Where does the Dot Product Come From?

R. Scott McIntire

Nov 6, 2024

1 Overview

One of the fundamental pieces of linear algebra is a bilinear function called the "dot-product" It is an important work horse, but where does it come from? Please don't give me a definition and tell me how to compute it. I would like to know how it came to be; or, at least, how might one have thought of it. The rest of this paper tries to do just that.

2 Vectors, Distance Relationships, and Projections

Let's start with the real line and solve a simple problem. If a car starts at point p_0 on the real line and moves at a velocity of v when will the car reach the point p_1? We assume that this problem is non-trivial:

- $v \neq 0$.
- $p_0 \neq p_1$.

That is, find a t^* so that

$$p_0 + t^* v = p_1$$

So t^* satisfies:

$$t^*\, v = p_1 - p_0$$

Since v is a (non-zero) number we can divide and solve for t^*:

$$t^* = \frac{p_1 - p_0}{v} \tag{1}$$

This is the solution *provided* t^* is not negative.

In other words, there is no solution if the car is moving away from q.

We note, however, in this simple example that there is an answer to the question: "What is the closest point to p_1 that the car can get to?" The answer is p_1 when the car moves towards p_1 at a non-zero velocity; otherwise, it is p_0.

We now look at the same problem in 2-dimensions. Points in two dimensions are described by a point in a plane which can be represented as a pair of two real numbers, (x, y). More generally, one can describe a point in 3-dimensions by a triple of numbers, (x, y, z). Our plan is to try and examine the "car" problem in high dimensions and look for a pattern. We will even look beyond three dimensions!

Let's try to mimic the "car" argument we had in 1-dimension. But we first note that to mimic it more consistently, we should have some way of adding/subtracting points(tuples) and the ability to divide by *numbers*.

We do this by replacing our "tuple" notation of (x, y) with the *vector* notation: $[x, y]$. Either way using tuple or vector notation is really a way to represent an ordered set of numbers. We will use the vector notation because that is what mathematicians use.

We also want to simplify the number of names we use, this is especially useful when looking at dimensions beyond 3. We will use bold font, $\mathbf{x}$, to mean a vector (for now in 2 dimensions) whose components are: x_1, x_2. But to mimic the 1-dimensional argument we need to know how to add/subtract vectors and how to "stretch" a vector with multiplication by a number. We will use the name that mathematicians use for number – *scalar*. But this is something we learn in high school – we simply add/subtract the components and for multiplication by numbers we just multiple each component separately.

We generalize this from 2 or 3 dimensions so that vectors may represent n dimensions. All this means is that a vector is an ordered set of numbers of a specific length. Here is the formal definition of how to add/subtract them and multiply them by scalars.

Theorem 1 *Let* $\mathbf{x}, \mathbf{y}$ *be vectors and let* a *be a number. We define vector addition and scalar multiplication as:*

$$\mathbf{x} + \mathbf{y} \equiv [x_1 + y_1,\, x_2 + y_2] \tag{2}$$
$$a * \mathbf{x} \equiv [ax_1,\, ax_2] \tag{3}$$

We drop the '*' operator when multiplying a vector by a number – just as we do in algebra, writing: $a\,\mathbf{x}$ instead of $a * \mathbf{x}$.

Keeping with the "car" example, a more realistic problem would be to drive on a surface – in two dimensions – and solve the problem above. To do so, we need a way to multiply a number (time), t, with a velocity vector, representing both a speed and a direction. And we need to be able to add this to the original position to get the current position. This is just what the above definitions of scalar multiplication and vector addition provide. Over time, t, the path swept out, or *spanned* by the car is a line in two dimensional space: $L = \{\mathbf{p} + t\,\mathbf{v} \,|\, t \in R\}$. Here, L represents a collection of "position" vectors representing the car's trajectory. Check that the *units* make sense.

These definitions make sense in that the new position of a "car" after time t when starting at position $\mathbf{p}$ is determined by taking each of the component positions and adjusting them by adding t times the corresponding velocity component.

