Linear Algebra

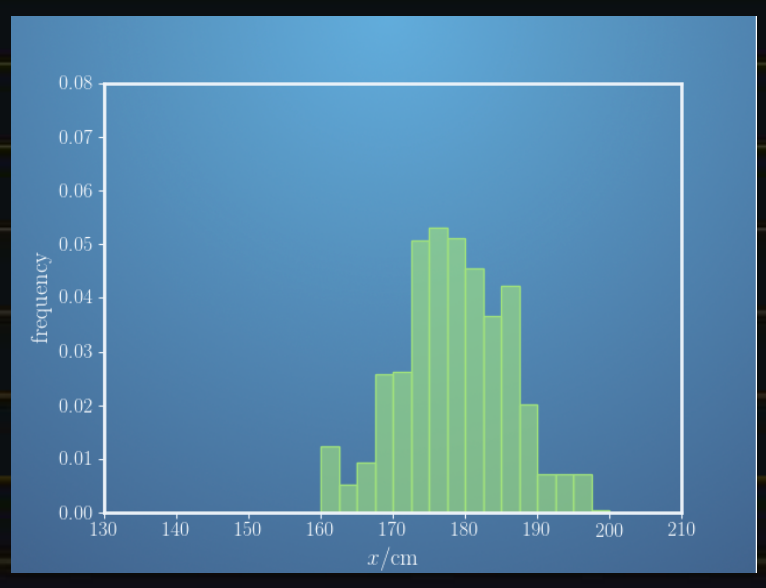
(*) *a = cost of 1 apple in $*

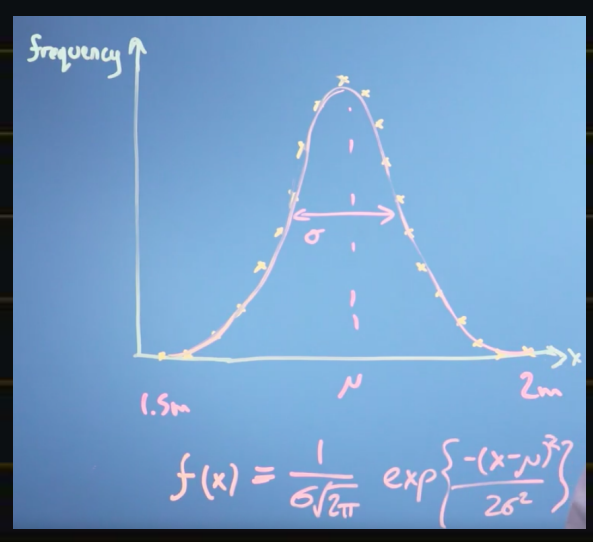
b = cost of 1 banana in $

$$2a + 3b = 8$$
$$10a + b = 13$$
$$\Rightarrow \underbrace{\begin{bmatrix} 2 & 3 \\ 10 & 1 \end{bmatrix}}_{\text{matrix}} \underbrace{\begin{bmatrix} a \\ b \end{bmatrix}}_{\text{vectors}} = \begin{bmatrix} 8 \\ 13 \end{bmatrix}$$

(*) *Fitting a curve to a set of data.*

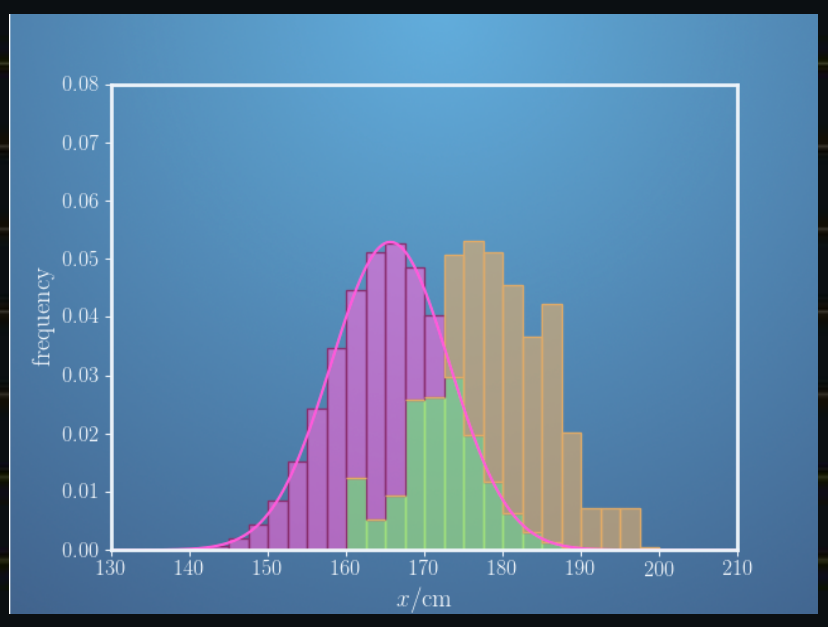
For eg., given the height frequencies of a population, we may use a normal distribution for fitting.





In order to fit a normal distribution to the above data, we need to find good enough values of μ *and* σ*, i.e. the mean and the standard deviation, respectively.*

(This normal distribution has nothing to do with a probability density function)



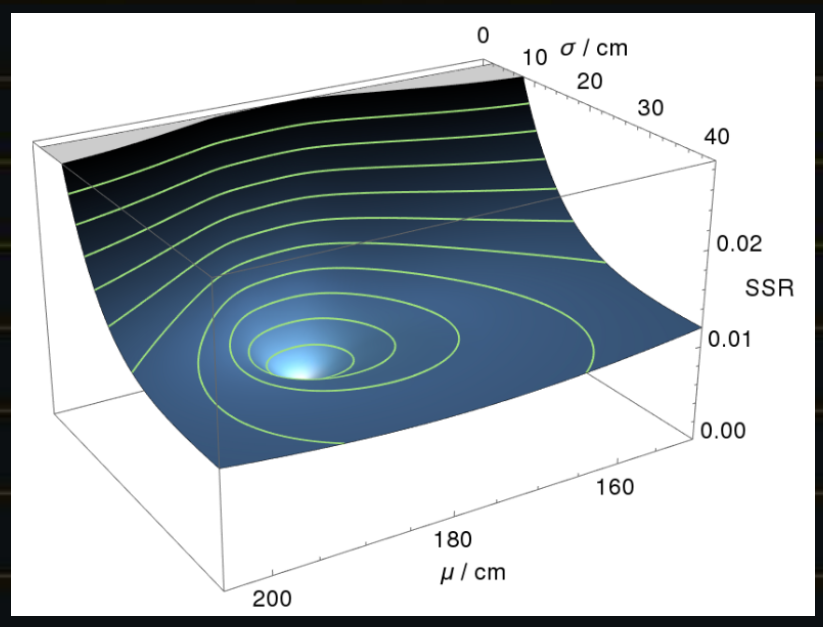
For eg., in this case, the pink bars represent the frequencies according to our assumption.

In order to make our model good enough, we need to minimize the residuals.

The performance of a model can be quantified in a single number. One measure we can use is the Sum of Squared Residuals (SSR). Here, we take all of the residuals (the differences between the measured and the predicted data), square them and add them together.

(*) *For a normal distribution, let the x-axis represent the mean, μ, as an independent variable, and the y-axis represent the standard deviation, σ, as another independent variable.*

Also, let the z-axis represent the corresponding SSR as the dependent variable.

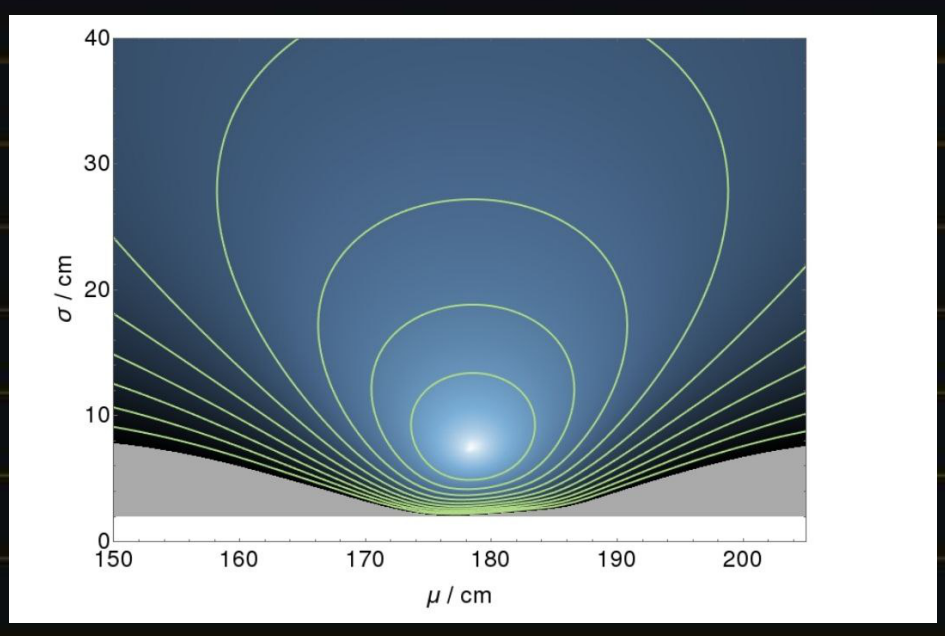


For every green curve, all points on it have the same value of SSR.

