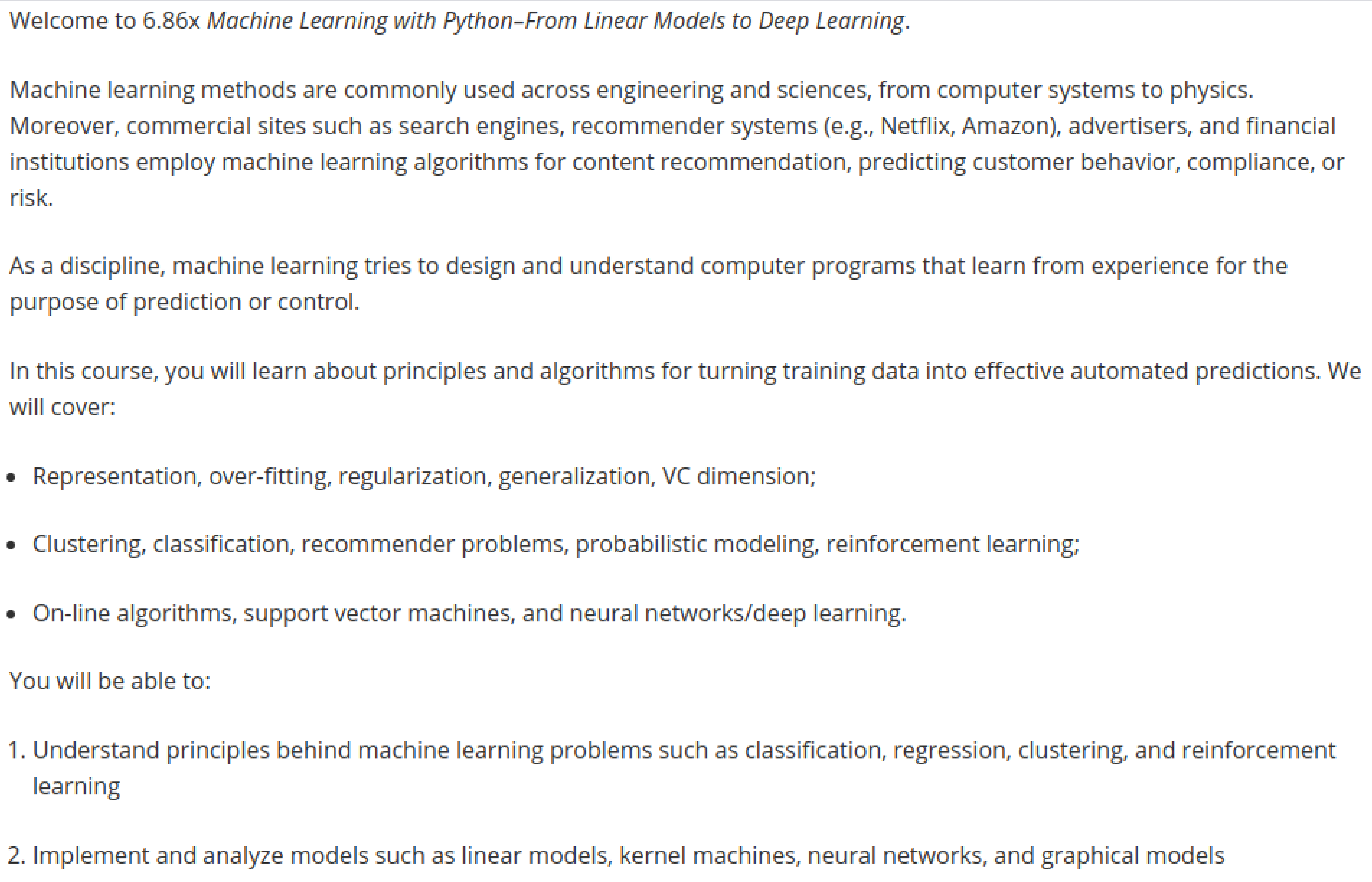


Welcome to 6.86x *Machine Learning with Python–From Linear Models to Deep Learning*.

