4-Linear-Algebra-Refresher

September 18, 2016

1 Linear Algebra Refresher

Today we'll review the essentials of linear algebra. Given the prerequisites for this course, I assume that you learned all of this once. What I want to do today is bring the material back into your mind fresh.

1.1 Vectors and Matrices

A **matrix** is a rectangular array of numbers, for example:

$$X = \begin{bmatrix} 1 & -2 & 1 \\ 0 & 2 & -8 \\ 6 & 5 & 9 \end{bmatrix}$$

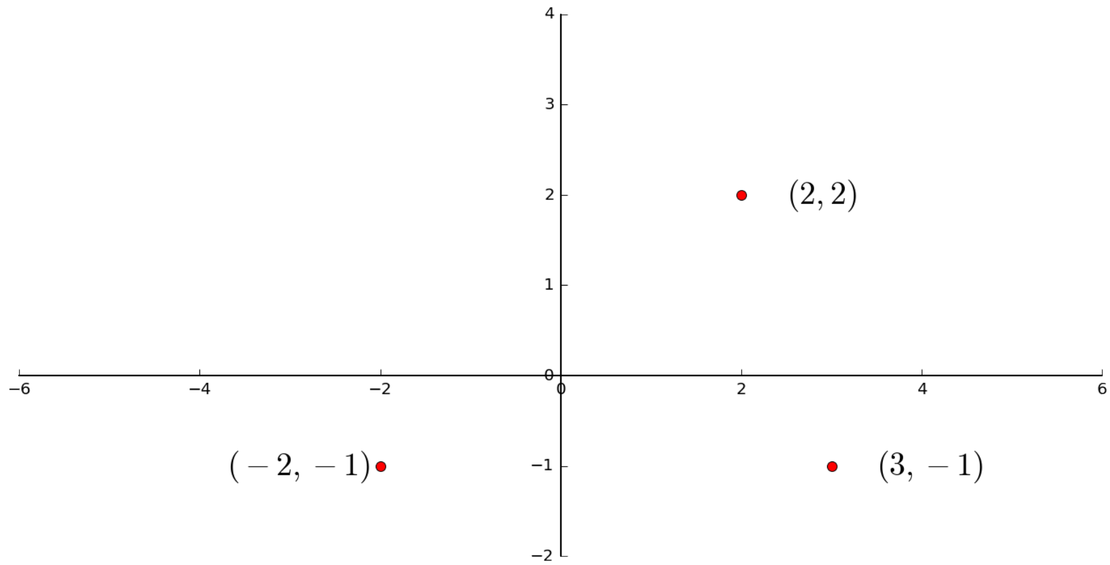
A matrix with only one column is called a **column vector**, or simply a **vector**.
Here are some examples.
These are vectors in $\mathbb{R}^2$:

$$\mathbf{u} = \begin{bmatrix} 3 \\ -1 \end{bmatrix} \quad \mathbf{v} = \begin{bmatrix} .2 \\ .3 \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

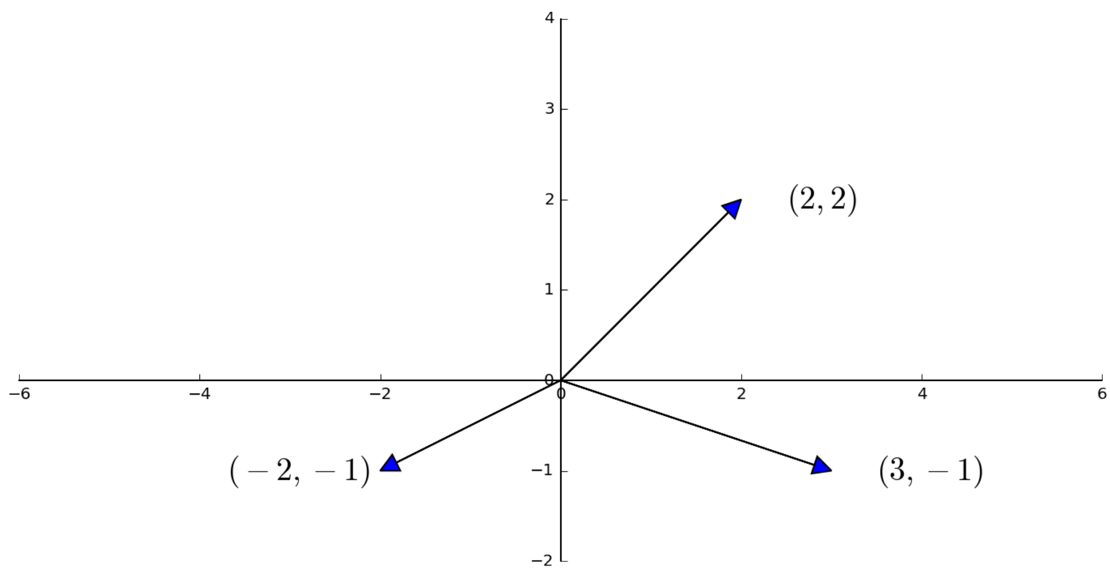
and these are vectors in $\mathbb{R}^3$:

$$\mathbf{u} = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} \quad \mathbf{v} = \begin{bmatrix} -1 \\ 0 \\ 2 \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}$$

We will use uppercase letters (X) for matrices and lowercase **bold** leters for vectors ($\mathbf{u}$).
A vector like $\begin{bmatrix} -2 \\ -1 \end{bmatrix}$ (also denoted $(-2, -1)$) can be thought of as a point on the plane.

