VMLS - Study Guide Chapter 2 Name: Ben Chen

Always include this title page with your PDF. Include your name above.

- Submit your work in Gradescope as a PDF - you will identify where your “questions are.”
- Identify the question number as you submit. Since we grade "blind" if the questions are NOT identified, the work WILL NOT BE GRADED and a 0 will be recorded. Always leave enough time to identify the questions when submitting.
- One section per page (if a page or less) - We prefer to grade the main solution in a single page, extra work can be included on the following page.
- Long instructions may be removed to fit on a single page.
- **Do not start a new question in the middle of a page.**
- Solutions to book questions are provided for reference.
- You may NOT submit given solutions - this includes minor modifications - as your own.
- Solutions that do not show individual engagement with the solutions will be marked as no credit and can be considered a violation of honor code.
- If you use the given solutions you must reference or explain how you used them, in particular...

For full credit, EACH book exercise in the Study Guides must use one or more of the following methods and FOR EACH QUESTION. Identify the number the method by number to ensure full credit.

Method 1 - Provide original examples which demonstrate the ideas of the exercise in addition to your solution.

Method 2 - Include and discuss the specific topics needed from the chapter and how they relate to the question.

Method 3 - Include original Python code, of reasonable length (as screenshot r text) to show how the topic or concept was explored.

Method 4 - Expand the given solution in a significant way, with additional steps and comments. All steps are justified. This is a good method for a proof for which you are only given a basic outline.

Method 5 - Attempt the exercise without looking at the solution and then the solution is used to check work. Words are used to describe the results.

Method 6 - Provide an analysis of the strategies used to understand the exercise, describing in detail what was challenging, who helped you or what resources were used. The process of understanding is described.

1. (10 pts) Select one page or section of Chapter Two of VMLS to annotate. Include a screenshot of your annotation here.

2.2 Taylor approximation

(Affine approximation... basically)

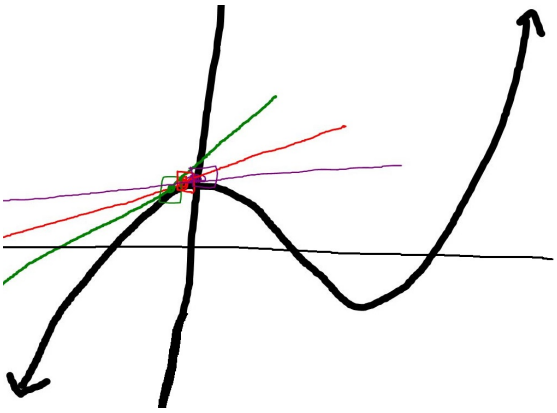
In many applications, scalar-valued functions of n variables, or relations between n variables and a scalar one, can be *approximated* as linear or affine functions. In these cases we sometimes refer to the linear or affine function relating the variables and the scalar variable as a *model*, to remind us that the relation is only an approximation, and not exact.

Differential calculus gives us an organized way to find an approximate affine model. Suppose that $f : \mathbf{R}^n \to \mathbf{R}$ is differentiable, which means that its partial derivatives exist (see §C.1). Let z be an n-vector. The (first-order) *Taylor approximation* of f near (or at) the point z is the function $\hat{f}(x)$ of x defined as

where linear regression comes from?

$$\hat{f}(x) = f(z) + \frac{\partial f}{\partial x_1}(z)(x_1 - z_1) + \cdots + \frac{\partial f}{\partial x_n}(z)(x_n - z_n),$$

where $\frac{\partial f}{\partial x_i}(z)$ denotes the partial derivative of f with respect to its ith argument, evaluated at the n-vector z. The hat appearing over f on the left-hand side is a common notational hint that it is an approximation of the function f. (The approximation is named after the mathematician Brook Taylor.)



This is what I imagine (visually) what a taylor approximation probably looks like. It looks like a bounding series of boxes around to approximate a curve. And it does this by deriving the slope at that given point to turn it into a linear function (lowering the exponent). I'm guesisng this is similar (fundamentally) to bounding boxes in computer vision.