Machine learning methods are commonly used across engineering and sciences, from computer systems to physics. Moreover, commercial sites such as search engines, recommender systems (e.g., Netflix, Amazon), advertisers, and financial institutions employ machine learning algorithms for content recommendation, predicting customer behavior, compliance, or risk.

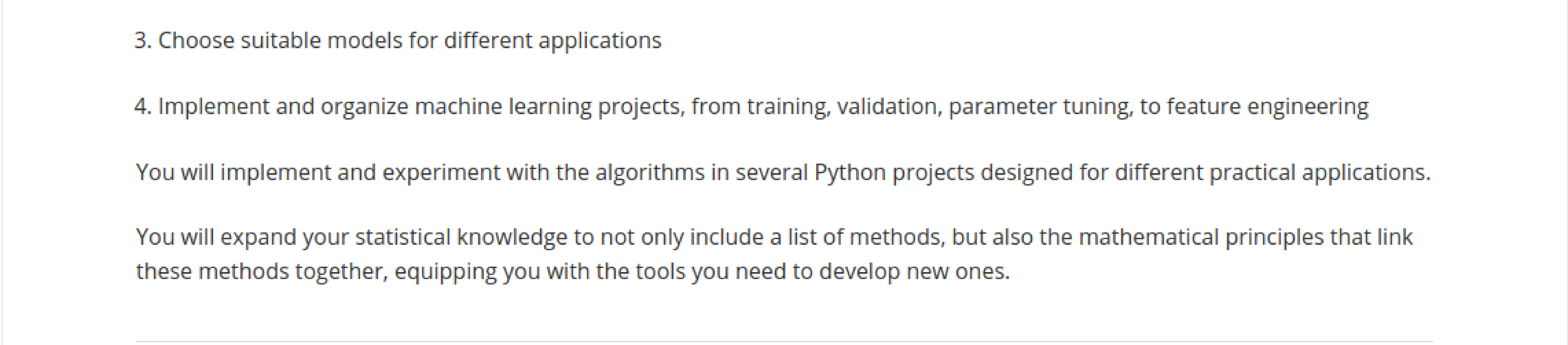
As a discipline, machine learning tries to design and understand computer programs that learn from experience for the purpose of prediction or control.

In this course, you will learn about principles and algorithms for turning training data into effective automated predictions. We will cover:

- Representation, over-fitting, regularization, generalization, VC dimension;
- Clustering, classification, recommender problems, probabilistic modeling, reinforcement learning;
- On-line algorithms, support vector machines, and neural networks/deep learning.

You will be able to:

1. Understand principles behind machine learning problems such as classification, regression, clustering, and reinforcement learning
2. Implement and analyze models such as linear models, kernel machines, neural networks, and graphical models