Let's repeat the problem we worked out in the scalar case: At what time, if any, will the car reach the point $\mathbf{p}_1$ starting at $\mathbf{p}_0$? So, we need to solve for some t^* so that:

$$\mathbf{p}_0 + t^*\mathbf{v} = \mathbf{p}_1$$

We see that t^* must satisfy

$$t^*\mathbf{v} = \mathbf{p}_1 - \mathbf{p}_0 \tag{4}$$

But now, we can't just divide by $\mathbf{v}$ to get the answer, t^*, as $\mathbf{v}$ is not a number.

We can think of a (4) as a shifted version of the origin problem. Instead of starting at a point, p_0, we are starting at the origin 0 and trying to "hit" $\mathbf{p}_1$ using the directional speed, $\mathbf{v}$.

In any event, if $\mathbf{v}$ is not pointing *exactly* to the destination, $\mathbf{p}_1 - \mathbf{p}_0$, there is no chance of finding a solution.

Here's a different question: Drive a car on one of the roads passing through $\mathbf{p}_0$ and drop someone off on this road at the point closest to the point $\mathbf{p}_1$.

In more mathematical terms: Starting at $\mathbf{p}_0$, drive with velocity $\mathbf{v}$ (go along some road at a fixed speed) and find the time t^* at which I can drop off a passenger so that I am as close as possible to the destination, $\mathbf{p}_1$.

Need to solve: Find t^* so that the distance between $\mathbf{p}_0 + t\,\mathbf{v}$ and $\mathbf{p}_1$ is minimized. This can be written as: find the time, t^*, which minimizes the length of the vector: $\mathbf{p}_0 + t\,\mathbf{v} - \mathbf{p}_1 = t\,\mathbf{v} - (\mathbf{p}_1 - \mathbf{p}_0)$. This is the same as finding the time to get to the closest point to the vector $\mathbf{p}_1 - \mathbf{p}_0$ using the velocity vector, $\mathbf{v}$, when starting at $\mathbf{0}$.

So, we look instead at the simpler problem: Starting at $\mathbf{0}$ and driving with a velocity of $\mathbf{v}$, find t^* so as to minimize the distance to a given vector, $\mathbf{p}$. We will also generalize this to n dimensions and *define* the length of a vector, $\mathbf{x}$, as a generalization of the distance formula for vectors in 2 and 3 dimensions; namely, $\text{length}(\mathbf{x}) = \sqrt{\sum_{i=1}^{n} x_i^2}$.[1]

$$\min_t \sum_{i=1}^{n} (p_i - t v_i)^2$$

Using calculus, the minimum distance is achieved when the derivative at t^* is 0.

$$-2 \sum_{i=1}^{n} (p_i - t^* v_i) v_i = 0$$

$$t^* = \frac{\sum_{i=1}^{n} p_i v_i}{\sum_{i=1}^{n} v_i^2}$$

The point that is closest along the line *spanned* by $\mathbf{v}$ is called the projection of $\mathbf{p}$ onto $\mathbf{v}$. Denoted as $P_{\mathbf{v}}(\mathbf{p}) = t^*\mathbf{v}$, which is:

$$P_{\mathbf{v}}(\mathbf{p}) = \frac{\sum_{i=1}^{n} p_i v_i}{\sum_{i=1}^{n} v_i^2} \mathbf{v}$$

We introduce a function that is a refactoring[2] of the essential element of the formula – as we see that we are repeating a sum in the numerator and the denominator.

Define the "Dot Product" between two vectors as:

Def 1 *The Dot Product of two vectors is defined by*

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i$$

[1] Minimizing the square of the distance is the same as minimizing the distance, and its also easier from the perspective of calculus.

[2] This is a common practice in the software engineering world. If we start repeating code we refactor it into a subroutine.