The first-order Taylor approximation $\hat{f}(x)$ is a very good approximation of $f(x)$ when all x_i are near the associated z_i. Sometimes $\hat{f}$ is written with a second vector argument, as $\hat{f}(x; z)$, to show the point z at which the approximation is developed. The first term in the Taylor approximation is a constant; the other terms can be interpreted as the contributions to the (approximate) change in the function value (from $f(z)$) due to the changes in the components of x (from z).

Evidently $\hat{f}$ is an affine function of x. (It is sometimes called the *linear approximation* of f near z, even though it is in general affine, and not linear.) It can be written compactly using inner product notation as

regressive! as predicted

$$\hat{f}(x) = f(z) + \nabla f(z)^T(x - z), \tag{2.5}$$

where $\nabla f(z)$ is an n-vector, the *gradient of* f (at the point z),

this is the change between different points in estimation/approximation?

$$\nabla f(z) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(z) \\ \vdots \\ \frac{\partial f}{\partial x_n}(z) \end{bmatrix}. \tag{2.6}$$

The first term in the Taylor approximation (2.5) is the constant $f(z)$, the value of the function when $x = z$. The second term is the inner product of the gradient of f at z and the *deviation* or *perturbation* of x from z, *i.e.*, $x - z$.

We can express the first-order Taylor approximation as a linear function plus a constant,

$$\hat{f}(x) = \nabla f(z)^T x + (f(z) - \nabla f(z)^T z),$$

but the form (2.5) is perhaps easier to interpret.

The first-order Taylor approximation gives us an organized way to construct an affine approximation of a function $f : \mathbf{R}^n \to \mathbf{R}$, near a given point z, when there is a formula or equation that describes f, and it is differentiable. A simple example, for $n = 1$, is shown in figure 2.3. Over the full x-axis scale shown, the Taylor approximation $\hat{f}$ does not give a good approximation of the function f. But for x near z, the Taylor approximation is very good.

Example. Consider the function $f : \mathbf{R}^2 \to \mathbf{R}$ given by $f(x) = x_1 + \exp(x_2 - x_1)$, which is not linear or affine. To find the Taylor approximation $\hat{f}$ near the point $z = (1, 2)$, we take partial derivatives to obtain

$$\nabla f(z) = \begin{bmatrix} 1 - \exp(z_2 - z_1) \\ \exp(z_2 - z_1) \end{bmatrix},$$

which evaluates to $(-1.7183, 2.7183)$ at $z = (1, 2)$. The Taylor approximation at $z = (1, 2)$ is then

$$\begin{aligned} \hat{f}(x) &= 3.7183 + (-1.7183, 2.7183)^T(x - (1, 2)) \\ &= 3.7183 - 1.7183(x_1 - 1) + 2.7183(x_2 - 2). \end{aligned}$$

Table 2.2 shows $f(x)$ and $\hat{f}(x)$, and the approximation error $|\hat{f}(x) - f(x)|$, for some values of x relatively near z. We can see that $\hat{f}$ is indeed a very good approximation of f, especially when x is near z.

The more points along the polynomial curve, the smoother the approximation is! The local best guess matches the function's behavior at a given point and that's done through the series of derivatives(?)

METHOD #5

2. (10 pts) Solve the Chapter 2 Random exercise from the video and Piazza in your own words here.

2.7 *General formula for affine functions.* Verify that formula (2.4) holds for any affine function $f : \mathbf{R}^n \to \mathbf{R}$. You can use the fact that $f(x) = a^T x + b$ for some n-vector a and scalar b.

$$f(x) = f(0) + x_1 \left(f(e_1) - f(0)\right) + \cdots + x_n \left(f(e_n) - f(0)\right)$$

What we are trying to accomplish is we are verifying that the formula above is true for any kind of affine function. In order to do this, we need to establish some form of base case, like what happens when x = 0?

f(0) = a transpose 0 + b and by property of transpose, anything that is transpose 0 is 0 so f(0) = b

Now, we need to abstract this to f(ei) to apply this to the i-th basis vector so that we know that f(ei) = a transpose ei + b. Since we know that a transpose ei can be just denoted as ai, f(ei) = ai + b.

