MATRIX CALCULUS NOTES

JAMES BURKE

1. Basics. Let $\mathbb{X}$ and $\mathbb{Y}$ be normed linear spaces with norms $\|\cdot\|_x$ and $\|\cdot\|_y$, respectively. A linear transformation (or mapping) from $\mathbb{X}$ to $\mathbb{Y}$ is any mapping $\mathcal{L} : \mathbb{X} \to \mathbb{Y}$ such that

$$\mathcal{L}(\alpha x + \beta z) = \alpha\mathcal{L}(x) + \beta\mathcal{L}(z) \quad \forall\, x, z \in \mathbb{X} \text{ and } \alpha, \beta \in \mathbb{R}.$$

Let $\mathbb{L}[\mathbb{X}, \mathbb{Y}]$ denote the normed space of continuous linear transformations from $\mathbb{X}$ to $\mathbb{Y}$ where the (operator) norm is given by

$$\|\mathcal{T}\| := \sup_{\|x\|_x \le 1} \|\mathcal{T}x\|_y \qquad \forall\, \mathcal{T} \in \mathbb{L}[\mathbb{X}, \mathbb{Y}].$$

The topological dual of the normed linear space $\mathbb{X}$ is

$$\mathbb{X}^* := \mathbb{L}[\mathbb{X}, \mathbb{R}]$$

with the *duality pairing* denoted by

$$\langle \phi,\, x\rangle = \phi(x) \quad \forall\, (\phi, x) \in \mathbb{X}^* \times \mathbb{X}.$$

The duality pairing gives rise to the notion of the *adjoint* of a linear operator: given $\mathcal{T} \in \mathbb{L}[\mathbb{X}, \mathbb{Y}]$, the adjoint of $\mathcal{T}$ is the unique linear operator $\mathcal{T}^* \in \mathbb{L}[\mathbb{Y}^*, \mathbb{X}^*]$ defined by

$$\langle y^*,\, \mathcal{T}(x)\rangle = \langle \mathcal{T}^*(y^*),\, x\rangle \quad \forall\, (y^*, x) \in \mathbb{Y}^* \times \mathbb{X}.$$

A Euclidean space is a finite dimensional inner product space. The space $\mathbb{R}^n$ can be endowed with an infinite variety of inner products, however, every inner product on $\mathbb{R}^n$ takes the form $\langle x,\, y\rangle_V = x^TVy$ for some symmetric positive definite matrix $V \in \mathbb{S}^n_{++}$ (= the cone of symmetric $n \times n$ positive definite matrices). On the space $\mathbb{R}^{m\times n}$, the standard inner product is the Frobenius inner product given by $\langle A,\, B\rangle_F = \langle A,\, B\rangle = \operatorname{tr}\left(A^TB\right)$. The duality pairing on a Euclidean space can be show to be equivalent to the inner product.

Let $\mathbb{X}$ and $\mathbb{Y}$ be finite dimensional real vector spaces, and let $\{x^j\}_{j=1}^n$ and $\{y^i\}_{i=1}^m$ be bases for $\mathbb{X}$ and $\mathbb{Y}$, respectively. Given $x = \sum_{j=1}^n a_jx^j \in \mathbb{X}$, the linear mapping

$$x \overset{\kappa}{\mapsto} (a_1, \ldots, a_n)^T$$

is a linear isomorphism between $\mathbb{X}$ and $\mathbb{R}^n$. The mapping κ is called the coordinate mapping from $\mathbb{X}$ to $\mathbb{R}^n$ associated with the basis $\{x^j\}_{j=1}^n$. Let η be the coordinate mapping for $\mathbb{Y}$ to $\mathbb{R}^m$ associated with the basis $\{y^i\}_{i=1}^m$. Given $\mathcal{T} \in \mathbb{L}[\mathbb{X}, \mathbb{Y}]$, there exist uniquely defined $\{t_{ij} \mid i = 1, \ldots, m,\ j = 1, \ldots, n\} \subset \mathbb{R}$ such that

$$\mathcal{T}x^j = \sum_{i=1}^m t_{ij}y^i,\ j = 1, \ldots, n.$$

Consequently, given $x \in \mathbb{X}$ we have $\eta(\mathcal{T}x) = T\kappa(x)$, where $T \in \mathbb{R}^{m\times n}$ is the matrix whose components are (t_{ij}). Using this kind of basis representation for a linear transformation between finite dimensional spaces can be very useful, but it can also obscure the action of a linear transformation and make its representation unnecessarily arduous. Nonetheless, one often simply identifies a linear transformation with its matrix representation associated with an agreed upon preferred pair of bases. All of this formalism and be extended to vector spaces over arbitrary fields. The primary fields of interest to us are the real and complex fields.

We will be interested in linear transformations on $\mathbb{R}^{m\times n}$. Two very useful tools in this context are the *Kronecker* and *Hadamard* products.

Definition 1.1 (The Kronecker and Hadamard Products). *Let $A, C \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{s\times t}$. The Kronecker product of A with B is the $ms \times mt$ matrix given by*

$$A \otimes B := \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix},$$

and the Hadamard product of A and C is the $m \times n$ matrix given by

$$A \odot C := (a_{ij}c_{ij}),$$

i.e., the componentwise product of A and C.