Now the numerator and denominator in the fraction of the projection formula can be written:

$$\begin{aligned} \sum_{i=1}^{n} p_i v_i &= \mathbf{p}\cdot\mathbf{v} \\ \sum_{i=1}^{n} v^2 &= \mathbf{v}\cdot\mathbf{v} \end{aligned}$$

With this refactoring, the projection formula can be simplified to:

$$P_{\mathbf{v}}(\mathbf{p}) = \frac{\mathbf{p}\cdot\mathbf{v}}{\mathbf{v}\cdot\mathbf{v}}\mathbf{v} \tag{5}$$

We use the function, $\|\cdot\|$ to denote length of a vector:

$$\|\mathbf{x}\|^2 = \sum_{i=1}^{n} x_i^2$$

But this can be written more succinctly in terms of the dot product:

$$\|\mathbf{x}\|^2 = \mathbf{x}\cdot\mathbf{x}$$

Getting back to the original problem, the closest point to $\mathbf{p}_1$ when starting at $\mathbf{p}_0$ and moving along $\mathbf{v}$ is:

$$P_{\mathbf{v}}(\mathbf{p}_1 - \mathbf{p}_0) = \frac{(\mathbf{p}_1 - \mathbf{p}_0)\cdot\mathbf{v}}{\mathbf{v}\cdot\mathbf{v}}\mathbf{v}$$

There is good reason to have the name "product" in the name "Dot Product". It has most of the properties of a multiplication operator.

"Multiplicative Properties" of the Dot Product:

- $\mathbf{p}\cdot\mathbf{v} = \mathbf{v}\cdot\mathbf{p}$ – Dot product is symmetric.
- $(\lambda_1\mathbf{p}_1 + \lambda_2\mathbf{p}_2)\cdot\mathbf{v} = \lambda_1\mathbf{p}_1\cdot\mathbf{v} + \lambda_2\mathbf{p}_2\cdot\mathbf{v}$ – Dot product distributes addition and scalar multiplication on the left. And since this operator is symmetric, it distributes on the right as well.
- $\|\mathbf{p}\| = \sqrt{\mathbf{p}\cdot\mathbf{p}}$ – The dot product *induces* the length function. Notice that the length of a vector $\mathbf{p}$ is non-negative if and only if $\mathbf{p} \neq \mathbf{0}$.

This means that we can redo the minimization problem that gave us a projection because the algebra of the dot product is like multiplication and the derivation proceeds using only the *multiplicative properties*.

$$\begin{aligned} \min_t \|\mathbf{p} - t\,\mathbf{v}\|^2 \\ \min_t (\mathbf{p} - t\,\mathbf{v})\cdot(\mathbf{p} - \mathbf{t}\,\mathbf{v}) \\ \min_t \mathbf{p}\cdot\mathbf{p} - 2t\,\mathbf{p}\cdot\mathbf{v} + t^2\mathbf{v}\cdot\mathbf{v} \end{aligned} \tag{6}$$

Notice that the optimization problem minimizes distance and we can write distance as a dot product of a vector with itself, the optimization is defined entirely with respect to scalar multiplication and dot product of *constants*. These constants do *NOT* affect the optimization; and so, the optimization becomes an optimization in one variable – that is, we only need single variable calculus.

Taking the derivative in t and solving for t when the derivative is 0 gives the minimal t which we will call t^*. The derivative is:

$$\begin{aligned} -2\mathbf{p}\cdot\mathbf{v} + 2t\mathbf{v}\cdot\mathbf{v} &= 0 \\ t^* &= \frac{\mathbf{p}\cdot\mathbf{v}}{\mathbf{v}\cdot\mathbf{v}} \end{aligned} \tag{7}$$

We will revisit the dot product later in this section. But for now, we will get back to the projection.