Note: I am just saying "ai" and "ei" but what I really mean in these scenarios is e sub-i and a sub-i.

Now that we know what f(ei) is, what is f(ei) - f(0)? Well we know f(0) = b so it's just..f(ei) - f(0) = (ai + b) - b) = ai, the b cancels out with -b.

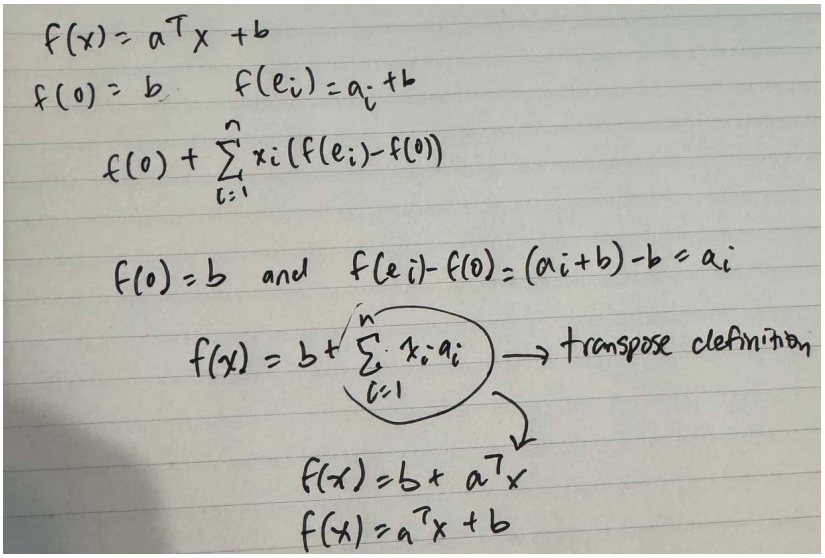
Now we substitute this back into the original formula and we get some kind of crazy blob that looks like:

f(x) = f(0) +x1(f(e1)-f(0)) + ... + xn(f(en) - f(0)) and this simplifies to... f(x) = b + x1a1 + x2a2 + ... + xnan

And this "x1a1 + x2a2 + ... + xnan" part is just the definition of transpose so we can substitute that as a transpose x or x transpose a (order doesn't matter here due to the fact that multiplication and addition are individually order agnostic (w + v = v + w, and v*w = w*v).

Therefore, this shows that the formula given in the problem holds true for any affine function since we can distill it to follow f(x) = (a transpose x) + b.

Another way to do this is by using the Summation Notation



METHOD #5

3. (20 pts) Explain the solution to 2.1 here in your own words. (Since you are given a solution, you will be graded on your ability to explain).

2.1 *Linear or not?* Determine whether each of the following scalar-valued functions of n-vectors is linear. If it is a linear function, give its inner product representation, *i.e.*, an n-vector a for which $f(x) = a^T x$ for all x. If it is not linear, give specific x, y, α, and β for which superposition fails, *i.e.*,

$$f(\alpha x + \beta y) \neq \alpha f(x) + \beta f(y).$$

(a) The spread of values of the vector, defined as $f(x) = \max_k x_k - \min_k x_k$.

(b) The difference of the last element and the first, $f(x) = x_n - x_1$.

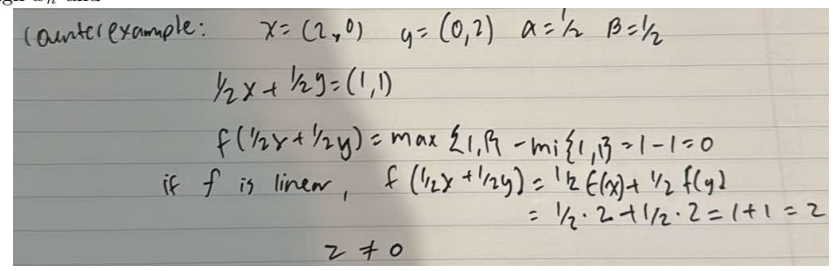
(c) The median of an n-vector, where we will assume $n = 2k + 1$ is odd. The median of the vector x is defined as the $(k + 1)$st largest number among the entries of x. For example, the median of $(-7.1, 3.2, -1.5)$ is -1.5.