Another useful tool is the *vec* operator: given $A \in \mathbb{R}^{m\times n}$,

$$\mathsf{vec}(A) := \begin{bmatrix} A_{\cdot 1} \\ A_{\cdot 2} \\ \vdots \\ A_{\cdot n} \end{bmatrix},$$

that is, $\mathsf{vec}(A)$ is the mn vector obtained by stacking the columns of A on top of each other. Clearly, $\langle A,\, B\rangle_F = \mathsf{vec}(A)^T\mathsf{vec}(B)$. Note that $\mathsf{vec} \in \mathbb{L}[\mathbb{R}^{m\times n},\, \mathbb{R}^{mn}]$.

Proposition 1.2 (Properites of the Kronecker Product).

1. $A \otimes B \otimes C = (A \otimes B) \otimes C = A \otimes (B \otimes C)$
2. $(A \otimes B)(C \otimes D) = AC \otimes BD$ *when AC and BD exist.*
3. $(A \otimes B)^T = (A^T \otimes B^T)$
4. $\mathrm{tr}\,(A \otimes B) = \mathrm{tr}\,(A)\,\mathrm{tr}\,(B)$
5. $(A^\dagger \otimes B^\dagger) = (A \otimes B)^\dagger$, *where $M^\dagger$ is the Moore-Penrose pseudo inverse of the matrix M.*
6. $\mathrm{rank}\,(A \otimes B) = \mathrm{rank}\,A\,\mathrm{rank}\,B$
7. *For vectors a and b, $\mathsf{vec}(ab^T) = b \otimes a$.*
8. *If AXB is a well defined matrix product, then*

$$\mathsf{vec}(AXB) = (B^T \otimes A)\mathsf{vec}(X).$$

In particular,

$$\mathsf{vec}(AX) = (I \otimes A)\mathsf{vec}(X) \quad \textit{and} \quad \mathsf{vec}(XB) = (B^T \otimes I)\mathsf{vec}(X),$$

where I is interpreted as the identity matrix of the appropriate dimension.

Examples 1.3 (Linear Operators and Matrix Representations).

1. *Consider the vector space $\mathbb{R}^{n\times n}$ and let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times k}$. We define the linear transformation $\mathcal{T} \in \mathbb{L}[\mathbb{R}^{n\times n}, \mathbb{R}^{m\times k}]$ by $\mathcal{T}(X) = AXB$:*

$$\mathcal{T}(\alpha X + \beta Y) = A(\alpha X + \beta Y)B = \alpha AXB + \beta AYB = \alpha\mathcal{T}(X) + \beta\mathcal{T}(Y).$$

On the linear space $\mathbb{R}^{m\times n}$ the set of matrices

$$\{E_{ij} \mid i = 1, \dots, m,\ j = 1, \dots, n\},$$

where E_{ij} is the matrix having a one in the ij position and zero elsewhere, is the standard unit coordinate basis for $\mathbb{R}^{m\times n}$. Observe that vec *is the coordinate mapping on $\mathbb{R}^{m\times n}$ associated with the standard unit coordinate matrices E_{ij}. Using* vec *one can compute a matrix representation for $\mathcal{T}$ with respect to the standard unit coordinate bases. This is done by applying the vec operator to the expression $\mathcal{T}(X) = AXB$ and then using properties of the the Kronecker product to get a matrix formula:*

$$\mathsf{vec}(T(X)) = \mathsf{vec}(AXB) = (B^T \otimes A)\mathsf{vec}(X).$$

That is, the matrix representation of T in the unit coordinate bases is $T = B^T \otimes A$.

2. *Again consider the vector space $\mathbb{R}^{n\times n}$, but now let $A, B \in \mathbb{R}^{n\times n}$. We define the linear transformation $\mathcal{T} \in \mathbb{L}[\mathbb{R}^{n\times n}, \mathbb{R}^{n\times n}]$ by $\mathcal{T}(X) = AX + XB$. Again, we can obtain a matrix representation for this operator in the unit coordinate basis by using* vec *and applying properties of the Kronecker product:*

$$\begin{aligned}\mathsf{vec}(T(X)) &= \mathsf{vec}(AX) + \mathsf{vec}(XB) = (I \otimes A)\mathsf{vec}(X) + (B^T \otimes I)\mathsf{vec}(X)\\ &= [(I_n \otimes A) + (B^T \otimes I_n)]\mathsf{vec}(X).\end{aligned}$$

That is, the matrix representation of T in the unit coordinate bases is $T = (I_n \otimes A) + (B^T \otimes I_n)$.

3. *Let $\mathbb{P}^n[t]$ be the linear space of real polynomials of degree n or less in the variable t. Given $\lambda \in \mathbb{R}$, the polynomials $\mathsf{e}_{(k,\lambda)}(t) := (t-\lambda)^k,\ k = 0, 1, \dots, n$ are known to form a basis for $\mathbb{P}^n$. Consider the linear transformation $D \in \mathbb{L}[\mathbb{P}^n, \mathbb{P}^{n-1}]$ given by $D(p) = p'$, where p' is the derivative of p with respect to t. Give the matrix representation of D in the bases $\mathsf{e}_{(k,0)}$. What about the bases $\mathsf{e}_{(k,\lambda)}$?*

2. Derivatives.

First the definitions.

DEFINITION 2.1. *Let $F : \mathcal{O} \to \mathbb{Y}$, where $\mathcal{O} \subset \mathbb{X}$ is open.*

1. *We say that F is Gateau differentiable at $x \in \mathcal{O}$ if there exists $J \in \mathbb{L}[\mathbb{X}, \mathbb{Y}]$ such that*

$$\lim t \to 0 \frac{F(x+td) - F(x) - tJd}{t} = 0 \quad \forall\, d \in \mathbb{X},$$

where we call J the Gateau derivative of F at x.