One way of thinking of $\mathbf{p}$ is as a combination of the $\mathbf{v}$ direction and another direction. We could think of the component of $\mathbf{p}$ in the $\mathbf{v}$ direction as the projection of $\mathbf{p}$ onto $\mathbf{v}$ as that is the best we can do using $\mathbf{v}$ alone to get to $\mathbf{p}$. Notice that the projection of $\mathbf{p}$ onto $\mathbf{v}$ is independent of the "length" of $\mathbf{v}$. If we replace $\mathbf{v}$ with any multiple of $\mathbf{v}$ say $\lambda\mathbf{v}$ and look at the projection of $\mathbf{p}$ onto $\lambda\mathbf{v}$ we have:

$$P_{\lambda\mathbf{v}}(\mathbf{p}) = \frac{\mathbf{p}\cdot\mathbf{v}_1}{\|\mathbf{v}_1\|^2}\mathbf{v}_1 = \frac{\mathbf{p}\cdot\lambda\mathbf{v}}{\lambda^2\|\mathbf{v}\|^2}\lambda\mathbf{v} = \frac{\mathbf{p}\cdot\mathbf{v}}{\|\mathbf{v}\|^2}\mathbf{v} = P_{\mathbf{v}}(\mathbf{p})$$

This says that the projection of $\mathbf{p}$ onto $\mathbf{v}$ is independent of the length of the vector $\mathbf{v}$. That is, it only depends on its "direction". (It doesn't matter how fast you drive me in a given direction, the place you drop me off to be as close to $\mathbf{p}$ is going to be the same.)

What about the projection of a multiple of $\mathbf{p}$ onto $\mathbf{v}$, or combinations of $\mathbf{p}$ with other vectors onto $\mathbf{v}$.

Using the "multiplicative properties" of the dot product, it is not hard to show the following:

- $P_{\mathbf{v}}(P_{\mathbf{v}}(\mathbf{p})) = P_{\mathbf{v}}(\mathbf{p})$.
- $P_{\mathbf{v}}(\lambda\mathbf{p}) = \lambda P_{\mathbf{v}}(\mathbf{p})$.
- $P_{\mathbf{v}}(\mathbf{p}_1 + \mathbf{p}_2) = P_{\mathbf{v}}(\mathbf{p}_1) + P_{\mathbf{v}}(\mathbf{p}_1)$.

We can combine these three into just two properties:

- $P_{\mathbf{v}}(P_{\mathbf{v}}(\mathbf{p})) = P_{\mathbf{v}}(\mathbf{p})$ – Makes sense: We shouldn't be able to get closer doing a second projection.
- $P_{\mathbf{v}}(\lambda_1\mathbf{p}_1 + \lambda_2\mathbf{p}_2) = \lambda_1 P_{\mathbf{v}}(\mathbf{p}_1) + \lambda_2 P_{\mathbf{v}}(\mathbf{p}_1)$ – Powerful: Can compute projections of expressions easily based on components.

We can think of a vector as having a direction and a length. How do we say this precisely? Well, we can think of the direction of $\mathbf{v}$ as being the thing that all vectors that are like $\mathbf{v}$ have. All vectors like $\mathbf{v}$ are really just multiples of $\mathbf{v}$:

$$\{t\,\mathbf{v} |\ \forall t \in R\}$$

That is, they are all scalar multiples of $\mathbf{v}$. In other words, vectors are like $\mathbf{v}$ if they have the same direction but the "speeds" may vary. (On the highway we are all going in the same direction, but with potentially differing speeds.)

One way to talk about the direction of this collection is to pick a *canonical* representative. The simplest thing to do is to pick the one in the same direction (orientation – it is a positive multiple of $\mathbf{v}$) with unit length. If $\mathbf{u}$ is this vector, then the projection onto the direction $\mathbf{v}$ is:

$$P_{\mathbf{u}}(\mathbf{p}) = (\mathbf{p}\cdot\mathbf{u})\,\mathbf{u}$$

If we continue to think of $P_{\mathbf{v}}(\mathbf{p})$ as the component in the $\mathbf{v}$ direction, then what is the other component that we add to this to get $\mathbf{p}$?