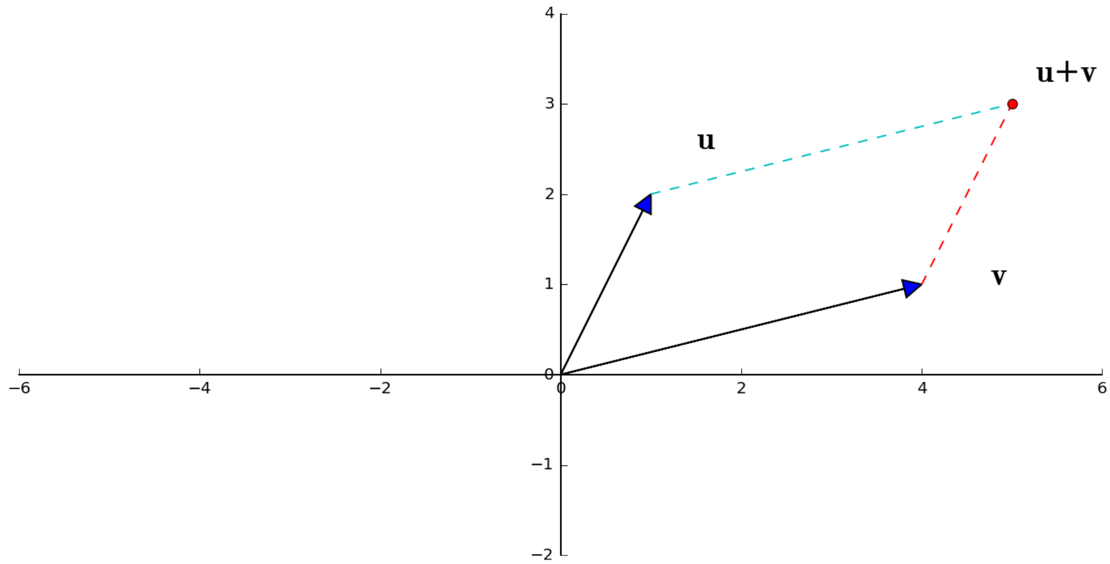


Sometimes we draw an arrow from the origin to the point. This comes from physics, but can be a helpful visualization in any case.



1.2 Vector Addition, Geometrically

A geometric interpretation of vector sum is as a parallelogram. If $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^2$ are represented as points in the plane, then $\mathbf{u} + \mathbf{v}$ corresponds to the fourth vertex of the parallelogram whose other vertices are $\mathbf{u}, 0$, and $\mathbf{v}$.



1.3 Matrix Multiplication

Using addition and multiplication by scalars, we can create equations using vectors.

Then we make the follow equivalence:

A vector equation

$$x_1\mathbf{a_1} + x_2\mathbf{a_2} + ... + x_n\mathbf{a_n} = \mathbf{b}$$

Can also be written as

$$A\mathbf{x} = \mathbf{b}$$

where

$$A = [\mathbf{a_1}\ \mathbf{a_2}\ ...\ \mathbf{a_n}]$$

and

$$\mathbf{b} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

We then extend this to define matrix multiplication as

Definition. If A is an $m \times n$ matrix and B is $n \times p$ matrix with columns $\mathbf{b_1}, \ldots, \mathbf{b_p}$, then the product AB is defined as the $m \times p$ matrix whose columns are $A\mathbf{b_1}, \ldots, A\mathbf{b_p}$. That is,

$$AB = A[\mathbf{b_1}\ \ldots\ \mathbf{b_p}] = [A\mathbf{b_1}\ \ldots\ A\mathbf{b_p}].$$

Continuing on, we can define the **inverse** of a matrix:
We have to recognize that this inverse does not exist for all matrices.

- It only exists for square matrices
- And not even for all square matrices – only those that are "invertible."

A matrix that is not invertible is called a **singular** matrix.
Definition. A matrix A is called **invertible** if there exists a matrix C such that

$$AC = I \text{ and } CA = I.$$

In that case C is called the *inverse* of A. Clearly, C must also be square and the same size as A. The inverse of A is denoted A^{-1}.