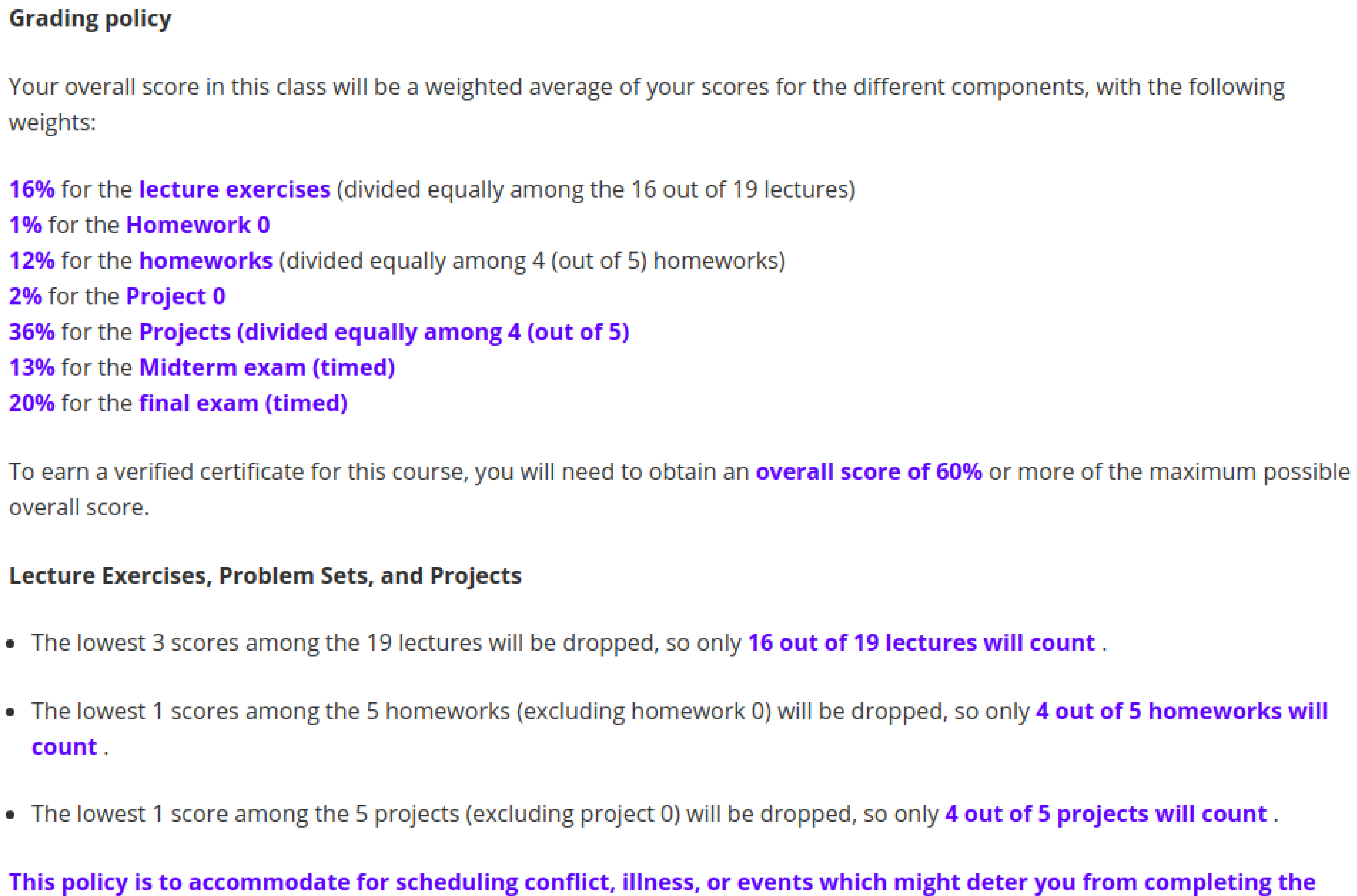


3. Choose suitable models for different applications

4. Implement and organize machine learning projects, from training, validation, parameter tuning, to feature engineering

You will implement and experiment with the algorithms in several Python projects designed for different practical applications.

You will expand your statistical knowledge to not only include a list of methods, but also the mathematical principles that link these methods together, equipping you with the tools you need to develop new ones.



Grading policy

Your overall score in this class will be a weighted average of your scores for the different components, with the following weights:

16% for the **lecture exercises** (divided equally among the 16 out of 19 lectures)
1% for the **Homework 0**
12% for the **homeworks** (divided equally among 4 (out of 5) homeworks)
2% for the **Project 0**
36% for the **Projects (divided equally among 4 (out of 5)**
13% for the **Midterm exam (timed)**
20% for the **final exam (timed)**

To earn a verified certificate for this course, you will need to obtain an **overall score of 60%** or more of the maximum possible overall score.