$$\mathbf{p} = P_{\mathbf{v}}(\mathbf{p}) + \mathbf{w}$$

By subtraction the other component is:

$$\mathbf{w} = \mathbf{p} - P_{\mathbf{v}}(\mathbf{p})$$

Does $\mathbf{w}$ have a component in the $\mathbf{v}$ direction or visa versa? The component of $\mathbf{w}$ onto $\mathbf{v}$ is:

$$P_{\mathbf{v}}(\mathbf{w}) = P_{\mathbf{v}}\left(\mathbf{p} - P_{\mathbf{v}}(\mathbf{p})\right) = P_{\mathbf{v}}(\mathbf{p}) - P_{\mathbf{v}}(P_{\mathbf{v}}(\mathbf{p})) = P_{\mathbf{v}}(\mathbf{p}) - P_{\mathbf{v}}(\mathbf{p}) = 0$$

Now, one can repeat this and show that there is no component of $\mathbf{v}$ onto $\mathbf{w}$. That is, $\mathbf{v}$ and $\mathbf{w}$ are as "independent" as possible.

We know that the length of a vector can be described in terms of the dot product and we now know that $\mathbf{p} = \mathbf{v} + \mathbf{w}$. What is the length of $\mathbf{p}$ when it is a sum of two vectors that share no components between them?

$$\begin{aligned} \|\mathbf{p}\|^2 &= \mathbf{p}\cdot\mathbf{p} = (\mathbf{v}+\mathbf{w})\cdot(\mathbf{v}+\mathbf{w}) \\ &= \mathbf{v}\cdot\mathbf{v} + 2\mathbf{v}\cdot\mathbf{w} + \mathbf{w}\cdot\mathbf{w} \\ &= \|\mathbf{v}\|^2 + 2 * 0 + \|\mathbf{w}\|^2 \\ &= \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 \end{aligned} \tag{8}$$

In two and three dimensions, the vectors $\mathbf{p}$, $\mathbf{v}$, and $\mathbf{w}$ form a triangle. When the lengths are a Pythagorean triple, then $\mathbf{v}$ is perpendicular to $\mathbf{w}$. That is, when $\mathbf{v}\cdot\mathbf{w} = 0$, $\mathbf{v}$ and $\mathbf{w}$ are perpendicular.

We generalize this, introducing the term "orthogonal". We define two vectors to be *orthogonal* if their dot product is zero. Given that two vectors, $\mathbf{w}_1$ and $\mathbf{w}_2$ are orthogonal and sum to $\mathbf{p}$, we have a generalization of the Pythagorean formula, with

$$\|\mathbf{p}\|^2 = \|\mathbf{w}_1\|^2 + \|\mathbf{w}_2\|^2$$

One can show using induction that if the vectors $\{\mathbf{w}_i\}_{i=1}^m$ are pairwise orthogonal and sum to $\mathbf{p}$, then

$$\begin{aligned} \sum_{i=1}^{m} \mathbf{w}_i &= \mathbf{p} \\ \mathbf{w}_i\cdot\mathbf{w}_j &= 0 \quad \forall i,j \in [1,m]\ i \neq j \\ \|\mathbf{p}\|^2 &= \sum_{i=1}^{m} \|\mathbf{w}_i\|^2 \end{aligned} \tag{9}$$

3 Standard Unit Vectors and their Generalization

The standard unit vectors are the so-called coordinate vectors. In R^2 these are: $\mathbf{e}_1 = [1, 0]$ and $\mathbf{e}_2 = [0, 1]$. In R^n they are: $\{\mathbf{e}_i\}_{i=1}^n$ where $\mathbf{e}_i = \overbrace{[0, \ldots, 1, 0, \ldots]}^{i-1 \text{ 0's followed by 1 followed by 0s}}$ Notice that this collection of vectors have unit length and are pair-wise orthogonal. And it is easy to write down the combination of $\mathbf{e}_i$ that produce a given $\mathbf{x} \in R^n$.

$$\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{e}_i$$

We claim that given another set of vectors with the same properties it is just as easy to write down the combination of these vectors which produces $\mathbf{x}$. Starting with R^2, given two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ with each of unit length and orthogonal to one another, assuming that we can write any $\mathbf{x}$ as some combination of the two, we claim that the values of u_1 and u_2 such that $\mathbf{x} = u_1\mathbf{v}_1 + u_2\mathbf{v}_2$ are easy to compute.