1.4 Inner Product

Now we introduce the **inner product**, also called the **dot product.**

The inner product is defined for two vectors, in which one is transposed to be a single row. It returns a single number. For example, the expression

$$[3\ 5\ 1]\begin{bmatrix} 2 \\ -1 \\ 4 \end{bmatrix}$$

is the same as

$$(3 \cdot 2) + (5 \cdot -1) + (1 \cdot 4)$$

which is the sum of the products of the corresponding entries, and yields the scalar value 5. This is the inner product of the two vectors.

Thus the inner product of $\mathbf{x}$ and $\mathbf{y}$ is $\mathbf{x}^T\mathbf{y} = \mathbf{y}^T\mathbf{x}$.

The general definition of the inner product is:

$$[x_1\ x_2\ \dots\ x_n]\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i$$

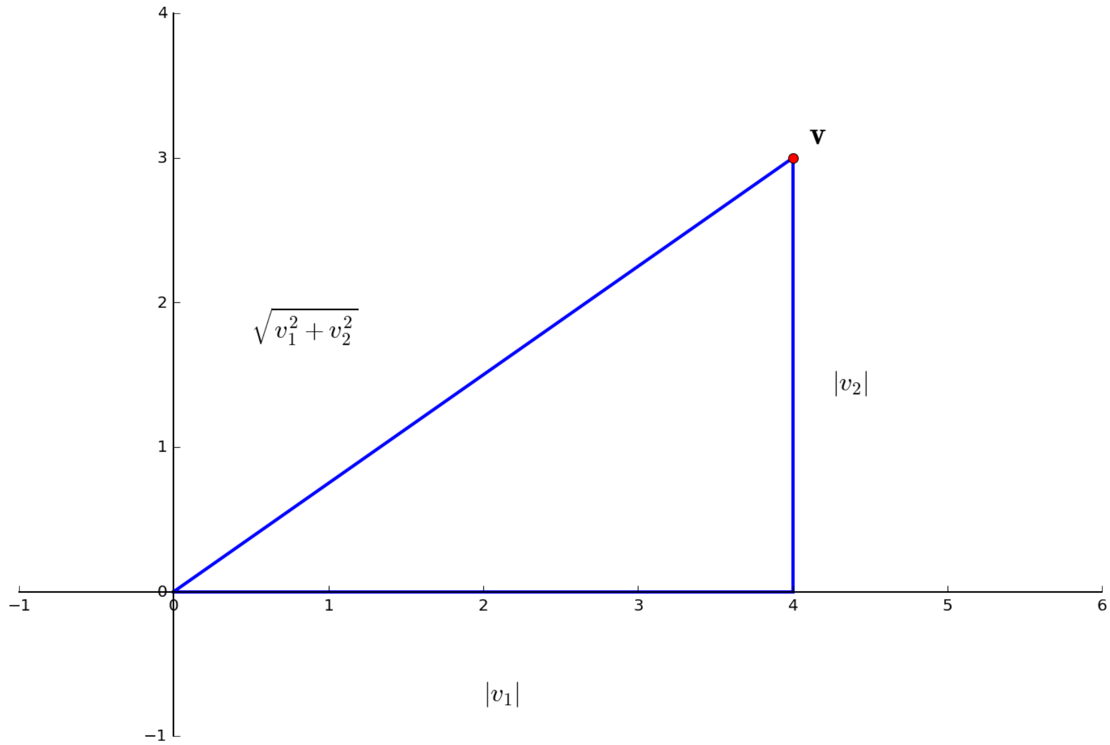
1.5 Vector Norm

If $\mathbf{v}$ is in $\mathbb{R}^n$, with entries $v_1, \dots, v_n$, then the square root of $\mathbf{v}^T\mathbf{v}$ is defined because $\mathbf{v}^T\mathbf{v}$ is nonnegative.

