Engsci 721 – Inverse Problems Supplement I

Oliver Maclaren (*oliver.maclaren@auckland.ac.nz*)

1 Overview

Some helpful facts on linear algebra and matrix calculus.

A lot is beyond the scope of this course, but by the end of the lecture material you should be able to answer the **'Test your understanding'** questions in each section. I might add another one on e.g. some basic probability theory later.

2 Spaces, norms, inner products etc

2.1 Mathematical spaces

To define continuity, distance, size etc in general mathematical spaces we need to abstract these concepts beyond their usual setting (e.g. beyond the real number line $\mathbb{R}$). This leads to ideas like (in decreasing order of generality) **topological spaces**, **metric spaces**, **vector spaces**, **normed vector spaces** and **inner product vector spaces**. Another common name for vector spaces is **linear spaces**. In general these 'spaces' can be thought of as **abstract sets with extra concrete structure**, usually introduced via functions on, or subsets of, these sets.

While we won't make much use of these distinctions, keeping in mind the available structure can be important. For example, in contrast to vector spaces, neither topological nor metric spaces come with any algebraic structure on the elements in general: it doesn't necessarily make sense to 'add' two elements of a general metric space together. For more detail on these spaces and their relationships, see e.g. courses on **functional analysis**.

In this course we work in the fairly familiar $\mathbb{R}^n$, which is a **normed vector space** (also an inner product space): we can add vectors together, multiply them by scalars and measure their 'size.' This latter notion is defined via the concept of a **norm**.

2.2 Norms - general definition

Given a vector space X, a **norm** is a function defined on elements $x \in X$ such that

$$
\begin{aligned}
&(1)\ \ ||x|| = 0 \text{ iff } x = 0 \\
&(2)\ \ ||ax|| = |a|\ ||x|| \ \forall x \in X, a \in \mathbb{R} \\
&(3)\ \ ||x + y|| \leq ||x|| + ||y|| \ \forall x, y \in X.
\end{aligned}
$$

Intuitively, a norm **measures the size of a vector**, and generalises the usual concept of **magnitude** $|a|$ for $a \in \mathbb{R}$ to $\mathbb{R}^n$.

2.3 Norms - examples

We consider the p-norms, also called L_p-norms, in $\mathbb{R}^n$:

$$
||x||_p = (|x_1|^p + ... + |x_n|^p)^{1/p}
$$

where $p \geq 1$. Particular cases include

- $p = 2$, the **Euclidean/$\mathbf{L_2}$ norm**.
- $p = 1$, the **taxicab/$\mathbf{L_1}$ norm**
- $p \to \infty$, the **max/$\mathbf{L_\infty}$ norm** where $||x||_\infty = \max_i |x_i|$

2.4 Norms - relationships

As discussed in class, there is a sense in which all norms in finite dimensional $\mathbb{R}^n$ are 'equivalent' and each can approximate the others. As n grows large, however, these approximations become much looser. Many difficulties of the infinite-dimensional case are inherited by the large finite n case due to e.g. numerical rounding.

Furthermore, the **solutions** to problems involving optimisation with respect to a norm are often **qualitatively different in character** even if they have **quantitatively similar sizes**. For example, minimising with respect to the L_1 norm tends to produce solutions with many **exactly zero** entries, while the L_2 norm produces entries with many **close-to-but-not-exactly-zero** entries.

- E.g. in an $\mathbb{R}^3$ problem we might get L_1 solutions of the form $(0,0,1)^T$ and L_2 solutions of the form $\frac{1}{\sqrt{3}}(1,1,1)^T$. Both have norm of 1 in their respective norms, but e.g. $||\frac{1}{\sqrt{3}}(1,1,1)^T||_1 = \sqrt{3} > ||(0,0,1)^T||_1 = 1$.

Thus if, for example, **true zeros** or **sparse** solutions are desired, the L_1 norm would be a more natural choice than the L_2 norm.

2.5 Inner products

The space $\mathbb{R}^n$ is also an **inner product** space. (If you're keeping track, inner product spaces are special cases of normed spaces, which are special cases of metric spaces, which are special cases of topological spaces…).

