Handout 03: Vector and Matrix Computations

Last time we started working with simple linear regression models that have only a single characteristic x that we can use to predict some response y. As you might imagine, most interesting examples have significantly more data inputs. To represent these, and to derive a comparable derivation in the multivariate case, we need some matrix theory and vector calculus. Today we are going to review (and in a few cases even derive the basic properties) of both today. If the material is completely new to you, I suggest taking some time this week to re-view the material again before next class.

Vectors

For us, a vector is simple an ordered collection of real numbers. The number of terms in the collection is called the *dimension* of the vector and we write $v \in \mathbb{R}^n$ to represent an n-dimensional vector v. You can think of the vector as a column of number:

$$v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \in \mathbb{R}^n. \tag{3.1}$$

We use the notation v_i to refer to the i'th term in the vector. Vector addition is defined *componentwise* such that:

$$v + u = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} v_1 + u_1 \\ v_2 + u_2 \\ \vdots \\ v_n + u_n \end{bmatrix} \tag{3.2}$$

Only vectors of the same dimension can be added together. We can also multiply a vector by a fixed scalar value $\alpha \in \mathbb{R}$ in a component-wise fashion:

$$\alpha \cdot v = \alpha \cdot \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} \alpha \cdot v_1 \\ \alpha \cdot v_2 \\ \vdots \\ \alpha \cdot v_n \end{bmatrix}. \tag{3.3}$$

It is possible to define a component-wise multiplication of two vectors, but this is rarely very useful and we will skip this for today.

We can also describe the size of a vector by thinking of the distance between the set of numbers in n-dimensional space and the origin. Specifically, the Euclidean-norm

(or ℓ_2-norm) of a vector is given and defined by:

$$||v||_2 = \sqrt{\sum_i v_i^2}. \tag{3.4}$$

This should correspond with other definitions you may have seen for distance measures. Finally, we will also define the inner product between two vectors of the same dimension as:

$$u \cdot v = \sum_i u_i v_i. \tag{3.5}$$

This form will be most useful for us, but its helpful to also visualize the dot product geometrically by its equivalent form:

$$u \cdot v = ||u||_2 \cdot ||v||_2 \cdot cos(\theta) \tag{3.6}$$

For the angle θ between the two vectors; of particular note, the dot product is zero for perpendicular vectors. Note that the Euclidean-norm can be defined by using the dot product of a vector with itself:

$$v \cdot v = \sum_i v_i v_i = ||v||_2^2 \tag{3.7}$$

There is a lot of other very interesting and useful geometric intuition behind these definitions that we don't have time to get into right now. Hopefully some of these will arise as we work through the next few weeks and you will see how they apply to linear regression theory.

Gradient

Assume that we have a real valued function f defined on n-dimensional vectors. In other words:

$$f : \mathbb{R}^n \to \mathbb{R}. \tag{3.8}$$

The gradient of f, denoted by ∇f, is given by the vector of partial derivatives with respect to each component:

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial v_1} \\ \frac{\partial f}{\partial v_2} \\ \vdots \\ \frac{\partial f}{\partial v_n} \end{bmatrix} \tag{3.9}$$

As with first derivatives, we can use the gradient to find the critical point of a multi-valued function. Understanding gradient is very important for statistical learning.

A typical workflow consists of computing the gradient of the loss function, $\nabla\mathcal{L}$, trying to set this to zero, and then evaluating the output. In today's lab you will work on deriving several important properties of the gradient function.

Matrices

Consider a function that takes as an input vectors of dimension n and returns as an output vectors of dimension m:

$$M : \mathbb{R}^m \to \mathbb{R}^n. \tag{3.10}$$

We say that M is a linear function if we can take scalar quantities outside of the function,

$$M(\alpha \cdot x) = \alpha \cdot M(x), \quad x \in \mathbb{R}^m, \alpha \in \mathbb{R}, \tag{3.11}$$

And we can split vector sums across the function,

$$M(x + y) = M(x) + M(y), \quad x, y \in \mathbb{R}^n. \tag{3.12}$$

It turns out that any such map can be described a grid of numbers with n rows and m columns:

$$M = \begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,m} \\ m_{2,1} & \ddots & \cdots & m_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n,1} & m_{n,2} & \cdots & m_{n,m} \end{bmatrix} \tag{3.13}$$

By defining:

$$M(v)_j = \sum_i m_{i,j} \cdot v_i \in \mathbb{R}. \tag{3.14}$$

You can think of this as taking the dot product of the j'th row of the matrix M and the input vector v. Notice that we are abusing notation by letting M be the grid of numbers *and* the function. This is intentional because we will use the notation:

$$M(v) = Mv \in \mathbb{R}^n, \quad v \in \mathbb{R}^m. \tag{3.15}$$

To represent the action of applying the function described by a matrix M to a vector v.