This means that multiple combinations of values of μ and σ can lead to the same value of SSR.

Instead of drawing a 3-d plot, we can also draw a 2-d contour plot to represent the same thing.

A contour plot is a graphical technique for representing a 3-dimensional surface by plotting constant z slices, called contours, on a 2-dimensional format. That is, given a value for z, lines are drawn for connecting the (x, y) coordinates where that value of z occurs.



(*) Practice Quiz: Exploring parameter space

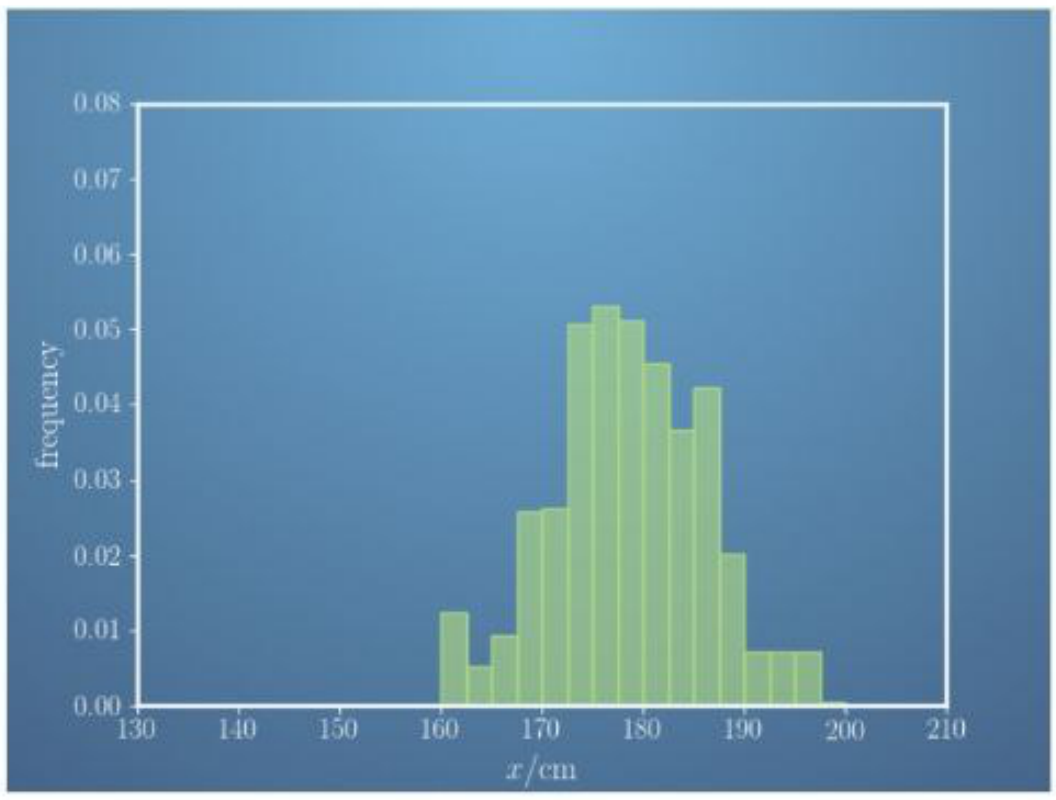
1. In this exercise, we shall see how it is often convenient to use vectors in machine learning. These could be in the form of data itself, or model parameters, and so on. **1.6 / 2 points**

The purpose of this exercise is to set the scene for Linear Algebra and the rest of the maths we will cover in the specialization. If this is confusing right now - stick with us! We'll build up your skills throughout the rest of the course. For this reason we've set a low pass mark for this quiz, but even if you don't pass in one go, reading the feedback from a wrong answer can often give more insight than guessing a correct answer!

* * *

The problem we shall focus on in this exercise is the distribution of heights in a population.

If we do a survey of the heights of people in a population, we may get a distribution like this:



This *histogram* indicates how likely it is for anyone in the survey to be in a particular height *range*. (6 ft is around 183 cm)

This histogram can also be represented by a vector, i.e. a list of numbers. In this case, we record the frequency of people with heights in little groups at 2.5 cm intervals, i.e. between 150 cm and 152.5 cm, between 152.5 cm and 155 cm, and so on. We can define this as the vector $\mathbf{f}$ with components,

$$\mathbf{f} = \begin{bmatrix} f_{150.0,\ 152.5} \\ f_{152.5,\ 155.0} \\ f_{155.0,\ 157.5} \\ f_{157.5,\ 160.0} \\ f_{160.0,\ 162.5} \\ \vdots \end{bmatrix}$$

These vector components are then the sizes of each bar in the histogram.

Of the following statements, select all that you think are true.

- [x] There are at least 10 elements in the frequency vector, $\mathbf{f}$.

  Correct
  The data has been grouped into around 20 bins each having a width of 2.5 cm, in the range between 150 cm and 210 cm. Around 15 of these have a non-zero frequency.

- [x] If another sample was taken under the same conditions, the frequencies should be broadly similar.

  Correct
  For a sufficiently large sample, the data will represent the population it is taken from.

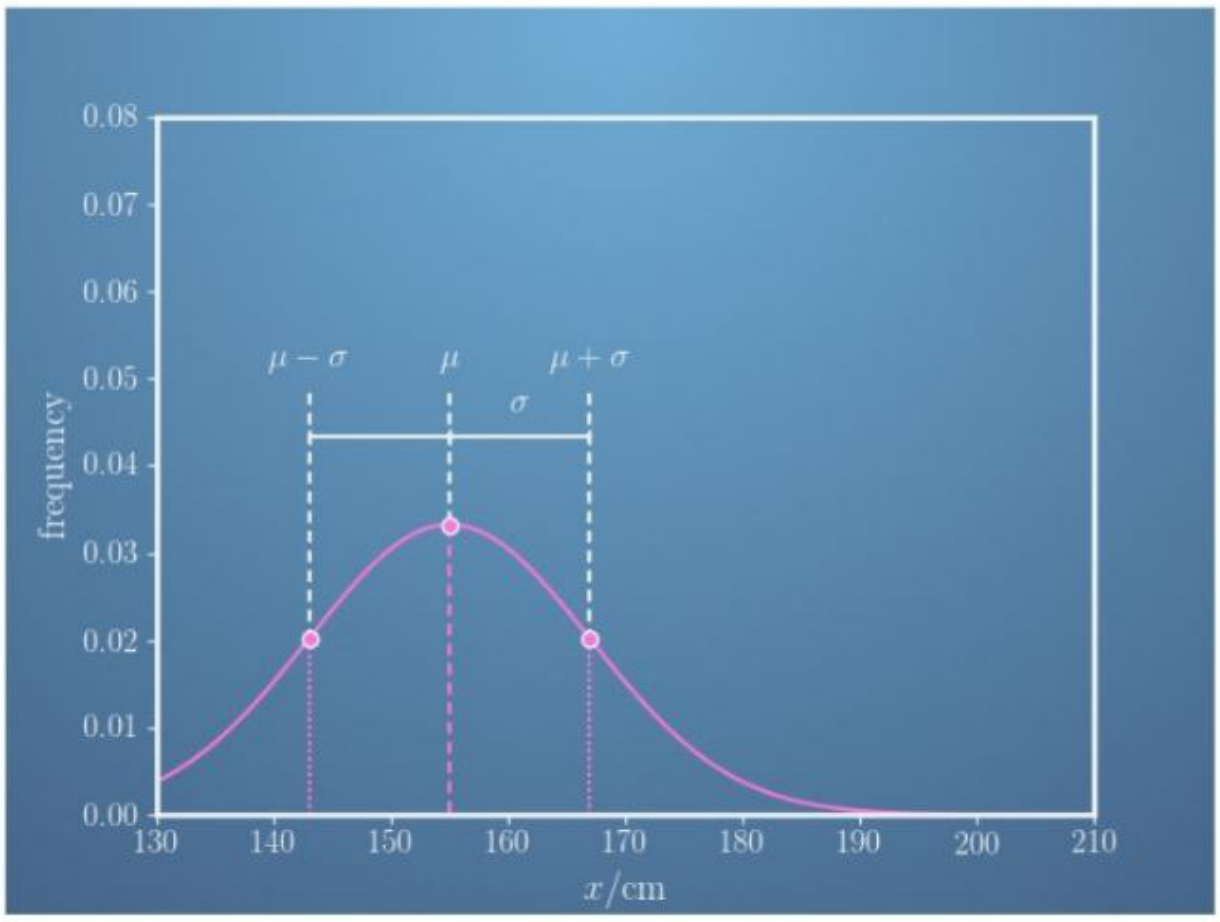
- [x] If another sample was taken under the same conditions, the frequencies would be exactly the same.