This means that we can take inner products $\langle x, y\rangle$ between vectors x, y, where an inner product $\langle \cdot, \cdot \rangle$ satisfies, for all x, y, z in an inner produce space X and $a, b \in \mathbb{R}$:

$$
\begin{aligned}
&(1)\quad \langle x, x\rangle \geq 0 \text{ and } \langle x, x\rangle = 0 \text{ iff } x = 0 \\
&(2)\quad \langle x, y\rangle = \langle y, x\rangle \\
&(3)\quad \langle ax + by, z\rangle = a\langle x, z\rangle + b\langle y, z\rangle
\end{aligned}
$$

Note that (2) and (3) imply that $\langle \cdot, \cdot \rangle$ is a **linear** function in both arguments (i.e. is bilinear).

Roughly, you can think of the inner product as first **projecting** one vector onto the other, giving something like two elements in $\mathbb{R}$ (i.e. two points lying along a common line), and then multiplying these two elements in the usual $\mathbb{R}$ way.

2.6 Inner products and dot products

In $\mathbb{R}^n$ we also have the **dot product** of two vectors x, y defined via

$$x \cdot y = x^T y = x_1 y_1 + ... + x_n y_n$$

where we take vectors to be **column vectors** by default. This leads to the **dot product version of the inner product**

$$\langle x, y\rangle = x \cdot y = x^T y$$

which we will assume by default. Note though that other inner products can be used.

2.7 Inner products and norms

Given an inner produce one can define the **inner product norm** via

$$||x|| = \langle x, x\rangle^{1/2}$$

We use this relationship often, especially in the form

$$||x||^2 = \langle x, x\rangle$$

e.g. when using the squared L_2 norm. Along with the previous note on dot products we have

$$||x||^2 = \langle x, x\rangle = x^T x$$

We will **usually convert norms to their expressions in terms of standard matrix/vector operations**, which we will learn to differentiate below.

2.8 Test your understanding of norms and inner products

- Use the usual norms and inner products on $\mathbb{R}^n$ to show that our least-squares objective function $||Ax-y||_2^2$ can be written as $x^TA^TAx - 2y^TAx - y^Ty$.
- Suppose you wanted to use a regularisation term that emphasised exactly piecewise linear solutions. This means you want small - and zero where possible - second derivatives, so small $||D_2x||_p^p$ for some p, and where D_2 is the second-order finite difference operator discussed in class. Which norm is appropriate here?
- In the above question, suppose you instead wanted 'smooth' solutions in the sense of having 'small but non-zero' second derivatives. Which norm is appropriate for this case?

3 Matrix calculus

The first subsection below establishes notation, conventions etc and touches on some of the more subtle aspects of vector, matrix and tensor calculus. Most of this is quite useful but not strictly needed – all you will actually need in this course is the **key rules** is the subsection after.

3.1 Conventions and background

We will assume all vectors are column vectors by default, but that e.g. derivatives of scalar-valued functions with respect to column vectors produce row vectors. Derivative operators will be denoted by d (occassionally D) with a subscript to indicate what the derivative is with respect to, e.g. d_xf denotes the derivative of f with respect to x. This is true regardless of whether f and x are scalars, vectors, matrices etc.

The convention that e.g. derivatives of scalar-valued functions with respect to column vectors produce row vectors is consistent with what is sometimes called the *numerator layout* or *Jacobian formulation*. The reason for the latter terminology is that it is consistent with the usual definition of the Jacobian d_xf when f is a *vector-valued* function of a vector x. This is usually taken to have components

$$[d_xf]_{ij} = \frac{\partial f_i}{\partial x_j}.$$

When f is scalar-valued, i.e. i only ranges over one value, then this reduces to a matrix with a *single row* with entries

$$[d_xf]_j = \frac{\partial f}{\partial x_j}.$$

It is common to introduce the *gradient* ∇_xf of a scalar-valued function f as a *column* vector by defining

$$\nabla_xf = (d_xf)^T$$

This relationship carries over to vector-valued f as well, so that the gradient matrix is the transpose of the derivative matrix (Jacobian). When e.g. setting a derivative to zero we should thus technically write either

$$d_xf = 0^T$$

or

$$\nabla_xf = 0,$$

where 0 is a column vector of zeros and 0^T is a row vector of zeros. Sometimes 'we' are lazy, though.

Another notation is fairly common in statistics and related areas is to write

$$d_xf = \frac{\partial f}{\partial x^T}$$

and

$$\nabla_xf = \frac{\partial f^T}{\partial x}$$

and so $(\frac{\partial f}{\partial x^T})^T = \frac{\partial f^T}{\partial x}$. This notation is suggestive in that it indicates e.g. that the Jacobian should be laid out so that it 'varies vertically' following the column vector f and 'varies horiztonally' following the row vector x^T, giving the usual Jacobian layout where $[d_x f]_{ij} = \frac{\partial f_i}{\partial x_j}$. Often the f^T part is dropped in the ∇ definition.