(d) The average of the entries with odd indices, minus the average of the entries with even indices. You can assume that $n = 2k$ is even.

(e) Vector extrapolation, defined as $x_n + (x_n - x_{n-1})$, for $n \geq 2$. (This is a simple prediction of what x_{n+1} would be, based on a straight line drawn through x_n and x_{n-1}.)

I apologize if the font from hereon out looks a little weird, my pdf editor (pdf gear) would NOT format my added text properly, so I had to write my answer on google docs and then screenshot it to paste in the pdf as an image rather than type out my answer as usual.

Figure A)



Let's make clear of what linear means in terms of additivity and homogeneity:

Additivity means f(x + y) = f(x) + f(y) for all x, y
Homogeneity means f((alpha)(x)) = (alpha)(f(x)) for all scalars alpha and x

This is under the assumption of real numbers there must also be some fixed vector alpha is a real number such that f(x) = (a transpose x) for all x

a) The spread is not linear, because taking linear combinations of the vectors doesn't produce the same spread. The spread function is the difference between the largest and smallest entries of x so as soon as you change the order of the two vectors, it will break linearity. I drew an example by hand in the picture above on the right (figure A).

b) It is a linear combination because of the coordinates x1, ... , xn. If we take a = (-1, 0, 0, ... 0, 1), then f(x) = a transpose x = -1*x1, + 0 + ... + 0 + 1*xn and so any function that can be written as a transpose x will guaranteed satisfy additivity and homogeneity

c) This is not linear because it's similar to problem a) where the median has a nonlinear interaction with vector addition and scalar multiplication. It will fail additivity because when you choose a certain pair of x, y and a scalar, it will fail the test. Say we let x = (1000, -2, 5) and y = (-2, 10, 11) in R^3 (real number). When sorted in ascending order, median(x) is 5 and median(y) is 10. If we consider alpha = 1, beta = 1, x + y = (1000 -2, -2 + 10, 5 + 11) = (998, 8, 16) and sorted is (8, 16, 998) which median(x + y) is 16, but incorporating alpha(f(x)) + beta(f(y)) = 5 + 10 = 15, and 15 =/= 16. Therefore, not linear.

d) This is linear because of the coordinates of x. We can verify this by letting n = 4. f(x1,x2,x3,x4) = (x1 +x3)/2 - (x2+x4)/2. If x = (2, 6, 4, 0) then the odd indices average = 3 and the even indices will be 3. So f(x) = 3 - 3 = 0. If we made y = (10, 2, 3, 5) then the odd indices will average = 6.5 and the even indices = 3.5 so f(y) = 6.5 - 3.5 = 3. Put this together... alphax + betay = x + 2y = (2 + 20, 6 + 4, 4 + 6, 0+10) = (22, 10, 10, 10). Odd indices average = (22+10)/2 = 16. Even indices = (10+10)/2 =10 so f(x+2y) = 16-10 = 6, which makes the two equivalent and therefore makes f linear.

e) This is linear because of xn-1 and xn, assuming x exists within R^n (real number). f(x) = -x sub(n-1) + 2x sub(n) all the other coordinates can be pretty much ignored because they're multiplied by 0. The function f can be written as a transpose x where a = (0, 0, ... , 0, -1, 2) so this passes both additive and homogeneity tests, therefore, linear.

METHOD #5

I apologize if the font from hereon out looks a little weird, my pdf editor (pdf gear) would NOT format my added text properly, so I had to write my answer on google docs and then screenshot it to paste in the pdf as an image rather than type out my answer as usual.

4. (20 pts) Explain the solution to 2.4 here in your own words. (Since you are given a solution, you will be graded on your ability to explain).

point "v" as in vector of real numbers

2.4 *Linear function?* The function $\phi : \mathbf{R}^3 \to \mathbf{R}$ satisfies

$$\phi(1,1,0) = -1, \qquad \phi(-1,1,1) = 1, \qquad \phi(1,-1,-1) = 1.$$

"x" is given in problem

Choose one of the following, and justify your choice: ϕ must be linear; ϕ could be linear; ϕ cannot be linear.

