Matrix Calculus Notes

James Burke

2000

Linear Spaces and Operators

$\mathbb{X}$ and $\mathbb{Y}$ – normed linear spaces with norms $\|\cdot\|_x$ and $\|\cdot\|_y$.

$\mathcal{L} : \mathbb{X} \to \mathbb{Y}$ is a linear transformations (or operators) if
$$\mathcal{L}(\alpha x + \beta z) = \alpha \mathcal{L}(x) + \beta \mathcal{L}(z) \quad \forall\, x, z \in \mathbb{X} \text{ and } \alpha, \beta \in \mathbb{R}.$$

$\mathbb{L}[\mathbb{X}, \mathbb{Y}]$ the normed space of continuous linear operators:
$$\|\mathcal{T}\| := \sup_{\|x\|_x \leq 1} \|\mathcal{T} x\|_y \quad \forall\, \mathcal{T} \in \mathbb{L}[\mathbb{X}, \mathbb{Y}].$$

$\mathbb{X}^* := \mathbb{L}[\mathbb{X}, \mathbb{R}]$ – topological dual of $\mathbb{X}$ with the *duality pairing*
$$\langle \phi,\, x \rangle = \phi(x) \quad \forall\, (\phi, x) \in \mathbb{X}^* \times \mathbb{X}.$$

The duality pairing gives rise to *adjoints* of a linear operator:
$\mathcal{T} \in \mathbb{L}[\mathbb{X}, \mathbb{Y}]$ defines $\mathcal{T}^* \in \mathbb{L}[\mathbb{Y}^*, \mathbb{X}^*]$ by
$$\langle y^*,\, \mathcal{T}(x) \rangle = \langle \mathcal{T}^*(y^*),\, x \rangle \quad \forall\, (y^*, x) \in \mathbb{Y}^* \times \mathbb{X}.$$

Matrix Representations

$\{x^j\}_{j=1}^n$ and $\{y^i\}_{i=1}^m$ are bases for $\mathbb{X}$ and $\mathbb{Y}$.

Given $x = \sum_{j=1}^n a_j x^j \in \mathbb{X}$, the linear mapping

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

is a linear isomorphism between $\mathbb{X}$ and $\mathbb{R}^n$ and is called the *coordinate mapping* from $\mathbb{X}$ to $\mathbb{R}^n$ associated with the basis $\{x^j\}_{j=1}^n$.

Matrix Representations

$\{x^j\}_{j=1}^n$ and $\{y^i\}_{i=1}^m$ are bases for $\mathbb{X}$ and $\mathbb{Y}$.

Given $x = \sum_{j=1}^n a_j x^j \in \mathbb{X}$, the linear mapping

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

is a linear isomorphism between $\mathbb{X}$ and $\mathbb{R}^n$ and is called the *coordinate mapping* from $\mathbb{X}$ to $\mathbb{R}^n$ associated with the basis $\{x^j\}_{j=1}^n$.

η – coordinate mapping from $\mathbb{Y}$ to $\mathbb{R}^m$ 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

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

Therefore, $\eta(\mathcal{T}x) = T\kappa(x)$ where where $(t_{ij}) = T \in \mathbb{R}^{m \times n}$.

The Kronecker Product

$A, C \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{s\times t}$. The Kronecker product of A with B is the $ms \times mt$ matrix given by

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

$$A = [2 \ \ -1], \ B = \begin{bmatrix} 0 & -1 \\ 2 & 1 \end{bmatrix}$$

$$A \otimes B = \begin{bmatrix} 0 & -2 & 0 & 1 \\ 4 & 2 & -2 & -1 \end{bmatrix}$$

The Matrix Coordinate Map "vec"

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, $\langle A, B \rangle_F = \mathsf{vec}(A)^T \mathsf{vec}(B)$ and $\mathsf{vec} \in \mathbb{L}[\mathbb{R}^{m \times n}, \mathbb{R}^{mn}]$.

Properties 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. $(A^\dagger \otimes B^\dagger) = (A \otimes B)^\dagger$, where $M^\dagger$ is the Moore-Penrose pseudo inverse of the matrix M.
5. For vectors a and b, $\mathsf{vec}(ab^T) = b \otimes a$.
6. If AXB is a well defined matrix product, then

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

In particular,

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

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

Spectrum of the Kronecker Product

$A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times m}$ have spectrum $\{\lambda_i\}_{i=1}^n$, $\{\mu_i\}_{i=1}^m$ with multiplicity, resp.ly.

