Module 2 Review Notes: Mathematical Review

Neural Networks

Dr. Mark Fleischer

2.1 Linear Algebra

Notation:

Summation: $$\sum_{j=1}^{n} a_{ij} = a_{i1} + a_{i2} + \cdots + a_{in}$$

Product: $$\prod_{j=1}^{n} a_{ij} = a_{i1} \cdot a_{i2} \cdot \ldots \cdot a_{in}$$

If the row vector $\vec{a} = (a_1, a_2, \ldots, a_n)$, then $\vec{a}^T = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$.

Vector Addition:

Given two vectors **a** and **b** *of the same size,* ***a*** **+ b =** ($a_1 + b_1$, $a_2 + b_2$, …, $a_n + b_n$)
For subtraction, analogous with the + substituted with a -.

Vector Multiplication (the inner product, the dot product)

$\vec{a} \cdot \vec{b} = (\vec{a}, \vec{b}) = \left\langle \vec{a}, \vec{b} \right\rangle = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$ = some scalar quantity.

In Matrix-Vector form we can write the inner product as

$$\mathbf{a}^T \mathbf{b} = (a_1, a_2, \ldots, a_n) \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

If this scalar = 0, that is, if (*a*, *b*) = 0, when both *a* and *b* are non-zero vectors (what is a non-zero vector?) (a zero vector is when *all* elements are 0 --- a non-zero vector is the negation of a zero vector —when not all elements are 0, *i.e.,* there exists some element that is non-zero), then the vectors *a* and *b* are said to be *orthogonal.*

The Outer Product:

$$\mathbf{a}^T\mathbf{b} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} (b_1, b_2, \ldots, b_n) = \begin{pmatrix} a_1b_1 & a_1b_2 & \cdots & a_1b_n \\ a_2b_1 & a_2b_2 & \cdots & a_2b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_nb_1 & a_nb_2 & \cdots & a_nb_n \end{pmatrix}$$

Linear Independence

A set of vectors $\mathbf{v}_i$ i = 1, …, *n* are said to be *linearly independent* if and only if

$\sum_{i=1}^{n} a_i \mathbf{v}_i = \mathbf{0}$ and a_i = 0 for all i. Provide orthonormal example (1,0,0,0), (0,1,0,0), etc.

Vector Magnitude (Length):

Desirable properties of a "length"(magnitude or norm):

1. Positivity: //**y**// ≥ 0 for all y and //**y**// = 0 if and only if **y** = **0**.
2. Homogeneity: //*c***y**// = |*c*| //**y**// for all scalars *c* and vectors **y**.
3. The Triangle Inequality: //**x** + **y**// ≤ //**x| |** + ||**y**// for all vectors **x** and **y**.

The following definitions satisfy these requirements:

$$|\vec{a}| \equiv \|\vec{a}\| = \left[\sum_{i=1}^{n} a_i^2\right]^{1/2} = (\vec{a},\vec{a})^{1/2}$$ This is called the *Euclidean Norm*.

There are other, useful (depending on the context) norms:

$$|\vec{a}|_p \equiv \|\vec{a}\|_p = \left[\sum_{i=1}^{n} a_i^p\right]^{1/p} = (\vec{a},\vec{a})_p^{1/p}$$ this is the *p*-norm.

Finally, another one (the Fleischer norm):

Let **y** be an *n* vector such that

$\mathbf{y} = \sum_{\mathrm{i}=1}^{n} \alpha_i \mathbf{v}_i$ for some set of *n* linearly independent vectors **v**$_i$. Then

$\|\mathbf{y}\|_\mathbf{V} \equiv \sum_{i=1}^{n} |\alpha_i|$ is a norm with respect to the matrix (**v**$_1$| **v**$_2$| $\cdots$|**v**$_n$).

Distance Between Vectors

Define the distance between two points denoted by vectors as:

$$\left\|\vec{a}-\vec{b}\right\| = \left[\sum_{i=1}^{n}(a_i - b_i)^2\right]^{1/2} = \left(\vec{a}-\vec{b},\vec{a}-\vec{b}\right)^{1/2}$$

Also, note that $\cos\theta = \frac{\langle \mathbf{a},\mathbf{b}\rangle}{\|\mathbf{a}\|\|\mathbf{b}\|}$. Now why is that? What are some properties that make this make sense?

Matrix Types

A matrix **A** is said to be positive definite if for all non-zero vectors **x**, $\mathbf{x}^T\mathbf{A}\mathbf{x} > \mathbf{0}$. A matrix **A** is said to be positive semi-definite if for all non-zero vectors **x**, $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq \mathbf{0}$.