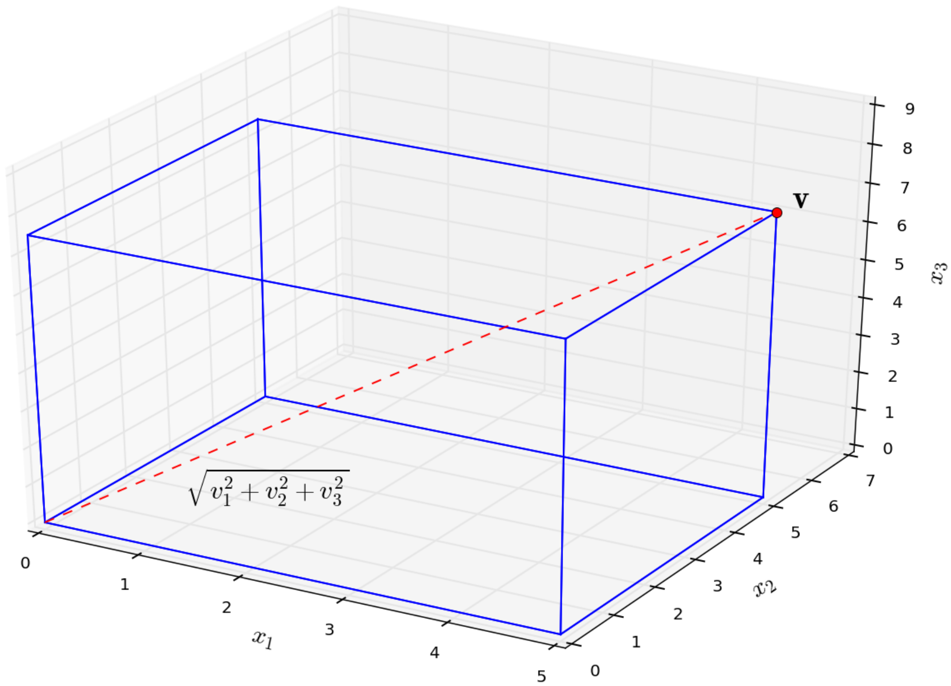
Definition. The ℓ_2 **norm** of $\mathbf{v}$ is the nonnegative scalar $\|\mathbf{v}\|_2$ defined by

$$\|\mathbf{v}\|_2 = \sqrt{\mathbf{v}^T\mathbf{v}} = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2} = \sqrt{\sum_{i=1}^{n} v_i{}^2}.$$

If we leave the $_2$ subscript off, it is implied.
The (ℓ_2) norm of $\mathbf{v}$ is its **length** in the usual sense.
This follows directly from the Pythagorean theorem:



4
3
2
1
0
−1
−1
0
1
2
3
4
5
6
$\mathbf{v}$
$\sqrt{v_1^2 + v_2^2}$
$|v_2|$
$|v_1|$



A vector of length 1 is called a **unit vector**.

If we divide a nonzero vector $\mathbf{v}$ by its length – that is, multiply by $1/\|\mathbf{v}\|$ – we obtain a unit vector $\mathbf{u}$.

We say that we have *normalized* $\mathbf{v}$, and that $\mathbf{u}$ is *in the same direction* as $\mathbf{v}$.

1.6 Distance in $\mathbb{R}^n$

It's very useful to be able to talk about the **distance** between two points (or vectors) in $\mathbb{R}^n$.

Definition. For $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$, the **distance between u and v,** written as $\text{dist}(\mathbf{u}, \mathbf{v})$, is the length of the vector $\mathbf{u} - \mathbf{v}$. That is,

$$\text{dist}(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|.$$

This definition agrees with the usual formulas for the Euclidean distance between two points. The usual formula is

$$\text{dist}(\mathbf{u}, \mathbf{v}) = \sqrt{(v_1 - u_1)^2 + (v_2 - u_2)^2 + \cdots + (v_n - u_n)^2}.$$

Which you can see is equal to

$$\|\mathbf{u} - \mathbf{v}\| = \sqrt{(\mathbf{u} - \mathbf{v})^T(\mathbf{u} - \mathbf{v})} = \sqrt{\begin{bmatrix} u_1 - v_1 & u_2 - v_2 & \dots & u_n - v_n \end{bmatrix} \begin{bmatrix} u_1 - v_1 \\ u_2 - v_2 \\ \vdots \\ u_n - v_n \end{bmatrix}}$$

There is another important reformulation of distance. Consider the squared distance:

$$\|\mathbf{u}-\mathbf{v}\|^2 = (\mathbf{u}-\mathbf{v})^T(\mathbf{u}-\mathbf{v})$$

Expanding this out, we get:

$$\|\mathbf{u}-\mathbf{v}\|^2 = \mathbf{u}^T\mathbf{u} + \mathbf{v}^T\mathbf{v} - 2\mathbf{u}^T\mathbf{v}.$$

