

# The Tensor Cookbook

Thomas Dybdahl Ahle

May 8, 2024

# Chapter 1

## Introduction

**What is this?** These pages are a guide to tensors, using the visual language of “tensor diagrams”. For illustrating the generality of the approach, I’ve tried to closely follow the legendary “Matrix Cookbook”. As such, most of the presentation is a collection of facts (identities, approximations, inequalities, relations, ...) about tensors and matters relating to them. You won’t find many results not in the original cookbook, but hopefully the diagrams will give you a new way to understand and appreciate them.

**It’s ongoing:** The Matrix Cookbook is a long book, and not all the sections are equally amenable to diagrams. Hence I’ve opted to skip certain sections and shorten others. Perhaps in the future, I, or others, will expand the coverage further.

For example, while we cover all of the results on Expectation of Linear Combinations and Gaussian moments, we skip the section on general multi-variate distributions. I have also had to rearrange the material a bit, to avoid having to introduce all the notation up front.

**Complex Matrices and Covariance** Tensor diagrams (or networks) are currently most often seen in Quantum Physics. Here most values are complex numbers, which introduce some extra complexity. In particular transposing a matrix now involves taking the conjugate (flipping the sign of the imaginary part), which introduces the need for co- and contra-variant tensors. None of this complexity is present with standard real valued matrices, as is common e.g. in Machine Learning applications. For simplicity I have decided to not include these complexities.

**Tensorgrad** The symbolic nature of tensor diagrams make them well suited for symbolic computation.

**Advantages of Tensor Diagram Notation:** Tensor diagram notation has many benefits compared to other notations:

Various operations, such as a trace, tensor product, or tensor contraction can be expressed simply without extra notation. Names of indices and tensors can often be omitted. This saves time and lightens the notation, and is especially useful for internal indices which exist mainly to be summed over. The order of the tensor resulting from

a complicated network of contractions can be determined by inspection: it is just the number of unpaired lines. For example, a tensor network with all lines joined, no matter how complicated, must result in a scalar.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Notation and Nomenclature . . . . .	5
1.2	Basics . . . . .	5
1.2.1	Trace . . . . .	5
<b>2</b>	<b>Simple Derivatives</b>	<b>8</b>
2.1	Derivatives of Matrices, Vectors and Scalar Forms . . . . .	8
2.1.1	First Order . . . . .	8
2.1.2	Second Order . . . . .	9
2.1.3	Higher-order and non-linear . . . . .	9
2.1.4	Gradient and Hessian . . . . .	9
2.2	Derivatives of Traces . . . . .	10
2.2.1	First Order . . . . .	10
2.2.2	Second Order . . . . .	10
2.2.3	Higher Order . . . . .	10
2.2.4	Other . . . . .	10
<b>3</b>	<b>Statistics and Probability</b>	<b>11</b>
3.0.1	Definition of Moments . . . . .	11
3.0.2	Expectation of Linear Combinations . . . . .	11
3.0.3	Weighted Scalar Variable . . . . .	12
3.0.4	Gaussian Moments . . . . .	12
<b>4</b>	<b>Solutions and Decompositions</b>	<b>13</b>
4.1	Kronecker and Vec Operator . . . . .	13
4.2	Solutions to linear equations . . . . .	14
4.2.1	The Lyapunov Equation . . . . .	14
4.3	Eigenvalues and Eigenvectors . . . . .	15
4.4	Decompositions . . . . .	15
4.4.1	Singular Value Decomposition . . . . .	15
4.4.2	Triangular Decomposition . . . . .	15
4.4.3	LDM Decomposition . . . . .	15
4.4.4	LDL Decomposition . . . . .	15