A linear function respects scaling so if you have a point v in R^3 and you multiply that point out by -1, then a linear function would have to multiply the output by -1 as well.

So using the points x = (-1, 1, 1) and -x = (1, -1, -1), you can see immediately that x is just -x(-1). Mathematically, it's like a "duh!"-moment.

But if we took a function called ϕ and said it was linear, ϕ(-x) = ϕ((-1)x) = (-1)ϕ(x). So if ϕ(x) = 1, then ϕ(-x) should be -1. But in the problem, ϕ(x) is given = 1, and ϕ(-x) is also 1. They are not negatives of each other like they're supposed to be.

ϕ(-x) = -ϕ(x) is what would make it linear. But given the output numbers, that's not possible (the negative sign did not flip).

Which means that this is a contradiction to the laws of linear functions. Therefore, ϕ is not and cannot linear.

I described my answer in this way because it was the lowest hanging fruit. The overall concept is just x (1) = x and x(-1) = -x and if it doesn't = -x then you know there's a problem. Imagine the function is a black box and you feed different inputs in, and it doesn't change the outcome. That's what this problem is probing for. It's asking for "prove that this blackbox is not working as intended".

Coincidentally, the book answered the question in the same way. I think because it's just the easiest way to describe it and since the elements in the question are all primitive, the answers given are going to be pretty primitive as well. I imagine other students probably answered a similar way.