So, given that $\mathbf{x}$ can be written as a combination of two we have:

$$\mathbf{x} = u_1\mathbf{v}_1 + u_2\mathbf{v}_2 \tag{10}$$

From this we can form two equations, each the dot product with $\mathbf{v}_1$ and $\mathbf{v}_2$ respectively:

$$\mathbf{x}\cdot\mathbf{v}_1 = u_1\mathbf{v}_1\cdot\mathbf{v}_1 + u_2\mathbf{v}_1\cdot\mathbf{v}_2 \tag{11}$$

$$\mathbf{x}\cdot\mathbf{v}_2 = u_1\mathbf{v}_2\cdot\mathbf{v}_1 + u_2\mathbf{v}_2\cdot\mathbf{v}_2 \tag{12}$$

But since the $\mathbf{v}_1$ and $\mathbf{v}_2$ have unit length, the dot product with themselves is 1; and since the dot product of one with another is 0, we may write these two equations as:

$$\mathbf{x}\cdot\mathbf{v_1} = u_1 \tag{13}$$

$$\mathbf{x}\cdot\mathbf{v_2} = u_2 \tag{14}$$

This is, we can instantly solve what would be a system of equations – which would get progressively harder in higher dimensions. We can write this more formally as:

Theorem If $\mathbf{x}$ is a vector in R^2, and $\mathbf{v}_1$ and $\mathbf{v}_2$ are two vectors of unit length which are orthogonal, then assuming that $\mathbf{x}$ is some combination of $\mathbf{v}_1$ and $\mathbf{v}_2$, then that combination is:

$$\mathbf{x} = (\mathbf{x}\cdot\mathbf{v}_1)\mathbf{v}_1 + (\mathbf{x}\cdot\mathbf{v}_2)\mathbf{v}_2$$

If it is the case that in R^n we have n vectors of unit length and which are all orthogonal to each other, then by the same reasoning (assuming that these n vectors can span the full space) we have:[3]

Theorem If $\mathbf{x}$ is a vector in R^n, and $\{\mathbf{v}_i\}_{i=1}^n$ vectors of unit length which are orthogonal with each other, then $\mathbf{x}$ may be written as:

$$\mathbf{x} = \sum_{i=1}^{n} (\mathbf{x} \cdot \mathbf{v}_i)\mathbf{v}_i \tag{15}$$

We give a name to collections of such orthogonal vectors.

Def A collection of m vectors, $\{\mathbf{v}_i\}_{i=1}^m$, such that each has unit length and are pair-wise orthogonal are called an *orthonormal* set.

4 Projection onto a Linear Space

We can repeat what we did in the last section, considering instead that we can travel on an incline representing a mountain in 3 space. We are free to drive anywhere on the side of the mountain and want to drop off someone so that they can be picked up by a helicopter based on an adjacent mountain. We want to drop the person off so that the helicopter flies the least distance – helicopter time and fuel are expensive.

We can proceed in the same way as before, but now rather than a single direction to drive in we have what amounts to two directions. We are saying that there is a direction $\mathbf{v}_1$ and a direction $\mathbf{v}_2$ so that we can get to any point on the mountain if we drive for sometime in the direction $\mathbf{v}_1$ and then drive in the direction $\mathbf{v}_2$.

The solution to the problem is similar to the last section:

$$\min_{u_1,u_2} \|\mathbf{p} - (u_1\mathbf{v}_1 + u_2\mathbf{v}_2)\|^2 \tag{16}$$

Just as before, we solve to find u_1^*, u_2^* that minimize the above. This occurs when the *partial* derivatives of u_1 and u_2 are 0. Since the distance can be written in terms of the dot product we may rewrite this minimization problem as:[4]

[3] It turns out, as you might imagine, that n orthogonal vectors (each pointing in a "different direction") are always able to span R^n.

