i,j}$$

and hence the components of the gradient are exactly the elementwise derivatives

$$(\nabla f)_{i,j} = \frac{\partial f}{\partial A_{i,j}}$$

6 Nonlinear Root finding, Optimization and Adjoint Differentiation

Problem 28

Suppose that $A(p)$ takes a vector $p \in \mathbb{R}^{n-1}$ and returns the $n \times n$ triadiagonal real symmetric matrix

$$A(p) = \begin{pmatrix} a_1 & p_1 & & & & & \\ p_1 & a_2 & p_2 & & & & \\ & p_2 & a_3 & p_3 & & & \\ & & p_3 & a_4 & p_4 & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & & a_{n-1} & p_{n-1} \\ & & & & & p_{n-1} & a_n \end{pmatrix}$$

where $a \in \mathbb{R}^n$ is some constant vector. Now define a scalar valued function $f(p)$ by

$$g(p) = (c^T A(p) b)^2$$

for some constant vectors $b, c in \mathbb{R}^n$. p, a are chosen such that A is invertible.

1. write down a formula for computing $\frac{\partial g}{\partial p_1}$
2. coding problem
3. coding problem

Solution. Consider

1. From the chain rule and the formula for the differential of a matrix inverse we have

$$dg = -2(c^T A^{-1} b) c^T A^{-1} dA A^{-1} b$$

but notice that $c^T A^{-1} b$ is a scalar

7 Derivtive of Matrix Determinant and Inverse(7)

Theorem 29

Given A is a square matrix we have

$$\nabla(\det A) = \text{cofactor}(A) = (\det A)(A^{-1})^T = \text{adj}(A^T) = \text{adj}(A)^T$$

Furthermore we have

$$d(\det A) = \text{tr}(\det(A) A^{-1} dA) = \text{tr}(\text{adj}(A) dA) = \text{tr}(\text{cofactor}(A)^T dA)$$

Proof. For the 1st line we have done it before in `CS229 Stanford Intro to ML`. Simply recall that

$$\frac{\partial \det A}{\partial A_{i,j}} = C_{i,j}$$

where $C_{i,j}$ as an element of the cofactor matrix C. Therefore we have that

$$\nabla(\det A) = C$$

As for the second line we may use linearization near the identity like so

$$\det(I + dA) - 1 = (1 + \lambda_1)(1 + \lambda_2) - 1 = \cancel{\det dA} + \operatorname{tr}(dA)$$

where λ_1, λ_2 are the eigenvalues of dA. Recall why this makes sense from our knowledge of characteristic polynomials for $n = 2$.Note that we dropped dA because the elements of dA are infinitesimally small and $\det dA$ is a product of its elements(in particular the eigenvalues). Therefore we have

$$\begin{aligned} d(\det(A)) = \det(A + A(A^{-1}dA)) - \det(A) &= \det(A)\det(I + A^{-1}dA - 1) \\ &= \det(A)\operatorname{tr}(A^{-1}dA) = \operatorname{tr}(\det(A)A^{-1}dA) \\ &= \operatorname{tr}(\operatorname{adj}(A)dA) \end{aligned}$$

Example 30

Let find $d(\det(xI - A))$

Solution. While this may be solved by writing this out as a product of eigenvalues then doing

$$\frac{d}{dx}\prod_i (x - \lambda_i) = \sum_i \prod_{j \neq i}(x - \lambda_j) = \prod_i (x - \lambda_i)\left\{\sum_i (x - \lambda_i)^{-1}\right\}$$

where the second line is essentially produce rule, that is for each i, differentiate $(x - \lambda_i)$ then hold the other brackets constant. Then sum up all the terms

$$\dots(x - \lambda_{i-1})\underbrace{\frac{d}{dx}(x - \lambda_{i-1})}_{=1}(x - \lambda_{i-1})\dots$$

Notice how the last term works. Essentially it sums up all products of brackets where each summand has bracket i excluded. However we could use our formula above to simplify this. Consider

$$d(\det(xI - A)) = \det(xI - A)\operatorname{tr}((xI - A)^{-1}d(xI - A)) = \det(xI - A)\operatorname{tr}(xI - A)^{-1}dx$$

Example 31

For another application consider that we may do

$$d(\log(\det(A))) = \frac{d(\det A)}{\det A} = \det(A^{-1})d(\det(A)) = \operatorname{tr}(A^{-1}dA)$$

8 Second Derivatives, Bilinear Maps and Hessian Matrices(12)

Definition 32

The **hessian** of f has entries

$$H_{i,j} = \frac{\partial^2 f}{\partial x_j \partial x_i} = \frac{\partial^2 f}{\partial x_i \partial x_j} = H_{j,i}$$

which is symmmetric assuming $f \in C^2$. So $H = (\nabla f)'$

See that

$$d(\nabla f) = (\nabla f)' dx = H dx$$

Fact 33

We first consider the quadratic approximation(2nd order taylor series)

$$f(x + \delta x) = f(x) + (\nabla f)^T \delta x + \frac{1}{2} \delta x^T H \delta x + o(||\delta x||^2)$$

To see why this makes sense consider dx(a column vector so each i in dx_i is a row so must take transpose to have valid matrix product). Considering individual values(that is $dx_i, H_{i,j}, dx_j$ which are all scalars)

$$(dx_i)^T H_{i,j} dx_j \in \mathbb{R}$$

which will be familiar single dimensional taylor expansion. This is how the 2nd order taylor approximation makes sense. Also note that clearly both

$$(\nabla f)^T \delta x \in \mathbb{R} \quad \text{and} \quad \frac{1}{2} \delta x^T H \delta x \in \mathbb{R}$$

We would like to express this approximation as

$$f(x + \delta x) = f(x) + f'(x)[\delta x] + \frac{1}{2} f''(x)[\delta x'][\delta x]$$

as before we know that we have $f'(x) = (\nabla f)^T$ and from here it seems that we are implying

$$f''(x)[dx'][dx] = dx'^T H dx \in \mathbb{R}$$

But how does this representation make sense? First consider

Proposition 34

$f''(x)[dx'][dx]$ is indeed a symmetric bilinear map $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ as reflected by the hessian.

Proof. Consider

$$
\begin{aligned}
\underbrace{f''(x)[dx'}_{df'(x)}, dx] &= f'(x+dx')[dx] - f'(x)[dx] \\
&= (f(x+dx'+dx)) + f(x+dx') - (f(x+dx) - f(x)) \\
&= f(x+dx+dx') + f(x) - f(x+dx) - f(x+dx') \\
&= (f(x+dx+dx') - f(x+dx)) - (f(x+dx') - f(x)) \\
&= f'(x+dx)[dx'] - f'(x)[dx'] \\
&= f''(x)[dx, dx']
\end{aligned}
$$

□

So our representation makes sense.

Example 35

Let $f(x) = x^T Ax$ for $x \in \mathbb{R}^n$ and A an $n \times n$ matrix. As above, f

Solution. recall as computed in an earlier example we have

$$f' = (\nabla f)^T = x^T(A + A^T)$$

This implies that $\nabla f = (A + A^T)x$ therefore

$$H = (\nabla f)' = (A + A^T)$$

Fact 36

Note that we have a special relationship here in this case. Consider

$$
\begin{aligned}
f(x) = x^T Ax &= (x^T Ax)^T \quad \text{since scalar=scalar}^T \\
&= x^T A^T x \\
&= \frac{1}{2}(x^T Ax + x^T A^T x) = \frac{1}{2}x^T(A + A^T)x \\
&= \frac{1}{2}x^T Hx = \frac{1}{2}f''[x, x]
\end{aligned}
$$