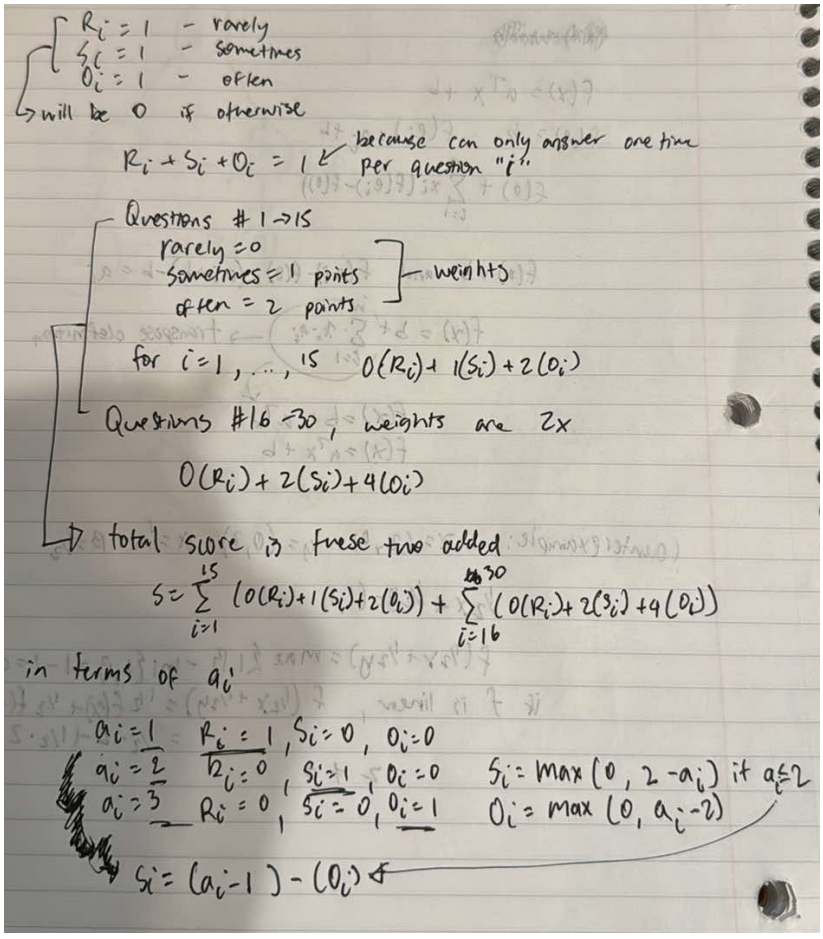
METHOD #5

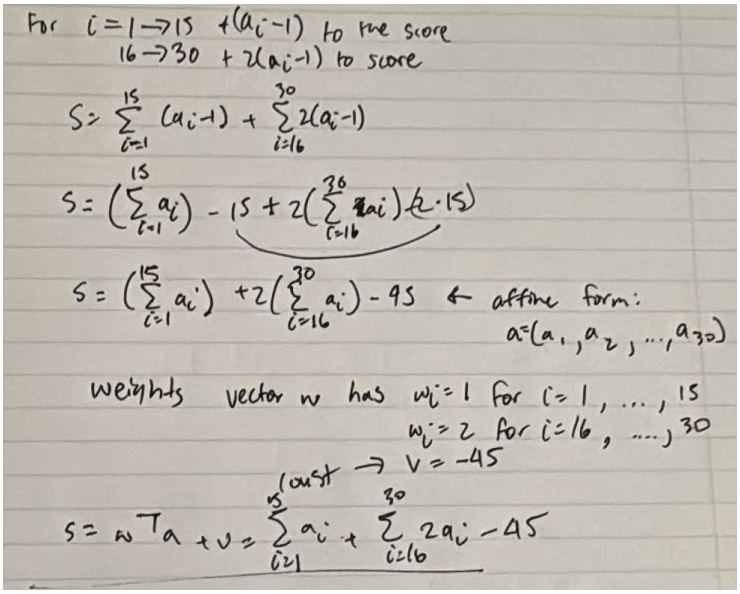
I apologize if the font from hereon out looks a little weird, my pdf editor (pdf gear) would NOT format my added text properly, so I had to write my answer on google docs and then screenshot it to paste in the pdf as an image rather than type out my answer as usual.

5. (20 pts) Explain the solution to 2.6 here in your own words. (Since you are given a solution, you will be graded on your ability to explain).

2.6 *Questionnaire scoring.* A questionnaire in a magazine has 30 questions, broken into two sets of 15 questions. Someone taking the questionnaire answers each question with 'Rarely', 'Sometimes', or 'Often'. The answers are recorded as a 30-vector a, with $a_i = 1, 2, 3$ if question i is answered Rarely, Sometimes, or Often, respectively. The total score on a completed questionnaire is found by adding up 1 point for every question answered Sometimes and 2 points for every question answered Often on questions 1–15, and by adding 2 points and 4 points for those responses on questions 16–30. (Nothing is added to the score for Rarely responses.) Express the total score s in the form of an affine function $s = w^T a + v$, where w is a 30-vector and v is a scalar (number).

1 = rarely
2 = sometimes
3 = often





My solution is a bit different than what was done by the book. The book seems to have kept the transpose format whereas I used summation (which is similar) to better understand things conceptually. I wanted to understand things relative to weights because there was an easy way to understand the relationship between i = 1 -> 15, and i = 16 -> 30 relative to each other

I apologize if the font from hereon out looks a little weird, my pdf editor (pdf gear) would NOT format my added text properly, so I had to write my answer on google docs and then screenshot it to paste in the pdf as an image rather than type out my answer as usual.

6. (20 pts)

a) Redo the proof on page 30 that an inner product satisfies superposition with 2 vectors of your own choosing (using length 3).

b) Is this still a proof?

c) Why is it useful?

Superposition and linearity. The inner product function f defined in (2.1) satisfies the property

$$\begin{aligned} f(\alpha x + \beta y) &= a^T(\alpha x + \beta y) \\ &= a^T(\alpha x) + a^T(\beta y) \\ &= \alpha(a^T x) + \beta(a^T y) \\ &= \alpha f(x) + \beta f(y) \end{aligned}$$

for all n-vectors x, y, and all scalars α, β. This property is called *superposition*. A function that satisfies the superposition property is called *linear*. We have just shown that the inner product with a fixed vector is a linear function.