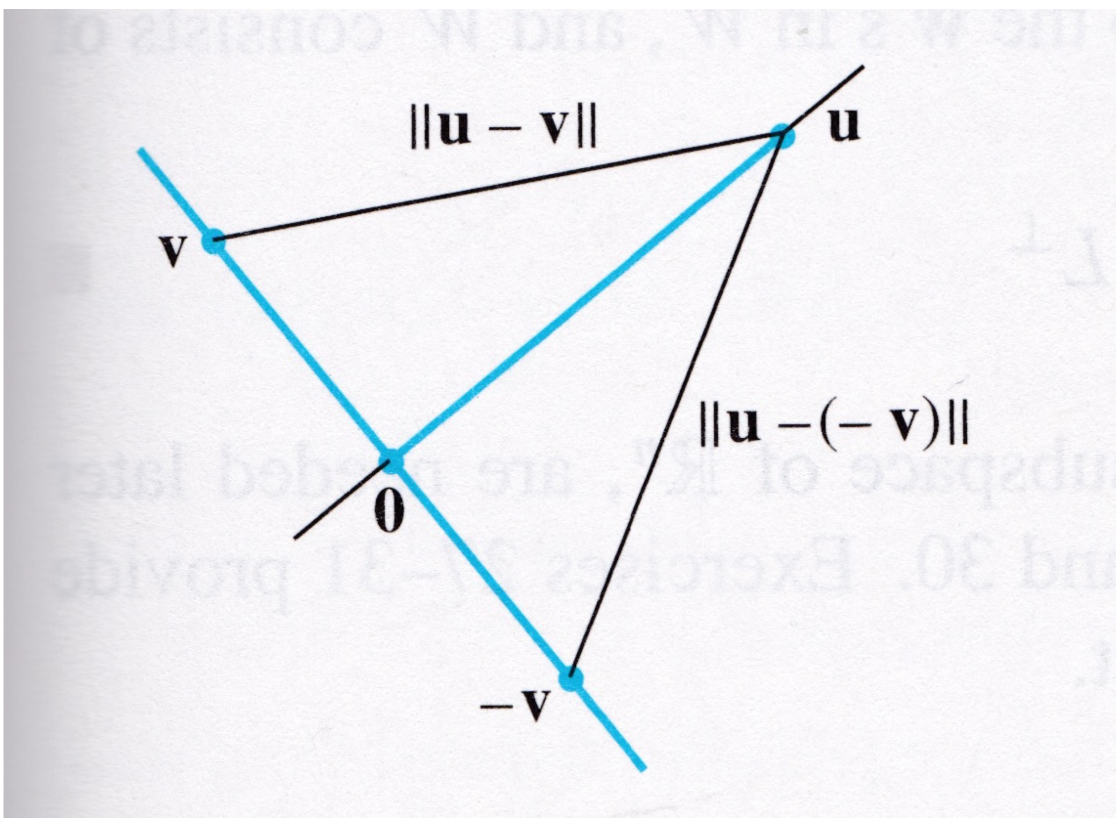
$$= \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - 2\mathbf{u}^T\mathbf{v}.$$

This is an important formula because it relates **distance** between two points to their **lengths** and **inner product.**

1.7 Orthogonal Vectors

Now we turn to another familiar notion from 2D geometry, which we'll generalize to $\mathbb{R}^n$: the notion of being **perpendicular.**

You may recall the classic method from Euclid for how to construct a line perpendicular to another line at a point:



One constructs an isoceles triangle centered at the point. Because the sides are equal, the two inner triangles are right triangles.

So the two blue lines are perpendicular if and only if the distance from $\mathbf{u}$ to $\mathbf{v}$ is equal to the distance from $\mathbf{u}$ to $-\mathbf{v}$.

This is the same as requiring the squares of their distances to be equal.

We have already seen that the distance from $\mathbf{u}$ to $\mathbf{v}$ is:

$$[\text{dist}(\mathbf{u},\mathbf{v})]^2 = \|\mathbf{u}-\mathbf{v}\|^2$$

$$= \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - 2\mathbf{u}^T\mathbf{v}$$

Similarly, the distance from $\mathbf{u}$ to $-\mathbf{v}$ is

$$= \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 + 2\mathbf{u}^T\mathbf{v}$$

So $\text{dist}(\mathbf{u},\mathbf{v}) = \text{dist}(\mathbf{u},-\mathbf{v})$ if and only if $\mathbf{u}^T\mathbf{v} = 0$.

So now we can define perpendicularity in $\mathbb{R}^n$:

Definition. Two vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$ are **orthogonal** to each other if $\mathbf{u}^T\mathbf{v} = 0$.

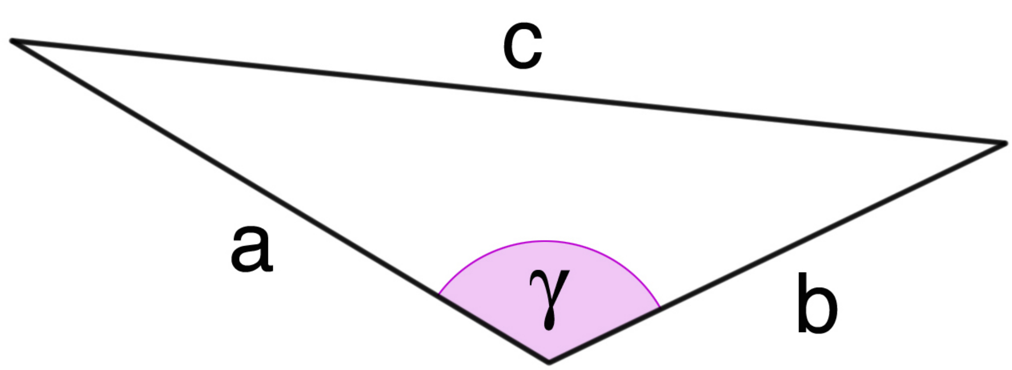
1.8 The Angle Between Two Vectors

There is an important connection between the inner product of two vectors and the **angle** between them.