Lecture Exercises, Problem Sets, and Projects

- The lowest 3 scores among the 19 lectures will be dropped, so only **16 out of 19 lectures will count** .
- The lowest 1 scores among the 5 homeworks (excluding homework 0) will be dropped, so only **4 out of 5 homeworks will count** .
- The lowest 1 score among the 5 projects (excluding project 0) will be dropped, so only **4 out of 5 projects will count** .

This policy is to accommodate for scheduling conflict, illness, or events which might deter you from completing the

work before the deadline with the best grades you can. However, we still fully expect you to learn the material for any dropped assignments, and the exams will cover everything.

Note that not every homework, set of lecture exercises, project will have the same number of raw points. For example, homework 0 may have 53 points and homework 1 may have 43 points. However, each homework receives the same weight for the purpose of calculating your overall score. Similar for lecture exercises and projects.

Under the “Progress" tab at the top, you can see your score broken down for each assignment, as well as a summary plot.

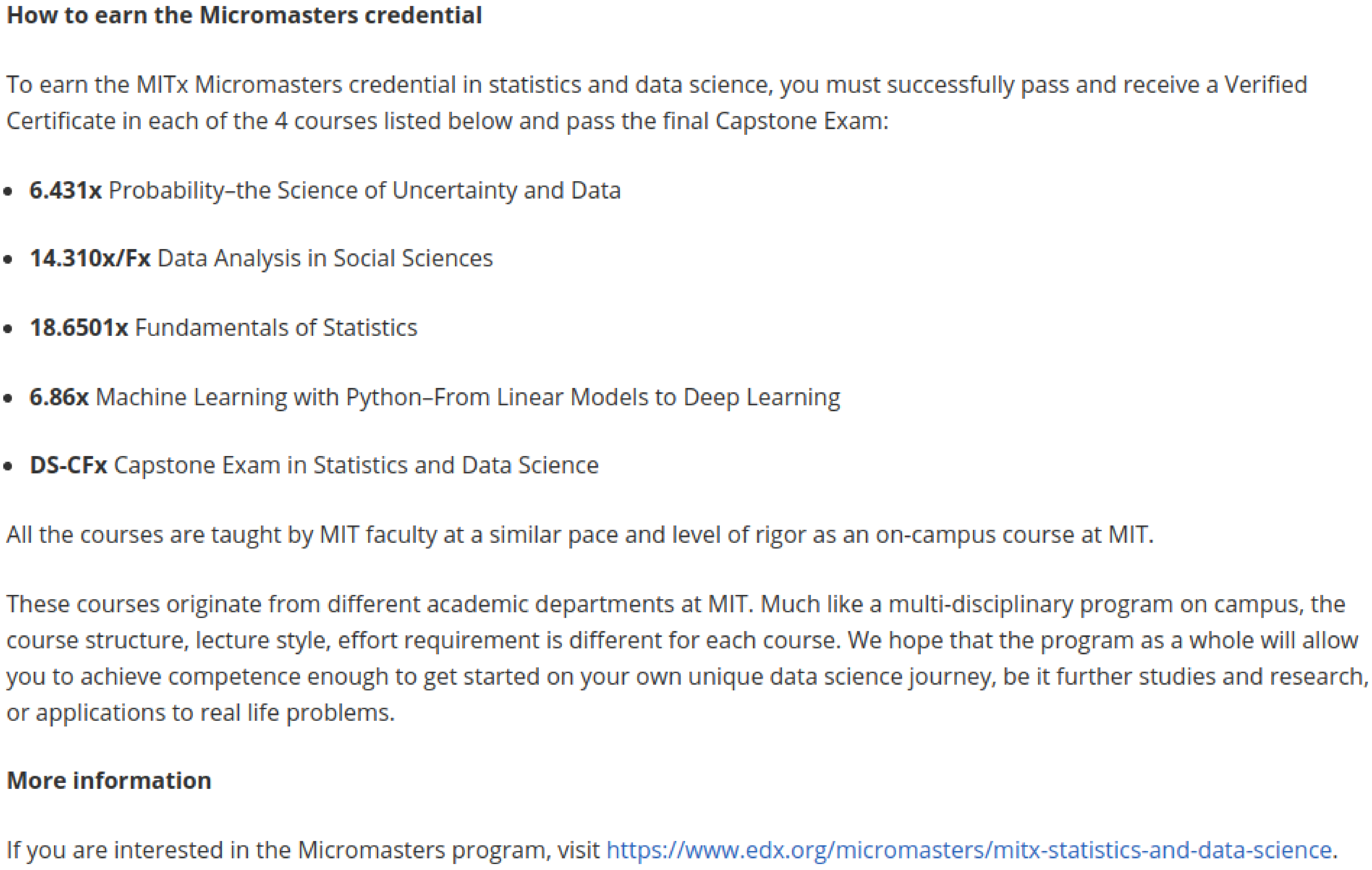
Timed Exams

The midterm exam and final exam are **timed exams** . This means that each exam is available for approximately a week, but once you open the exam, there is a limited amount of time (48 hours), counting from when you start, within which you must complete the exam. Please plan in advance for the exams. If you do not complete the whole exam during the allowed time, you will miss the points associated with the questions that have not been answered. The exams are designed to assess your knowledge. **There are no extensions granted to these deadlines**. You can find the exam dates on the calendar on the previous page.