2. *We say that F is Frechét differentiable at $x \in \mathcal{O}$ if there exists $J \in \mathbb{L}[\mathbb{X}, \mathbb{Y}]$ such that*

$$\lim y \to x \frac{\|F(y) - F(x) - J(y-x)\|}{\|y - x\|} = 0,$$

where we call J the Frechét derivative of F at x. The Frechét derivative of F at x is denoted by $F'(x)$.

The notions of Gateaux and Frechét differentiability coincide in finite dimensions. This equivalence is often useful in computing derivatives since it reduces the computation over the scalars.

For mappings $\psi : \mathbb{R}^n \to \mathbb{R}$, we call the embedding of the matrix representation for $\psi'(x)$ in the standard unit coordinate bases into $\mathbb{R}^n$ the gradient of ψ at x and write $\nabla\psi(x)$. When the partial derivatives of ψ exist and are continuous, $\nabla\psi(x)$ is just the vector of these partial derivatives. More generally, since $\psi'(x) \in \mathbb{L}[\mathbb{R}^n, \mathbb{R}] = (\mathbb{R}^n)^*$ and $\mathbb{R}^n$ is an inner product space, we can identify $\psi'(x)$ with an element of $\mathbb{R}^n$. We call this element the gradient of ψ at x and write it as $\nabla\psi(x)$.

For mappings $F : \mathbb{R}^n \to \mathbb{R}^m$, we can decompose F into its coordinate functions

$$F(x) = \begin{bmatrix} F_1(x) \\ F_2(x) \\ \vdots \\ F_m(x) \end{bmatrix},$$

where $F_i : \mathbb{R}^n \to \mathbb{R},\ i = 1, \ldots, m$. Consequently, the matrix representation of $F'(x)$ in the standard unit coordinate vectors is given by

$$\nabla F(x) = \begin{bmatrix} \nabla F_1(x)^T \\ \nabla F_2(x)^T \\ \vdots \\ \nabla F_m(x)^T \end{bmatrix},$$

which is called the Jacobian of F at x (some authors call $\nabla F(x)^T$ the Jacobian of F at x in order the preserve the consistency in the use of the transpose with ∇).

To compute Frechét derivatives, it is extremely useful to use what is called little-o notation:

$$F(y) = F(x) + F'(x)(y - x) + o(\|y - x\|),$$

where $o(t)$ is an element of the class of functions satisfying $\lim_{t\to 0} \frac{o(t)}{t} = 0$. Observe that if $\mathcal{T} \in \mathbb{L}[\mathbb{X}, \mathbb{Y}]$, then, for all $x \in \mathbb{X}$ and $y \in \mathbb{Y}$, we have

$$\mathcal{T}y = \mathcal{T}(x + (y - x)) = \mathcal{T}x + \mathcal{T}(y - x),$$

and so $\mathcal{T}'(x) = \mathcal{T}$ for all $x \in \mathbb{X}$. That is, for any pair of topological vector spaces $\mathbb{X}$ and $\mathbb{Y}$, the derivative of an element of $\mathbb{L}[\mathbb{X}, \mathbb{Y}]$ is itself.

Example 2.2.

1. *Let $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and define the function $f : \mathbb{R}^n \to \mathbb{R}$ by $f(x) := \frac{1}{2}\|Ax - b\|_2^2$. Then*

$$\begin{aligned} f(x + \Delta x) &= \tfrac{1}{2}\|A(x + \Delta x) - b\|_2^2 \\ &= \tfrac{1}{2}\langle (Ax - b) + A\Delta x,\ (Ax - b) + A\Delta x\rangle \\ &= \tfrac{1}{2}\langle Ax - b,\ Ax - b\rangle + \langle Ax - b,\ A\Delta x\rangle + \tfrac{1}{2}\langle A\Delta x,\ A\Delta x\rangle \\ &= f(x) + \left\langle A^T(Ax - b),\ \Delta x\right\rangle + o(\|\Delta x\|_2^2). \end{aligned}$$

Hence $f'(x)u = (Ax - b)^T Au$ for all $u \in \mathbb{R}^n$. Consequently, in the standard unit coordinate bases, the matrix representation for $f'(x)$ is $(Ax - b)^T A$, so $\nabla f(x) = A^T(Ax - b)$. Note that this computation is elementary and we did not need to compute the individual partial derivatives first.

2. *Consider the function $h(x) := \frac{1}{2}\|F(x)\|_2^2$, where $F : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable. The same technique as in the example above can be applied to show that $\nabla h(x) = \nabla F(x)^T F(x)$ without the need to compute the individual partial derivatives first: in the computation to follow (1) the fact that the sum, product, inner product, or norm of two or more little-o functions is another little-o function and (2) the computation in the first example is also used;*