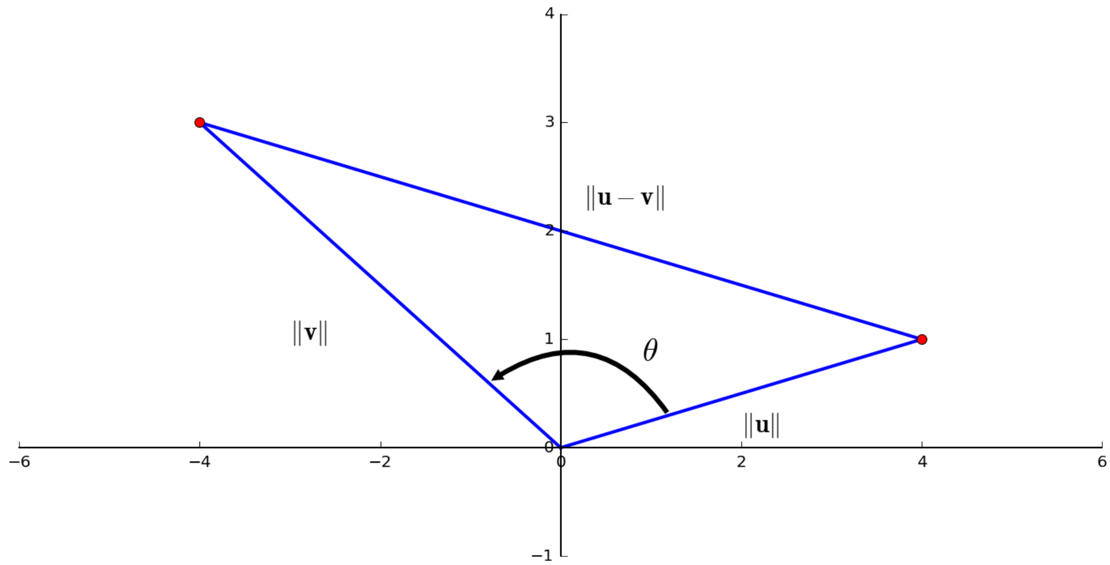
This connection is very useful (eg, in visualizing data mining operations).

We start from the **law of cosines:**

$$c^2 = a^2 + b^2 - 2ab\cos\gamma$$



Now let's interpret this law in terms of vectors $\mathbf{u}$ and $\mathbf{v}$:



Applying the law of cosines we get:

$$\|\mathbf{u}-\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - 2\|\mathbf{u}\|\|\mathbf{v}\|\cos\theta$$

And by definition:

$$\|\mathbf{u}-\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - 2\mathbf{u}^T\mathbf{v}$$

So

$$\mathbf{u}^T\mathbf{v} = \|\mathbf{u}\|\|\mathbf{v}\|\cos\theta$$

This is a **very** important connection between the notion of inner product and trigonometry. One implication in particular concerns **unit vectors.**

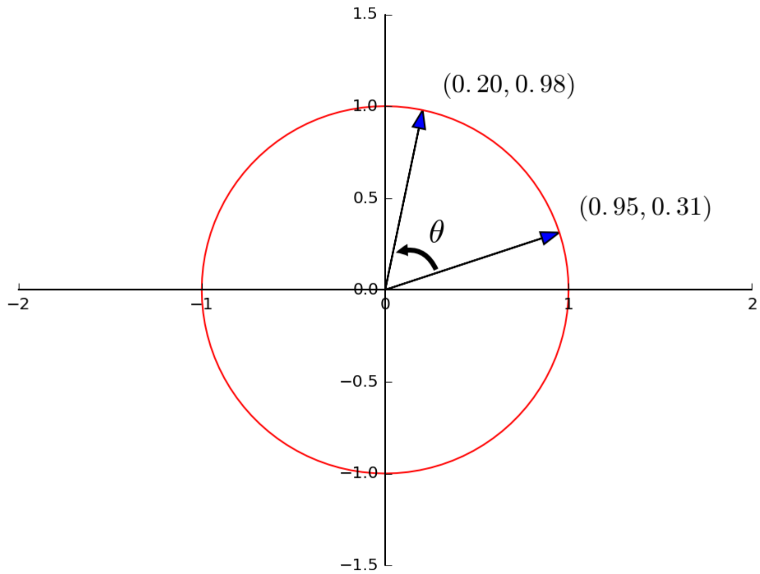
$$\mathbf{u}^T\mathbf{v} = \|\mathbf{u}\|\|\mathbf{v}\|\cos\theta$$

So

$$\frac{\mathbf{u}^T\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} = \cos\theta$$

$$\frac{\mathbf{u}^T}{\|\mathbf{u}\|}\frac{\mathbf{v}}{\|\mathbf{v}\|} = \cos\theta$$

So we have the very simple rule, that for two unit vectors, their inner product is the cosine of the angle between them!