  This should not be selected
  The data is just a sample of the population. If the sample is large enough it will be representative of the population it is taken from, but there will always be fluctuations around the population distribution.

- [ ] None of the other statements.
- [ ] No one in the world is less than 160 cm tall.

2. One of the tasks of machine learning is to fit a model to data in order to represent the underlying distribution. **1 / 1 point**

For the heights of a population, a model we may use to predict frequencies is the Normal (or Gaussian) distribution. This is a model for a bell-shaped curve, which looks like this,



It has the slightly complicated equation,

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

the exact form of which is unimportant, except that it is dependent on two parameters, the *mean*, μ, where the curve is centred, and the *standard deviation*, σ, which is the characteristic width of the bell curve (measured from the mean).

We can put these two parameters in a vector, $\mathbf{p} = \begin{bmatrix} \mu \\ \sigma \end{bmatrix}$.

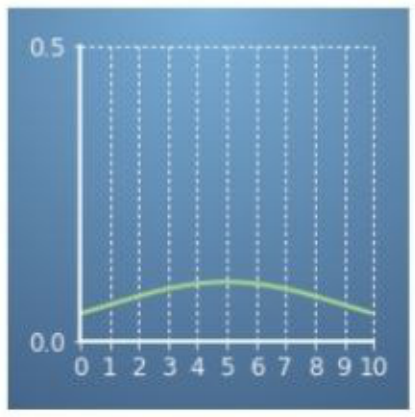
Pick the parameter vector $\mathbf{p}$ which best describes the distribution pictured.

- ○ $\mathbf{p} = \begin{bmatrix} 167 \\ 24 \end{bmatrix}$
- ○ $\mathbf{p} = \begin{bmatrix} 143 \\ 167 \end{bmatrix}$
- ○ $\mathbf{p} = \begin{bmatrix} 167 \\ 12 \end{bmatrix}$
- ◉ $\mathbf{p} = \begin{bmatrix} 155 \\ 12 \end{bmatrix}$
- ○ $\mathbf{p} = \begin{bmatrix} 155 \\ 3 \end{bmatrix}$

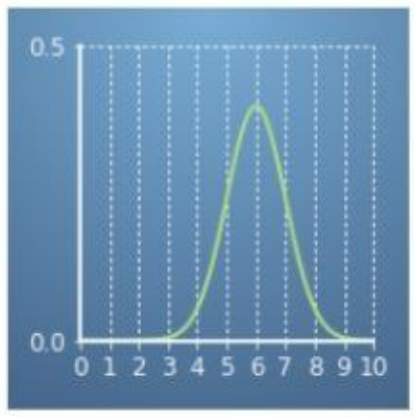
Correct
The mean is 155 cm and the standard deviation is 12 cm.

3. Pick the Normal distribution that corresponds the closest to the parameter vector $\mathbf{p} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$.

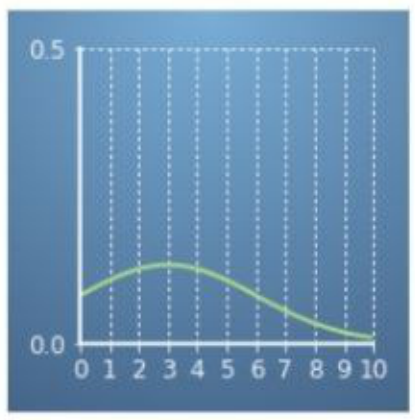
1 / 1 point



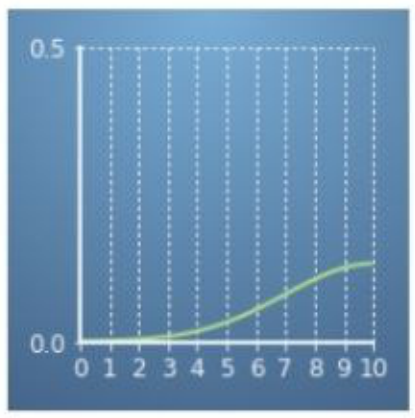
$$$$



$$$$



$$$$



$$$$

Correct

This distribution has parameters, $\mathbf{p} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$.

4. A model allows us to predict the data in a distribution. In our example we can start with a parameter vector $\mathbf{p}$ and convert it to a vector of expected frequencies $\mathbf{g_p}$, for example,

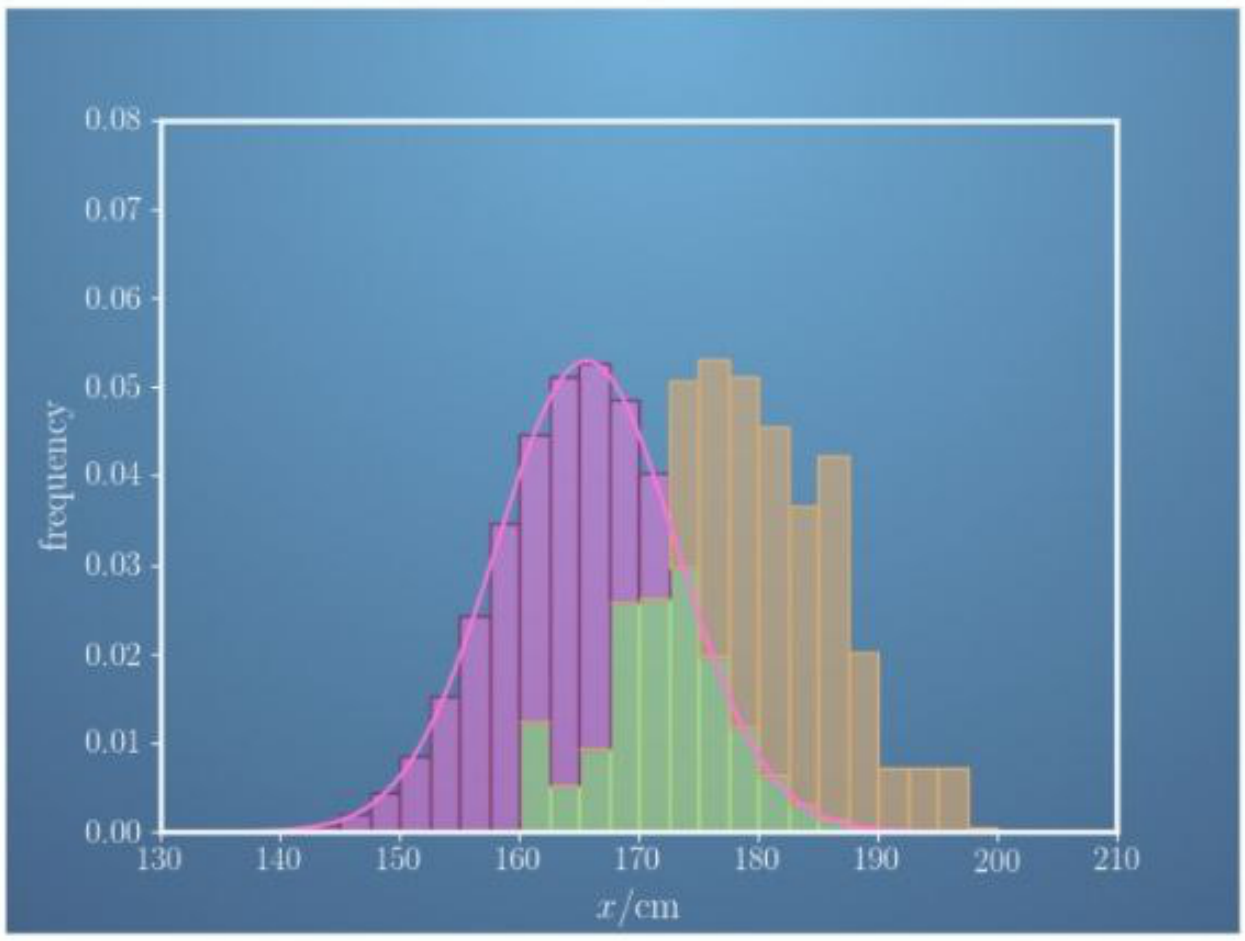
1 / 1 point

$$\mathbf{g_p} = \begin{bmatrix} g_{150.0,\,152.5} \\ g_{152.5,\,155.0} \\ g_{155.0,\,157.5} \\ g_{157.5,\,160.0} \\ g_{160.0,\,162.5} \\ \vdots \end{bmatrix}$$

A model is only considered good if it fits the measured data well. Some specific values for the parameters will be better than others for a model. We need a way fit a model's parameters to data and quantify how good that fit is.

One way of doing so is to calculate the "residuals", which is the difference between the measured data and the modelled prediction for each histogram bin.

This is illustrated below. The model is shown in pink, the measured data is shown in orange and where they overlap is shown in green. The height of the pink and orange bars are the residuals.