$$\begin{aligned}
h(x+\Delta x) &:= \tfrac{1}{2}\left\|F(x) + \nabla F(x)\Delta x + o(\|\Delta x\|_2^2)\right\|_2 \\
&= \tfrac{1}{2}\left\langle (F(x) + \nabla F(x)\Delta x) + o(\|\Delta x\|_2^2),\, (F(x) + \nabla F(x)\Delta x) + o(\|\Delta x\|_2^2)\right\rangle \\
&= \tfrac{1}{2}\left\langle F(x) + \nabla F(x)\Delta x,\, F(x) + \nabla F(x)\Delta x\right\rangle + \\
&\qquad \left\langle F(x) + \nabla F(x)\Delta x,\, o(\|\Delta x\|_2^2)\right\rangle + \tfrac{1}{2}\left\langle o(\|\Delta x\|_2^2),\, o(\|\Delta x\|_2^2)\right\rangle \\
&= \tfrac{1}{2}\|F(x) + \nabla F(x)\Delta x\|_2^2 + o(\|\Delta x\|_2^2) \\
&= h(x) + \langle F(x),\, \nabla F(x)\Delta x\rangle + \tfrac{1}{2}\|\nabla F(x)\Delta x\|_2^2 + o(\|\Delta x\|_2^2) \\
&= h(x) + \left\langle \nabla F(x)^T F(x),\, \Delta\right\rangle + o(\|\Delta x\|_2^2).
\end{aligned}$$

Note that this is simply the chain rule applied to h.

Theorem 2.3 (Chain Rule). *Let $\mathbb{X}$, $\mathbb{Y}$ and $\mathbb{Z}$ be normed linear spaces and consider the mappings $H : \mathbb{Y} \to \mathbb{Z}$ and $F : \mathbb{X} \to \mathbb{Y}$. Define the composition of F with H to be the mapping $H \circ F : \mathbb{X} \to \mathbb{Z}$ given by $(H \circ F)(x) := H(F(x))$. If $x \in \mathbb{X}$ is such that F' exists and is continuous at x and H' exists at $F(x)$, then $(H \circ F)'$ exists at x and is given by*

$$(H \circ F)'(x) = H'(F(x)) \circ F'(x).$$

Remark 2.4. *When $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$ and $Z = \mathbb{R}$, then the gradient of $H \circ F$ at x is given by*

$$\nabla(H \circ F)(x) = \nabla F(x)^T \nabla H(F(x)).$$

This is validated in Example 2.2.

Example 2.5 (The Derivative of Linear Operators and the Chain Rule).

1. *Consider the linear transformation $\mathcal{T} \in \mathbb{L}[\mathbb{R}^{n\times n}, \mathbb{R}^{n\times n}]$ given by $\mathcal{T}(X) = AX + XB$ in Example 1.3, and let $F : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is given by $F(x) :=$ $\mathrm{diag}\,(x)$, where the linear transformation $\mathrm{diag}\,(\cdot) \in \mathbb{L}[\mathbb{R}^n, \mathbb{R}^{n\times n}]$ maps x to the $n \times n$ matrix whose diagonal is x. Then*

$$(\mathcal{T} \circ \mathrm{diag}\,(\cdot))'(x)(d) = A\mathrm{diag}\,(d) + \mathrm{diag}\,(d)\,B$$

for all $x \in \mathbb{R}^n$.

2. *Consider the linear operator $\mathcal{T} \in \mathbb{L}[\mathbb{R}^{n\times n}, \mathbb{R}^{m\times k}]$ given by $\mathcal{T}(X) = AXB$ discussed in Example 1.3. We have $\mathcal{T}'(X)(D) = ADB$ for all X. Consequently, if $\mathbb{Y}$ is a normed linear space and $F : \mathbb{Y} \to \mathbb{R}^{n\times n}$ is (Gateaux) Frechét differentiable at $y \in \mathbb{Y}$, then the derivative of the mapping $(\mathcal{T} \circ F)$ is given by $(\mathcal{T} \circ F)'(y)(d) = A(F'(y)d)B$. For example, if $F : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is given by $F(x) := \exp(-\frac{1}{2}\|x - x^0\|_2^2)I_n$ for some $x^0 \in \mathbb{R}^n$, then*

$$\nabla F(x) = I_n \exp(-\tfrac{1}{2}\|x - x^0\|_2^2)(x - x^0)^T \in \mathbb{L}[\mathbb{R}^n, \mathbb{R}^{n\times n}],$$

and so

$$\begin{aligned}(\mathcal{T}\circ F)'(x)(d) &= A(F'(x)d)B\\ &= A(\exp(-\tfrac{1}{2}\left\|x-x^0\right\|_2^2)\left\langle x-x^0,\, d\right\rangle I_n)B\\ &= \exp(-\tfrac{1}{2}\left\|x-x^0\right\|_2^2)(x-x^0)^T d\, AB.\end{aligned}$$

There are any forms of the *product rule* depending on the structure of the product. For example, the product of two scalar-valued function, or to matrix-valued functions, or one scalar-valued and one matrix-valued function. But all of these product rules have the same general form that can easily be derived from the definition of the derivative. We give one such product rule for two matrix-valued functions.

THEOREM 2.6 (Product Rule for Matrix-valued Functions). *Let $\mathbb{Z}$ be a normed linear space and let $F_i : \mathbb{Z} \to \mathbb{R}^{n\times n},\ i=1,2$, be such that both are differentiable at $z\in\mathbb{Z}$. Then the mapping $G : \mathbb{Z}\to\mathbb{R}^{n\times n}$ given by $G(z) := F_1(z)F_2(z)$ is differentiable at z with*

$$G'(z)(D) = [F_1'(z)(D)]F_2(z) + F_1(z)[F_2'(z)(D)].$$

Proof.

$$\begin{aligned}&F_1(z+\Delta z)F_2(z+\Delta z)\\ &= (F_1(z)+F_1'(z)(\Delta z)+o_1(\|\Delta z\|))(F_2(z)+F_2'(z)(\Delta z)+o_2(\|\Delta z\|))\\ &= F_1(z)F_2(z)+F_1'(z)(\Delta z)F_2(z)+F_1(z)F_2'(z)(\Delta z)+o(\|\Delta z\|).\end{aligned}$$