<b>5</b>	<b>Determinant and Inverses</b>	<b>16</b>
5.1	Determinant . . . . .	16
5.2	Inverses . . . . .	17
<b>6</b>	<b>Functions</b>	<b>18</b>
6.0.1	Pseudo-linear form . . . . .	18
6.1	Taylor Series . . . . .	18
<b>7</b>	<b>Advanced Derivatives</b>	<b>19</b>
7.1	Derivatives of vector norms . . . . .	19
7.1.1	Two-norm . . . . .	19
7.2	Derivatives of matrix norms . . . . .	19
7.3	Derivatives of Structured Matrices . . . . .	19
7.3.1	The Chain Rule . . . . .	19
7.3.2	The Hessian Chain Rule . . . . .	19
7.3.3	Symmetric . . . . .	20
7.3.4	Diagonal . . . . .	20
7.3.5	Toeplitz . . . . .	20
7.4	Derivatives of a Determinant . . . . .	20
7.5	General forms . . . . .	20
7.6	Linear forms . . . . .	20
7.7	Square forms . . . . .	20
7.8	Derivatives of an Inverse . . . . .	20
7.9	Derivatives of Eigenvalues . . . . .	20
<b>8</b>	<b>Special Matrices</b>	<b>21</b>
8.0.1	Block matrices . . . . .	21
8.0.2	The Discrete Fourier Transform Matrix . . . . .	21
8.0.3	Hermitian Matrices and skew-Hermitian . . . . .	21
8.0.4	Idempotent Matrices . . . . .	21
8.0.5	Orthogonal matrices . . . . .	21
8.0.6	Positive Definite and Semi-definite Matrices . . . . .	21
8.0.7	Singleentry Matrix, The . . . . .	21
8.0.8	Symmetric, Skew-symmetric/Antisymmetric . . . . .	22
8.0.9	Toeplitz Matrices . . . . .	22
8.0.10	Units, Permutation and Shift . . . . .	22
8.0.11	Vandermonde Matrices . . . . .	22
<b>9</b>	<b>Machine Learning Applications</b>	<b>23</b>
9.1	Least Squares . . . . .	23
9.2	Hessian of Cross Entropy Loss . . . . .	23
9.3	Convolutional Neural Networks . . . . .	23
9.4	Transformers / Attention . . . . .	23
9.5	Tensor Sketch . . . . .	23

<b>10 Tensorgrad</b>	<b>24</b>
10.1 Simplification Rules . . . . .	24
10.2 Functions . . . . .	24
10.3 Isomorphisms . . . . .	24
<b>11 Appendix</b>	<b>25</b>

## 1.1 Notation and Nomenclature

Dot product	$a-b$	$y = \sum_i a_i b_i$	$[\cdot \cdot \cdot \cdot] \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$	$= y$
Outer product	$-a \quad b-$	$Y_{i,j} = a_i b_j$	$\begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} [\cdot \cdot \cdot \cdot]$	$= -Y-$
Matrix-Vector	$-A-b$	$y_i = \sum_j A_{i,j} b_j$	$\begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$	$= -y$
Matrix-Matrix	$-A-B-$	$Y_{i,k} = \sum_j A_{i,j} B_{j,k}$	$\begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}$	$= -Y-$

The rank-4 copy tensor,  $C_{i,j,k,l} = [i = j = k = l]$ ,  $\succ\diagup\diagdown\prec$ , differs from the outer product of two identity matrices (in the Cookbook denoted  $J$ ), which satisfies  $J_{i,j,k,l} = [i = k][j = l]$  and which we'd write as  $J = \begin{smallmatrix} \circ & \circ \\ \hline \circ & \circ \end{smallmatrix}$ , and satisfies, for example,  $\frac{dX}{dX} = J$ .

## 1.2 Basics

$-A- = -Q-\underset{\lambda}{\circ}-Q^{-1}-$  where  $\lambda_i$  is the  $i$ th eigenvalue of  $A$ .

More general principles:

1. You can contract edges in any order.
2. You can always contract connected subgraphs of Copy tensors.
3. distributive law of sums/products.
4. Sums and broadcasting.

### 1.2.1 Trace

The “trace” of a square matrix is defined  $\text{Tr}(A) = \sum_i A_{i,i}$ . In tensor diagram notation,

that corresponds to a self-edge:  $\begin{array}{c} \curvearrowright \\ A \end{array}$ .