A better fit would have as much overlap as it can, reducing the residuals as much as possible.

How could the model be improved to give the best fit to the data?

- [x] Increase the mean, μ.

Correct
The mean of the model is too low.

- [x] Keep the standard deviation, σ, approximately the same.

Correct
The model has a width similar to the data.

- [ ] Keep the mean, μ, approximately the same.
- [ ] Decrease the mean, μ.
- [ ] Increase the standard deviation, σ.
- [ ] Decrease the standard deviation, σ.

5. The performance of a model can be quantified in a single number. One measure we can use is the *Sum of Squared Residuals*, SSR. Here we take all of the residuals (the difference between the measured and predicted data), square them and add them together. **1 / 1 point**

In the language of vectors we can write this as, $\mathrm{SSR}(\mathbf{p}) = |\mathbf{f} - \mathbf{g}_\mathbf{p}|^2$, which will be explained further on in this course.

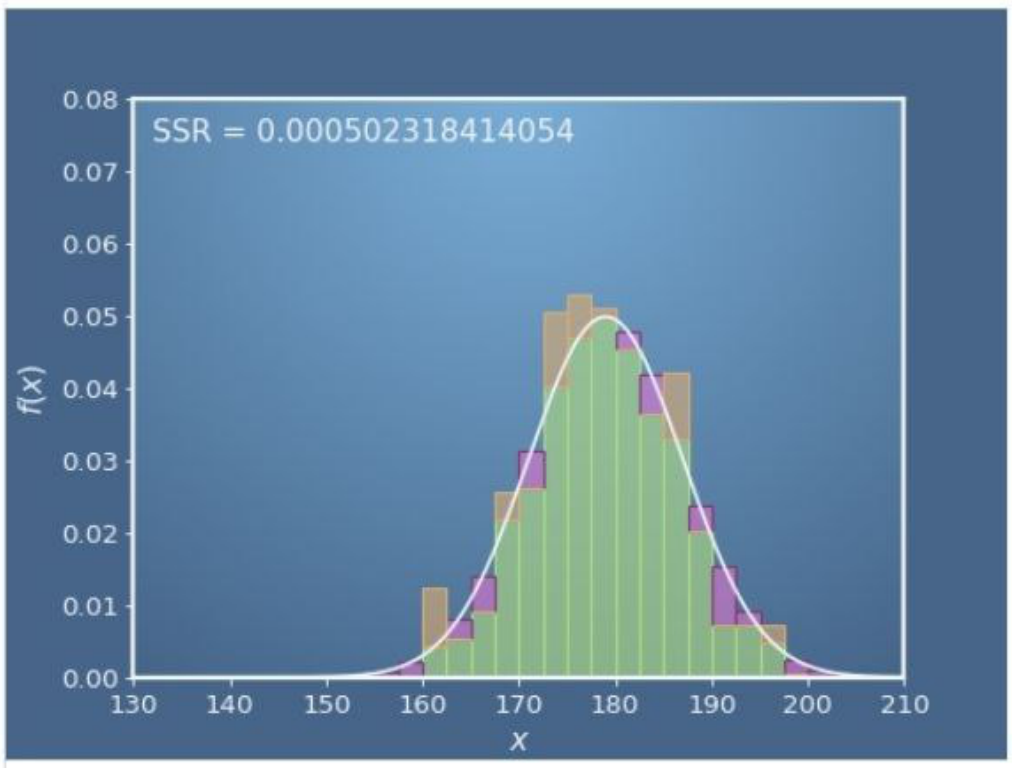
Use the following code block to play with parameters of a model, and try to get the best fit to the data.

```
# Play with values of μ and σ to find the best fit.
μ = 179 ; σ = 8
p = [μ, σ]
histogram(p)

```

Run

Reset



Find a set of parameters with a fit $\mathrm{SSR} \leq 0.00051$

Input your fitted parameters into the code block below.

```
# Replace μ and σ with values that minimise the SSR.
p = [179, 8]

```

Run

Reset

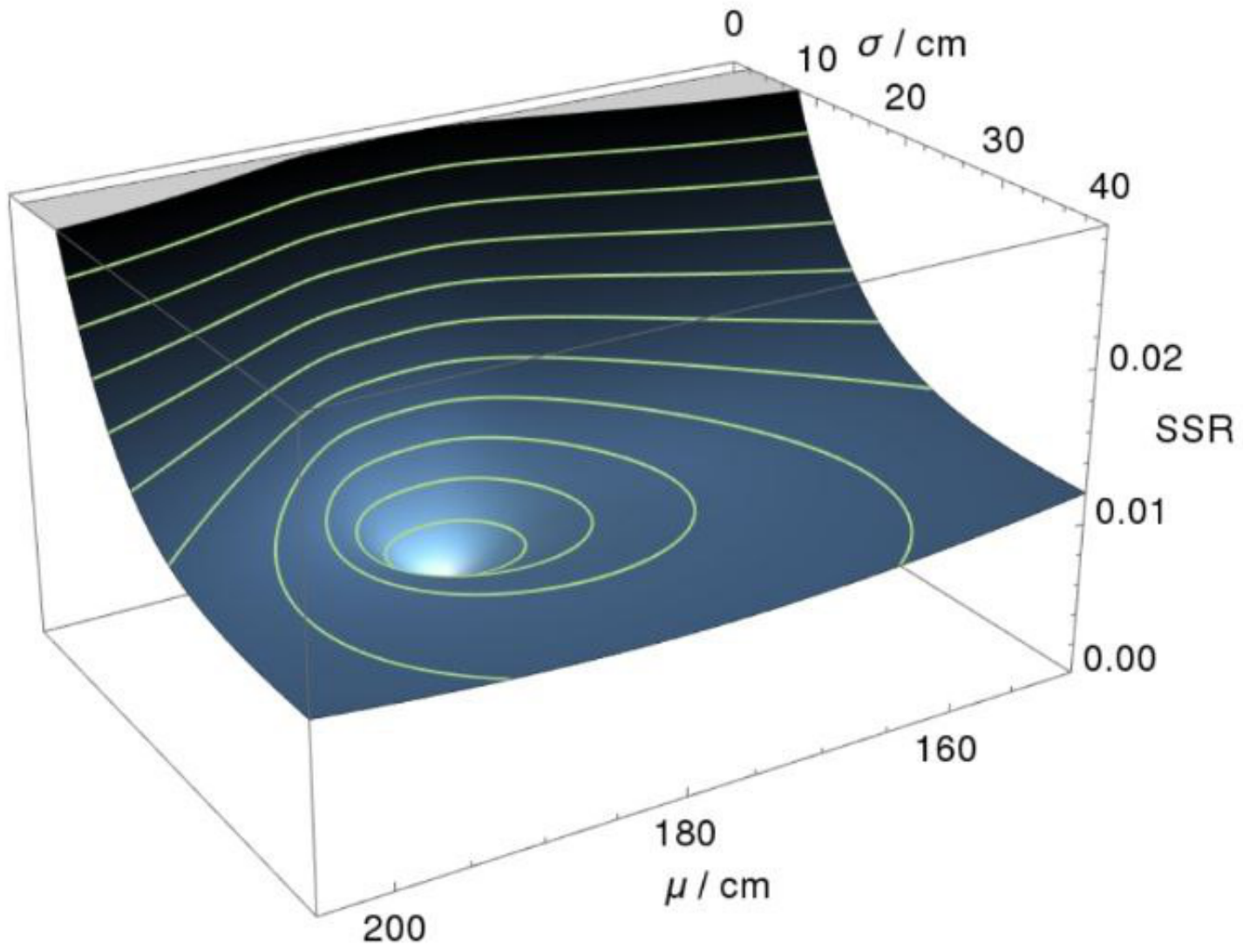
Correct

Well done! You found a model that fits the data acceptably well according the the criterion defined for SSR.

6. Since each parameter vector $\mathbf{p}$ represents a different bell curve, each with its own value for the sum of squared residuals, SSR, we can draw the surface of SSR values over the space spanned by $\mathbf{p}$, such as μ and σ in this example.