[4] We are assuming that the mountain is represented by a plane that goes through the origin.

$$\min_{u_1,u_2} (\mathbf{p} - (u_1\mathbf{v}_1 + u_2\mathbf{v}_2)) \cdot (\mathbf{p} - (u_1\mathbf{v}_1 + u_2\mathbf{v}_2)) \tag{17}$$

Taking the derivatives of this with respect to u_1 and u_2 we have:[5]

$$-2\mathbf{v}_1 \cdot (\mathbf{p} - (u_1\mathbf{v}_1 + u_2\mathbf{v}_2)) = 0 \tag{18}$$
$$-2\mathbf{v}_2 \cdot (\mathbf{p} - (u_1\mathbf{v}_1 + u_2\mathbf{v}_2)) = 0 \tag{19}$$

If we use vectors, $\mathbf{v}_1$ and $\mathbf{v}_2$ which are orthonormal, then this simplifies greatly to:

$$u_1 = \mathbf{p} \cdot \mathbf{v_1} \tag{20}$$
$$u_2 = \mathbf{p} \cdot \mathbf{v_2} \tag{21}$$

If we use the notation, $[\mathbf{v}_1, \mathbf{v}_2]$ to indicate the set of points *spanned* (swept out) by $\mathbf{v}_1$, and $\mathbf{v}_2$, then when $\mathbf{v}_1$ and $\mathbf{v}_2$ are orthonormal we have:

$$P_{[\mathbf{v}_1,\mathbf{v}_2]}(\mathbf{x}) = (\mathbf{x} \cdot \mathbf{v}_1)\mathbf{v}_1 + (\mathbf{x} \cdot \mathbf{v}_2)\mathbf{v}_2 \tag{22}$$

Def. We define $P_{[\mathbf{v_1},\ldots,\mathbf{v_m}]}$) to be the projection onto the set of points *spanned* (swept out) by the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$.

Formula (22) can generalize this so that if $\{\mathbf{v}_i\}_{i=1}^m$ is an orthonormal collection of vectors then the projection of a given vector $\mathbf{x} \in R^n$ onto the space spanned by collection is given by:

$$P_{[\mathbf{v}_1,\ldots,\mathbf{v}_m]}(\mathbf{x}) = \sum_{i=1}^{m} (\mathbf{x} \cdot \mathbf{v}_i)\mathbf{v}_i \tag{23}$$

Note that when $n = m$, we reproduce formula (15), as $P_{[\mathbf{v}_1,\ldots,\mathbf{v}_n]}(\mathbf{x}) = \mathbf{x}$.

[5]This follows since the dot product behaves like multiplication and the derivative of $f(x) * f(x)$ is $f'(x) * f(x) + f(x) * f'(x) = 2f'(x) * f(x)$. Since dot product behaves the same way: the derivative of $f(t, \mathbf{a}) \cdot f(t, \mathbf{b})$ is $2f'(t, \mathbf{a}) \cdot f(t, \mathbf{a})$.

5 Generalization of Projection based on the essence of the "Dot Product"

One can repeat the above notion of projection using a distance measure created from a generalization of the dot product. The generalization is called an *inner product*. We will use an "infix" form of such a function and use the notation $\langle \mathbf{x}, \mathbf{y} \rangle$. An inner product is any function of two vectors; i.e., $F : V \times V \to R$, with the following properties:

- $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ – Symmetry.
- $\langle \lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2, \mathbf{y} \rangle = \lambda_1 \langle \mathbf{x}_1, \mathbf{y} \rangle + \lambda_2 \langle \mathbf{x}_2, \mathbf{y} \rangle$ – Distributes addition and scalar multiplication.
- $\langle \mathbf{x}, \mathbf{x} \rangle = 0 \quad \text{iff } \mathbf{x} = \mathbf{0}$ – Non-degeneracy.
- $\|\mathbf{x}\| \equiv \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ – The inner product induces a *norm* – a notion of length of vectors.

The induced "length" has the properties of a length.

