Statistical Machine Learning Notes 1

Background

Instructor: Justin Domke

Contents

1 Probability

This section is largely inspired by Larry Wasserman's *All of Statistics* and Arian Melki and Tom Do's Review of Probability Theory.

1.1 Sample Spaces

The basic element of probability theory is that of a **sample space** Ω. We can think of this as "the set of all the things that can happen". For example, if we are flipping a coin, we might have

$$\Omega = \{H, T\},$$

or if we are flipping three coins, we might have

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

An **event** A is defined to be a subset $\mathcal{F} \subset \Omega$ of the sample space. **Probability measures** are defined as a function p from events to real numbers

$$P : \mathcal{F} \to \mathbb{R}.$$

In order to be considered a valid probability distribution[1], these functions have to obey some natural axioms:

- $P(A) \geq 0$ - "Probabilities aren't negative."
- $P(\Omega) = 1$ - "*Something* happens."
- If $A_1, A_2, ..., A_N$ are disjoint, then $P(A_1 \cup A_2) = P(A_1) + P(A_2)$.

Given these axioms, many other properties can be *derived*:

- $P(\emptyset) = 0$
- $A \subset B \to p(A) \leq p(B)$
- $P(A) = 1 - P(A^c)$ (where $A^c = \{\omega : \omega \in \Omega, \omega \notin A\}$
- $P(A) \leq 1$

[1] Or probability density, if variables are continuous.

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Conditional probabilities are defined as

$$P(A|B) = \frac{P(A, B)}{P(B)}.$$

Though samples spaces are the foundation of probability theory, we rarely need to deal with them explicitly. This is because they usually disappear into the background, while we work with random variables.

1.2 Random Variables

A **random variable** X is a function from a sample space to the real numbers

$$X : \Omega \to \mathbb{R}.$$

Technically, we need to impose some mild regularity conditions on the function in order for it to be a valid random variable, but we won't worry about such niceties. Given a random variable X, we define X^{-1} to be the function that maps a real number to all the events corresponding to that number, i.e.

$$X^{-1}(A) = \{\omega : X(\omega) \in A\}.$$

Given this, we can assign probabilities to random variables as

$$P(X \in A) = P(X^{-1}(A)).$$

Note here that we are "overloading" the notation for a probability distribution by allowing it to take either events or conditions on random variables as input. We will also use (fairly self-explanatory) notations like

$$P(X = 5), \text{ or } P(3 \leq X \leq 7).$$

Now, why introduce random variables? Basically, because we usually don't care about every detail of the sample space. If we are flipping a set of 5 coins, it is awkward to deal with the set

$$\Omega = \{HHHHH, HHHHT, ..., TTTTT\},$$

which would involve specifying probabilities for each of the possible 2^5 outcomes. However, if we define a random variable

$$X(\omega) = \text{number of } H \text{ in } \omega,$$

we must only define probabilities $P(X = 0)$, $P(X = 1)$, ..., $P(X = 5)$.

Example: Let $\Omega = \{(x, y) : x^2 + y^2 \leq 1\}$ be the unit disk. Take a random point $\omega = (x, y)$ from Ω. Some valid random variables are:

- $X(\omega) = x$
- $Y(\omega) = y$
- $Z(\omega) = x + y$
- $W(\omega) = \sqrt{x^2 + y^2}$

Now, we will commonly distinguish between two types of random variables.

- A **discrete random variable** is one in which $X(A)$ takes a finite or countably infinite number of values.
- A **continuous random variable** is one for which a probability density function exists. (Defined below!)

For a discrete random variable, we can define a **probability mass function** such that

$$P(X = x) = p_X(x).$$

Be careful! Do you understand how p is different from P? Do you understand how X is different from x?

For a continuous random variable, a **probability density function** can be defined[2] such that

$$P(a \leq X \leq b) = \int_a^b p_X(x)dx.$$

Note that we must integrate a probability density function to get a probability. For continuous random variables, $p_X(x) \neq P(X = x)$ (in general). It is quite common, in fact, for probability densities to be higher than one.

[2]Technically, we should first define the cumulative distribution function.

1.3 Multiple Random Variables

We will often be interested in more than one random variable at a time. If we take two discrete random variables X and Y, the **joint probability mass function** is defined as

$$P(X = x, Y = y) = p_{X,Y}(x, y).$$

For continuous random variables, the **joint probability density function** is defined such that

$$P((X, Y) \in A) = \int \int_A p_{X,Y}(x, y) dx dy.$$

Again, we need to integrate the joint probability density function in order to get a probability. Again, densities can be higher than zero.

The **marginal distribution** is defined by

$$p_X(x) = \begin{cases} \sum_y p_{X,Y}(x, y) & \text{if } Y \text{ is discrete} \\ \int_{-\infty}^{\infty} p_{X,Y}(x, y) & \text{if } Y \text{ is continuous} \end{cases}.$$

This is called "marginalization" because it was originally done by accountants literally writing numbers in the margins of tables. Reassuringly, it does work out that one obtains the same probability mass/density by deriving p_X directly from probabilities on the sample space or by first deriving $p_{X,Y}$ and then marginalizing. (Otherwise our choice of using the same notation p_X in both cases would be quite odd.)

For either type of variable, we define the **conditional probability mass/density function** as

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x, y)}{p_Y(y)},$$

assuming that $p_X(x) \neq 0$.