□

3. The Derivative of the Determinant. The determinant mapping gives us our first nontrivial mapping from $\mathbb{R}^{n\times n}$ to $\mathbb{R}$. The formula for the derivative is called *Jacobi's* formula. In a sense, this derivative should be "easy" since the determinant of a matrix is just a polynomial in the entries of the matrix. The standard way to compute the derivative of the determinant is to use Laplace's formula: given $X = (x_{ij}) \in \mathbb{R}^{n\times n}$, we have for all $i_0, j_0 \in \{1,2,\dots,n\}$ that

$$\det(X) = \sum_{i=1}^n x_{ij_0}(-1)^{i+j_0}\det(X(i,j_0)) = \sum_{j=1}^n x_{i_0j}(-1)^{i_0+j}\det(X(i_0,j))\,,$$

where $X(i,j)\in\mathbb{R}^{(n-1)\times(n-1)}$ is the matrix obtained from X by deleting the i^{th} row and j^{th} column. This formula immediately tells us that

$$\frac{\partial\det(X)}{\partial x_{ij}} = (-1)^{i+j}\det(X(i,j)) \quad \forall\, i,j\in\{1,2,\dots,n\}.$$

Consequently, the derivative of the determinant can be written in terms of the *classical adjoint* of X:

$$\mathsf{adj}(A) := \left((-1)^{i+j}\det(X(i,j))\right)^T.$$

That is,

$$(\det(\cdot))'(X)(D) = \operatorname{tr}\left(\mathsf{adj}(X)D\right) = \left\langle \mathsf{adj}(X)^T,\, D\right\rangle_F.$$

This can also be written in differential notation as

$$d\det(X) = \left\langle \mathsf{adj}(X)^T,\, dX \right\rangle_F,$$

which more explicitly describes how to apply the chain rule. The differential notation also tells us that the gradient of det at X is $\mathsf{adj}(X)^T$, i.e. $\nabla\det(X) = \mathsf{adj}(X)^T$. That is, $\mathsf{adj}(X)^T$ is the embedding of $\det(X)'$ into $\mathbb{R}^{n\times n}$ under the duality pairing given by the Frobenious inner product $\langle A,\, B\rangle_F = \operatorname{tr}\left(A^T B\right)$.

The determinate is the unique multilinear form on the columns (or rows) whose value at the identity is 1. Determinants have a much longer history than do matrices themselves since they were derived to solve linear systems of equations long before the invention of matrices. The culmination of this effort is what we now call *Cramer's rule*. Cramer's rule tells us that for any $A \in \mathbb{R}^{n\times n}$ we have

$$A\,\mathsf{adj}(A) = \mathsf{adj}(A)A = \det(A)I_n.$$

Consequently, when $\det(A) \neq 0$, then A^{-1} exists and we have

$$A^{-1} = \frac{1}{\det(A)}\mathsf{adj}(A) = \det(A^{-1})\,\mathsf{adj}(A) \qquad \text{and}$$
$$\mathsf{adj}(A) = \det(A)\,A^{-1}.$$

In particular, when A^{-1} exists, we have

$$\nabla\det(A) = \mathsf{adj}(A)^T = \det(A)\,A^{-T}.$$

As a brief note of caution, it is well established that the computation of a determinate is, in general, a highly unstable numerical process as is the computation of the inverse. Nonetheless, it is an extremely valuable theoretical tool.

The determinant has long been the key theoretical tool of understanding the eigenvalues (principal, proper, ... values) of a matrix. We now use this connection to give an alternative derivation of the derivative of the determinant. This alternative derivation nicely illustrates a powerful approach to computing derivatives using the connection between the Gateaux and Frechét derivatives in finite dimensions.

Recall that the characteristic polynomial of a matrix $A \in \mathbb{R}^{n\times n}$ are given by

$$\det(\lambda I_n - A) = \lambda^n - \operatorname{tr}(A)\,\lambda^{(n-1)} + \cdots + (-1)^n\det(A).$$

We use this fact to derive the derivative of the determinate on the nonsingular matrices using the formula for the Gateaux derivative.

Let $A \in \mathbb{R}^{n\times n}$ be nonsingular. Given $D \in \mathbb{R}^{n\times n}$ and $t \in \mathbb{R}\setminus\{0\}$, set $\lambda := t^{-}1$. The characteristic polynomial formula tells us that

$$\begin{aligned}
\det(A + tD) &= \det(tA(t^{-1}I - (-A^{-1}D)) \\
&= t^n\det(A)\det(\lambda I - (-A^{-1}D)) \\
&= t^n\det(A)(\lambda^n + \operatorname{tr}\left(A^{-1}D\right)\lambda^{(n-1)} + ... \\
&= \det(A)(1 + \operatorname{tr}\left(A^{-1}D\right)t + t^2(\text{stuff}) \\
&= \det(A) + \det(A)\operatorname{tr}\left(A^{-1}D\right)t + t^2(\text{stuff}) \\
&= \det(A) + t\left\langle \det(A)A^{-T},\, D\right\rangle + t^2(\text{stuff}).
\end{aligned}$$

Consequently, the Gateaux derivative, and hence the derivative, is given by

$$\nabla \det(A) = \det(A)A^{-T}.$$

An interesting consequence of this representation is that if $A_k \to A$ with the members of the sequence $\{A_k\}$ all nonsingular, then

$$\mathsf{adj}(A) = \lim_k \det(A_k)\, A_k^{-1}$$

even when A is singular! In this way we recover from our the second derivation of the derivative the full representation of the derivative even at singular matrices.