As with vectors, we could spend a whole year just talking about matrices. Rather than an exhaustive treatment, I want to instead quickly describe a few properties and notations that we will most useful. First, matrix multiplication is defined by function composition. The matrix product $A \cdot B$ is defined as the matrix that corresponds to applying the linear function defined by B and then applying the linear function implied by A. Note that this can only be defined with the number of

columns in A matches the number of rows in B (why?). If we set $C = A \cdot B$, then the following formula corresponds to this functional interpretation:

$$c_{i,j} = \sum_k a_{i,k} \cdot b_{k,j} \tag{3.16}$$

Where lower case letters refer to the elements in the corresponding uppercase matrices. The matrix product distributes,

$$A(B + C) = AB + AC, \tag{3.17}$$

But in general does not commute,

$$AB \neq BA. \tag{3.18}$$

The identity matrix I_n is given by ones on the diagonal and zeros elsewhere. For example

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3.19}$$

For a square matrix A we have:

$$AI_n = I_n A = A. \tag{3.20}$$

Finally, the matrix inverse A^{-1} of a square matrix is defined such that

$$A^{-1}A = AA^{-1} = I_n, \tag{3.21}$$

However the matrix A^{-1} is not guaranteed to exist.

The final notation we need is the matrix transpose, denoted by A^t and defined simply as flipping the matrix rows and columns. It has the property that it can be distributed within a summation:

$$(A + B)^t = A^t + B^t. \tag{3.22}$$

The transpose of a matrix product can also be distributed but the order of the matrices is flipped:

$$(AB)^t = B^t A^t. \tag{3.23}$$

The transpose is quite useful because we can use it to compute the dot product in an interesting way, as you will see in today's lab.

LAB QUESTIONS

1. Consider the vectors:

$$v = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad u = \begin{bmatrix} 2 \\ 4 \\ 8 \end{bmatrix}. \tag{3.24}$$

   Compute the values (a) $2 \cdot u$, (b) $u + v$, (c) the dot product $u \cdot v$, and (d) the squared norm $||v||_2^2$.

2. Take the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined by:

$$f(x, y, z) = x^2 + x \cdot y + sin(y) \cdot z. \tag{3.25}$$

   Write down the function for ∇f.

3. Take the matrix A defined as:

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 3 & 0 & 1 \end{bmatrix} \tag{3.26}$$

   What is the dimension of the output Av using v from question 1? Compute the value Av.

4. Take the matrix B defined as:

$$B = \begin{bmatrix} 2 & 3 \\ 7 & 1 \end{bmatrix}. \tag{3.27}$$

   Compute the value of $C = B \cdot A$.

5. Let $v \in \mathbb{R}^n$ be a vector of dimension n. We can actually view this vector as a matrix with 1 column and n rows. Similarly, v^t is a vector with n columns and 1 row. What is the value of $v^t v$ in terms of the components v_i? Can you write this in terms of the Euclidean-norm?

6. Consider the function g defined by:

$$g(b) = b^t x. \tag{3.28}$$

   For $b, x \in \mathbb{R}^n$. Note that we consider b to be variable but x to be fixed. Compute the partial derivative

$$\frac{\partial g}{\partial b_k}. \tag{3.29}$$

7. Write the gradient $\nabla_b g$ (the subscript just reinforces that this is a function of b and not x) in a compact format. In other words, do not just write the gradient component-wise.

8. Let U be a square matrix such that $Q^t = Q^{-1}$. We can any matrix with this property an *orthogonal* matrix. An orthogonal matrix corresponds to a linear map that is simply a rotation of an n-dimensional space, a property that we will try to get some insight on here. Show that the Euclidean-norm is unchanged when applying Q. In other words, show that:

$$||v||_2^2 = ||Qv||_2^2. \tag{3.30}$$

Then, show that the dot product between Qv and Qu is the same as the dot product between u and v. Use these two properties to argue that the action of Q appears to behave like a rotation in n-dimensional space.

9. Take a square matrix A such that $A^t = A$ and define the function g where:

$$g(b) = b^t A b. \tag{3.31}$$

We will assume that A is fixed and only b is variable. Convince yourself that:

$$g(b) = \sum_i \sum_j a_{i,j} \cdot b_i \cdot b_j. \tag{3.32}$$

10. Now, compute the partial derivative

$$\frac{\partial g}{\partial b_k}. \tag{3.33}$$

Note that is quite a bit more difficult that the other gradient questions I asked so please be careful. Also note that you can look ahead one question to check your answer.

11. Finally, take the answer to your last question the prove that

$$\nabla_b g(b) = \nabla_b \left(b^t A b\right) = 2Ab. \tag{3.34}$$

12. (Extra) If we remove the assumption that A is symmetric (i.e., $A = A^t$) then the more general form is given by:

$$\nabla_b g(b) = \nabla_b \left(b^t A b\right) = Ab + A^t b. \tag{3.35}$$

Try to prove this more general form as well.