The superposition equality

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y) \tag{2.2}$$

looks deceptively simple; it is easy to read it as just a re-arrangement of the parentheses and the order of a few terms. But in fact it says a lot. On the left-hand side, the term $\alpha x + \beta y$ involves *scalar-vector* multiplication and *vector addition*. On the right-hand side, $\alpha f(x) + \beta f(y)$ involves ordinary *scalar multiplication* and *scalar addition*.

If a function f is linear, superposition extends to linear combinations of any number of vectors, and not just linear combinations of two vectors: We have

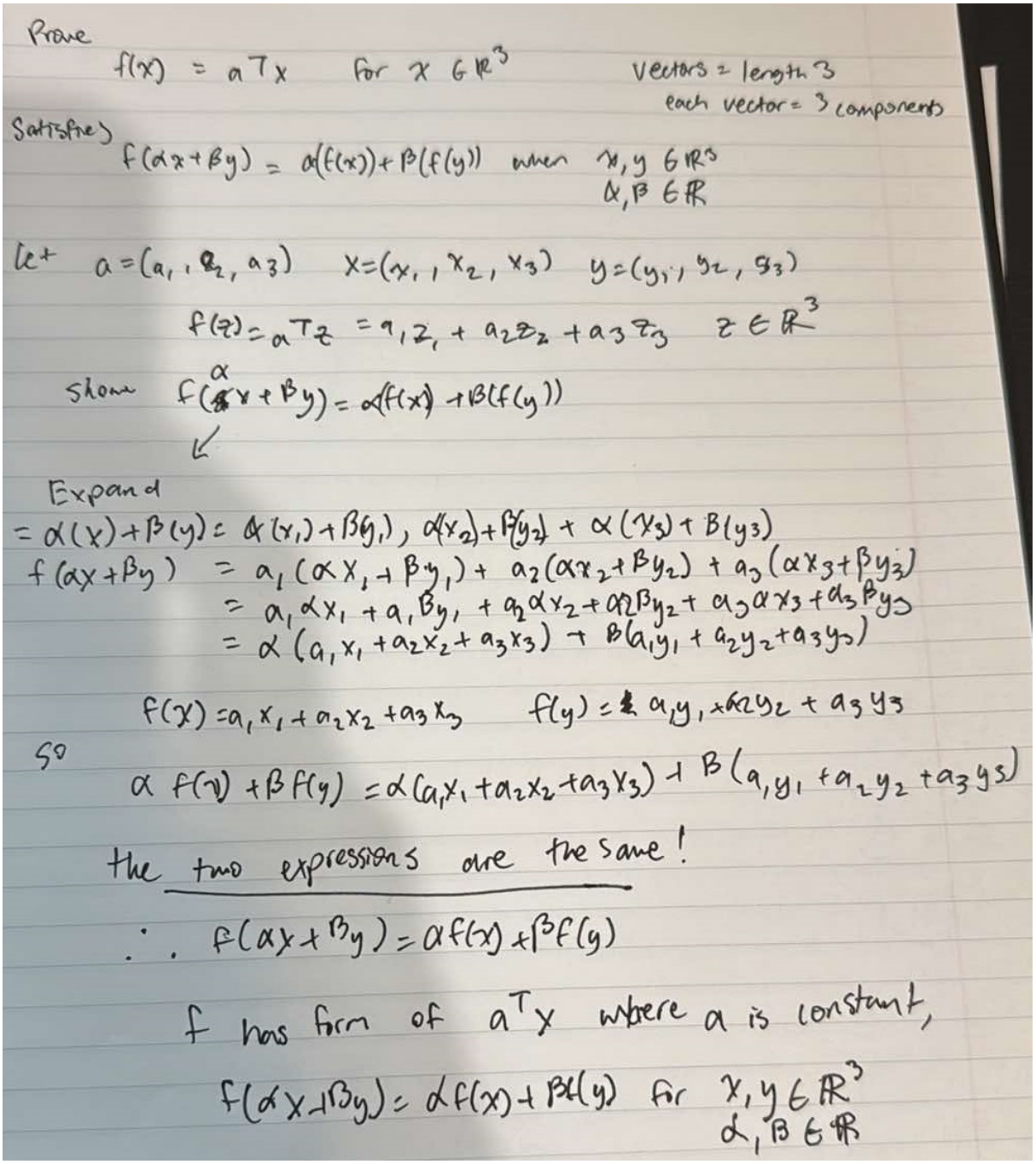
$$f(\alpha_1 x_1 + \cdots + \alpha_k x_k) = \alpha_1 f(x_1) + \cdots + \alpha_k f(x_k),$$