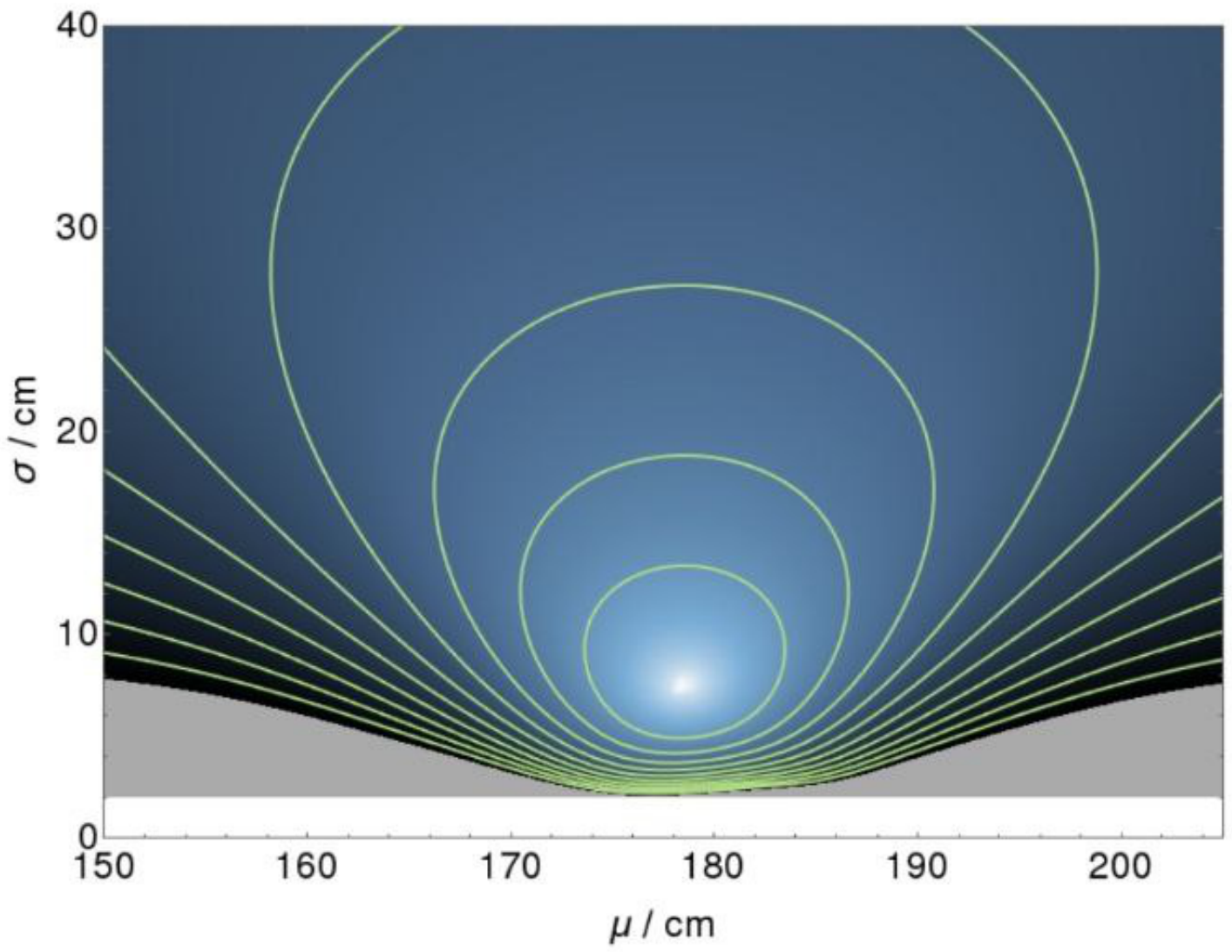
1 / 1 point

Here is an illustration of this surface for our data.



Every point on this surface represents the SSR of a choice of parameters, with some bell curves performing better at representing the data than others.

We can take a 'top-down' view of the surface, and view it as a contour map, where each of the contours (in green here) represent a constant value for the SSR.



The goal in machine learning is to find the parameter set where the model fits the data as well as it possibly can. This translates into finding the lowest point, the global minimum, in this space.

Select all true statements below.

- [ ] None of the other statements.
- [x] Each point on the surface represents a set of parameters $\mathbf{p} = \begin{bmatrix} \mu \\ \sigma \end{bmatrix}$.

Correct
This means each point in the space will generate a different histogram of expected data, which will perform better or worse against the measured data.

- [ ] You get the same model by following along a contour line.
- [ ] At the minimum of the surface, the model exactly matches the measured data.
- [x] Moving at right angles to contour lines in the parameter space will have the greatest effect on the fit than moving in other directions.

Correct
For example, moving along contour lines has no affect on the SSR (by definition). However moving perpendicular to them can significantly improve or reduce the quality of the fit.

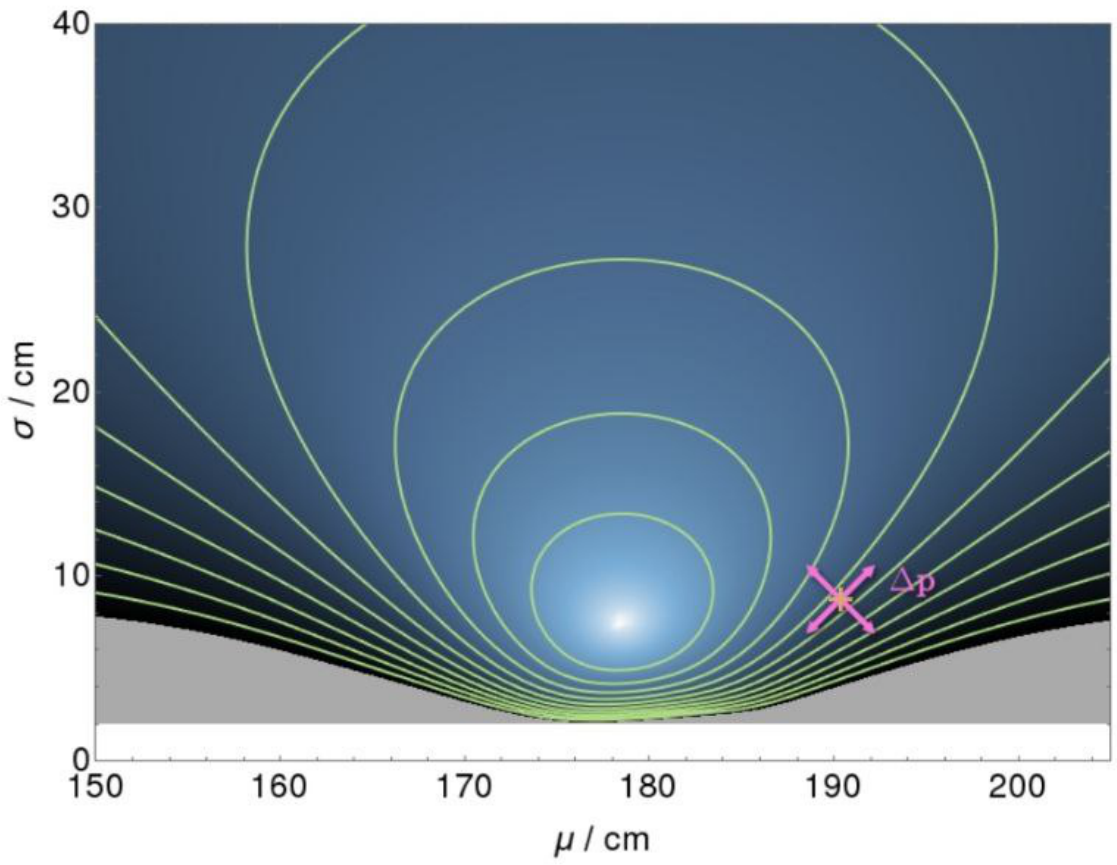
7. Often we can't see the whole parameter space, so instead of just picking the lowest point, we have to make educated guesses where better points will be. **1 / 1 point**

We can define another vector, $\Delta\mathbf{p}$, in the same space as $\mathbf{p}$ that tells us what change can be made to $\mathbf{p}$ to get a better fit.

For example, a model with parameters $\mathbf{p}' = \mathbf{p} + \Delta\mathbf{p}$ will produce a better fit to data, if we can find a suitable $\Delta\mathbf{p}$.

The second course in this specialisation will detail how to calculate these changes in parameters, $\Delta\mathbf{p}$.

Given the following contour map,



What $\Delta\mathbf{p}$ will give the best improvement in the model?

- ○ $\Delta\mathbf{p} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$
- ○ $\Delta\mathbf{p} = \begin{bmatrix} -2 \\ -2 \end{bmatrix}$
- ◉ $\Delta\mathbf{p} = \begin{bmatrix} -2 \\ 2 \end{bmatrix}$
- ○ $\Delta\mathbf{p} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$

Correct
This direction will decrease the SSR making the fit better.

(*) *Practice Quiz: Solving some simultaneous equations*

1. In this quiz you'll be reminded of how to solve linear simultaneous equations as a way to practice some basic linear algebra. Some of the ideas presented here will be relevant later in the course. **1 / 1 point**

Solving simultaneous equations is the process of finding the values of the variables (here x and y) that satisfy the system of equations. Let's start with the simplest type of simultaneous equation, where we already know all but one of the variables:

$3x - y = 2$

$x = 4$

Substitute the value of x into the first equation to find y, then select the correct values of x and y below.