4. More Derivative Examples. In this section we provide a few more examples of derivative computations.

4.1. X^{-1}. Since the eigenvalues of a matrix are continuous functions of the matrix entries, the set of nonsingular matrices is open in $\mathbb{R}^{n\times n}$. Hence one can apply the standard approach to computing the derivative of the inverse. However, I will take a more powerful approach that highlights some of the ideas I am trying to illustrate in these notes. This approach uses the *Banach Lemma* which makes use of the *spectral radius*: given $A \in \mathbb{R}^{n\times n}$, the spectral radius of A is the maximum modulus of its spectrum,

$$\rho(A) := \max\{|\lambda| \mid \det(\lambda - A) = 0\}.$$

Lemma 4.1 (Banach Lemma). *Given $A \in \mathbb{R}^{n\times n}$, if $\rho(A) < 1$, then $(I-A)^{-1}$ exists and is given by the geometric series*

$$(I-A)^{-1} = I + A + A^2 + A^3 + \dots.$$

In addition, we have

$$\frac{1}{1+\rho(A)} \le \rho((I-A)^{-1}) \le \frac{1}{1-\rho(A)}.$$

The derivative of the inverse mapping now easily follows. Let $GL_n(\mathbb{R})$ denote the set of real nonsingular $n \times n$ matrices. This set is called the *general linear group of degree n over* $\mathbb{R}$. It is an open subset of $\mathbb{R}^{n\times n}$. Define $\Phi : GL_n(\mathbb{R}) \to GL_n(\mathbb{R})$ by $\Phi(A) := A^{-1}$. Let $A \in GL_n(\mathbb{R})$ and $\Delta A \in \mathbb{R}^{n\times n}$ be such that $\rho(A^{-1}\Delta A) < 1$, then

$$\begin{aligned}(A+\Delta A)^{-1} &= (A(I + A^{-1}\Delta A))^{-1} \\ &= (I + A^{-1}\Delta A)^{-1}A^{-1} \\ &= (I - A^{-1}\Delta A) + o(\|\Delta A\|))A^{-1} \quad \text{(Banach Lemma)} \\ &= A^{-1} - A^{-1}\Delta A A^{-1} + o(\|\Delta A\|).\end{aligned} \tag{4.1}$$

Consequently,

$$\Phi'(A)(D) = -A^{-1}DA^{-1} \tag{4.2}$$

and

$$d\Phi(A) = -A^{-1}dAA^{-1}.$$

This result has numerous applications. For example, let $X \in \mathbb{R}^{m\times n}$ consider the mapping $\phi : \mathbb{S}^m \to \mathbb{S}^n$

$$\phi(V) := X^T V^{-1} X.$$

Since that map $W \mapsto X^T W X$ is linear, the derivative formula above tells us that

$$\phi'(V)(D) = -X^T V^{-1} D V^{-1} X \tag{4.3}$$

and

$$d\phi(V) = -X^T V^{-1}(dV)V^{-1}X.$$

At the outset of this section it was stated that the Banach Lemma approach was more powerful than the standard approach to the computation of the derivative of the inverse mapping. The reason for this is that it gives a power series expansion in terms of ΔA and consequently it shows that the inverse mapping is infinitely differentiable and gives a formulas for all of its derivatives. Let's see how we obtain the second derivative.

Let $\mathbb{X}$ and $\mathbb{Y}$ be two normed linear spaces. A mapping $\mathcal{Q} : \mathbb{X} \times \mathbb{X} \to \mathbb{Y}$ is said to be a bilinear form from $\mathbb{X}$ to $\mathbb{Y}$ if it is linear in each argument separately: for all $(x^i, z^j) \in \mathbb{X} \times \mathbb{X},\ i = 1, 2,$ and $\alpha, \beta, \gamma, \delta \in \mathbb{R}$

$$\begin{aligned}\mathcal{Q}(\alpha x^1 + \beta x^2, \gamma z^1 + \delta z^2) &= \alpha \mathcal{Q}(x^1, \gamma z^1 + \delta z^2) + \beta \mathcal{Q}(x^2, \gamma z^1 + \delta z^2)\\ &= \gamma \mathcal{Q}(\alpha x^1 + \beta x^2, z^1) + \delta \mathcal{Q}(\alpha x^1 + \beta x^2, z^2).\end{aligned}$$

The bilinear form $\mathcal{Q}$ is said to be symmetric if $\mathcal{Q}(x, z) = \mathcal{Q}(z, x)$. Let $\mathbb{B}[\mathbb{X}, \mathbb{Y}]$ denote the set of all continuous bilinear forms from $\mathbb{X}$ to $\mathbb{Y}$. If $\mathbb{Y} = \mathbb{R}$, the bilinear forms are called quadratic forms.

Example 4.2. *Given $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times n}$, and $C \in \mathbb{R}^{n\times k}$, the mapping $\mathcal{Q} : \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n} \to \mathbb{R}^{m\times k}$ given by*

$$\mathcal{Q}(X, Z) = AXBZC$$

is a bilinear form in $\mathbb{B}[\mathbb{R}^{n\times n}, \mathbb{R}^{m\times k}]$. This bilinear form is a a quadratic form if $m = k = 1$, and it is symmetric if $m = k = 1$, $A = C$ and $B \in \mathbb{S}^n$.