Two random variables X and Y are said to be **independent** if for all A and B,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

This is equivalent to either of the following:

- $p_{X,Y}(x, y) = p_X(x)p_Y(y)$ for all x and y.
- $p_{Y|X}(y|x) = p_Y(y)$ whenever $p_X(x) \neq 0$.

The random variables X and Y are said to be **conditionally independent given** Z if

$$P(X \in A, Y \in B | Z \in C) = P(X \in A | Z \in C) P(Y \in B | Z \in C).$$

This is equivalent to saying:

- $p_{X,Y|Z}(x, y|z) = p_{X|Z}(x|z) p_{Y|Z}(y|z)$ for all x, y, z.
- $p_{Y|X,Z}(y|x, z) = p_{Y|Z}(y|z)$ whenever $p_{X,Z}(x, z) \neq 0$.

1.4 Expectations

The **expected value** of a random variable is defined to be

$$\mathbb{E}[X] = \begin{cases} \sum_x x p_X(x) & \text{if } X \text{ is discrete} \\ \int x p_X(x) dx & \text{if } X \text{ is continuous} \end{cases}.$$

Note that in certain cases (which aren't even particularly weird), the integral might fail to converge, and we will say that the expected value does not exist. A couple important results about expectations follow.

Theorem 1. *If $X_1, ..., X_n$ are random variables, and $a_1, ..., a_n$ are constants, then*

$$\mathbb{E}[\sum_i a_i X_i] = \sum_i a_i \mathbb{E}[X_i].$$

Do not miss the word "independent" in the following theorem!

Theorem 2. *If $X_1, ..., X_n$ are independent random variables, then*

$$\mathbb{E}[\prod_i X_i] = \prod_i \mathbb{E}[X_i].$$

The **variance** of a random variable is defined by

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Roughly speaking, this measures how "spread out" the variable X is.

Some other useful results, which you should prove for "fun" if you've never seen them before follow. Again, note the word "independent" in the third result.

Theorem 3. *Assuming that the variance exists, then*

- $\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
- If a and b are constants, then $\mathbb{V}(aX + b) = a^2\mathbb{V}(X)$.
- If $X_1, ..., X_n$ are independent, and $a_1, ..., a_n$ are constants, then

$$\mathbb{V}[\sum_{i=1}^{n} a_i X_i] = \sum_{i=1}^{n} a_i \mathbb{V}[X_i].$$

Given two random variables X and Y, the **covariance** is defined by

$$\text{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Given two random variables X and Y, the **conditional expectation** of X given that $Y = y$ is

$$\mathbb{E}[X|Y = y] = \begin{cases} \sum x p_{X|Y}(x|y) & \text{if } X \text{ is discrete} \\ \int x p_{X|Y}(x|y) dx & \text{if } X \text{ is continuous} \end{cases}.$$

Given a bunch of random variables $X_1, X_2, ..., X_n$ it can be convenient to write them together in a vector $\mathbf{X}$. We can then define things such the expected value of the random vector

$$\mathbb{E}[\mathbf{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix},$$

or the **covariance matrix**

$$\text{Cov}[\mathbf{X}] = \begin{bmatrix} \text{Cov}[X_1, X_1] & \dots & \text{Cov}[X_1, X_n] \\ \vdots & \ddots & \vdots \\ \text{Cov}[X_n, X_1] & \dots & \text{Cov}[X_n, X_n] \end{bmatrix}.$$

1.5 Empirical Expectations

Suppose that we have some dataset $\{(\mathbf{x}^1, y^1), (\mathbf{x}^2, y^2), ..., (\mathbf{x}^N, y^N)\}$. Suppose further that we have some vector of weights $\mathbf{w}$. We might be interested in the sum of squares fit of y to $\mathbf{w} \cdot \mathbf{x}$.

$$R = \frac{1}{N} \sum_{i=1}^{N} (y^i - \mathbf{w} \cdot \mathbf{x}^i)^2.$$

Instead, we will sometimes convenient to define a discrete distribution that assigns $1/N$ probability to each of the values in the dataset. We will write expectations with respect to this distribution as $\hat{\mathbb{E}}$. Thus, we would write the above sum of squares error as

$$R = \hat{\mathbb{E}}[(Y - \mathbf{w} \cdot \mathbf{X})^2].$$

Once you get used to this notation, it often makes expressions simpler by getting rid of Σ, i and N.

2 Linear Algebra

2.1 Basics

A **vector** is a bunch of real numbers stuck together

$$\mathbf{x} = (x_1, x_2, ..., x_N).$$

Given two vectors, we define the inner product two be the sum of the products of corresponding elements, i.e.

$$\mathbf{x} \cdot \mathbf{y} = \sum_i x_i y_i.$$

(Obviously, $\mathbf{x}$ and $\mathbf{y}$ must be the same length in order for this to make sense.)