- ○ $x = 4, y = -10$
- ○ $x = 4, y = 2$
- ◉ $x = 4, y = 10$
- ○ $x = 4, y = 14$

✓ **Correct**
When you know one of the variables, substituting it into one of the equations is a good way to find the other variable.

x = 4

3x - y = 2

So, 3(4) - y = 2

=> 12 - y = 2

=> y = 10

2. The first goal when solving simple simultaneous equations should be to isolate one of the variables. For example, try taking the second equation away from the first to solve the following pair of equations: **1 / 1 point**

$3x - 2y = 7$

$2x - 2y = 2$

What value did you find for x? Now substitute x into one of the equations to find y, and select the correct pair below:

- ○ $x = 1, y = -4$
- ○ $x = 3, y = 1$
- ◉ $x = 5, y = 4$
- ○ $x = 7, y = 7$

✓ **Correct**
Elimination can be a useful method to solve a simple system of linear equations.

3x - 2y = 7 ... (i)

2x - 2y = 2 ... (ii)

(i) - (ii),

x = 5 ... (iii)

So, from (i) and (iii), 3(5) - 2y = 7 => 15 - 2y = 7 => 2y = 8 => y = 4

3. This method is called elimination, and you can use it even when the coefficients, the numbers in front of x and y, aren't the same. **1 / 1 point**

For example, to solve the following equations try multiplying both sides of the first equation by 2, then solve using the same method as the last question.

$3x - 2y = 4$

$6x + 3y = 15$

Select the correct values of x and y below:

- ○ $x = 3, y = 1$
- ○ $x = 4, y = -2$
- ◉ $x = 2, y = 1$
- ○ $x = 1, y = 2$

Correct

We've seen that elimination can be a useful method to solve a simple system of linear equations.

3x – 2y = 4 ... (i)

6x + 3y = 15 ... (ii)

(ii) – (2 x (i)),

(6x + 3y) – (6x – 4y) = 15 – 8 => 7y = 7 => y = 1 ... (iii)

So, from (i) and (iii), 3x – 2(1) = 4 => 3x = 6 => x = 2

4. A very similar technique can be used to find the inverse of a matrix, which you will learn about in week three of this course. **1 / 1 point**

There is also the substitution method, where we rearrange one of the equations to the form $x = ay + b$ or $y = cx + d$ and then substitute x or y into the other equation. Use any method you'd like to solve the following simultaneous equations:

$-2x + 2y = 20$

$5x + 3y = 6$

Select the correct values of x and y below:

- ◉ $x = -3, y = 7$
- ○ $x = -5, y = 5$
- ○ $x = 3, y = 13$
- ○ $x = 5, y = 15$

Correct

Substitution and elimination are useful techniques for solving simple systems of linear equations.

–2x + 2y = 20 => –x + y = 10 => x = y – 10 ... (i)

5x + 3y = 6 ... (ii)

So, from (i) & (ii), 5 (y – 10) + 3y = 6 => 5y – 50 + 3y = 6 => 8y = 56 => y = 7 ... (iii)

So, from (i) & (iii), x = 7 – 10 => x = –3

5. Systems of simultaneous equations can have more than two unknown variables. Below there is a system with three; x, y and z. First try to find one of the variables by elimination or substitution, which will lead to two equations and two unknown variables. Continue the process to find all of the variables.

1 / 1 point

Which values of x, y and z solve the following equations?

$3x - 2y + z = 7$

$x + y + z = 2$

$3x - 2y - z = 3$

Before you move on you might like to think about how many equations you would need to uniquely determine four, five, or more variables. Are there are any other rules for how the equations have to be related? In week two of this course you will learn about linear independence, which is very closely related to this.

- ○ $x = -1, y = -3, z = 4$
- ○ $x = 2, y = -2, z = 2$
- ◉ $x = 1, y = -1, z = 2$
- ○ $x = 1, y = -1, z = -2$

Correct

Substitution and elimination can be extended to more than two variables.

3x − 2y + z = 7 ... (i)

x + y + z = 2 ... (ii)

3x − 2y − z = 3 ... (iii)

(i) − (iii), 2z = 4 => z = 2 ... (iv)

So, from (ii) and (iv), x + y = 0 ... (v)

(i) − (ii), 2x − 3y = 5 ... (vi)

(2 x (v)) − (vi), 5y = −5 => y = −1 ... (vii)

So, from (ii), (iv) and (vii), x − 1 + 2 = 2 => x = 1

(*) Operations with vectors

A vector can be thought of as an arrow that can move about a space without changing its magnitude or direction.

In Physics, we treat the space as a physical space, but in other fields such as Data Science, we can think about a vector moving in a space of data (parameter space).

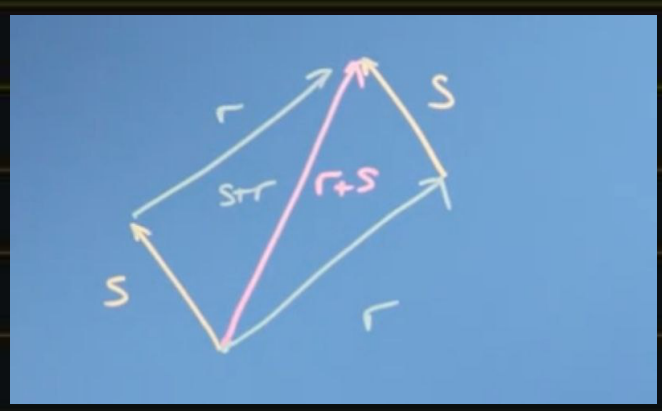
So, in this context, a vector can be thought of as a list of attributes of an object.

For eg., let a few of the attributes of a house be: an area of 120 sq. m, 2 bedrooms, 1 bathroom and a price of $150,000.

Then, we can represent this house as a vector like $\begin{bmatrix} 120 \\ 2 \\ 1 \\ 150 \end{bmatrix}$.

Now, vectors obey 2 rules – (i) addition, and (ii) multiplication by a scalar.

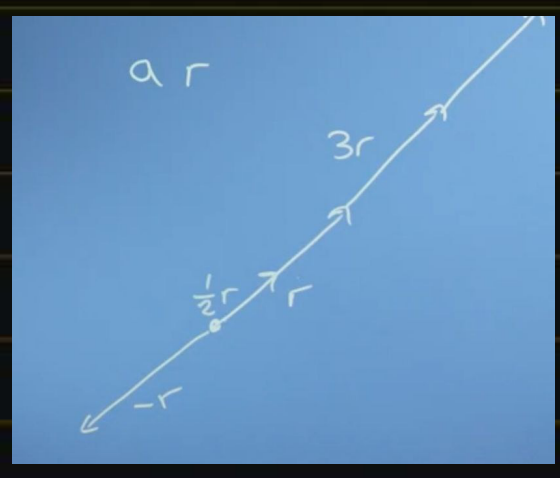
Addition of 2 vectors $\vec{r}$ and $\vec{s}$ –



$\vec{r} + \vec{s} = \vec{s} + \vec{r}$

Multiplication of a vector $\vec{r}$ by a scalar a –

Here, we have taken a to be 1/2, 3 and –1.

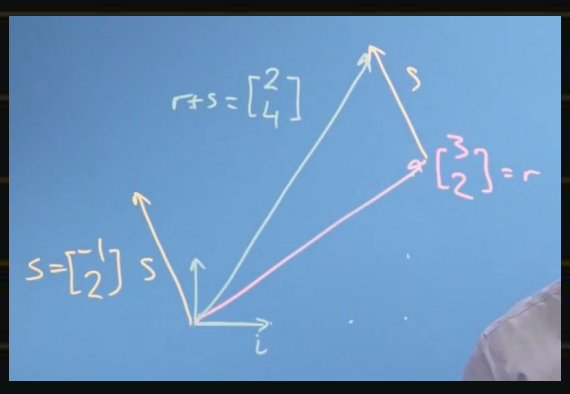


We take $\hat{i}$, $\hat{j}$ and $\hat{k}$ to be unit vectors along the x, y and z axes, respectively.

Here, the axes may represent different parameters, instead of representing the dimensions of a physical space.

So, in 2-d, a vector $\vec{r} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$ means $3\hat{i} + 2\hat{j}$.

Now, if we have another vector $\vec{s} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$, then in order to find $\vec{r} + \vec{s}$, we can simply add up the components, i.e. $\vec{r} + \vec{s} = \begin{bmatrix} 3 \\ 2 \end{bmatrix} + \begin{bmatrix} -1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 + (-1) \\ 2 + 2 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$.



Vector addition is commutative as well as associative, i.e.

