Congratulations! You passed!

TO PASS 40% or higher

Keep Learning

GRADE 100%

Exploring parameter space

TOTAL POINTS 8

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.

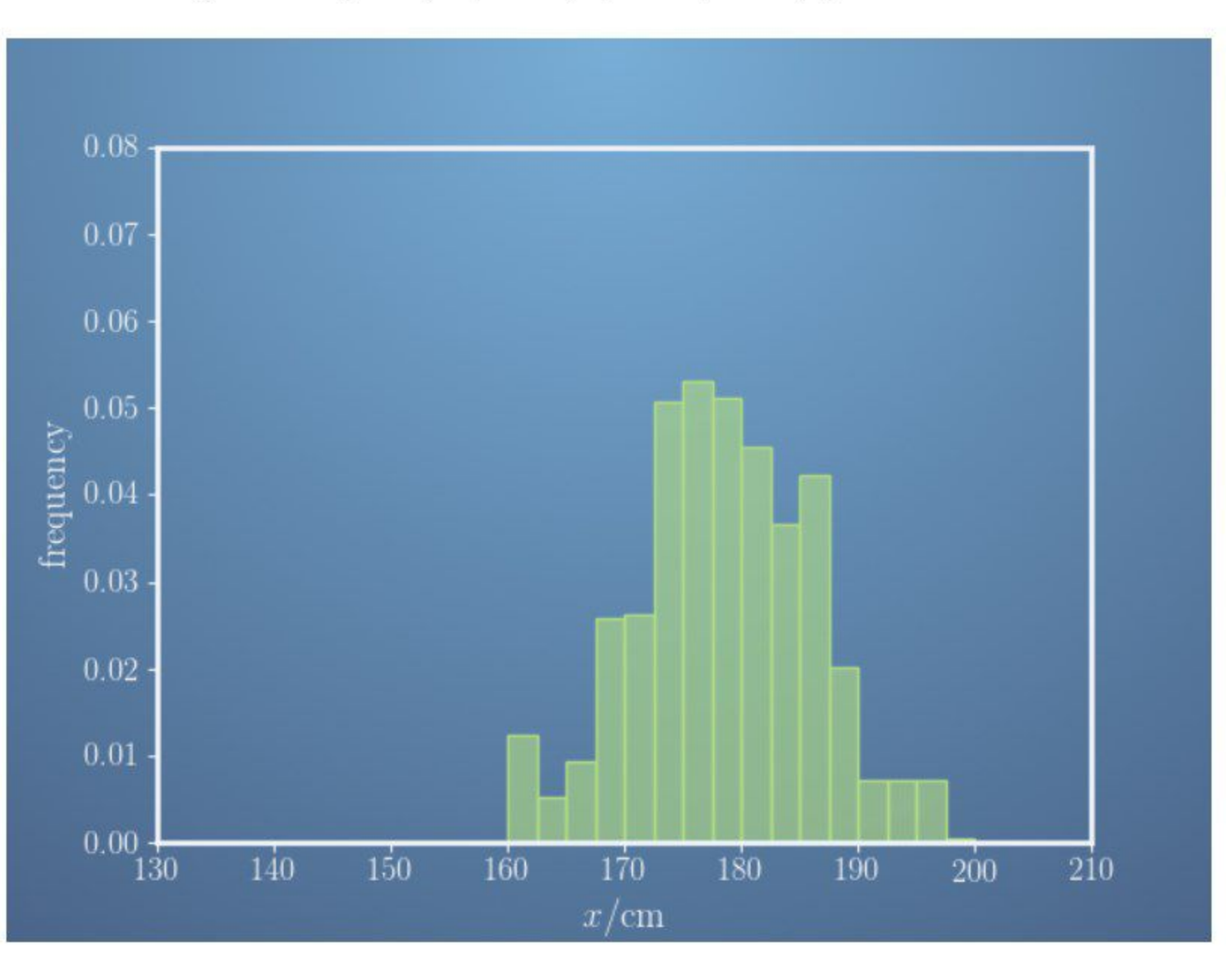
2 / 2 points

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.

- [ ] No one in the world is less than 160 cm tall.
- [ ] If another sample was taken under the same conditions, the frequencies would be exactly the same.
- [ ] None of the other statements.
- [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.