Finally, in physics and engineering, people often use tensor index notation. The distinction between column vectors and row vectors is usually made with upper or lower indices, e.g. x^i is a column vector and x_j is a row vector. A matrix (second order tensor) is written A^i_j and matrix multiplication $y = Ax$ is written

$$y^i = A^i_j x^j$$

where pairs of upper and lower indices are 'summed over.' The derivative is then written as e.g.

$$d_x f = \frac{\partial f}{\partial x^T} \leftrightarrow \frac{\partial f^i}{\partial x_j}.$$

A key result is that

$$\frac{\partial x^i}{\partial x_j} = \delta^i_j = \begin{cases} 0, \text{if } i \neq j, \\ 1, \text{if } i = j \end{cases}$$

i.e.

$$d_x x = \frac{\partial x}{\partial x^T} = I$$

.

Note however that this is not $d_x x^T$, which requires more careful treatment that we will avoid where possible. A tool that is useful for this is that the derivative of a *scalar* is the same as derivative of its transpose, since $a^T = a$ for scalars (this is not true for vectors!). This gives

$$d_x(a) = d_x(a^T)$$

if a is a scalar (possibly depending on x). For example,

$$d_x(x^T A x) = d_x(x^T A^T x)$$

since $x^T Ax$ is scalar-valued.

Importantly, matrix calculus obeys the multivariate chain rule in the form

$$d_x(f(g(x), h(x))) = d_g f d_x g + d_h f d_x h = \frac{\partial f}{\partial g^T}\frac{\partial g}{\partial x^T} + \frac{\partial f}{\partial h^T}\frac{\partial h}{\partial x^T}$$

.

The above two rules can be shown to imply the following form of the product rule for the inner product of vector-valued functions $g(x)$, $h(x)$ of a vector variable:

$$d_x(g^T h) = g^T d_x h + h^T d_x g.$$

Note the form of the second term in particular. This can be remembered via the rule

$$d_x(g^T h) = d_x(\bar{g^T} h)) + d_x(g^T \bar{h}))$$

where the overbar means 'treated as constant,' and using

$$d_x(g^T \bar{h})) = d_x(\bar{h^T} g)) = \bar{h^T} d_x g,$$

where the first equality follows since the bracket is scalar valued and the second from linearity, which avoids us trying to differentiate g^T.

3.2 Key rules

The three key rules of matrix calculus as far as we need are as follows. These can be shown/verified directly e.g. using index notation or more basic concepts like the multivariate chain rule, as in the lecture 2 notes, but **you can take them for granted**. Here x and y are vectors, A is a matrix and y and A are assumed to be independent of x.

3.2.1 Rule 1: differentiating a constant y (independent of x) wrt x

$$d_x(y) = 0^T.$$

Equivalently,

$$\nabla_x(y) = 0.$$

3.2.2 Rule 2: differentiating Ax and $\mathbf{y^T x}$ wrt x

$$d_x(Ax) = A.$$

Equivalently,

$$\nabla_x(Ax) = A^T$$

Treating y^T as a 1 by n matrix, we get

$$d_x(y^T x) = y^T.$$

Equivalently,

$$\nabla_x(y^T x) = y.$$

3.2.3 Rule 3: differentiating $\mathbf{x^T Ax}$ wrt x

$$\begin{aligned} d_x(x^T Ax) &= x^T(A + A^T) \\ &= 2x^T A \quad \text{if A is symmetric.} \end{aligned}$$

Equivalently,

$$\begin{aligned} \nabla_x(x^T Ax) &= (d_x(x^T Ax))^T \\ &= (A^T + A)x \\ &= 2Ax \quad \text{if A is symmetric.} \end{aligned}$$

3.3 Test your understanding of the key rules

- Derive the normal equations $A^T Ax = A^T y$ for least squares problems using matrix calculus. Hint: use what you learned in the previous section about how to write out $||Ax - y||^2$
- What is the derivative of $x^T x$? Hint: the identity matrix I is symmetric!

4 Tall/wide linear systems

4.1 Tall and wide definitions

- We call an $m \times n$ matrix A **tall** if $m > n$, i.e. it has more rows than columns.
- We call an $m \times n$ matrix A **wide** if $m < n$, i.e. it has more columns than rows.