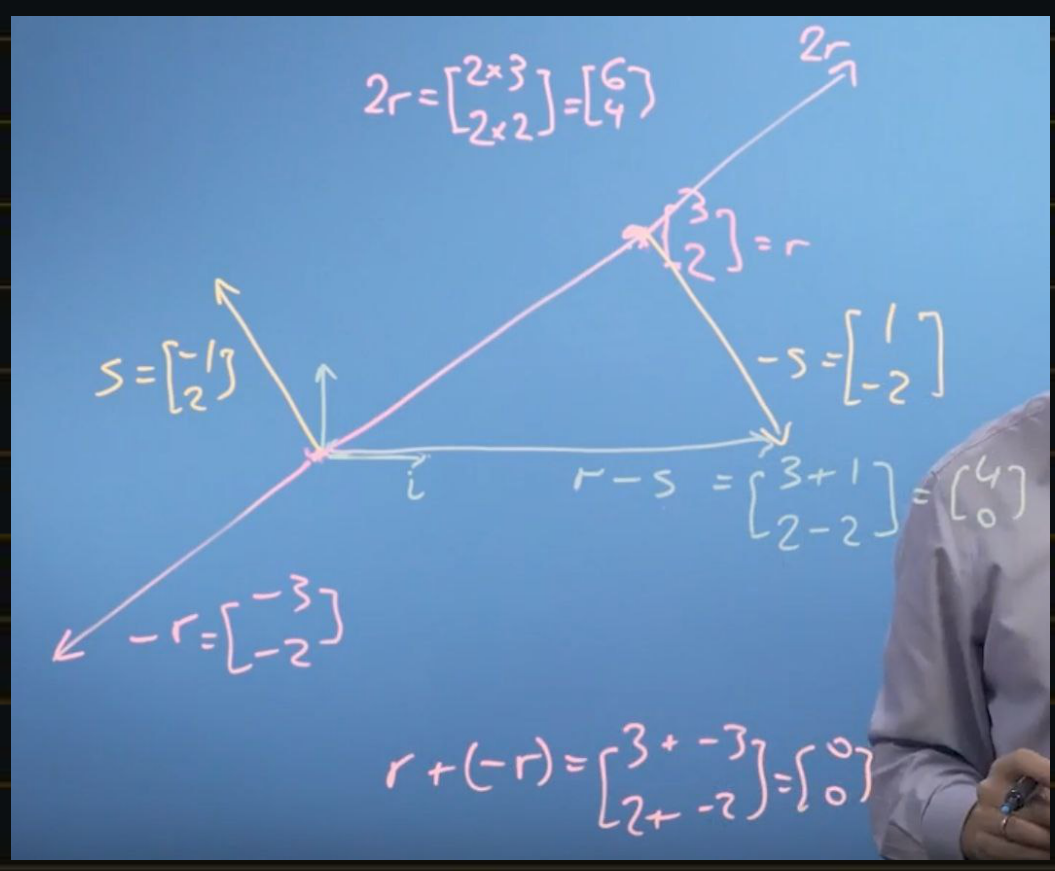
(1) $\vec{r} + \vec{s} = \vec{s} + \vec{r}$, and

(2) $(\vec{r} + \vec{s}) + \vec{t} = \vec{r} + (\vec{s} + \vec{t})$

If $\vec{r} = x\hat{i} + y\hat{j} = \begin{bmatrix} x \\ y \end{bmatrix}$, then $a\vec{r} = ax\hat{i} + ay\hat{j} = \begin{bmatrix} ax \\ ay \end{bmatrix}$, where a is a scalar.

Vector subtraction is essentially addition with a minus sign.

For eg., $\vec{r} + (-\vec{r}) = 0$.



Regarding the example of a house, 2 houses can now be represented as follows :-

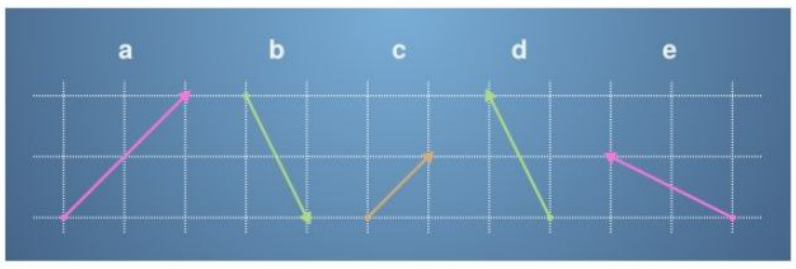
$$\begin{bmatrix} 120 \\ 2 \\ 1 \\ 150 \end{bmatrix} + \begin{bmatrix} 120 \\ 2 \\ 1 \\ 150 \end{bmatrix} = 2 \cdot \begin{bmatrix} 120 \\ 2 \\ 1 \\ 150 \end{bmatrix} = \begin{bmatrix} 240 \\ 4 \\ 2 \\ 300 \end{bmatrix}$$

i.e. 2 houses together take up 300 sq. m of area, have 4 bedrooms & 2 bathrooms and cost $300,000.

(*) Practice Quiz: Doing some vector operations

1. This aim of this quiz is to familiarise yourself with vectors and some basic vector operations. **1 / 1 point**

For the following questions, the vectors $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, $\mathbf{d}$ and $\mathbf{e}$ refer to those in this diagram:

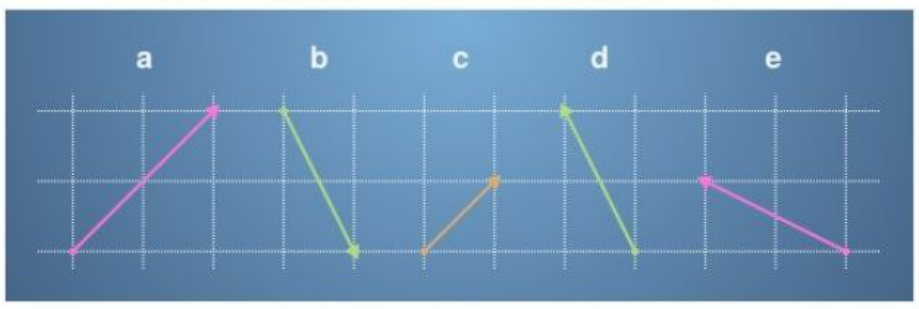


The sides of each square on the grid are of length 1. What is the numerical representation of the vector $\mathbf{a}$?

- ○ $\begin{bmatrix}2\\1\end{bmatrix}$
- ◉ $\begin{bmatrix}2\\2\end{bmatrix}$
- ○ $\begin{bmatrix}1\\2\end{bmatrix}$
- ○ $\begin{bmatrix}1\\1\end{bmatrix}$

Correct
You can get the numerical representation by following the arrow along the grid.

2. **1 / 1 point**



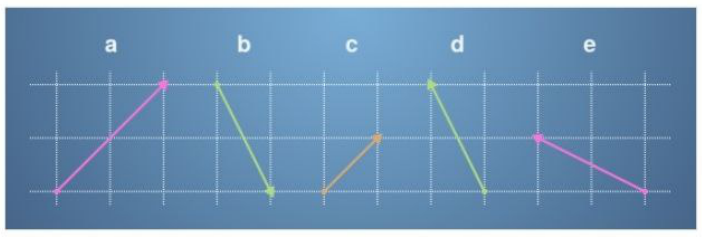
Which vector in the diagram corresponds to $\begin{bmatrix}-1\\2\end{bmatrix}$?

- ○ Vector $\mathbf{a}$
- ○ Vector $\mathbf{b}$
- ○ Vector $\mathbf{c}$
- ◉ Vector $\mathbf{d}$

Correct
You can get the numerical representation by following the arrow along the grid.

3.

1 / 1 point

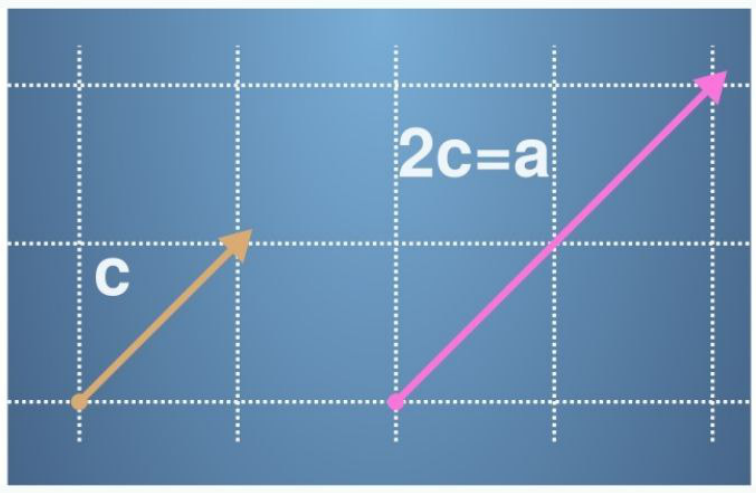


What vector is $2\mathbf{c}$?

Please select all correct answers.

- [ ] $\begin{bmatrix}-2\\2\end{bmatrix}$
- [x] $\mathbf{a}$

Correct
Multiplying by a positive scalar is like stretching out a vector in the same direction.

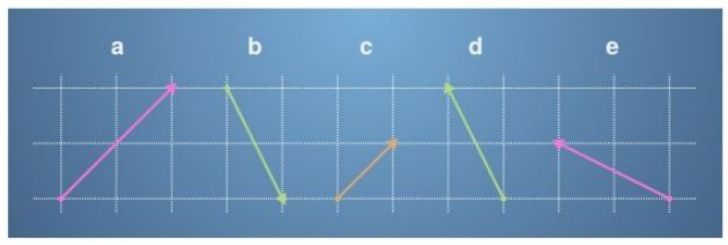


- [ ] $\mathbf{e}$
- [x] $\begin{bmatrix}2\\2\end{bmatrix}$

Correct
A scalar multiple of a vector can be calculated by multiplying each component.