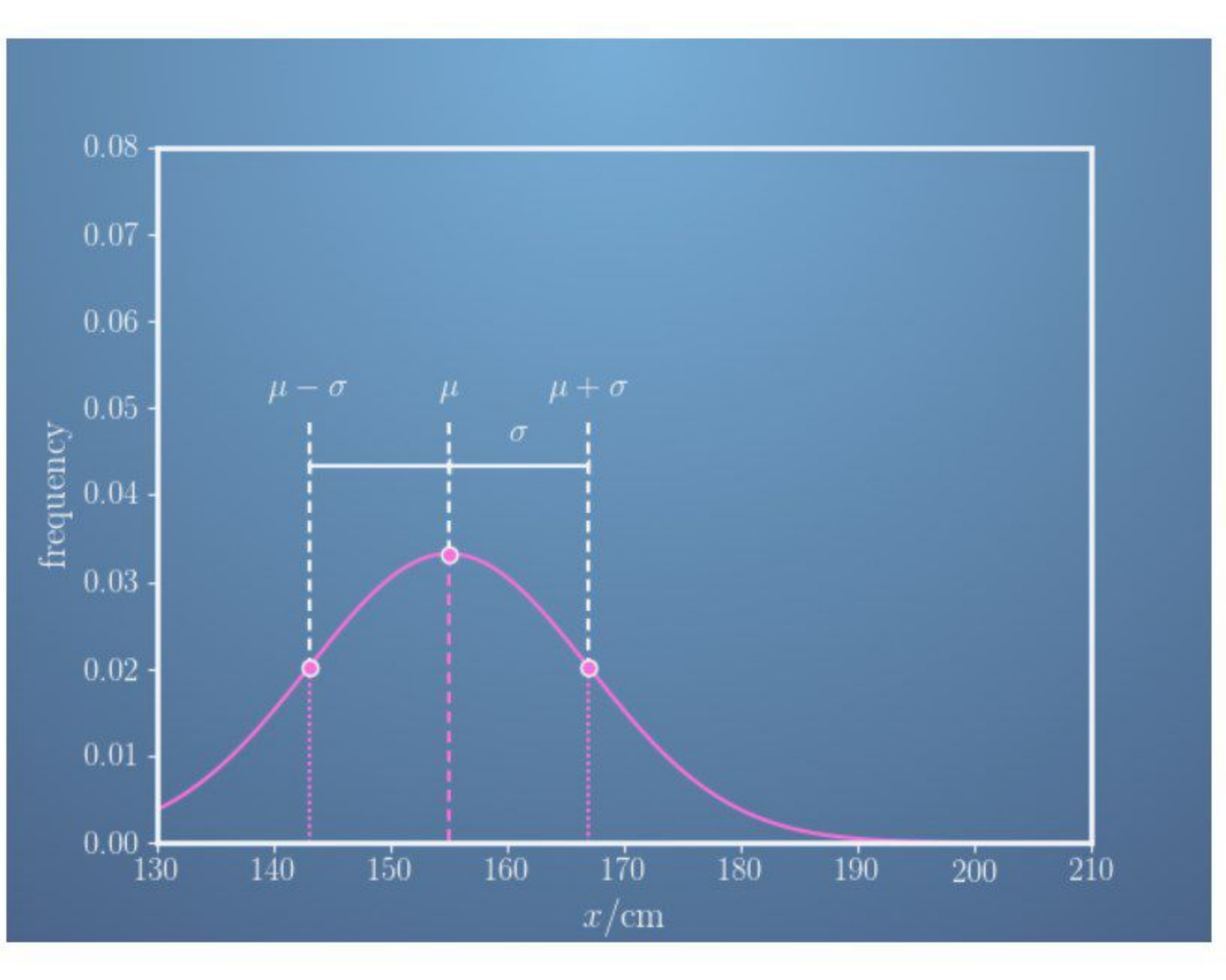
- [x] If another sample was taken under the same conditions, the frequencies should be broadly similar.

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

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.