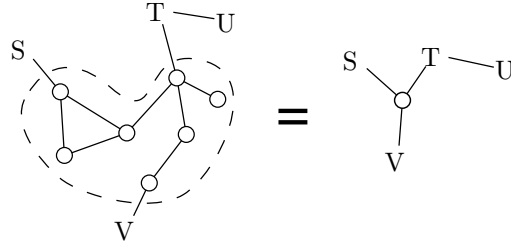


Figure 1.1: Contract spiders

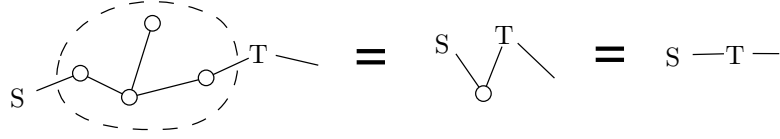


Figure 1.2: Eliminate Identity

$$\sum_{i=1}^n A_{ii} = \text{Tr}(A) = \text{Tr}(AI) \quad \text{A} = \text{A} \text{---} \circ \quad (11)$$

$$\text{Tr}(A) = \sum_i \lambda_i = \langle 1, \lambda \rangle \quad \text{A} = \text{Q} \text{---} \circ \text{---} \text{Q}^{-1} \quad (12)$$

$$\text{Tr}(A) = \text{Tr}(A^T) \quad \text{A} = \text{A} \quad (13)$$

$$\text{Tr}(AB) = \text{Tr}(BA) \quad \text{A} \text{---} \text{B} = \text{B} \text{---} \text{A} \quad (14)$$

$$\text{Tr}(A + B) = \text{Tr}(A) + \text{Tr}(B) \quad \text{(A + B)---} \circ = \text{A---} \circ + \text{B---} \circ \quad (15)$$

$$\begin{aligned} \text{Tr}(ABC) &= \text{Tr}(BCA) \\ &= \text{Tr}(CAB) \end{aligned} \quad \begin{aligned} \text{A---B---C} &= \text{B---C---A} \\ &= \text{C---A---B} \end{aligned} \quad (16)$$

$$\begin{aligned} a^T a &= \text{Tr}(aa^T) \\ a \text{---} a &= \text{Tr}(-a \text{---} a \text{---}) \\ &= \text{a---a} \end{aligned} \quad (17)$$

$$\left( \begin{array}{c} \text{Diagram 1} \\ \text{Diagram 2} \end{array} \right) = \begin{array}{c} \text{Diagram 3} \\ \text{Diagram 4} \end{array}$$

Figure 1.3: Distributive law



## Chapter 2

# Simple Derivatives

### 2.1 Derivatives of Matrices, Vectors and Scalar Forms

#### 2.1.1 First Order

The Matrix Cookbook defines the single-entry matrix  $J^{i,j} \in R^{n \times n}$  as the matrix which is zero everywhere except in the entry  $(i, j)$  in which it is 1. Alternatively we could write  $J_{n,m}^{i,j} = [i = n][j = m]$ .

$$\frac{\partial x^T a}{\partial x} = \frac{\partial a^T x}{\partial x} = a \quad (x - a) \begin{smallmatrix} \nearrow \\ \bullet \end{smallmatrix} = (x) \begin{smallmatrix} \nearrow \\ \bullet \end{smallmatrix} - a = \nearrow a \quad (69)$$

$$\frac{\partial a^T X b}{\partial X} = ab^T \quad (a - X - b) \begin{smallmatrix} \nearrow \\ \bullet \end{smallmatrix} = a - (X) \begin{smallmatrix} \nearrow \\ \bullet \end{smallmatrix} - b = a - \nearrow b \quad (70)$$

$$\frac{\partial X}{\partial X_{i,j}} = J^{i,j} \quad (-X - \begin{smallmatrix} \nearrow \\ \bullet \\ \nearrow \end{smallmatrix} \begin{smallmatrix} i \\ j \end{smallmatrix}) = -\begin{smallmatrix} \nearrow \\ \bullet \\ \nearrow \end{smallmatrix} \begin{smallmatrix} i \\ j \end{smallmatrix} \quad (73)$$

$$\begin{aligned} \frac{\partial (XA)_{i,j}}{\partial X_{m,n}} &= (J^{m,n} A)_{i,j} & (\begin{smallmatrix} i \\ - \end{smallmatrix} X - A \begin{smallmatrix} \nearrow \\ \bullet \\ \nearrow \end{smallmatrix} \begin{smallmatrix} m \\ n \end{smallmatrix}) &= - (X) \begin{smallmatrix} \nearrow \\ \bullet \end{smallmatrix} - A - \\ & & &= \begin{smallmatrix} i \\ \nearrow \end{smallmatrix} \begin{smallmatrix} m \\ \nearrow \end{smallmatrix} A \begin{smallmatrix} \nearrow \\ \bullet \end{smallmatrix} \end{aligned} \quad (74)$$