for any n vectors $x_1, \ldots, x_k$, and any scalars $\alpha_1, \ldots, \alpha_k$. (This more general k-term form of superposition reduces to the two-term form given above when $k = 2$.) To see this, we note that

$$\begin{aligned} f(\alpha_1 x_1 + \cdots + \alpha_k x_k) &= \alpha_1 f(x_1) + f(\alpha_2 x_2 + \cdots + \alpha_k x_k) \\ &= \alpha_1 f(x_1) + \alpha_2 f(x_2) + f(\alpha_3 x_3 + \cdots + \alpha_k x_k) \\ &\vdots \\ &= \alpha_1 f(x_1) + \cdots + \alpha_k f(x_k). \end{aligned}$$

METHOD #5 - I don't think there's a solution for this problem in the book.....?

A)



Prove

$f(x) = a^T x$ for $x \in \mathbb{R}^3$ vectors = length 3, each vector = 3 components

Satisfies $f(\alpha x + \beta y) = \alpha(f(x)) + \beta(f(y))$ when $x, y \in \mathbb{R}^3$, $\alpha, \beta \in \mathbb{R}$

let $a = (a_1, a_2, a_3)$ $x = (x_1, x_2, x_3)$ $y = (y_1, y_2, y_3)$

$f(z) = a^T z = a_1 z_1 + a_2 z_2 + a_3 z_3$ $z \in \mathbb{R}^3$

show $f(\alpha x + \beta y) = \alpha(f(x)) + \beta(f(y))$

Expand

$= \alpha(x) + \beta(y) = \alpha(x_1) + \beta y_1), \alpha(x_2) + \beta(y_2) + \alpha(x_3) + \beta(y_3)$

$f(\alpha x + \beta y) = a_1(\alpha x_1 + \beta y_1) + a_2(\alpha x_2 + \beta y_2) + a_3(\alpha x_3 + \beta y_3)$

$= a_1 \alpha x_1 + a_1 \beta y_1 + a_2 \alpha x_2 + a_2 \beta y_2 + a_3 \alpha x_3 + a_3 \beta y_3$

$= \alpha(a_1 x_1 + a_2 x_2 + a_3 x_3) + \beta(a_1 y_1 + a_2 y_2 + a_3 y_3)$

$f(x) = a_1 x_1 + a_2 x_2 + a_3 x_3$ $f(y) = a_1 y_1 + a_2 y_2 + a_3 y_3$

So

$\alpha f(x) + \beta f(y) = \alpha(a_1 x_1 + a_2 x_2 + a_3 x_3) + \beta(a_1 y_1 + a_2 y_2 + a_3 y_3)$

the two expressions are the same!

$\therefore f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$

f has form of $a^T x$ where a is constant,

$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$ for $x, y \in \mathbb{R}^3$, $\alpha, \beta \in \mathbb{R}$

B) Yes, this is still a proof, even if it's only proving by expanding. I showed step by step how f(alphax + betay) expanded to alpha(a1x1 + a2x2 + a3x3) + beta(a1y1 + a2y2 + a3y3) which is the same as alpha(f(x)) + beta(f(y))

C) Direct coordinate expansion proof is useful because you can see exactly how terms match up to both sides of the equation. This breaks down the nitty gritty and stops any handwavey-ness that might occur when you follow high-level principles like "a transpose x is linear". This is the kind of proof that says "well, how?". I did the proof in 3 dimensions, but because it's done in length-3, it can also be done in length-4, and 5, and 6, and so on. Fundamentally doesn't change, you just need to add more x and y and alphas and betas. I also made it very clear how each coordinate term corresponds to something like vector-scalar multiplication and dot product based on this linearity.