A **matrix** is a bunch of real numbers written together in a table

$$A = \begin{bmatrix} A_{11} & A_{12} & ... & A_{1N} \\ A_{21} & A_{22} & ... & A_{2N} \\ \vdots & \vdots & & \vdots \\ A_{M1} & A_{M2} & ... & A_{MN} \end{bmatrix}.$$

We will sometimes refer to the ith column or row of A as $A_{\cdot i}$ or $A_{i\cdot}$, respectively. Thus, we could also write

$$A = \begin{bmatrix} A_{\cdot 1} & A_{\cdot 2} & ... & A_{\cdot N} \end{bmatrix},$$

or

$$A = \begin{bmatrix} A_{1\cdot} \\ A_{2\cdot} \\ \vdots \\ A_{M\cdot} \end{bmatrix}.$$

Given a matrix A and a vector $\mathbf{x}$, we define the **matrix-vector product** by

$$\mathbf{b} = A\mathbf{x} \leftrightarrow b_i = \sum_j A_{ij}x_j = A_{i\cdot}\mathbf{x}.$$

There are a couple of other interesting ways of looking at matrix-vector multiplication. One is that we can see the matrix-vector product as a vector of the inner-product of each row of A with $\mathbf{x}$.

$$\mathbf{b} = A\mathbf{x} = \begin{bmatrix} A_{1\cdot}\mathbf{x} \\ A_{2\cdot}\mathbf{x} \\ \vdots \\ A_{M\cdot}\mathbf{x} \end{bmatrix}$$

Another is that we can write the result as the sum of the columns of A, weighted by the components of $\mathbf{x}$.

$$\mathbf{b} = A\mathbf{x} = A_{\cdot 1}x_1 + A_{\cdot 2}\mathbf{x}_2 + ... + A_{\cdot N}x_N.$$

Now, suppose that we have two matrices, A and C. We define the **matrix-matrix product** by

$$C = AB \leftrightarrow C_{ij} = \sum_k A_{ik}B_{kj} = A_{i\cdot}B_{\cdot j}.$$

Where does this definition come from? The motivation is as follows.

Theorem 4. *If $C = AB$, then*

$$C\mathbf{x} = A(B\mathbf{x}).$$

Proof.

$$\begin{aligned}(C\mathbf{x})_i &= \sum_j C_{ij}x_j \\ &= \sum_j \sum_k A_{ik}B_{kj}x_j \\ &= \sum_k A_{ik} \sum_j B_{kj}x_j \\ &= \sum_k A_{ik}(B\mathbf{x})_k \\ &= (A(B\mathbf{x}))_i\end{aligned}$$

□

Now, as with matrix-vector multiplication, there are a bunch of different ways to look at matrix-matrix multiplication. Here, we will look at A and B in two different ways.

$$\begin{aligned}A &= \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1N} \\ A_{21} & A_{22} & \dots & A_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ A_{M1} & A_{M2} & \dots & A_{MN} \end{bmatrix} \text{ (a table)} \\ &= \begin{bmatrix} A_{1\cdot} \\ A_{2\cdot} \\ \vdots \\ A_{M\cdot} \end{bmatrix} \text{ (a bunch of row vectors)} \\ &= \begin{bmatrix} A_{\cdot 1} & A_{\cdot 2} & \dots & A_{\cdot N} \end{bmatrix} \text{ (a bunch of column vectors)}\end{aligned}$$

Theorem 5. *If $C = AB$, then*

- $C = \begin{bmatrix} A_{1\cdot}B_{\cdot 1} & A_{1\cdot}B_{\cdot 2} & \dots & A_{1\cdot}B_{\cdot N} \\ A_{2\cdot}B_{\cdot 1} & A_{2\cdot}B_{\cdot 2} & \dots & A_{2\cdot}B_{\cdot N} \\ \vdots & \vdots & \ddots & \vdots \\ A_{M\cdot}B_{\cdot 1} & A_{M\cdot}B_{\cdot 2} & \dots & A_{M\cdot}B_{\cdot N} \end{bmatrix}$ (A matrix of inner products)
- $C = \sum_i A_{\cdot i}B_{i\cdot}$ (The sum out outer products)
- $C = \begin{bmatrix} AB_{\cdot 1} & AB_{\cdot 2} & \dots & AB_{\cdot N} \end{bmatrix}$ (A bunch of matrix-vector multiplies)

- $C = \begin{bmatrix} A_{1\cdot}B \\ A_{2\cdot}B \\ \vdots \\ A_{M\cdot}B \end{bmatrix}$ (A bunch of vector-matrix multiplies)

In a sense, matrix-matrix multiplication generalizes matrix-vector multiplication, if you think of a vector as a $N \times 1$ matrix.

Interesting aside. Suppose that A is $N \times N$ and B is $N \times N$. How much time will it take to compute $C = AB$? Believe it or not, this is an open problem! There is an obvious solution of complexity $\mathcal{O}(N^3)$. However, algorithms have been invented with complexities of roughly $\mathcal{O}(N^{2.807})$ and $\mathcal{O}(N^{2.376})$, though with weaknesses in terms of complexity of implementation, numerical stability and computational constants. (The idea of these algorithms is to manually find a way of multiplying $k \times k$ matrices for some chosen k (e.g. 2) while using less than k^3 multiplications, then using this recursively on large matrices.) It is conjectured that algorithms with complexity of essentially $\mathcal{O}(N^2)$ exist.

There are a bunch of properties that you should be aware of with matrix-matrix multiplication.

- $A(BC) = (AB)C$
- $A(B + C) = AB + AC$

Notice that $AB = BA$ isn't on this list. That's because it isn't (usually) true.