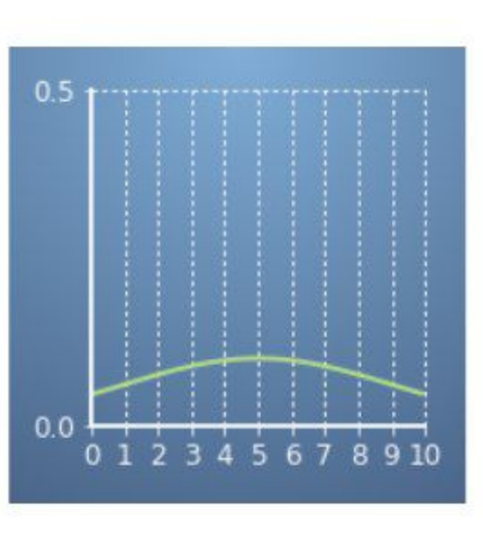
- ( ) $\mathbf{p} = \begin{bmatrix} 155 \\ 3 \end{bmatrix}$
- ( ) $\mathbf{p} = \begin{bmatrix} 143 \\ 167 \end{bmatrix}$
- (•) $\mathbf{p} = \begin{bmatrix} 155 \\ 12 \end{bmatrix}$
- ( ) $\mathbf{p} = \begin{bmatrix} 167 \\ 12 \end{bmatrix}$
- ( ) $\mathbf{p} = \begin{bmatrix} 167 \\ 24 \end{bmatrix}$

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

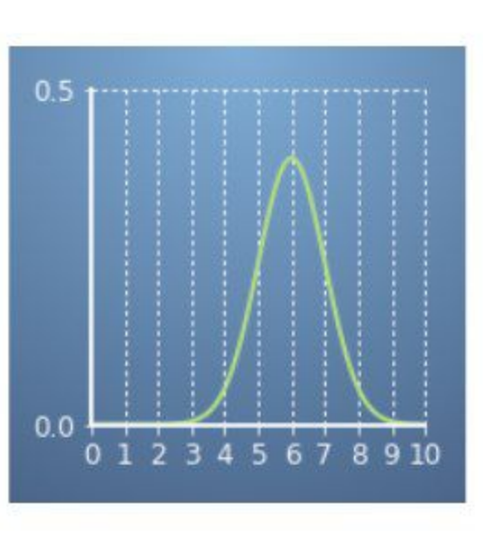
3. Pick the Normal distribution that corresponds the closest to the parameter vector $\mathbf{p} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$.

1 / 1 point

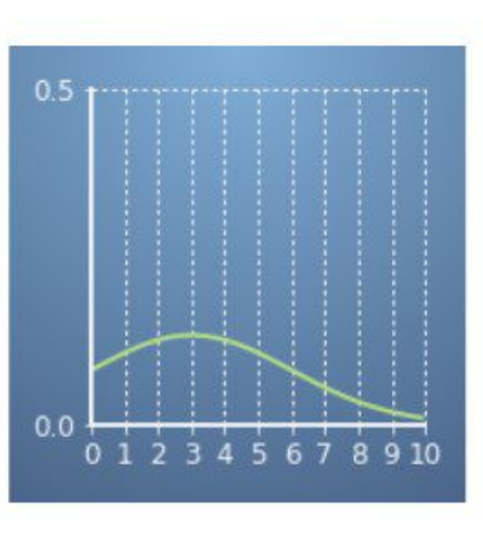
( )



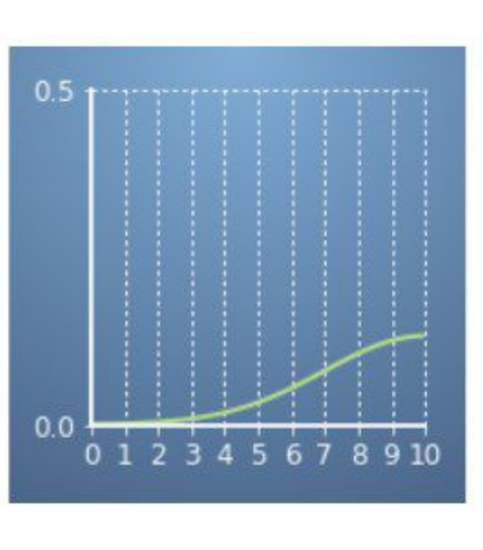
( )