### 2.1.2 Second Order

$$\begin{aligned} \frac{\partial}{\partial X_{i,j}} \sum_{k,l,m,n} X_{k,l} X_{m,n} &= \frac{\partial \|X\|_F^2}{\partial X_{i,j}} \\ &= 2 \sum_{k,l} X_{k,l} \end{aligned} \quad \begin{aligned} \left( \begin{array}{c} \circ - X - \circ \\ \circ - X - \circ \end{array} \right) \begin{array}{c} \nearrow \\ \searrow \end{array} \begin{array}{c} i \\ j \end{array} &= \begin{array}{c} \circ - X - \circ \\ \circ - (X) - \circ \end{array} \begin{array}{c} \nearrow \\ \searrow \end{array} \begin{array}{c} i \\ j \end{array} + \begin{array}{c} \circ - (X) - \circ \\ \circ - X - \circ \end{array} \begin{array}{c} \nearrow \\ \searrow \end{array} \begin{array}{c} i \\ j \end{array} \\ &= 2 \begin{array}{c} \circ - X - \circ \\ \circ - \begin{array}{c} \nearrow \\ \searrow \end{array} \begin{array}{c} i \\ j \end{array} - \circ \end{array} \end{aligned} \quad (76)$$

$$\begin{aligned} \frac{\partial b^T X^T X c}{\partial X} &= X(b c^T + c b^T) \\ (b - X^T - X - c) \begin{array}{c} \nearrow \\ \searrow \end{array} \begin{array}{c} i \\ j \end{array} &= b - X^T - (X) \begin{array}{c} \nearrow \\ \searrow \end{array} \begin{array}{c} i \\ j \end{array} - c \\ &+ b - (X^T) \begin{array}{c} \nearrow \\ \searrow \end{array} \begin{array}{c} i \\ j \end{array} - X - c \\ &= b - X^T - \begin{array}{c} \nearrow \\ \searrow \end{array} \begin{array}{c} i \\ j \end{array} - c \\ &+ b - \begin{array}{c} \nearrow \\ \searrow \end{array} \begin{array}{c} i \\ j \end{array} - X - c \\ &= -X - (-b c' + -c b') \end{aligned} \quad (77)$$

$$\begin{aligned} \frac{\partial}{\partial x} (Bx + b)^T C (Dx + d) &= B^T C (Dx + d) \\ &+ D^T C^T (Bx + b) \end{aligned} \quad (78)$$

$$\begin{aligned} \frac{\partial}{\partial X_{i,j}} (X^T B X)_{k,l} &= \delta_{l,j} (X^T B)_{k,i} \\ &+ \delta_{k,j} (B X)_{i,l} \end{aligned} \quad \begin{aligned} (\begin{array}{c} \circ - X^T - B - X - \circ \\ \circ - X^T - B - X - \circ \end{array}) \begin{array}{c} \nearrow \\ \searrow \end{array} \begin{array}{c} i \\ j \end{array} &= \begin{array}{c} \circ - X^T - B - \begin{array}{c} \nearrow \\ \searrow \end{array} \begin{array}{c} i \\ j \end{array} - \circ \\ &+ \begin{array}{c} \circ - \begin{array}{c} \nearrow \\ \searrow \end{array} \begin{array}{c} i \\ j \end{array} - B - X - \circ \end{array} \end{aligned} \quad (79)$$

$$\frac{\partial}{\partial X_{i,j}} X^T B X = X^T B J^{i,j} + J^{j,i} B X \quad \dots \quad (80)$$