The **identity matrix** I is an $N \times N$ matrix, with

$$I_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}.$$

It is easy to see that $I\mathbf{x} = \mathbf{x}$.

The **transpose** of a matrix is obtained by flipping the rows and columns.

$$(A^T)_{ij} = A_{ji}$$

It isn't hard to show that

- $(A^T)^T = A$
- $(AB)^T = B^T A^T$

$(A + B)^T = A^T + B^T$

3 Matrix Calculus

(This section is based on Minka's Old and New Matrix Algebra Useful for Statistics and Magnus and Neudecker's *Matrix Differential Calculus with Applications in Statistics and Econometrics*.)

Here, we will consider functions that have three types of inputs:

1. Scalar input
2. Vector input
3. Matrix input

We can also look at functions with three types of outputs

1. Scalar output
2. Vector output
3. Matrix output

So, the possible types of functions could be written like

input \ output	scalar	vector	matrix
scalar	$f(x)$	$\mathbf{f}(x)$	$F(x)$
vector	$f(\mathbf{x})$	$\mathbf{f}(\mathbf{x})$	
matrix	$f(X)$		

We might try listing the derivatives we might like to calculate in a table.

input \ output	scalar	vector	matrix
scalar	$\frac{df}{d\mathbf{x}}$	$\frac{d\mathbf{f}}{dx}$	$\frac{dF}{dx}$
vector	$\frac{df}{d\mathbf{x}}$	$\frac{d\mathbf{f}}{d\mathbf{x}^T}$	
matrix	$\frac{df}{dX}$		

The three entries that aren't listed are awkward to represent with matrix notation.

Now, we could compute any of the derivatives above via partials. For example, instead of directly trying to compute the matrix

$$\frac{d\mathbf{f}}{d\mathbf{x}^T},$$

we could work out each of the scalar derivatives

$$\frac{df_i}{dx_j}.$$

However, this can get very tedious. For example, suppose, we have the simple case

$$\mathbf{f}(\mathbf{x}) = A\mathbf{x}.$$

Then

$$f_i(\mathbf{x}) = \sum_j A_{ij}x_j.$$

whence

$$\frac{df_i}{dx_j} = A_{ij}.$$

Armed with this knowledge, we can get the simple expression

$$\frac{d\mathbf{f}}{d\mathbf{x}^T} = A.$$

Surely there is a simpler way!

3.1 Checking derivatives with finite differences

In practice, after deriving a complicated gradient, practitioners almost always check the gradient numerically. The key idea is the following. For a scalar function

$$f'(x) = \lim_{dx \to 0} \frac{f(x + dx) - f(x)}{dx}.$$

Thus, just by picking a very small number ϵ, (Say, $\epsilon = 10^{-7}$), we can approximate

$$f'(x) \approx \frac{f(x + \epsilon) - f(x)}{\epsilon}.$$

We can generalize this trick to more dimensions. If $f(\mathbf{x})$ is a scalar valued function of a vector,

$$\frac{df}{dx_i} \approx \frac{f(\mathbf{x} + \epsilon \hat{\mathbf{e}}_\mathbf{i}) - f(\mathbf{x})}{\epsilon}.$$

While if $f(X)$ is a scalar-valued function of a matrix,

$$\frac{df}{dX_{ij}} \approx \frac{f(X + \epsilon \hat{E}_{ij}) - f(X)}{\epsilon}.$$

If $\mathbf{f}(\mathbf{x})$ is a vector-valued function of a vector,

$$\frac{d\mathbf{f}}{dx_i} \approx \frac{\mathbf{f}(\mathbf{x} + \epsilon \hat{\mathbf{e}}_\mathbf{i}) - \mathbf{f}(\mathbf{x})}{\epsilon}.$$

Now, why don't we always just compute derivatives this way? Two reasons:

1. It is numerically unstable. The constant ϵ needs to be picked carefully, and accuracy can still be limited.
2. It is expensive. We need to call the function a number of times depending on the size of the input.

Still, it is highly advised to check derivatives this way. For most of the difficult problems below, I did this, and often found errors in my original derivation!

3.2 Differentials

Consider the regular old derivative

$$f'(x) = \lim_{dx \to 0} \frac{f(x + dx) - f(x)}{dx}.$$

This means that we can represent f as

$$f(x + dx) = f(x) + dx\, f'(x) + r_x(dx)$$

where the remainder function goes to zero quickly as h goes to zero, i.e.

$$\lim_{dx \to 0} \frac{r_x(dx)}{dx} = 0.$$

Now, think of the point x as being fixed. Then, we can look at the difference $f(x+dx)-f(x)$ as consisting of two terms:

- The linear term $dx\, f'(x)$.
- The "error" $r_x(dx)$, which gets very small as dx goes to zero.

For a simple scalar/scalar function, the **differential** is defined to be

$$df(x; dx) = dx\, f'(x).$$

The differential is a linear function of dx.