(•)



( )



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,

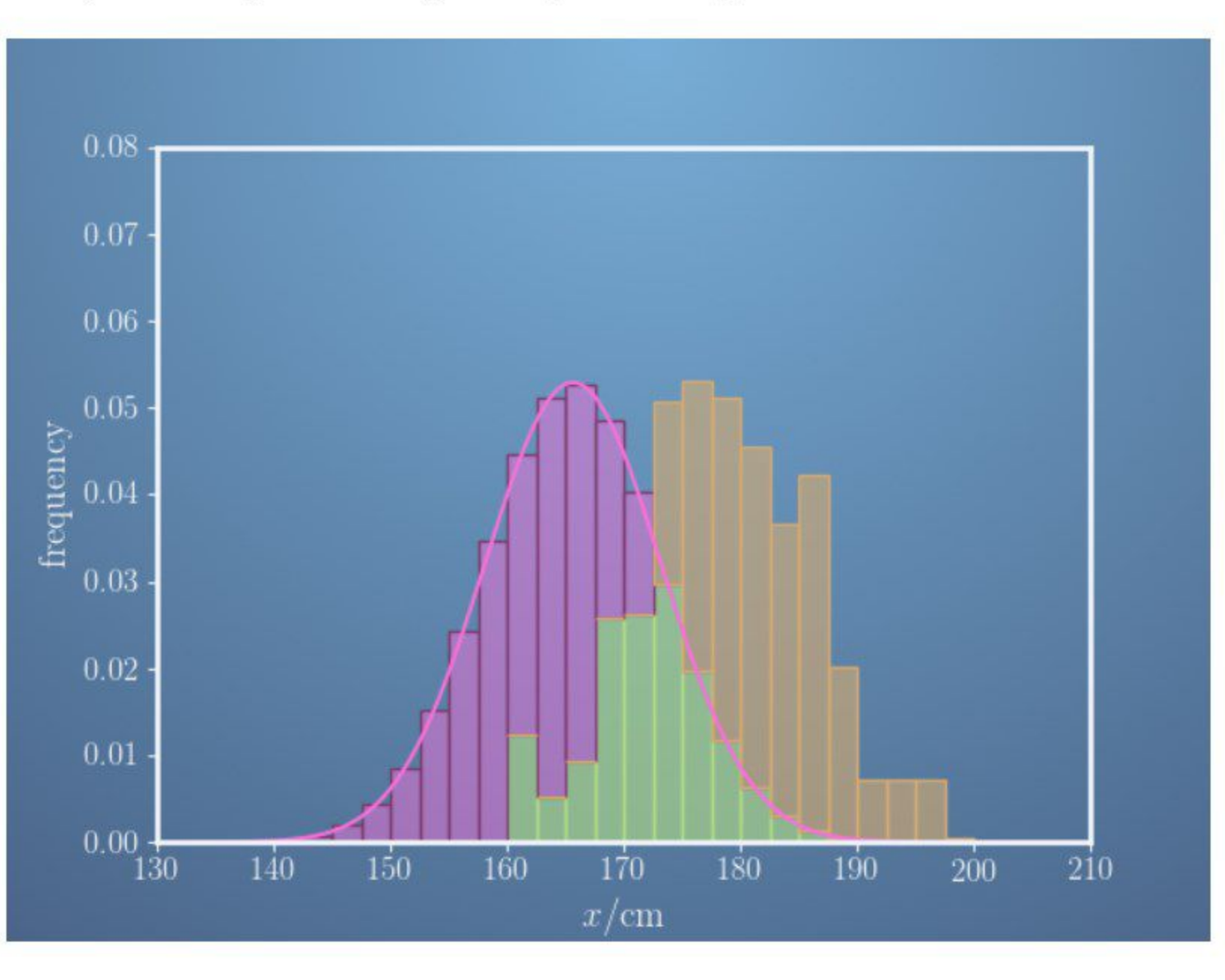
1 / 1 point

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

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.