Here $\mathbf{u} = \begin{bmatrix} 0.95 \\ 0.31 \end{bmatrix}$, and $\mathbf{v} = \begin{bmatrix} 0.20 \\ 0.98 \end{bmatrix}$.
So $\mathbf{u}^T\mathbf{v} = (0.95 \cdot 0.20) + (0.31 \cdot 0.98) = 0.5$
So $\cos\theta = 0.5$.
So $\theta = 60$ degrees.

1.9 Least Squares

In many cases we have a linear system

$$A\mathbf{x} = \mathbf{b}$$

that has no solution – perhaps due to noise or measurement error.

In such as case, we generally look for an $\mathbf{x}$ such that $A\mathbf{x}$ makes a good **approximation** to $\mathbf{b}$.

We can think of the quality of the approximation of $A\mathbf{x}$ to $\mathbf{b}$ as the distance from $A\mathbf{x}$ to $\mathbf{b}$, which is

$$\|A\mathbf{x} - \mathbf{b}\|.$$

The **general least-squares problem** is to find an $\mathbf{x}$ that makes $\|A\mathbf{x} - \mathbf{b}\|$ as small as possible.

Just to make this explicit: say that we denote $A\mathbf{x}$ by $\mathbf{y}$. Then

$$\|A\mathbf{x} - \mathbf{b}\|^2 = \sum_i (y_i - b_i)^2$$

Where we interpret y_i as the *estimated value* and b_i as the *measured value.*

So this expression is the **sum of squared error.** This is the most common measure of error used in statistics. Note that we could also say we are minimizing the ℓ_2 norm of the error.

This is a key principle!

Minimizing the length of $A\mathbf{x} - \mathbf{b}$ is the same as minimizing the sum of squared error.

An equivalent (and more common) way to express this is:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|.$$

which emphasizes that this is a minimization problem, also called an *optimization* problem. We can find $\hat{\mathbf{x}}$ using either

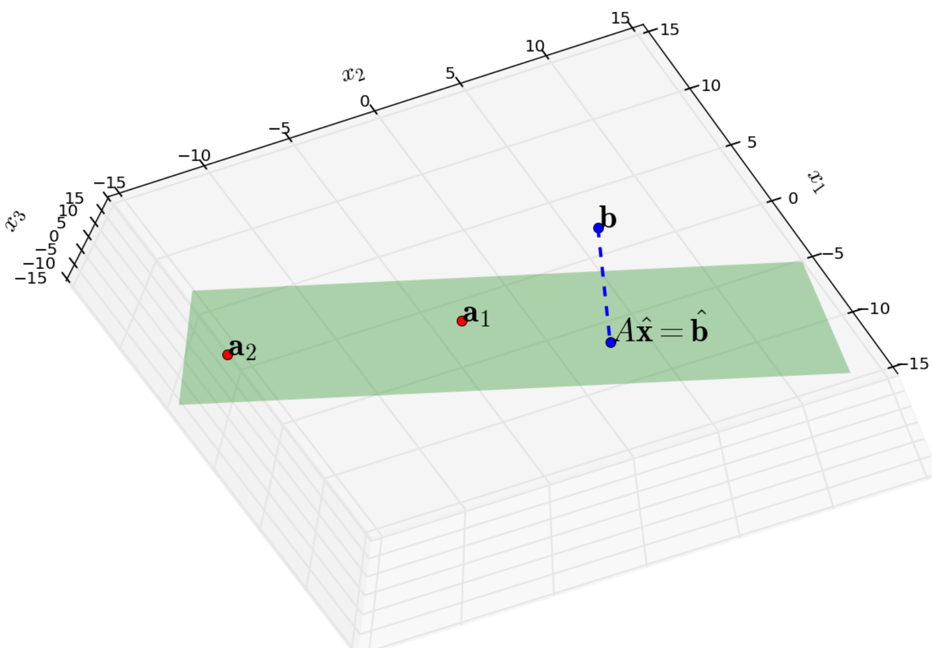
- geometric arguments (based on projecting $\mathbf{b}$ on the column space of A), or
- by calculus (taking the derivative of the rhs expression above and setting it equal to zero).