4.

1 / 1 point

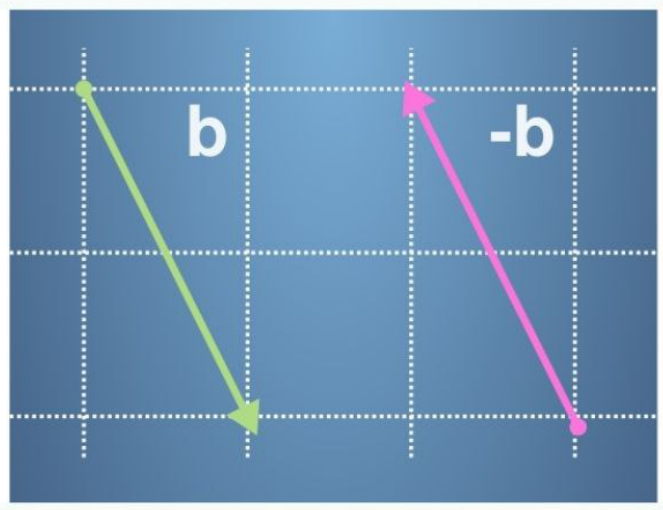


What vector is $-\mathbf{b}$?

Please select all correct answers.

- [ ] $\mathbf{e}$
- [x] $\mathbf{d}$

Correct
Multiplying by a negative number points the vector in the opposite direction.

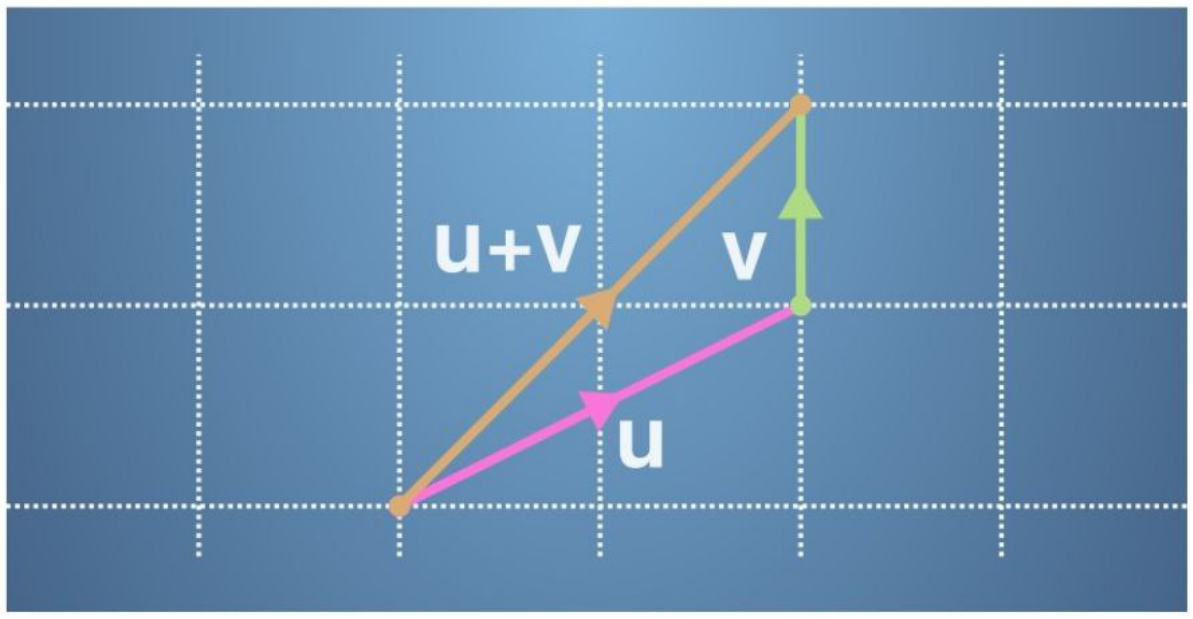


- [ ] $\begin{bmatrix}-2\\1\end{bmatrix}$
- [x] $\begin{bmatrix}-1\\2\end{bmatrix}$

Correct
A scalar multiple of a vector can be calculated by multiplying each component.

5. In the previous videos you saw that vectors can be added by placing them start-to-end. For example, the following diagram represents the sum of two new vectors, $\mathbf{u} + \mathbf{v}$:

1 / 1 point



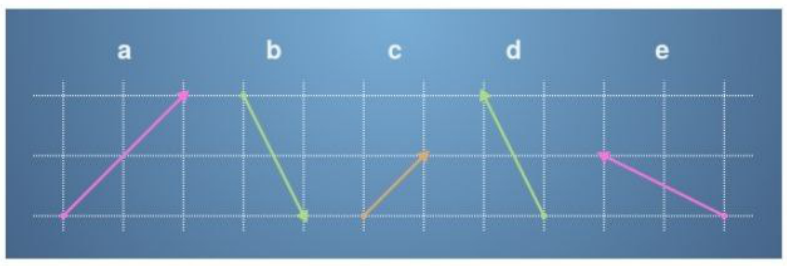
The sides of each square on the grid are still of length 1. Which of the following equations does the diagram represent?

- ○ $\begin{bmatrix}1\\2\end{bmatrix} + \begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}2\\2\end{bmatrix}$
- ○ $\begin{bmatrix}1\\2\end{bmatrix} + \begin{bmatrix}1\\0\end{bmatrix} = \begin{bmatrix}2\\2\end{bmatrix}$
- ○ $\begin{bmatrix}1\\1\end{bmatrix} + \begin{bmatrix}1\\0\end{bmatrix} = \begin{bmatrix}2\\1\end{bmatrix}$
- ◉ $\begin{bmatrix}2\\1\end{bmatrix} + \begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}2\\2\end{bmatrix}$

Correct
We can see that summing the vectors by adding them start-to-end and adding up the individual components gives us the same answer.

6. Let's return to our vectors defined by the diagram below:

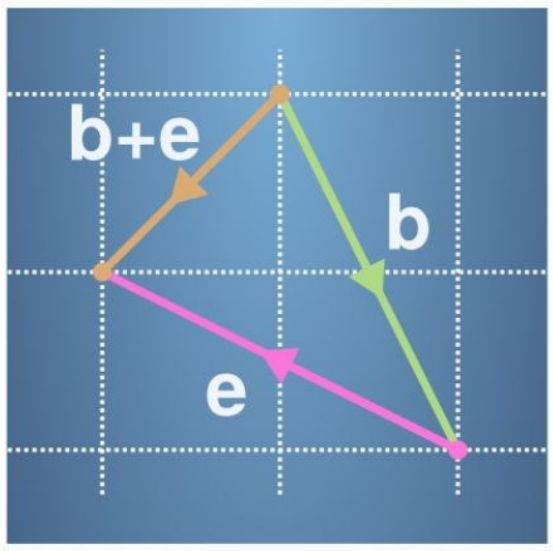
1 / 1 point



What is the vector $\mathbf{b} + \mathbf{e}$?

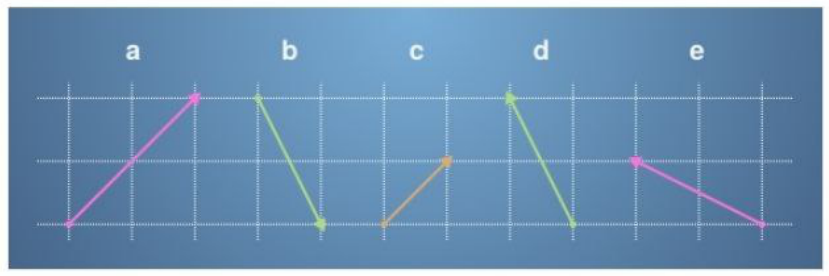
- ○ $\begin{bmatrix}-1\\2\end{bmatrix}$
- ○ $\begin{bmatrix}1\\3\end{bmatrix}$
- ○ $\begin{bmatrix}2\\-1\end{bmatrix}$
- ◉ $\begin{bmatrix}-1\\-1\end{bmatrix}$

Correct
Vectors are added together entry by entry. They can also be thought of as adding start to end, like in the following diagram:



7.

1 / 1 point



What is the vector $\mathbf{d} - \mathbf{b}$?

- ○ $\begin{bmatrix} -4 \\ 2 \end{bmatrix}$
- ○ $\begin{bmatrix} 2 \\ -4 \end{bmatrix}$
- ○ $\begin{bmatrix} 4 \\ -2 \end{bmatrix}$
- ◉ $\begin{bmatrix} -2 \\ 4 \end{bmatrix}$

Correct

Remember that vectors add by attaching the end of one to the start of the other, and that multiplying by a negative number points the vector in the opposite direction.