Then the eigenvalues of $A \otimes B$ are

$$\lambda_i \mu_j, \;\; i, j = 1, \ldots, n$$

and the eigenvalues of $(I_n \otimes A) + (B \otimes I_m)$ are

$$\lambda_i + \mu_j, \;\; i, j = 1, \ldots, n.$$

$(I_n \otimes A) + (B \otimes I_m)$ is called the *Kronecker sum* of A and B.

In particular,
$\operatorname{tr}(A \otimes B) = \operatorname{tr}(A) \operatorname{tr}(B)$ and $\det(A \otimes B) = \det A^n \det(B)^m$.

Singular Values of the Kronecker Product

$A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{s \times t}$ have singular value decompositions $U_A \Sigma_A V_A^T$ and $U_B \Sigma_B V_V^T$. Then, after reordering, the singular value decomposition of $A \otimes B$ is

$$(U_A \otimes U_B)(\Sigma_A \otimes \Sigma_B)(V_A^T \otimes V_B^T).$$

In particular, the nonzero singular values of $A \otimes B$ are

$$\sigma_i(A)\sigma_j(B), \ i = 1, \ldots, \operatorname{rank} A, \ j = 1, \ldots, \operatorname{rank} B.$$

Matrix Representations for $\mathbb{L}[\mathbb{R}^{m\times n}, \mathbb{R}^{s\times t}]$

On $\mathbb{R}^{m\times n}$ the matrices

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

where E_{ij} is the matrix having a one in the ij position and zero elsewhere, form 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 this basis, where the coordinates are ordered by columns.

We show how to use vec and $\otimes$ to compute a matrix representation for of elements $\mathbb{L}[\mathbb{R}^{m\times n}, \mathbb{R}^{s\times t}]$ with respect to the standard unit coordinate bases on $\mathbb{R}^{m\times n}$ and $\mathbb{R}^{s\times t}$.

Example 1

Let $A \in \mathbb{R}^{s\times m}$ and $B \in \mathbb{R}^{n\times t}$ and define $\mathcal{T} \in \mathbb{L}[\mathbb{R}^{m\times n}, \mathbb{R}^{s\times t}]$ by $\mathcal{T}(X) = AXB$.

Then, using the coordinate mapping vec we get

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

Hence, the matrix representation of $\mathcal{T}$ in the coordinate bases is

$$T = (B^T \otimes A) \ .$$

Example 2

Define $\mathcal{T} \in \mathbb{L}[\mathbb{R}^{n\times n}, \mathbb{R}^{n\times n}]$ by $\mathcal{T}(X) = AX + XB$, where $A, B \in \mathbb{R}^{n\times n}$. Then

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

That is, the matrix representation of $\mathcal{T}$ in the unit coordinate bases is

$$T = (I_n \otimes A) + (B^T \otimes I_n)$$

the Kronecker sum of A and B^T.

The Derivative of det

The standard way to compute the derivative of the determinant is to use Laplace's formula: $\forall\ i_0, j_0 \in \{1, 2, \ldots, n\}$

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

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

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

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,

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

In differential notation, $d\det(X) = \left\langle \mathsf{adj}(X)^T, dX \right\rangle_F$, which more explicitly describes how to apply the chain rule.

The Determinant

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 .They were derived to solve linear systems 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

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

So, 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

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

The Banach Lemma

The spectral radius of $A \in \mathbb{R}^{n\times n}$ is the maximum modulus of its spectrum,

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

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

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

Derivatives of A^{-1}: $d(X)^{-1} = -X^{-1}(dX)X^{-1}$

The *general linear group of degree* n *over* $\mathbb{R}$, $GL_n(\mathbb{R})$, is the set of real nonsingular $n \times n$ matrices.

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

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

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

$$\mathsf{vec}(\Phi'(A)(D)) = -\mathsf{vec}(A^{-1}DA^{-1}) = -(A^{-T} \otimes A^{-1})\mathsf{vec}(D)$$
$$\implies \nabla\Phi(A) = -A^{-T} \otimes A^{-1}$$

This procedure shows that Φ is $\mathcal{C}^\infty$ and all these derivatives are easily computed from the Banach Lemma.

Chain Rule Example

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

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

$$\implies \nabla\psi(V) = -V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1} \ \in \mathbb{S}^n.$$