Either way, we get the result that:
$\hat{\mathbf{x}}$ is the solution of:

$$A^T A\hat{\mathbf{x}} = A^T\mathbf{b}$$

And we can then prove that $A^T A$ is always invertible, and so solve $A^T A\mathbf{x} = A^T\mathbf{b}$ as

$$\hat{\mathbf{x}} = (A^T A)^{-1} A^T\mathbf{b}$$

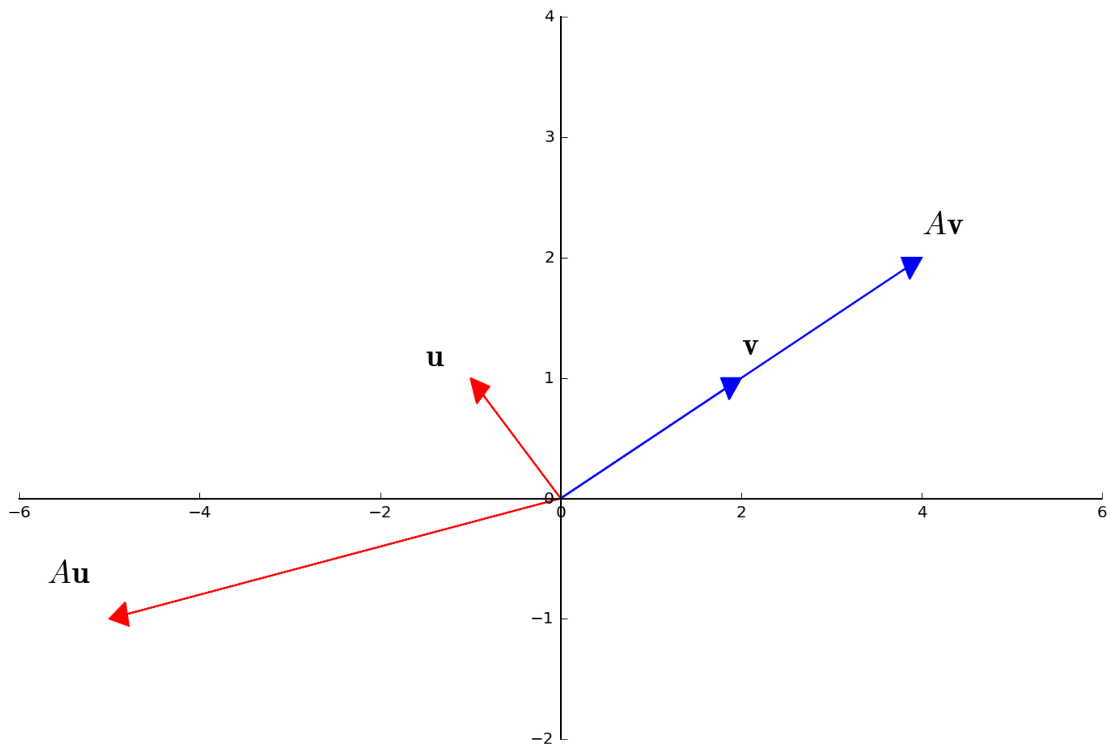


1.10 Eigendecomposition

Consider a square matrix A. An **eigenvector** of A is a special vector which **does not change its direction** when multiplied by A.

Example.

Let $A = \begin{bmatrix} 3 & -2 \\ 1 & 0 \end{bmatrix}$, $\mathbf{u} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$, $\mathbf{v} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$. The images of $\mathbf{u}$ and $\mathbf{v}$ under multiplication by A are shown here:



Definition.

An **eigenvector** of an $n \times n$ matrix A is a nonzero vector $\mathbf{x}$ such that $A\mathbf{x} = \lambda\mathbf{x}$ for some scalar λ.

A scalar λ is called an **eigenvalue** of A if there is a nontrivial solution $\mathbf{x}$ of $A\mathbf{x} = \lambda\mathbf{x}$.

Such an $\mathbf{x}$ is called an *eigenvector corresponding to* λ.

An $n \times n$ matrix has at most n distinct eigenvectors and at most n distinct eigenvalues.

In some cases, the matrix may be **factorable** using it eigenvectors and eigenvalues.

This is called **diagonalization.**

A square matrix A is said to be **diagonalizable** if it can be expressed as:

$$A = PDP^{-1}$$

for some invertible matrix P and some diagonal matrix D.

In fact, $A = PDP^{-1}$, with D a diagonal matrix, if and only if the columns of P are n linearly independent eigenvectors of A.

In that case, D holds the corresponding eigenvalues of A.

That is,

$$D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

and

$$P = [\mathbf{v}_1 \; \mathbf{v}_2 \; \cdots \; \mathbf{v}_n].$$

A special case occurs when A is **symmetric,** i.e., $A = A^T$.

In that case, the eigenvectors of A are all mutually **orthogonal.**

We can scale each eigenvector to have unit norm, in which case the eigenvectors are **orthonormal.**

So, that means that $PP^T = I$.

So $P^{-1} = P^T$.

So, in the special case of a symmetric matrix, we can decompose A as:

$$A = PDP^T.$$

These facts will be key to understanding the **Singular Value Decomposition,** which we will study later on.