$$\begin{aligned} \frac{\partial}{\partial x} x^T B x &= (B + B^T) x \\ (x - B - x) \begin{array}{c} \nearrow \\ \searrow \end{array} \begin{array}{c} i \\ j \end{array} &= -B - x + x - B - \begin{array}{c} \nearrow \\ \searrow \end{array} \begin{array}{c} i \\ j \end{array} \\ &= x - \begin{array}{c} \nearrow \\ \searrow \end{array} \begin{array}{c} i \\ j \end{array} \left( \begin{array}{c} \begin{array}{c} \nearrow \\ \searrow \end{array} \begin{array}{c} i \\ j \end{array} B - \begin{array}{c} \nearrow \\ \searrow \end{array} \begin{array}{c} i \\ j \end{array} \end{array} \right) \end{aligned} \quad (81)$$

$$\frac{\partial}{\partial X} b^T X^T D X c = D^T X b c^T + D X c b^T \quad \dots \quad (82)$$

$$\frac{\partial}{\partial X} (Xb + c)^T D (Xb + c) = (D + D^T) (Xb + c) b^T \quad \dots \quad (83)$$

Assume  $W$  is symmetric, then... (84) - (88)

### 2.1.3 Higher-order and non-linear

...

### 2.1.4 Gradient and Hessian

...



## Chapter 3

# Statistics and Probability

### 3.0.1 Definition of Moments

Let  $x \in \mathbb{R}^n$  is a random variable. We write  $m = E[x] \in \mathbb{R}^n$  for the expectation and  $M = \text{Var}[x] = E[(x - m)(x - m)^T]$  for the covariance (when these quantities are defined.)

In tensor diagrams, we will use square brackets:

$$m = [-x] \quad \text{and} \quad M = [-(x \ominus m) \quad (x \div m) -]$$

Note we used the German minus,  $\div$ , to distinguish subtraction from contraction edges.

We can also define the third and fourth centralized moment tensors

$$M_3 = \begin{bmatrix} (x \div m) - \\ (x \div m) - \\ (x \div m) - \end{bmatrix} \quad \text{and} \quad M_4 = \begin{bmatrix} (x \div m) - \\ (x \div m) - \\ (x \div m) - \\ (x \div m) - \end{bmatrix}.$$

### 3.0.2 Expectation of Linear Combinations

General principle: The “linearity of expectation” lets you pull out all parts of the graph not involving  $X$ .

#### Linear Forms

$$E[AXB + C] = AE[X]B + C \quad \begin{bmatrix} -A-X-B- \\ + -C- \end{bmatrix} = \begin{bmatrix} -A-[X]-B- \\ + -C- \end{bmatrix} \quad (312)$$

$$\begin{aligned} \text{Var}[Ax] &= A\text{Var}[x]A^T \quad \begin{bmatrix} A-x \div [A-x] \\ A-x \div [A-x] \end{bmatrix} = \begin{bmatrix} A-(x \div m) \\ A-(x \div m) \end{bmatrix} \quad (313) \\ &= A \begin{bmatrix} (x \div m) \\ (x \div m) \end{bmatrix} \\ &= -A-M_2-A- \end{aligned}$$

### Quadratic Forms

$$\begin{aligned}
 E[x^T A x] &= \text{Tr}(A \Sigma) + \mu^T A \mu \\
 [x - A - x] &= [(x \div \mu) - A - (x \div \mu) + \mu - A - \mu] \\
 &= \left[ \begin{matrix} (x \div m) \\ (x \div m) \end{matrix} \right] A + \mu - A - \mu \\
 &= \overbrace{\Sigma - A} + \mu - A - \mu
 \end{aligned}$$

### Cubic Forms

#### 3.0.3 Weighted Scalar Variable

Let  $y = w^T x$ , and let  $m = E[y]$ , then

$$\begin{aligned}
 E[y] &= m = w^T \mu \\
 E[(y - m)^2] &= w - M_2 - w \\
 E[(y - m)^3] &= w - M_3 - w \\
 E[(y - m)^4] &= w - M_4 - w
 \end{aligned}$$

For specific distributions, like  $x$  Gaussian, we can often reduce the moment tensors further. Khintchine's inequality also gives a way to bound all of these in terms of  $E[(y - m)^2]$ .