Similarly, consider a vector/vector function $f : \mathbb{R}^n \to \mathbb{R}^m$. If there exists a matrix $A \in \mathbb{R}^{m \times n}$ such that

$$\mathbf{f}(\mathbf{x} + d\mathbf{x}) = \mathbf{f}(\mathbf{x}) + A(\mathbf{x})d\mathbf{x} + \mathbf{r}_\mathbf{x}(d\mathbf{x}),$$

where the remainder term goes quickly to zero, i.e.

$$\lim_{d\mathbf{x} \to \mathbf{0}} \frac{\mathbf{r}_\mathbf{x}(d\mathbf{x})}{||d\mathbf{x}||} = \mathbf{0},$$

then the matrix $A(\mathbf{x})$ is said to be the derivative of $\mathbf{f}$ at $\mathbf{x}$ and

$$d\mathbf{f}(\mathbf{x}; d\mathbf{x}) = A(\mathbf{x})d\mathbf{x}$$

is said to be the **differential** of $\mathbf{f}$ at $\mathbf{x}$. The differential is again a linear function of $d\mathbf{x}$.

Now, since differentials are always functions of $(\mathbf{x}; d\mathbf{x})$, below we will stop writing that. That is, we will simply say

$$d\mathbf{f} = A(\mathbf{x})d\mathbf{x}.$$

However, we would always remember that $d\mathbf{f}$ depends on both $\mathbf{x}$ and $d\mathbf{x}$.

More generally still, we could write $dF(X; dX)$ to be the linear part of $F(X + dX) - F(x)$, but we won't define this formally.

3.3 How to differentiate

Suppose we have some function $f(X)$ or $f(\mathbf{x})$ for $F(x)$ or whatever. How do we calculate the derivative? There is an "algorithm" of sorts for this:

1. Calculate the differential df or df or dF.
2. Manipulate the differentiatial into an appropriate form
3. Read off the result.

Remember, below, we will drop the argument of the differential, and write it simply as df, $d\mathbf{f}$ or dF.

input\output	scalar	vector	matrix
scalar	if $df = a(x)dx$ then $\frac{df}{dx} = a(x)$	if $d\mathbf{f} = \mathbf{a}(x)dx$ then $\frac{d\mathbf{f}}{dx} = \mathbf{a}(x)$	if $dF = A(x)dx$ then $\frac{dF}{dx} = A(x)$
vector	if $df = \mathbf{a}(\mathbf{x}) \cdot d\mathbf{x}$ then $\frac{df}{d\mathbf{x}} = \mathbf{a}(\mathbf{x})$	if $d\mathbf{f} = A(\mathbf{x})d\mathbf{x}$ then $\frac{d\mathbf{f}}{d\mathbf{x}^T} = A(\mathbf{x})$	
matrix	if $df = A(X) \cdot dX$ then $\frac{df}{dX} = A(X)$		

3.4 Traces and inner products

Traditionally, the matrix\scalar rule above is written in the (in my opinion less transparent) form

$$\text{if } df = \text{tr}(A(X)^{\cdot T} dX) \text{ then } \frac{df}{dX} = A(X). \tag{3.1}$$

This is explained by the following identity.

Theorem 6.

$$A \cdot B = tr(A^T B)$$

Proof.

$$\text{tr}(A^T B) \;=\; \sum_i (A^T B)_{ii} = \sum_i \sum_j A_{ji} B_{ji} = A \cdot B.$$

□

A crucial rule for the manipulation of traces is that the matrices can be "cycled". If one will use the rule in Eq. 3.1, this identity is crucial for getting differentials into the right form to read off the result.

$$\text{tr}(AB) = \text{tr}(BA)$$

$$\begin{aligned} \text{tr}(ABC...Z) &= \text{tr}(BC...ZA) \\ &= \text{tr}(C...ZAB). \end{aligned}$$

We can leverage this rule to derive some new results for manipulating inner products. First, we have results for three matrices.

Theorem 7. *If A, B and C are real matrices such that $A \cdot (BC)$ is well defined, then*

$$\begin{aligned} A \cdot (BC) &= B^T \cdot (CA^T) \\ &= B \cdot (AC^T) \\ &= C^T \cdot (A^T B) \\ &= C \cdot (B^T A) \end{aligned}$$

Proof.

$$\begin{aligned} A \cdot (BC) &= \mathrm{tr}(A^T BC) \\ &= \mathrm{tr}(BCA^T) \\ &= B^T \cdot (CA^T) \\ &= B \cdot (AC^T) \\ &= \mathrm{tr}(CA^T B) \\ &= C^T \cdot (A^T B) \\ &= C \cdot (B^T A) \end{aligned}$$

□

The way to remember this is that you can swap the position of A with either B or C, but then you transpose the one (of B or C) you didn't swap with.

Now, we consider results for four. For higher numbers of matrices, just put some together in a group and apply one of the above rules. (These can be proven in similar ways by converting to the trace form, permuting the matrices, and then converting back.)