- [ ] Increase the standard deviation, σ.
- [ ] Decrease the mean, μ.
- [ ] Keep the mean, μ, approximately the same.
- [ ] 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, $\mathrm{SSR}(\mathbf{p}) = |\mathbf{f} - \mathbf{g_p}|^2$, 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.

```
1  # Play with values of μ and σ to find the best fit.
2  μ = 179 ; σ = 8
3  p = [μ, σ]
4  histogram(p)
5
```

Run

Reset

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

Input your fitted parameters into the code block below.

```
1  # Replace μ and σ with values that minimise the SSR.
2  μ = 179 ; σ = 8
3  p = [μ, σ]
4
```

Run

Reset

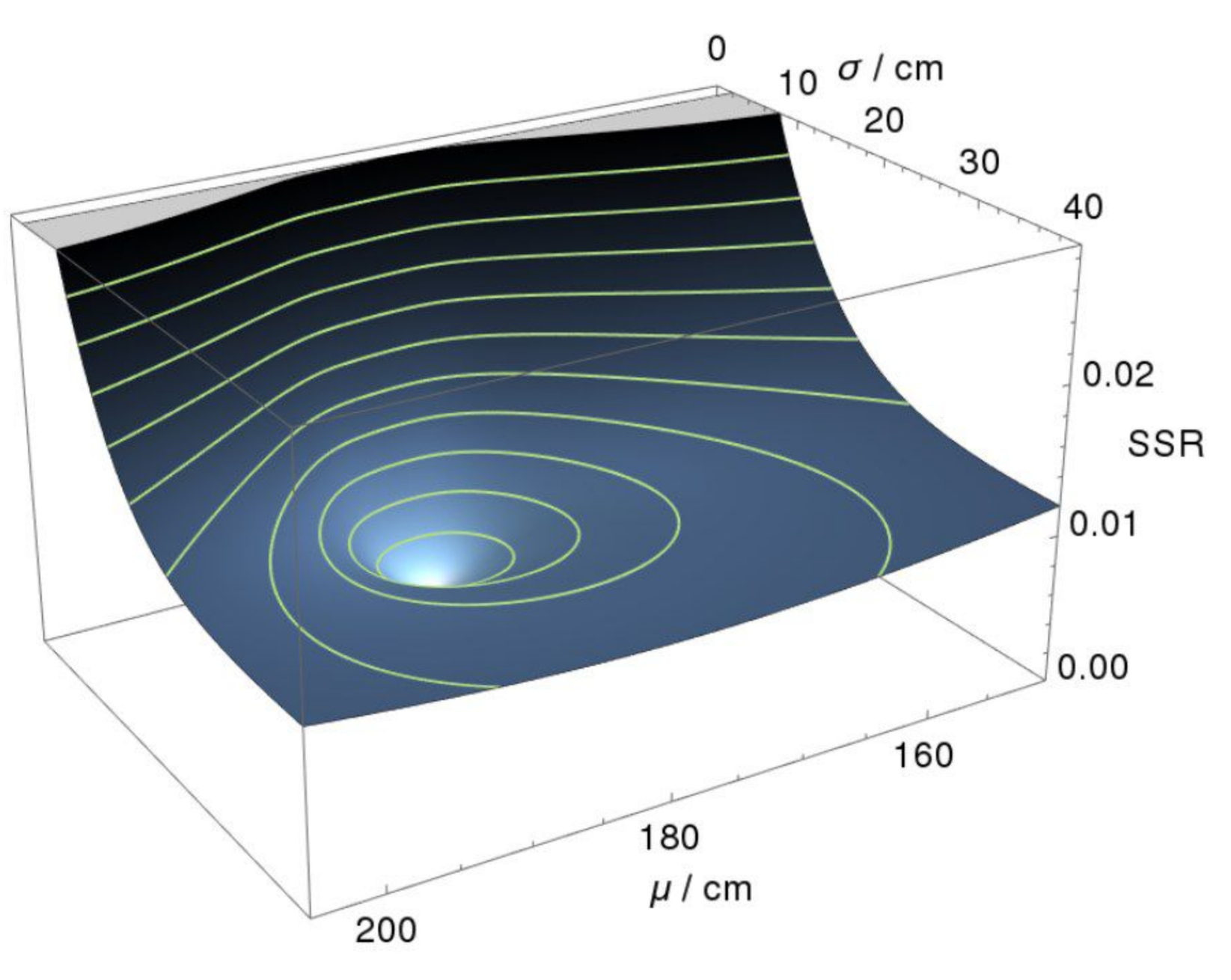
```
[179, 8]
```

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.