#### 3.0.4 Gaussian Moments

Mean and covariance of linear forms

Mean and variance of square forms

Cubic forms

Mean of Quartic Forms

Gaussian Integration by Parts

General principle for Gaussian expectations.

## Chapter 4

# Solutions and Decompositions

### 4.1 Kronecker and Vec Operator

The Kronecker Product and Vec Operator are ways to avoid dealing with tensors, and generally best avoided.

$$\underbrace{\begin{array}{c} \text{A} \\ \diagup \quad \diagdown \\ \text{X} \quad \quad \quad \\ \diagdown \quad \diagup \\ \text{B} \end{array}}_{\text{vec}(A^T X B)} = \underbrace{\text{X} \begin{array}{c} \text{A} \\ \diagup \quad \diagdown \\ \text{B} \end{array}}_{\text{vec}(X)^T (A \otimes B)}$$

Some basic properties:

$$\begin{array}{c} i \\ \diagdown \\ \diagup \\ j \end{array} \begin{array}{c} \diagup \\ \text{---} \triangle \text{---} \\ \diagdown \end{array} \begin{array}{c} k \\ \text{---} \end{array} = [i + j \text{ } n = k]$$

Opposing triangles cancel:

$$\begin{array}{c} \text{---} \triangle \text{---} \\ \text{---} \triangle \text{---} \end{array} = \text{---}$$

and

$$\text{---} \triangle \text{---} \triangle \text{---} = \text{---}$$

### The Kronecker Product

The Kronecker product of an  $m \times n$  matrix  $A$  and an  $r \times q$  matrix  $B$ , is an  $mr \times nq$  matrix,  $A \otimes B$  defined as

$$A \otimes B = \begin{bmatrix} A_{1,1}B & A_{1,2}B & \cdots & A_{1,n}B \\ A_{2,1}B & A_{2,2}B & \cdots & A_{2,n}B \\ \vdots & \vdots & \ddots & \vdots \\ A_{m,1}B & A_{m,2}B & \cdots & A_{m,n}B \end{bmatrix}$$

The Kronecker product has the following properties:

$$A \otimes (B + C) = A \otimes B + A \otimes C \quad (4.1)$$

$$A \otimes (B \otimes C) = (A \otimes B) \otimes C \quad (4.2)$$

$$aA \otimes bB = ab(A \otimes B) \quad (4.3)$$

$$(A \otimes B)^T = A^T \otimes B^T \quad (4.4)$$

$$(A \otimes B)(C \otimes D) = AC \otimes BD \quad (4.5)$$

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1} \quad (4.6)$$

$$\text{Tr}(A \otimes B) = \text{Tr}(A)\text{Tr}(B) \quad (4.7)$$

### The Vec Operator

A few more good properties here.

$$\text{vec}(AXB) = (B^T \otimes A)\text{vec}(X) \quad (4.8)$$

$$\text{Tr}(A^T B) = \text{vec}(A)^T \text{vec}(B) \quad (4.9)$$

$$\text{vec}(A + B) = \text{vec}(A) + \text{vec}(B) \quad (4.10)$$

$$\text{vec}(aA) = a\text{vec}(A) \quad (4.11)$$

$$a^T X B X^T c = \text{vec}(X)^T (B \otimes ca^T) \text{vec}(X) \quad (4.12)$$

## 4.2 Solutions to linear equations

### 4.2.1 The Lyapunov Equation

This is a nice trick that can be well explained. Though it requires some talk about Kronecker products. Also “(5.1.11) Encapsulating Sum” which is a generalization.

### 4.3 Eigenvalues and Eigenvectors

### 4.4 Decompositions

#### 4.4.1 Singular Value Decomposition

#### 4.4.2 Triangular Decomposition

#### 4.4.3 LDM Decomposition

#### 4.4.4 LDL Decomposition



## Chapter 5

# Determinant and Inverses

### 5.1 Determinant

It's convenient to write the determinant in tensor notation as

$$\det(A) = \frac{1}{n!} \overline{\overline{A \cdots A}}$$

where  $\overline{\overline{i_1 \ i_2 \ \cdots \ i_n}} = \varepsilon_{i_1, \dots, i_n}$  is the rank- $n$  Levi-Civita tensor defined by

$$\varepsilon_{i_1, \dots, i_n} = \begin{cases} \text{sign}(\sigma) & \sigma = (i_1, \dots, i_n) \text{ is a permutation} \\ 0 & \text{otherwise.} \end{cases}$$