Definition 4.3 (Second Derivative). *Let $F : \mathbb{X} \to \mathbb{Y}$ we say that F is twice differentiable at x if F is differentiable at x and there is a bilinear form $\mathcal{Q} \in \mathbb{Q}[\mathbb{X}, \mathbb{Y}]$ such that*

$$\lim_{z\to x} \frac{\|F(z) - (F(z) + F'(x)(z - x) + \frac{1}{2}\mathcal{Q}(z - x, z - x))\|}{\|y - x\|^2} = 0.$$

We call $\mathcal{Q}$ the second derivative of F at x and write $\mathcal{Q} = F''(x)$.

One can also define the second derivative by defining it to be the derivative of the derivative. For this approach, recall that $F'(x) \in \mathbb{L}[\mathbb{X}, \mathbb{Y}]$ and so $F' : \mathbb{X} \to \mathbb{L}[\mathbb{X}, \mathbb{Y}]$. Consequently, $F''(x) \in \mathbb{L}\left[\mathbb{X}, \mathbb{L}[\mathbb{X}, \mathbb{Y}]\right]$, $F''(x)(D) \in \mathbb{L}[\mathbb{X}, \mathbb{Y}]$, and $F''(x)(D_1)(D_2) \in \mathbb{Y}$. In particular, $F''(x)$ yields a bilinear form from $\mathbb{X}$ to $\mathbb{Y}$ and so the two approaches are equivalent in this sense, but they may give different representations for the same mapping which we call second derivative. In practice, one usually takes the derivative of the derivative to get the second derivative, but there are cases where the appropriate quadratic form presents itself for free due to the way the function is defined.

By applying the Banach Lemma, we obtain an instance where the quadratic form associated with the second derivatve is obtained for "free". Following the derivative computations in (4.1), we have

$$\begin{aligned}(A+\Delta A)^{-1} &= (A(I+A^{-1}\Delta A))^{-1}\\ &= (I+A^{-1}\Delta A)^{-1}A^{-1}\\ &= (I - A^{-1}\Delta A) + A^{-1}\Delta A) + o(\|\Delta A\|^2))A^{-1} \quad \text{(Banach Lemma)}\\ &= A^{-1} - A^{-1}\Delta A A^{-1} + A^{-1}\Delta A A^{-1}\Delta A A^{-1} + o(\|\Delta A\|^2).\end{aligned} \tag{4.4}$$

Consequently,

$$\Phi''(A)(D,D) = 2A^{-1}DA^{-1}DA^{-1}.$$

We now compute the second derivative by differentiating the derivative. We do this by applying the product rule to the derivative formula. Recall that the derivative of $\Phi(A) = A^{-1}$ is given in (4.2). We can write this expression as the product of two functions $\Phi'(A)(D_1) = F_1(A)F_2(A)$, where $F_1(A) := -A^{-1}D_1$ and $F_2(A) := A^{-1}$ with the variable D_1 considered to be fixed. We have

$$\begin{aligned}F_1'(A)(D_2) &= A^{-1}D_2A^{-1}D_1 \qquad \text{and}\\ F_2'(A)(D_2) &= -A^{-1}D_2A^{-1}.\end{aligned}$$

Consequently,

$$\Phi''(A)(D_1)(D_2) = (A^{-1}D_2A^{-1}D_1)(A^{-1}) + (-A^{-1}D_1)(-A^{-1}D_2A^{-1}),$$

so that

$$\Phi''(V)(D)(D) = 2A^{-1}DA^{-1}DA^{-1}$$

as desired.

4.2. $\ln\det(X)$**.** The natural log of the determinant, or log-det function, appears in many application particularly in statistics. The first concern is choosing a suitable domain for the log-det function. Since the logarithm is only defined over the real number, one must restrict X to have positive determinant, and, since the eigenvalues of a matrix are continuous functions of the matrix entries, the set of matrices with positive determinant is open. In addition, the component functions of the classical adjoint are polynomials in the matrix entries so that the partials of the gradient are continuous functions of the matrix entries. Therefore, one can directly apply the chain rule to the log-det at any matrix having positive determinant to obtain the derivative of the log-det at that point: for $A \in \mathbb{R}^{n\times n}$ with $\det(A) > 0$, we have

$$\nabla(\ln\det(\cdot))(A) = A^{-T}.$$

In many applications, the ambient space is $\mathbb{S}^n$ rather than $\mathbb{R}^{n\times n}$. This is also an inner product space under the Frobenious inner product, and if A symmetric so is $\mathsf{adj}(A)$. Hence the derivative of $\det(A)$ is again $\mathsf{adj}(A)$, and now $\nabla(\ln\det(\cdot))(A) = A^{-1}$ whenever $\det(A) > 0$ due to symmetry. In most applications of log-det on $\mathbb{S}^n$, one restricts the application of log-det to the open convex cone $\mathbb{S}^n_{++}$ of real symmetric positive definite matrices whose closure is $\mathbb{S}^n_+$ the cone of real symmetric positive semidefinite matrices. The log-det is well defined on $\mathbb{S}^n_{++}$.