2.2 Partial Derivatives

$$\frac{\partial f(\mathbf{x})}{\partial x_1} = \lim_{h\to 0}\frac{f(x_1+h, x_2, \ldots, x_n) - f(x_1, x_2, \ldots, x_n)}{h}$$

Example:

If $f(x,y) = ax^2+bxy+cx+dy+ey^3+g$, then

$$\frac{\partial f(x,y)}{\partial x} = 2ax + by + c$$ and

$$\frac{\partial f(x,y)}{\partial y} = bx + d + 3ey^2$$

2.3 Gradient Vectors

We define the analogy of a derivative in vector spaces using the gradient vector defined by

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_n}\right)$$

2.4 The Directional Derivative

Consider a multivariable function $f(x_1, x_2; \ldots x_n) = f(\mathrm{x})$ and we want to find its derivative. We all know that we can write the *partial* derivatives by holding all the variables constant except for the one we are taking the derivative with respect to as in the examples above. In this case, it is as if we are moving, taking a step, in the direction of that particular variable. What if we move in a different direction that is not along a particular axis? How can we formulate a derivative expression for a given direction?

Example:

Let $f(x,y,z) = 2x^3y - 3y^2z$ and we are considering the directional derivative at point **A** = (1, 2, -1) in the direction of point **B** = (3, -1, 5). We can define the *direction vector* **d = B – A** and this is the vector (2, -3, 6). Thus, we "move" in the direction **d** = (2, -3, 6) from point **A** = (1, 2, -1). We can normalize **d** by defining another vector in the same direction but with unit length. **How can we do this?**

$||\mathbf{d}|| = (2^2 + (-3)^2 + 6^2)^{1/2} = (4 + 9 + 36)^{1/2} = (49)^{1/2} = 7$. Thus, redefine the direction vector as a vector of unit length in the given direction. We get $\mathbf{d} = \left(\frac{2}{7}, -\frac{3}{7}, \frac{6}{7}\right)$ and $||\mathbf{d}|| = 1$. Now lets "parameterize" this direction vector to get $\mathbf{d}(t) = \left(\frac{2}{7}t, -\frac{3}{7}t, \frac{6}{7}t\right) = t\mathbf{d}$. Thus, now $\boldsymbol{x(t) = 2t/7}$, similarly for y and z. *Using* the basic definition of a derivative we get the following definition for a *directional derivative*:

$$\frac{d\,f(x,y,z)}{d\mathbf{d}} = \lim_{t\to 0+} \frac{f(x+\frac{2}{7}t, y+\frac{-3}{7}t, z+\frac{6}{7}t) - f(x,y,z)}{t}$$

Chain Rule for Partial Derivatives

This is based very directly on the chain rule for derivatives of single variable functions. One can derive this expression by decomposing the total change in a function from moving in one direction.

$$\begin{aligned} \frac{d\,f(x,y,z)}{d\mathbf{d}} &= \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt} + \frac{\partial f}{\partial z}\frac{dz}{dt} \\ &= \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right) \cdot \left(\frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt}\right) \\ &= \nabla f(x,y,z) \cdot \mathbf{d} \end{aligned}$$

Now we can calculate the directional derivative —the rate change in $f(\mathbf{x})$ per unit change in direction **d**. First, we calculate the value of $\nabla f(x,y,z)$—the gradient vector at point A:

$$\begin{aligned} \nabla f(x,y,z) &= \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right) \\ &= \left(6x^2y, 2x^3 - 6yz, -3y^2\right) \end{aligned}$$

Which at the point (1 ,2, -1) yields $\nabla f(1,2,-1) = (12,14,-12)$. Thus, the directional derivative is

$$\nabla f(x,y,z) \cdot \mathbf{d} = (12,14,-12) \times \left(\tfrac{2}{7}, \tfrac{-3}{7}, \tfrac{6}{7}\right) = -\tfrac{90}{7}.$$

2.5 The Direction of Steepest Descent

It turns out that this directional derivative can be expressed as follows:

$$\nabla f(x,y,z)\cdot \mathbf{d} = \|\nabla f(x,y,z)\| \times \|\mathbf{d}\| \times \cos\theta$$

where θ is the angle between the gradient vector $\nabla f(x,y,z)$ and the direction vector **d**. This is maximized when $\cos\theta = 1$, *i.e.,* when θ = 0. Thus, when the angle between the gradient vector and the direction vector is 0, we have the largest change in *f*. In other words, if the direction vector **d** is the same as the gradient vector, we get the biggest bang per buck! Thus, the direction of steepest descent/ascent is the same as the gradient vector.