To see that the definition makes sense, let's first consider

$$\det(I) = \frac{1}{n!} \overline{\overline{1 \cdots 1}} = \frac{1}{n!} \sum_{i_1, \dots, i_n, j_1, \dots, j_n} \varepsilon_{i_1, \dots, i_n} \varepsilon_{j_1, \dots, j_n} [i = j] = \frac{1}{n!} \sum_{i_1, \dots, i_n} \varepsilon_{i_1, \dots, i_n}^2 = 1.$$

In general we get from the permutation definition of the determinant:

$$\begin{aligned} \overline{\overline{A \cdots A}} &= \sum_{i_1, \dots, i_n, j_1, \dots, j_n} \varepsilon_{i_1, \dots, i_n} \varepsilon_{j_1, \dots, j_n} A_{i_1, j_1} \cdots A_{i_n, j_n} \\ &= \sum_{\sigma, \tau} \text{sign}(\sigma) \text{sign}(\tau) A_{\sigma_1, \tau_1} \cdots A_{\sigma_n, \tau_n} \\ &= \sum_{\sigma} \text{sign}(\sigma) \sum_{\tau} \text{sign}(\tau) A_{\sigma_1, \tau_1} \cdots A_{\sigma_n, \tau_n} \\ &= \sum_{\sigma} \text{sign}(\sigma)^2 \det(A) \\ &= n! \det(A). \end{aligned}$$

The definition generalizes to Cayley's "hyper determinants" by . . . .

A curious property is that

$$\begin{array}{|c|} \hline A \\ \hline \end{array} \cdots \begin{array}{|c|} \hline A \\ \hline \end{array} = \begin{array}{|c|} \hline \cdots \\ \hline \end{array} \begin{array}{|c|} \hline A \\ \hline \end{array} \cdots \begin{array}{|c|} \hline A \\ \hline \end{array}$$

$$(18) \quad \det(A) = \prod_i \lambda_i \quad \dots$$

$$(19) \quad \det(cA) = c^n \det(A) \quad \begin{array}{|c|} \hline cA \\ \hline \end{array} \cdots \begin{array}{|c|} \hline cA \\ \hline \end{array} = c^n \begin{array}{|c|} \hline A \\ \hline \end{array} \cdots \begin{array}{|c|} \hline A \\ \hline \end{array}$$

$$(20) \quad \det(A) = \det(A^T) \quad \dots$$

$$(21) \quad \det(AB) = \det(A)\det(B) \quad \begin{array}{|c|} \hline A \\ \hline \end{array} \cdots \begin{array}{|c|} \hline A \\ \hline \end{array} \begin{array}{|c|} \hline B \\ \hline \end{array} \cdots \begin{array}{|c|} \hline B \\ \hline \end{array} = \begin{array}{|c|} \hline A \\ \hline \end{array} \cdots \begin{array}{|c|} \hline A \\ \hline \end{array} \begin{array}{|c|} \hline B \\ \hline \end{array} \cdots \begin{array}{|c|} \hline B \\ \hline \end{array}$$

$$(22) \quad \det(A^{-1}) = 1/\det(A) \quad \dots$$

$$(23) \quad \det(A^n) = \det(A)^n \quad \dots$$

$$(24) \quad \det(I + uv^T) = 1 + u^T v \quad \dots$$

## 5.2 Inverses

Might be reduced, unless cofactor matrices have a nice representation?

## Chapter 6

# Functions

Stuff about analytical matrix functions. Such as Exponential Matrix Function. I'd rather talk about my general function notation. And maybe about taylor series?

### 6.0.1 Pseudo-linear form

Maybe this should just be an example in a table?