Theorem 8.

$$\begin{aligned} A \cdot (BCD) &= C^T \cdot DA^T B \\ &= C \cdot B^T AD^T \end{aligned}$$

3.5 Rules for Differentials

In the following, α is a real constant, and f and g are real-valued functions (we write them as taking a matrix input for generality, but these rules also work if the input is scalar or vector-valued.), and $r : \mathbb{R} \to \mathbb{R}$.

$$
\begin{aligned}
d(\alpha) &= 0 \\
d(\alpha f) &= \alpha df \\
d(f+g) &= df + dg \\
d(f-g) &= df - dg \\
d(fg) &= df\, g(X) + f(X)\, dg \\
d(\frac{f}{g}) &= \frac{df\, g(X) - f(X)\, dg}{g^2(X)} \\
d(f^{-1}) &= -f^{-2}(X) df \\
d(f^{\alpha}) &= \alpha f^{\alpha-1}(X) df \\
d(\log f) &= \frac{df}{f(X)} \\
d(e^f) &= e^{f(X)} df \\
d(\alpha^f) &= \alpha^{f(X)} \log(\alpha)\, df \\
d(r(f)) &= r'(f(X))\, df
\end{aligned}
$$

On the other hand, suppose that A is a constant matrix and F and G are matrix-valued functions. There are similar, but slightly more complex rules.

$$
\begin{aligned}
d(A) &= 0 \\
d(\alpha F) &= \alpha dF \\
d(F+G) &= dF + dG \\
d(F-G) &= dF - dG \\
d(FG) &= dF\, G(X) + F(X)\, dG \\
d(AF) &= A\, dF \\
d(F^T) &= (dF)^T \\
d(F^{-1}) &= -F(X)^{-1}(dF)F(X)^{-1} \\
d(|F|) &= |F(X)|F(X)^{-1} \cdot dF \\
d(F \cdot G) &= d(F) \cdot G + F \cdot d(G) \\
d(A \cdot F) &= A \cdot d(F)
\end{aligned}
$$

Another important rule is for "elementwise" functions. Suppose that $R : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ is a function that operates elementwise. (e.g. exp or sin.) Then

$$d(R(F)) = R'(F(X)) \odot dF,$$

where $\odot$ denotes elementwise multiplication.

A useful rule for dealing with elementwise multiplication is that

$$\mathbf{x} \odot \mathbf{y} = \text{diag}(\mathbf{x})\mathbf{y}.$$

Also note that

$$\text{diag}(\mathbf{x})\mathbf{y} = \text{diag}(\mathbf{y})\mathbf{x},$$

and, similarly,

$$\mathbf{x}^T \text{diag}(\mathbf{y}) = \mathbf{y}^T \text{diag}(\mathbf{x}).$$

3.6 Examples

Lets try using these rules to compute some interesting derivatives.

Example 9. Consider the function $f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$. This has the differential

$$\begin{aligned} df &= d(\mathbf{x}^T A\mathbf{x}) \\ &= d((\mathbf{x}^T A)\mathbf{x}) \\ &= d(\mathbf{x}^T A)\,\mathbf{x} + (\mathbf{x}^T A)d(\mathbf{x}) \\ &= (d(A^T\mathbf{x}))^T\mathbf{x} + (\mathbf{x}^T A)d(\mathbf{x}) \\ &= (A^T\,d\mathbf{x})^T\mathbf{x} + (\mathbf{x}^T A)d\mathbf{x} \\ &= d(\mathbf{x}^T)A\mathbf{x} + \mathbf{x}^T A d\mathbf{x} \\ &= \mathbf{x}^T(A + A^T)d\mathbf{x} \\ &= (A + A^T)\mathbf{x} \cdot d\mathbf{x} \end{aligned}$$

From which we can conclude that

$$\frac{df}{d\mathbf{x}} = (A + A^T)\mathbf{x}.$$

Example 10. Consider the function $f(\mathbf{X}) = \mathbf{a}^T X\mathbf{a}$. This has the differential

$$\begin{aligned} df &= d(\mathbf{a}^T X\mathbf{a}) \\ &= d(\mathbf{a}^T X)\mathbf{a} + \mathbf{a}^T X d(\mathbf{a}) \\ &= \mathbf{a}^T dX\mathbf{a} \\ &= (\mathbf{a}\mathbf{a}^T) \cdot X \end{aligned}$$

This gives the result

$$\frac{df}{dX} = \mathbf{a}\mathbf{a}^T.$$

Example 11. Consider the function $\mathbf{f}(\mathbf{x}) = A\mathbf{x}$. This has the differential

$$d\mathbf{f} = Ad\mathbf{x}$$

From which we can conclude that

$$\frac{d\mathbf{f}}{d\mathbf{x}^T} = A.$$

Example 12. Consider the function

$$f(X) = |X|.$$

This has the differential

$$df = |X|X^{-1} \cdot dX.$$

whence we can conclude that

$$\frac{df}{dX} = |X|X^{-1}.$$

Example 13. Consider the function $f(X) = |X^T X|$. This has the differential

$$\begin{aligned} df &= |X^T X|(X^T X)^{-1} \cdot d(X^T X) \\ &= |X^T X|(X^T X)^{-1} \cdot (X^T dX + d(X^T)\, X) \\ &= 2|X^T X|(X^T X)^{-1} \cdot X^T\, dX \\ &= dX \cdot 2X|X^T X|(X^T X)^{-1}. \end{aligned}$$

and so we have that

$$\frac{df}{dX} = 2X|X^T X|(X^T X)^{-1}.$$

(I believe that Minka's notes contain a small error in this example.)

