

ML & LinAlg Math Cheat Sheet

November 18, 2017

Contents

1	Notation	1
2	Derivative	2
2.a	Vector Gradient	2
3	Determinant Operator	2
3.a	Random Properties	2
4	Trace Operator	2
4.a	Properties	2
4.a.i	$\text{tr}(c\mathbf{A} + d\mathbf{B}) = c\text{tr}(\mathbf{A}) + d\text{tr}(\mathbf{B})$	2
4.a.ii	$\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA}) = \text{tr}(\mathbf{A}^\top \mathbf{B}) = \text{tr}(\mathbf{AB}^\top) = \sum_{i,j} a_{ij}b_{ij}$	2
4.b	Derivatives	2
4.b.i	$\nabla_{\mathbf{x}} \text{tr}(\mathbf{xx}^\top \mathbf{A}) = \mathbf{x}^\top (\mathbf{A} + \mathbf{A}^\top)$	2
4.c	Relation to Determinant	2
5	Expected Values	2
A	Proofs	3
A.a	Trace	3
A.a.i	$\nabla_{\mathbf{x}} \text{tr}(\mathbf{xx}^\top \mathbf{A}) = \mathbf{x}^\top (\mathbf{A} + \mathbf{A}^\top)$	3

1 Notation

Vectors are column vectors denoted by lower-case bolded variables, such that

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}.$$

A row vector is denoted $\mathbf{x}^\top = [x_1 \dots x_N]$. A matrix is indicated by a bolded upper-case variable, such that an $N \times M$ matrix is

$$\mathbf{A} = \{a_{ij}\} = [\mathbf{a}_1 \dots \mathbf{a}_M] = \begin{bmatrix} \mathbf{a}_1^\top \\ \vdots \\ \mathbf{a}_N^\top \end{bmatrix} = \begin{bmatrix} a_{1,1} & \dots & a_{1,M} \\ \vdots & \ddots & \vdots \\ a_{N,1}^\top & \dots & a_{N,M} \end{bmatrix}.$$

For some random variable x , let $\mathbb{E}[x]$ denote its expected value.

2 Derivative

2.a Vector Gradient

$$\nabla_{\mathbf{x}} \mathbf{y} = [\frac{\partial \mathbf{y}}{\partial x_1}, \dots, \frac{\partial \mathbf{y}}{\partial x_N}] \quad (1)$$

3 Determinant Operator

3.a Random Properties

For scalar c and $N \times N$ identity matrix I ,

$$\det(cI) = c^N.$$

4 Trace Operator

Defined for $N \times N$ square matrix \mathbf{A} as

$$\text{tr}(\mathbf{A}) \stackrel{\text{def}}{=} \sum_i^N a_{ii} \quad (2)$$

4.a Properties

$$4.a.i \quad \text{tr}(c\mathbf{A} + d\mathbf{B}) = c \text{tr}(\mathbf{A}) + d \text{tr}(\mathbf{B})$$

For scalars c and d , square matrices \mathbf{A} and \mathbf{B} .

$$4.a.ii \quad \text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA}) = \text{tr}(\mathbf{A}^\top \mathbf{B}) = \text{tr}(\mathbf{AB}^\top) = \sum_{i,j} a_{ij} b_{ij}$$

And clearly, also $\text{tr}(\mathbf{B}^\top \mathbf{A}) = \text{tr}(\mathbf{BA}^\top) = \sum_{i,j} a_{ij} b_{ij} = \text{tr}(\mathbf{AB})$.

4.b Derivatives

$$4.b.i \quad \nabla_{\mathbf{x}} \text{tr}(\mathbf{x}\mathbf{x}^\top \mathbf{A}) = \mathbf{x}^\top (\mathbf{A} + \mathbf{A}^\top)$$

For square matrix \mathbf{A} . Note that $\mathbf{x}^\top (\mathbf{A} + \mathbf{A}^\top) = 2\mathbf{x}^\top \mathbf{A}$ for symmetric \mathbf{A} .

See appendix [A.a.i](#) for proof.

4.c Relation to Determinant

5 Expected Values

For $\mathbf{x} \in \mathbb{R}^d$, with expected value $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\top]$,

$$\mathbb{E}[x_i^2] = \Sigma_{i,i} + \mu_i^2 \quad (3)$$

$$\mathbb{E}[\mathbf{x}^\top \mathbf{A} \mathbf{x}] = \text{tr}(\mathbf{A} \boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{\mu} \quad (4)$$

$$\mathbb{E}_{\mathbf{x}} [(y - \mathbf{x}^\top \mathbf{w})^2] = (y - \mathbf{w}^\top \boldsymbol{\mu})^2 + \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}. \quad (5)$$

A Proofs

A.a Trace

A.a.i $\nabla_{\mathbf{x}} \text{tr}(\mathbf{x}\mathbf{x}^\top \mathbf{A}) = \mathbf{x}^\top (\mathbf{A} + \mathbf{A}^\top)$

This proof can likely be generalized to non-square matrixes (and possibly some communicativeness, given the flexibility afforded by the trace), but the restricted case is presented here.

For square $N \times N$ matrix \mathbf{A} ,

$$\nabla_{\mathbf{x}} \text{tr}(\mathbf{x}\mathbf{x}^\top \mathbf{A}) = \frac{d}{d\mathbf{x}} \text{tr}(\mathbf{x}\mathbf{x}^\top \mathbf{A}) = \frac{d}{d\mathbf{x}} \sum_i^N \sum_k^N x_i x_k a_{ik}.$$

Recall eq. (1), and consider for any $j \in \{1, \dots, N\}$:

$$\begin{aligned} \frac{\partial}{\partial x_j} \sum_i^N \sum_k^N x_i x_k a_{ik} &= [x_1 a_{1,j} + x_2 a_{2,j} + \dots + x_{j-1} a_{j-1,j} + x_{j+1} a_{j+1,j} + \dots + x_N a_{N,j}] \\ &\quad + \frac{\partial}{\partial x_j} \sum_k^N x_j x_k a_{jk} \\ &= \left[\sum_i^N x_i a_{ij} - x_j a_{jj} \right] + \sum_k^N x_k a_{jk} - x_j a_{jj} + \frac{\partial}{\partial x_j} x_j x_j a_{jj} \\ &= \sum_i^N x_i a_{ij} + \sum_k^N x_k a_{jk} - 2x_j a_{jj} + 2x_j a_{jj} \\ &= \mathbf{x}^\top \mathbf{a}_j + \mathbf{x}^\top [\mathbf{a}^\top]_j, \end{aligned}$$

where $[\mathbf{a}^\top]_j$ is the j th column of \mathbf{A}^\top .

This equally applies for any j in $1 \dots N$, and so for the full gradient:

$$\begin{aligned} \nabla_{\mathbf{x}} \text{tr}(\mathbf{x}\mathbf{x}^\top \mathbf{A}) &= \frac{d}{d\mathbf{x}} \sum_i^N \sum_k^N x_i x_k a_{ik} = [\mathbf{x}^\top \mathbf{a}_1 \dots \mathbf{x}^\top \mathbf{a}_N] + [\mathbf{x}^\top [\mathbf{a}^\top]_1 \dots \mathbf{x}^\top [\mathbf{a}^\top]_N] \\ &= \mathbf{x}^\top (\mathbf{A} + \mathbf{A}^\top). \end{aligned}$$