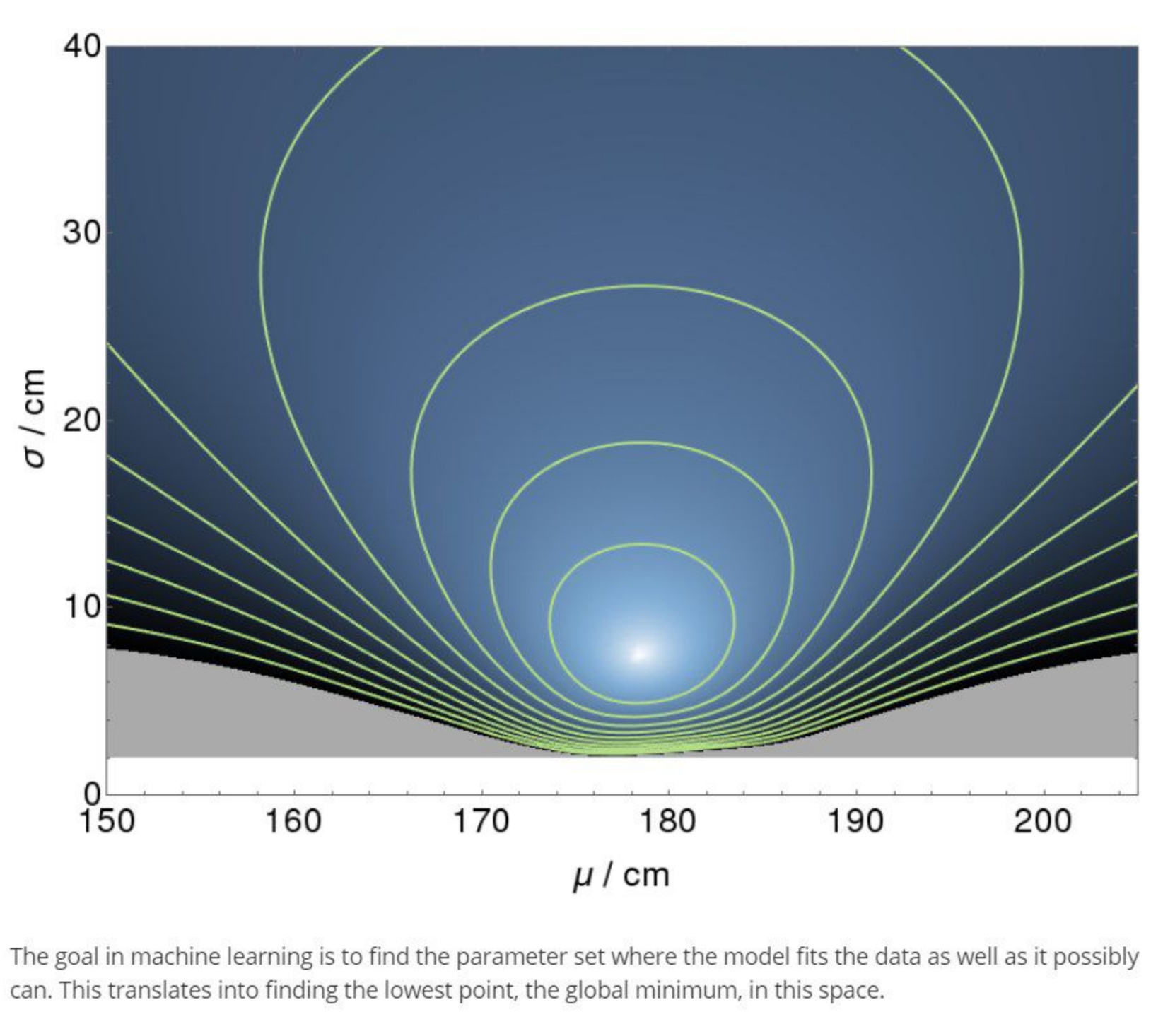
1 / 1 point

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.