In this context, an important application is the map $\psi : \mathbb{S}^m \to \mathbb{R} \cup \{+\infty\}$ given by

$$\psi(V) := \begin{cases} \ln\det(X^TV^{-1}X) & , \ V \in \mathbb{S}^n_{++} \\ +\infty & , \ \text{otherwise,} \end{cases}$$

where $X \in \mathbb{R}^{m\times n}$ is such that $\ker(X) = \{0\}$. Since $\ker(X) = \{0\}$, it is easily seen that $X^TV^{-1}X \in \mathbb{S}^n_{++}$ whenever $V \in \mathbb{S}^m_{++}$. By combining the results of this section and those of the previous section and using the fact that $\operatorname{tr}\left(A^TB\right) = \operatorname{tr}\left(BA^T\right)$ for all $A, B \in \mathbb{R}^{m\times n}$, the chain rule yields

$$\begin{aligned} \psi'(V)(D) &= \left\langle \nabla(\ln\det(\cdot))(X^TV^{-1}X),\, (X^T(\cdot)^{-1}X)'(V)(D) \right\rangle \\ &= \left\langle (X^TV^{-1}X)^{-1},\, -X^TV^{-1}DV^{-1}X \right\rangle \\ &= -\operatorname{tr}\left((X^TV^{-1}X)^{-1}X^TV^{-1}DV^{-1}X\right) \\ &= -\operatorname{tr}\left(V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}D\right) \\ &= \left\langle -V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1},\, D \right\rangle, \end{aligned}$$

which tells us that

$$\nabla\psi(V) = -V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}.$$

Observe that $\nabla\psi(V) \in \mathbb{S}^m$ as it should be.

An expression for the second derivative is obtained by applying the product rule to $\nabla\psi(V)$: for $V \in \mathbb{S}^n_{++}$ and $D \in \mathbb{S}^n$,

$$\begin{aligned} (\nabla\psi(\cdot))'(V)(D) = {} & V^{-1}DV^{-1}X(X^TV^{-1}X)^{-1}XV^{-1} + V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}DV^{-1} \\ & - V^{-1}X(X^TV^{-1}X)^{-1}XV^{-1}DV^{-1}X(X^TV^{-1}X)^{-1}XV^{-1}. \end{aligned}$$

Hence,

$$\begin{aligned} \psi''(V)(D)(D) &= \langle (\nabla\psi(\cdot))'(V)(D),\, D \rangle \\ &= \operatorname{tr}\left(V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}DV^{-1}D\right) \\ &\quad + \operatorname{tr}\left(V^{-1}DV^{-1}X(X^TV^{-1}X)^{-1}XV^{-1}D\right) \\ &\quad - \operatorname{tr}\left(V^{-1}X(X^TV^{-1}X)^{-1}XV^{-1}DV^{-1}X(X^TV^{-1}X)^{-1}XV^{-1}D\right) \\ &= \operatorname{tr}\big(2V^{-1}X(X^TV^{-1}X)^{-1}XV^{-1}DV^{-1}D \\ &\qquad -V^{-1}X(X^TV^{-1}X)^{-1}XV^{-1}DV^{-1}X(X^TV^{-1}X)^{-1}XV^{-1}D\big) \\ &= \operatorname{tr}\Big(2(V^{-1/2}QV^{-1/2}DV^{-1/2})(V^{-1/2}D) - (V^{-1/2}QV^{-1/2}DV^{-1/2})Q(V^{-1/2}D)\Big) \\ &= \operatorname{tr}\Big((V^{-1/2}QV^{-1/2}DV^{-1/2})[2I - Q](V^{-1/2}D)\Big), \end{aligned}$$

where $Q := V^{-1/2}X(X^TV^{-1}X)^{-1}XV^{-1/2}$ is easily seen to be the orthogonal projection onto Ran $\left(V^{-1/2}X\right)$. Hence

$$\begin{aligned}\psi''(V)(D)(D) &= \operatorname{tr}\Big((V^{-1/2}QV^{-1/2}DV^{-1/2})[2I-Q](V^{-1/2}D)\Big)\\ &= \operatorname{tr}\Big((V^{-1/2}Q^2V^{-1/2}DV^{-1/2})[2I-Q](V^{-1/2}D)\Big)\\ &= \operatorname{tr}\Big((V^{-1/2}DV^{-1/2}Q)^T[2I-Q](V^{-1/2}DV^{-1/2}Q)\Big)\\ &= \operatorname{tr}\left(Z^T[2I-Q]Z\right)\\ &= \sum_{j=1}^{n} z_j^T[2I-Q]z_j,\end{aligned}$$

where $Z := V^{-1/2}DV^{-1/2}Q$ whose j^{th} column is $z_j,\ j=1,\dots,n$. Since Q is an orthogonal projection, for all $z\in\mathbb{R}^n$,

$$z^T[2I-Q]z = 2\left\|z\right\|_2^2 - \left\|Qz\right\|_2^2 = 2\left\|(I-Q)z\right\|_2^2 + \left\|Qz\right\|_2^2 = \left\|(I-Q)z\right\|_2^2 + \left\|z\right\|_2^2 \geq 0$$

with equality if and only if $z=0$. That is, $2I-Q\in\mathbb{S}^n_{++}$. Hence,

$$\psi''(V)(D)(D)\geq 0 \qquad \forall\, D\in\mathbb{S}^n$$

with equality if and only if $D=0$, or equivalently, the quadratic form $\psi''(V)(\cdot)(\cdot)$ is positive definite on $\mathbb{S}^n_{++}$. In particular, this implies that the function ψ is strictly convex on $\mathbb{S}^n$.

Remark 4.4. *For further illusrtations of matrix differentiation , see my notes* Meta-Analysis Variance Estimators in Mixed Effects Models*.*