Example 14. Consider the function

$$f(\mathbf{x}) = e^{-\mathbf{x}^T A \mathbf{x}}$$

We have that

$$\begin{aligned} d(e^{-\mathbf{x}^T A \mathbf{x}}) &= e^{-\mathbf{x}^T A \mathbf{x}} d(-\mathbf{x}^T A \mathbf{x}) \\ &= -e^{-\mathbf{x}^T A \mathbf{x}} (A + A^T)\mathbf{x} \cdot d\mathbf{x} \\ \frac{df}{d\mathbf{x}} &= -e^{-\mathbf{x}^T A \mathbf{x}} (A + A^T)\mathbf{x}. \end{aligned}$$

Example 15. Consider the function

$$f(\mathbf{x}) = \sin(\mathbf{x} \cdot \mathbf{x}).$$

$$\begin{aligned} df &= \cos(\mathbf{x} \cdot \mathbf{x}) d(\mathbf{x} \cdot \mathbf{x}) \\ &= 2\cos(\mathbf{x} \cdot \mathbf{x})(\mathbf{x} \cdot d\mathbf{x}) \\ \frac{df}{d\mathbf{x}} &= 2\cos(\mathbf{x} \cdot \mathbf{x})\mathbf{x} \end{aligned}$$

Example 16. Consider the (Normal distribution) function

$$f(\mathbf{x}) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp(-\frac{1}{2}\mathbf{x}^T \Sigma^{-1} \mathbf{x}).$$

The differential is

$$\begin{aligned} df &= \frac{1}{|2\pi\Sigma|^{1/2}} \exp(-\frac{1}{2}\mathbf{x}^T \Sigma^{-1} \mathbf{x}) d(-\frac{1}{2}\mathbf{x}^T \Sigma^{-1} \mathbf{x}) \\ &= \frac{1}{|2\pi\Sigma|^{1/2}} \exp(-\frac{1}{2}\mathbf{x}^T \Sigma^{-1} \mathbf{x})(-\frac{1}{2})(\Sigma^{-1} + \Sigma^{-T})\mathbf{x} \cdot d\mathbf{x} \\ &= -f(\mathbf{x})\Sigma^{-1}\mathbf{x} \cdot d\mathbf{x} \end{aligned}$$

From which we can conclude that

$$\frac{df}{d\mathbf{x}} = -f(\mathbf{x})\Sigma^{-1}\mathbf{x}.$$

Example 17. Consider again the Normal distribution, but now as a function of Σ.

$$f(\Sigma) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}).$$

The differential is

$$df = \frac{1}{|2\pi\Sigma|^{1/2}} \exp(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}) d(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}) + d(\frac{1}{|2\pi\Sigma|^{-1}}) \exp(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}).$$

Let's attack the two difficult parts in turn.

$$\begin{aligned} d(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}) &= -\frac{1}{2}\mathbf{x}^T d(\Sigma^{-1})\mathbf{x} \\ &= \frac{1}{2}\mathbf{x}^T\Sigma^{-1}d\Sigma\Sigma^{-1}\mathbf{x} \\ &= \frac{1}{2}(\Sigma^{-1}\mathbf{x})(\Sigma^{-1}\mathbf{x})^T \cdot d\Sigma. \end{aligned}$$

Next, we can calculate

$$\begin{aligned} d(\frac{1}{|2\pi\Sigma|^{1/2}}) &= d(|2\pi\Sigma|^{-1/2}) \\ &= -\frac{1}{2}|2\pi\Sigma|^{-3/2} d(|2\pi\Sigma|) \\ &= -\frac{1}{2}|2\pi\Sigma|^{-3/2}|2\pi\Sigma|(2\pi\Sigma)^{-1} \cdot d(2\pi\Sigma) \\ &= -\frac{1}{2}\frac{1}{|2\pi\Sigma|^{1/2}}\Sigma^{-1} \cdot d(\Sigma) \end{aligned}$$

Putting this all together, we obtain that

$$\begin{aligned} df &= \frac{1}{|2\pi\Sigma|^{1/2}} \exp(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}) d(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}) + d(\frac{1}{|2\pi\Sigma|^{-1}}) \exp(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}). \\ &= \frac{1}{|2\pi\Sigma|^{1/2}} \exp(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}) \frac{1}{2}(\Sigma^{-1}\mathbf{x})(\Sigma^{-1}\mathbf{x})^T \cdot d\Sigma - \frac{1}{2}\frac{1}{|2\pi\Sigma|^{1/2}}\Sigma^{-1} \cdot d(\Sigma) \exp(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}). \\ &= \frac{1}{2}f(\Sigma)(\Sigma^{-1}\mathbf{x})(\Sigma^{-1}\mathbf{x})^T \cdot d\Sigma - \frac{1}{2}f(\mathbf{x})\Sigma^{-1} \cdot d(\Sigma) \\ &= \frac{1}{2}f(\Sigma)\big((\Sigma^{-1}\mathbf{x})(\Sigma^{-1}\mathbf{x})^T - \Sigma^{-1}\big) \cdot d\Sigma \end{aligned}$$

So, finally, we obtain

$$\frac{df}{d\Sigma} = \frac{1}{2} f(\Sigma)\big((\Sigma^{-1}\mathbf{x})(\Sigma^{-1}\mathbf{x})^T - \Sigma^{-1}\big).$$

Example 18. Consider the function

$$f(\mathbf{x}) = \mathbf{1}^T \exp(A\mathbf{x}) + \mathbf{x}^T B\mathbf{x}.$$

Let's compute the gradient and Hessian both. First the gradient.