Warning: Note that the timed exams **CANNOT** be completed using the edX mobile app.

Exam Access: Note that the midterm and final exams will not be available to audit learners. These exams are assessment for learners interested in the course certificates.

MITx Committment to Accessibility



How to earn the Micromasters credential

To earn the MITx Micromasters credential in statistics and data science, you must successfully pass and receive a Verified Certificate in each of the 4 courses listed below and pass the final Capstone Exam:

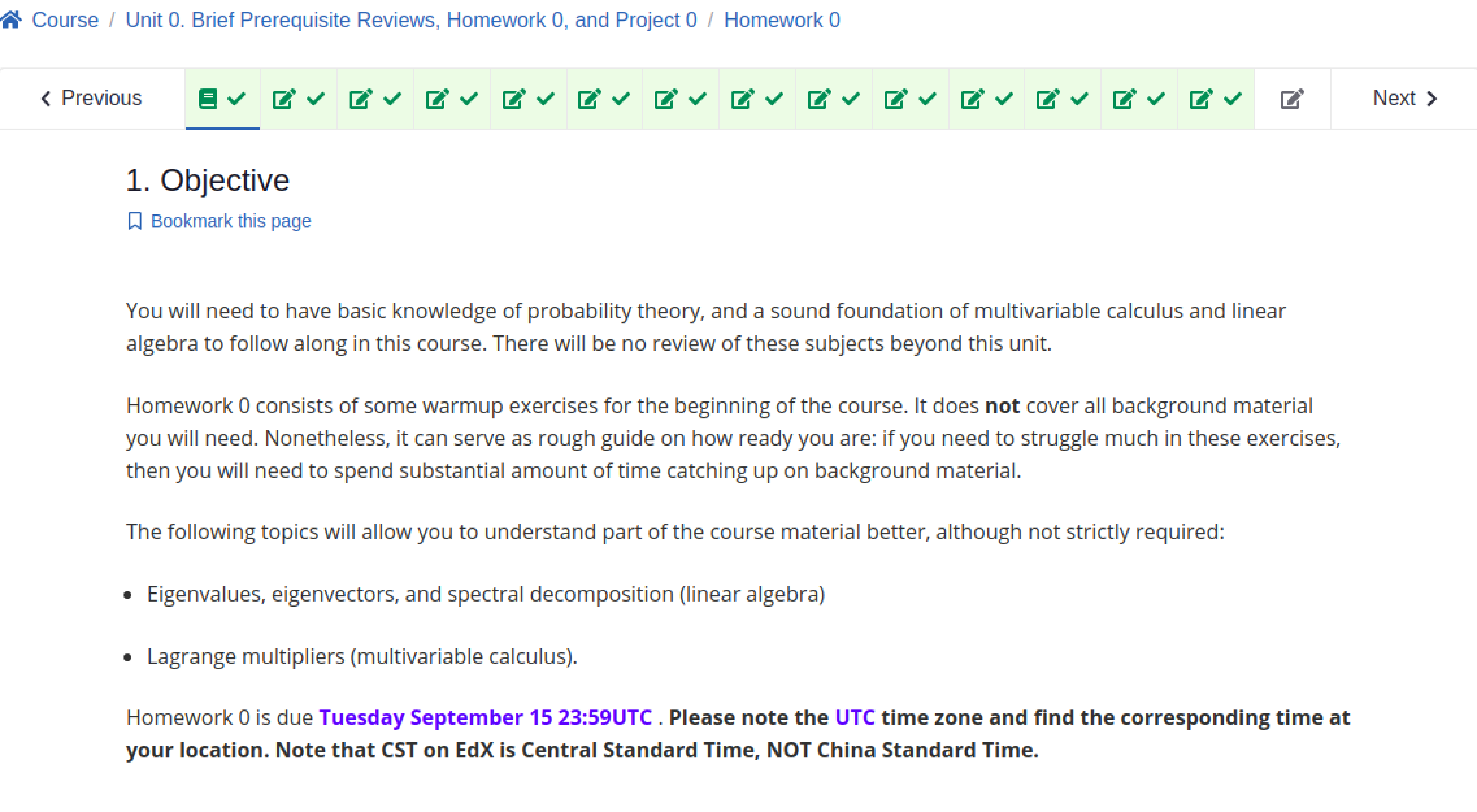
- **6.431x** Probability–the Science of Uncertainty and Data
- **14.310x/Fx** Data Analysis in Social Sciences
- **18.6501x** Fundamentals of Statistics
- **6.86x** Machine Learning with Python–From Linear Models to Deep Learning
- **DS-CFx** Capstone Exam in Statistics and Data Science