Derivation of Peyman Milanfar's gradient

$$\begin{aligned}d[\mathbf{f}(\mathbf{x})] &= d[\mathbf{A}(\mathbf{x})\mathbf{x}] \\&= d[\mathbf{A}(\mathbf{x})]\mathbf{x} + \mathbf{A}(\mathbf{x})d\mathbf{x} \\&= \text{vec}\{d[\mathbf{A}(\mathbf{x})]\mathbf{x}\} + \mathbf{A}(\mathbf{x})d\mathbf{x} \\&= \text{vec}\{\mathbf{Id}[\mathbf{A}(\mathbf{x})]\mathbf{x}\} + \mathbf{A}(\mathbf{x})d\mathbf{x} \\&= (\mathbf{x}^T \otimes \mathbf{I}) \text{vec}\{d[\mathbf{A}(\mathbf{x})]\} + \mathbf{A}(\mathbf{x})d\mathbf{x} \\&= (\mathbf{x}^T \otimes \mathbf{I}) D \text{vec}[\mathbf{A}(\mathbf{x})]d\mathbf{x} + \mathbf{A}(\mathbf{x})d\mathbf{x} \\&= [(\mathbf{x}^T \otimes \mathbf{I}) D \text{vec}[\mathbf{A}(\mathbf{x})] + \mathbf{A}(\mathbf{x})] d\mathbf{x}\end{aligned}$$

### 6.1 Taylor Series

I've always wanted to write these out properly with tensors.

## Chapter 7

# Advanced Derivatives

### 7.1 Derivatives of vector norms

#### 7.1.1 Two-norm

$$\frac{d}{dx} \|x - a\|_2 = \frac{x - a}{\|x - a\|_2} \quad (7.1)$$

$$\frac{d}{dx} \frac{x - a}{\|x - a\|_2} = \frac{I}{\|x - a\|_2} - \frac{(x - a)(x - a)^T}{\|x - a\|_2^3} \quad (7.2)$$

$$\frac{d}{dx} \|x\|_2^2 = \frac{d}{dx} \|x^T x\|_2 = 2x \quad (7.3)$$

### 7.2 Derivatives of matrix norms

### 7.3 Derivatives of Structured Matrices

#### 7.3.1 The Chain Rule

Sometimes the objective is to find the derivative of a matrix which is a function of another matrix.

#### 7.3.2 The Hessian Chain Rule

See main.tex

7.3.3 Symmetric

7.3.4 Diagonal

7.3.5 Toeplitz

7.4 Derivatives of a Determinant

7.5 General forms

7.6 Linear forms

7.7 Square forms

7.8 Derivatives of an Inverse

7.9 Derivatives of Eigenvalues

## Chapter 8

# Special Matrices

### 8.0.1 Block matrices

Stuff like Schur complements is interesting. But can we say anything useful using tensor diagrams?

### 8.0.2 The Discrete Fourier Transform Matrix

I think FFT can be nicely described with diagrams

Let's start with the Hadamard matrix:  $H_n = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{\otimes n}$ . Hm. It's just a bunch of matrices below each other, kinda boring.

What about the FFT? Does that require a bit more?

### 8.0.3 Hermitian Matrices and skew-Hermitian

Complex. Skip

### 8.0.4 Idempotent Matrices

Skip

### 8.0.5 Orthogonal matrices

Skip

### 8.0.6 Positive Definite and Semi-definite Matrices

Skip

### 8.0.7 Singleentry Matrix, The

Describes the matrix  $J$ . All of this is trivial with diagrams.

### **8.0.8 Symmetric, Skew-symmetric/Antisymmetric**

Could introduce Penrose's symmetric tensors here?

### **8.0.9 Toeplitz Matrices**

Could talk about the convolution tensor here...

### **8.0.10 Units, Permutation and Shift**

Not that interesting...

### **8.0.11 Vandermonde Matrices**

Does this have a nice description? Not a lot of properties are given in the Cookbook.

## Chapter 9

# Machine Learning Applications

9.1 Least Squares

9.2 Hessian of Cross Entropy Loss

9.3 Convolutional Neural Networks

9.4 Transformers / Attention

9.5 Tensor Sketch



## Chapter 10

# Tensorgrad

Implementation details

### 10.1 Simplification Rules

### 10.2 Functions

### 10.3 Isomorphisms

## Chapter 11

# Appendix

Contains some proofs, such as of equation 524 or 571. They are pretty long and could be useful for contrasting with the diagram proofs.