$$\begin{aligned} df &= \mathbf{1}^T d(\exp(A\mathbf{x})) + (B+B)^T\mathbf{x} \cdot d\mathbf{x} \\ &= \mathbf{1}^T(\exp(A\mathbf{x}) \odot d(A\mathbf{x})) + (B+B)^T\mathbf{x} \cdot d\mathbf{x} \\ &= \mathbf{1}^T \mathrm{diag}(\exp(A\mathbf{x}))Ad\mathbf{x} + (B+B)^T\mathbf{x} \cdot d\mathbf{x} \\ &= \exp(\mathbf{A}\mathbf{x})^T \mathrm{diag}(\mathbf{1})Ad\mathbf{x} + (B+B)^T\mathbf{x} \cdot d\mathbf{x} \\ &= \exp(\mathbf{A}\mathbf{x})^T Ad\mathbf{x} + (B+B)^T\mathbf{x} \cdot d\mathbf{x} \\ &= A^T \exp(A\mathbf{x}) \cdot d\mathbf{x} + (B+B)^T\mathbf{x} \cdot d\mathbf{x} \end{aligned}$$

So, we have

$$\frac{df}{d\mathbf{x}} = A^T \exp(A\mathbf{x}) + (B + B^T)\mathbf{x}.$$

Now, what about the Hessian? The key idea is, define

$$\mathbf{g}(\mathbf{x}) = A^T \exp(A\mathbf{x}) + (B + B^T)\mathbf{x}.$$

Then,

$$\begin{aligned} d\mathbf{g} &= A^T d(\exp(A\mathbf{x})) + (B + B^T)d\mathbf{x}. \\ &= A^T(\exp(A\mathbf{x}) \odot d(A\mathbf{x})) + (B + B^T)d\mathbf{x}. \\ &= A^T \mathrm{diag}(\exp(A\mathbf{x}))Ad\mathbf{x} + (B + B^T)d\mathbf{x}. \end{aligned}$$

Hence, we have

$$\frac{d\mathbf{g}}{d\mathbf{x}^T} = \frac{d^2 f}{d\mathbf{x} d\mathbf{x}^T} = A^T \mathrm{diag}(\exp(A\mathbf{x}))A + (B + B^T).$$

3.7 Cheatsheet

input\output	scalar	vector	matrix
scalar	if $df = a(x)dx$ then $\frac{df}{dx} = a(x)$	if $d\mathbf{f} = \mathbf{a}(x)dx$ then $\frac{d\mathbf{f}}{dx} = \mathbf{a}(x)$	if $dF = A(x)dx$ then $\frac{dF}{dx} = A(x)$
vector	if $df = \mathbf{a}(\mathbf{x}) \cdot d\mathbf{x}$ then $\frac{df}{d\mathbf{x}} = \mathbf{a}(\mathbf{x})$	if $d\mathbf{f} = A(\mathbf{x})d\mathbf{x}$ then $\frac{d\mathbf{f}}{d\mathbf{x}^T} = A(\mathbf{x})$	
matrix	if $df = A(X) \cdot dX$ then $\frac{df}{dX} = A(X)$		

$$\begin{aligned} A \cdot (BC) &= B^T \cdot (CA^T) \\ &= B \cdot (AC^T) \\ &= C^T \cdot (A^T B) \\ &= C \cdot (B^T A) \end{aligned}$$

$$\begin{aligned} A \cdot (BCD) &= C^T \cdot DA^T B \\ &= C \cdot B^T A D^T \end{aligned}$$

$$\mathbf{x} \odot \mathbf{y} = \text{diag}(\mathbf{x})\mathbf{y}.$$

$$\text{diag}(\mathbf{x})\mathbf{y} = \text{diag}(\mathbf{y})\mathbf{x},$$

$$\mathbf{x}^T \text{diag}(\mathbf{y}) = \mathbf{y}^T \text{diag}(\mathbf{x}).$$

$$
\begin{aligned}
d(\alpha) &= 0 \\
d(\alpha f) &= \alpha df \\
d(f+g) &= df + dg \\
d(f-g) &= df - dg \\
d(fg) &= df\, g(X) + f(X)\, dg \\
d(\frac{f}{g}) &= \frac{df\, g(X) - f(X)\, dg}{g^2(X)} \\
d(f^{-1}) &= -f^{-2}(X) df \\
d(f^{\alpha}) &= \alpha f^{\alpha-1}(X) df \\
d(\log f) &= \frac{df}{f(X)} \\
d(e^f) &= e^{f(X)} df \\
d(\alpha^f) &= \alpha^{f(X)} \log(\alpha)\, df \\
d(r(f)) &= r'(f(X))\, df
\end{aligned}
$$

$$
\begin{aligned}
d(A) &= 0 \\
d(\alpha F) &= \alpha dF \\
d(F+G) &= dF + dG \\
d(F-G) &= dF - dG \\
d(FG) &= dF\, G(X) + F(X)\, dG \\
d(AF) &= A\, dF \\
d(F^T) &= (dF)^T \\
d(F^{-1}) &= -F(X)^{-1}(dF)F(X)^{-1} \\
d(|F|) &= |F(X)|F(X)^{-1} \cdot dF \\
d(F \cdot G) &= d(F) \cdot G + F \cdot d(G) \\
d(A \cdot F) &= A \cdot d(F)
\end{aligned}
$$

$$
d(R(F)) = R'(F(X)) \odot dF,
$$