- $\|\mathbf{x}\| = 0 \quad \text{iff} \quad \mathbf{x} = \mathbf{0}.$ – Non-degeneracy.
- $\|\lambda \mathbf{x}\| = |\lambda| \|\mathbf{x}\| \quad \forall \lambda \in R \quad \forall \mathbf{x} \in V.$ – Homogeneity.
- $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \quad \mathbf{x}, \mathbf{y} \in V.$ – Triangle inequality.

If we repeat the exercise to determine the projection of one vector onto another by the length induced by an inner product, then all of the formulas above are duplicated with the only change being the replacement of $\mathbf{x} \cdot \mathbf{y}$ with $\langle \mathbf{x}, \mathbf{y} \rangle$.

The length and projection formulas becomes:

$$\|\mathbf{p}\|^2 = \langle \mathbf{p}, \mathbf{p} \rangle \tag{24}$$

$$P_{\mathbf{v}}(\mathbf{p}) = \frac{\langle \mathbf{p}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle} \mathbf{v} \tag{25}$$

One of the things we needed was that we could do calculus (we needed to minimization a length in order to find the projection) and for this we needed a notion of length. The inner product *defines* a notion of length automatically and all of the calculations proceed as

before. The only change is that wherever we see a dot product we can replace it with the inner product.

Example: We can define the following inner product on R^n. Let $\mathbf{w}$ be a given vector with the following properties:

1. $\sum_{i=1}^{n} w_i = 1$
2. $w_i > 0 \quad \forall i \in [1, n]$

Then we can define an inner product, $\langle \cdot, \cdot \rangle$ as:

$$\langle \mathbf{x}, \mathbf{y} \rangle \equiv \sum_{i=1}^{n} w_i x_i y_i$$

This modified "dot product" allows us to favour or discount some components over others. This is often used in data science.

The generalization is powerful and we can apply the geometric intuition of projections to sets of objects like functions and produce remarkable results. For instance, we can define a function set [6]

$$L^2 = \{\mathbf{f} | f \text{ is Integrable with } \int_{\Omega} f(x)^2 \, dx < \infty\}$$

L_2 is a set of functions where addition and scalar multiplication exist just like numeric vectors. Addition and scalar multiplication are defined by:[7]

$$(\mathbf{f} + \mathbf{g})(x) \equiv f(x) + g(x) \tag{26}$$
$$(\lambda * \mathbf{f})(x) \equiv \lambda f(x) \tag{27}$$

To understand the above, keep in mind what one must do. You must define a "+" operator so that adding two functions produces another function. But to describe what this function $f + g$ is, you must show what it does when applied to any "x" – which we've done. The new function $f + g$ when applied to "x" is what one would expect; it is the sum of the values of f and g applied to x. In a similar way we define scalar multiplication. And just like with ordinary vectors, we suppress the multiplication operator and write $\lambda \, \mathbf{f}$, instead of $\lambda * \mathbf{f}$.

[6] The exact notion of "Integrable" requires further explanation which we won't provide.

[7] This is like vectors where "x" is the continuous "index" into the "vector"(function). And just like vectors, to define one you need to explain what the value of the vector is on each "index".

With these definitions this space of function behaves like vectors of numbers.

Define $\langle \mathbf{f}, \mathbf{g} \rangle \equiv \int_\Omega f(x)g(x)\,dx$. This induces the length $\|f\| = \sqrt{\langle f, f \rangle} = \sqrt{\int_\Omega f(x)^2\,dx}$

So, this generalization can apply to things that are very different looking from typical vectors.

Given two functions $\mathbf{v}$ and $\mathbf{p}$, the projection formula becomes:

$$P_{\mathbf{v}}(\mathbf{p}) \quad = \quad \frac{\int_\Omega p(x)\,v(x)\,dx}{\int_\Omega v(x)^2\,dx}\mathbf{v} \tag{28}$$

Notice the projection is a new vector (function) that is a multiple of the function $\mathbf{v}$. You can evaluate it on a number just like any other function:

$$\left(P_{\mathbf{v}}(\mathbf{p})\right)(10) = \frac{\int_\Omega p(x)\,v(x)\,dx}{\int_\Omega v(x)^2\,dx}v(10)$$