- [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.

- [ ] None of the other statements.
- [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.

- [ ] At the minimum of the surface, the model exactly matches the measured data.
- [ ] You get the same model by following along a contour line.

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.

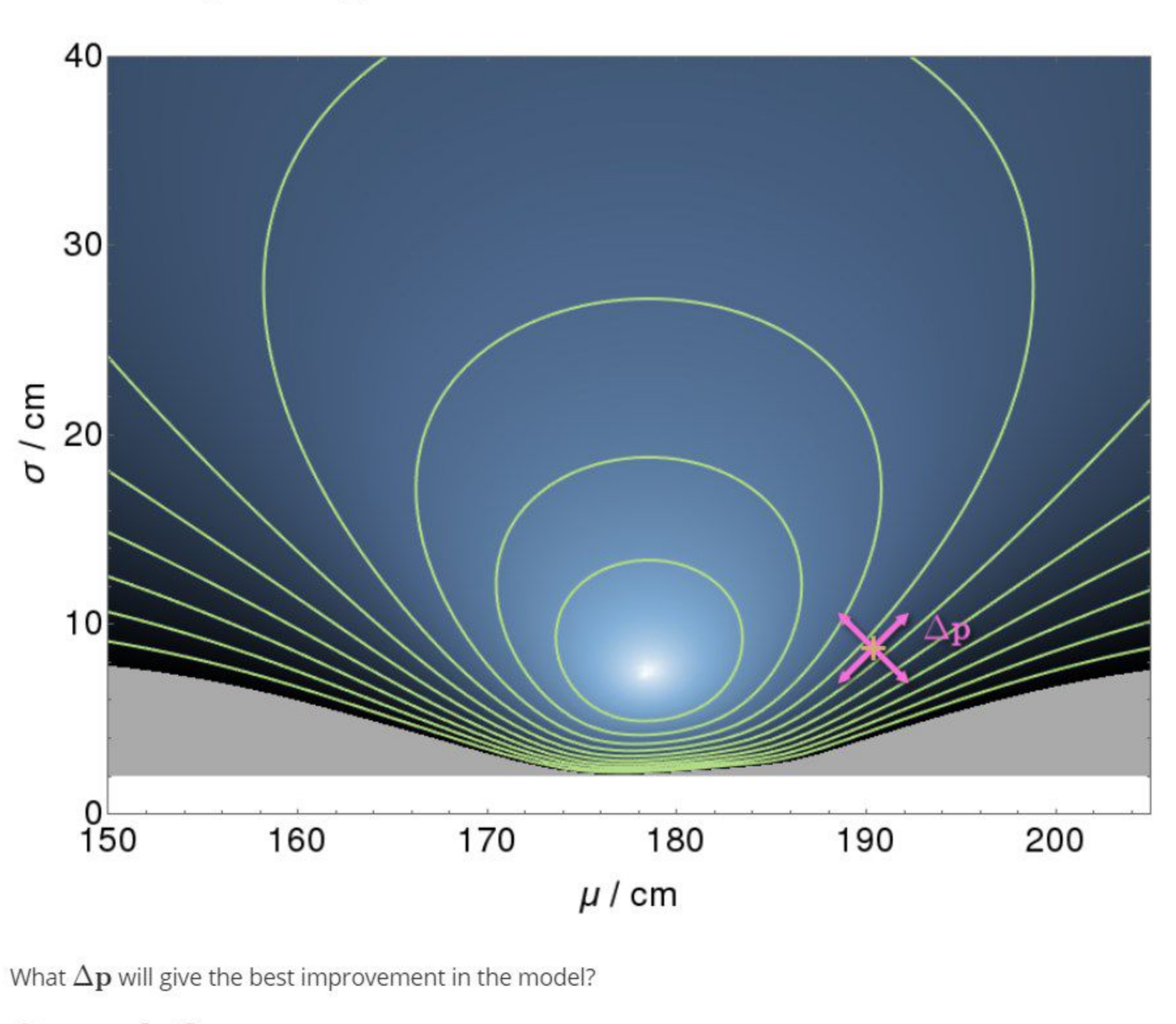
1 / 1 point

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.