2.6 Taylor's Theorem

For single variables we have

$$\begin{aligned} f(x) &= a_0 + a_1(x-x_0) + a_2(x-x_0)^2 \ldots \\ &= f(x_0) + (x-x_0)f'(x_0) + \frac{1}{2!}(x-x_0)^2 f''(x_0)\ldots \end{aligned}$$

This same approach may be taken with vectors and even vector-valued functions.
Thus,

$$\begin{aligned} f(\mathrm{x}) &= a_0 + a_1(\mathrm{x}-\mathrm{x}_0) + a_2(\mathrm{x}-\mathrm{x}_0)^2 \ldots \\ &= f(\mathrm{x}_0) + (\mathrm{x}-\mathrm{x}_0)\nabla f(\mathrm{x}_0) + \frac{1}{2!}(\mathrm{x}-\mathrm{x}_0)^{\mathrm{T}}\mathbf{H}(\mathrm{x}-\mathrm{x}_0)\ldots \end{aligned}$$

2.7 First Order Necessary Conditions

The First Order Necessary Conditions refer to a set of conditions that indicate that a point in *n* dimensional space is a local minimum (for a maximum it is easy to determine equivalent conditions). For functions of one variable, recall that a local min or max occurs when the derivative is 0. In the context of n variables, the analogous condition is $\nabla f(\mathbf{x}) = \mathbf{0} = (0,0,\ldots,0)$ *i.e.,* when all partial derivatives are 0.

Another way of stating this same idea is for all feasible directions **d**, $\nabla f \cdot \mathbf{d} \geq 0$. This means that moving in any direction **d**, the directional derivative is always nonnegative. In other words, the function *f* increases at some positive rate no matter what direction you move in. It is analogous to being at the bottom of some valley or at the south pole—moving in any direction will send you north.

Finally, a third way of stating this is that if $\mathbf{x}^*$ is a local min and *f*(**x**) is differentiable at $\mathbf{x}^*$, then for all directions **d**, there exists a *t*' > 0 such that for all $0 < t < t'$; $f(\mathbf{x}^*) < f(\mathbf{x}^* + t\mathbf{d})$.
Stated this way, moving in any direction some non-zero amount will *increase* the objective function value.

2.8 Second Order Sufficiency Conditions

So, when is a local minimum a global minimum? We know from the calculus for functions of a single variable that a local is a global if the *second* derivative is always positive. This means that any tangent line or slope or derivative is always increasing. These conditions can be stated in several equivalent ways (for minimization problems):

1. All second derivatives are positive.
2. Any points in a tangent plane or hyper-plane have objective functions less than or equal to the objective function value.
3. The objective function is convex.

These three elements can be stated mathematically:

1. The Hessian Matrix $\mathbf{H}(\mathbf{x})$ is positive semi-definite, *i.e.,* for all R^{n}, $\mathbf{x}^{\mathbf{T}}\mathbf{H}\mathbf{x} >= 0$:

2. $\forall \mathbf{x}, \mathbf{x}^* \in R^n$,

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*).$$

3. $\forall \mathbf{x}, \mathbf{x}^* \in R^n$ and $0 \leq \lambda \leq 1$,

$$\lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{x}^*) \geq f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{x}^*).$$

An Example

Suppose we want to minimize the function: $f(x_1, x_2) = x_1 x_2 - 6x_1 - 3x_2 + 18$.
Taking all partial derivatives, we see that:

$$\frac{\partial f}{\partial x_1} = x_2 - 6 = 0$$

$$\frac{\partial f}{\partial x_2} = x_1 - 3 = 0$$

Therefore, $x_1 = 3$; $x_2 = 6$ for a point where the partial derivatives are all zero. This may correspond to a local max or min. To show that it is a minimum, it is *sufficient* to show that one of the three second-order sufficiency conditions holds. Using the condition in item 2, we see that the equation for the tangent plane at say point (4,3) is:

$$\begin{aligned} f_P(x_1, x_2) &= f(4,3) + \nabla f(4,3) \begin{pmatrix} x_1 - 4 \\ x_2 - 3 \end{pmatrix} \\ &= -3 + (-3, 1) \begin{pmatrix} x_1 - 4 \\ x_2 - 3 \end{pmatrix} \\ &= -3 - 3(x_1 - 4) + 1(x_2 - 3) \\ &= -3x_1 + x_2 + 6 \end{aligned}$$

Thus, the equation of the tangent plane at point (4,3) is $f_P\,(x_1, x_2) =$ $-3\,x_1 + x_2 + 6$. Now consider any point (x_1, x_2) and compare the value of f_P with the objective function value f. For example, at point (0,0) the value of $f_P = 6$. For the objective function, $f(0; 0) = 18$. Since the relationship between the objective function and tangent plane hold as item 2 above, it *suggests* that the second-order conditions hold. To establish this however requires that we prove this relation holds for *all* points (x_1, x_2). Can you prove that they do?