4.2 Invertibility

- When the **columns** of a **tall** (or square) matrix are **linearly independent** then it has a **left** inverse (also called a *retraction*) $LA = I$. The converse is also true - the existence of a left inverse implies that the columns of a tall (or square) matrix are linearly independent.
- When the **rows** of a **wide** (or square) matrix are **linearly independent** then it has a **right** inverse (also called a *section*). The converse is also true - the existence of a right inverse implies that the rows of a wide (or square) matrix are linearly independent.
- It follows that a matrix has **both** a left inverse and a right inverse iff it is **square** with **linearly independent** rows and columns. In this case the left and right inverses are the same matrix, and just called the **inverse**.
- Any matrix A has a **generalised/pseudo-inverse** A^+. The most general algebraic property defining a generalised inverse A^+ is $AA^+A = A$. The (Moore-Penrose) pseudoinverse is a slightly more special case of a generalised inverse, and can be defined either by additional algebraic properties or via optimisation problems. In particular, given we want to 'solve' $Ax = y$ (even if a traditional unique solution does not exist), it is the mapping A^+ from any given y to a corresponding x solution, defined by first fitting the data in the least-squares sense, $\min_x \|y - Ax\|^2$, and then choosing the 'smallest' solution $\min_x \|x\|^2$.
- The generalised inverse is a left and/or right inverse under the relevant conditions given above for left and right inverses to exist, but if none hold then it is **not a true inverse in general**. Instead, in the case of the Moore-Penrose pseudoinverse anyway, and as mentioned above, it gives the least-squares data reduction, least-squares model reduction solution.

4.3 Test your understanding of tall/wide systems

- A typical least squares data approximation problem requires one to 'solve' $Ax = y$ when A is $m \times n$ with $m > n$ and (say) linearly independent columns. In what sense is this 'solvable?' What sort of inverse, if any, 'solves' this problem?
- A typical least squares model reduction problem requires one to 'solve' $Ax = y$ when A is $m \times n$ with $m < n$ and (say) linearly independent rows. In what sense is this 'solvable?' What sort of inverse, if any, 'solves' this problem?
- What are the expressions for the *data resolution* R_D and *model (parameter) resolution* R_M (or R_θ etc) operators in terms of A and A^+? In the two 'least squares' problems above (data approximation, model reduction), which of R_D and R_M do you expect to be (near) identity and which not?

5 SVD

- The (full) **SVD** of a matrix A, with dimensions $m \times n$, is given by $A = U\Sigma V^T$ where U is the $m \times m$ matrix with orthonormal columns consisting of **left** singular vectors, V is the $n \times n$ matrix with orthonormal columns consisting of **right** singular vectors and Σ is a $m \times n$ 'diagonal' matrix of singular values (i.e. only the so-called 'main diagonal' has non-zero entries). This has r positive entries, where r is the **rank** of A.
- The **reduced SVD** of a matrix A is given by $A = U_r\Sigma_r V_r^T$ where r is the **rank** of the matrix, U_r is the $m \times r$ matrix consisting of the **first** r **columns** of U and V_r is the $n \times r$ matrix consisting of the **first** r **columns** of V. Σ_r is the $r \times r$ matrix of positive singular values.
- Since the rank $r \leq \max\{m, n\}$, the matrices U_r (dimensions $m \times r$) and V_r (dimensions $n \times r$) are **both tall or square** (i.e. never wide).
- By construction the **columns** of both U_r and V_r are **linearly independent**.
- The two previous points imply that U_r and V_r both **always have left inverses**, though not necessarily right inverses. Their left inverses are given by $L_U = U^T$ and $L_V = V^T$ respectively.
- When rank $r = m$ then $U_r = U$ and U^T is both a left and a right inverse for U.
- When rank $r = n$ then $V_r = V$ and V^T is both a left and a right inverse for V.

5.1 Test your understanding of the SVD

- Sketch out what the matrices for the **full** and the **reduced** SVD look like and hence indicate their relationship.
- Derive general expressions for the **resolution operators** in terms of U_r and V_r. Why is it important in this derivation that left inverses always exist for U_r and V_r?
- When rank $r = m$ which **resolution operator** is the **identity**? What does the other do? Note that $n \geq r = m$ so here the original matrix A is **wide**.
- When rank $r = n$ which **resolution operator** is the **identity**? What does the other do? Note that $m \geq r = n$ so here the original matrix A is **tall**.