All the courses are taught by MIT faculty at a similar pace and level of rigor as an on-campus course at MIT.

These courses originate from different academic departments at MIT. Much like a multi-disciplinary program on campus, the course structure, lecture style, effort requirement is different for each course. We hope that the program as a whole will allow you to achieve competence enough to get started on your own unique data science journey, be it further studies and research, or applications to real life problems.

More information

If you are interested in the Micromasters program, visit https://www.edx.org/micromasters/mitx-statistics-and-data-science.


Course / Unit 0. Brief Prerequisite Reviews, Homework 0, and Project 0 / Homework 0
Previous
✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓
Next
Bookmark this page

1. Objective

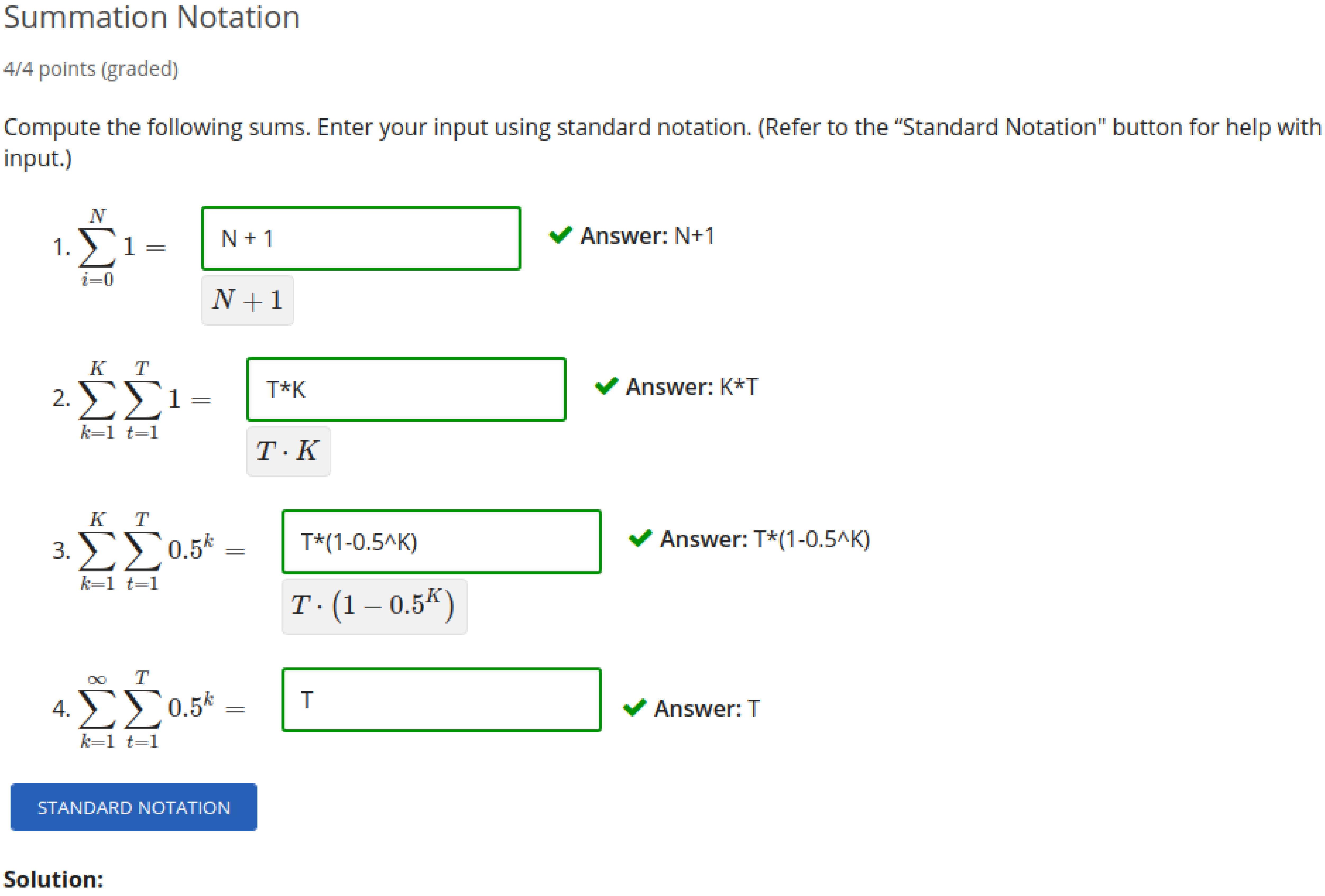
You will need to have basic knowledge of probability theory, and a sound foundation of multivariable calculus and linear algebra to follow along in this course. There will be no review of these subjects beyond this unit.

Homework 0 consists of some warmup exercises for the beginning of the course. It does **not** cover all background material you will need. Nonetheless, it can serve as rough guide on how ready you are: if you need to struggle much in these exercises, then you will need to spend substantial amount of time catching up on background material.

The following topics will allow you to understand part of the course material better, although not strictly required:

- Eigenvalues, eigenvectors, and spectral decomposition (linear algebra)
- Lagrange multipliers (multivariable calculus).

Homework 0 is due **Tuesday September 15 23:59UTC** . **Please note the UTC time zone and find the corresponding time at your location. Note that CST on EdX is Central Standard Time, NOT China Standard Time.**



Summation Notation

4/4 points (graded)

Compute the following sums. Enter your input using standard notation. (Refer to the “Standard Notation" button for help with input.)

1. $\sum_{i=0}^{N} 1 =$ N + 1 ✓ **Answer:** N+1

   $N+1$

2. $\sum_{k=1}^{K}\sum_{t=1}^{T} 1 =$ T*K ✓ **Answer:** K*T

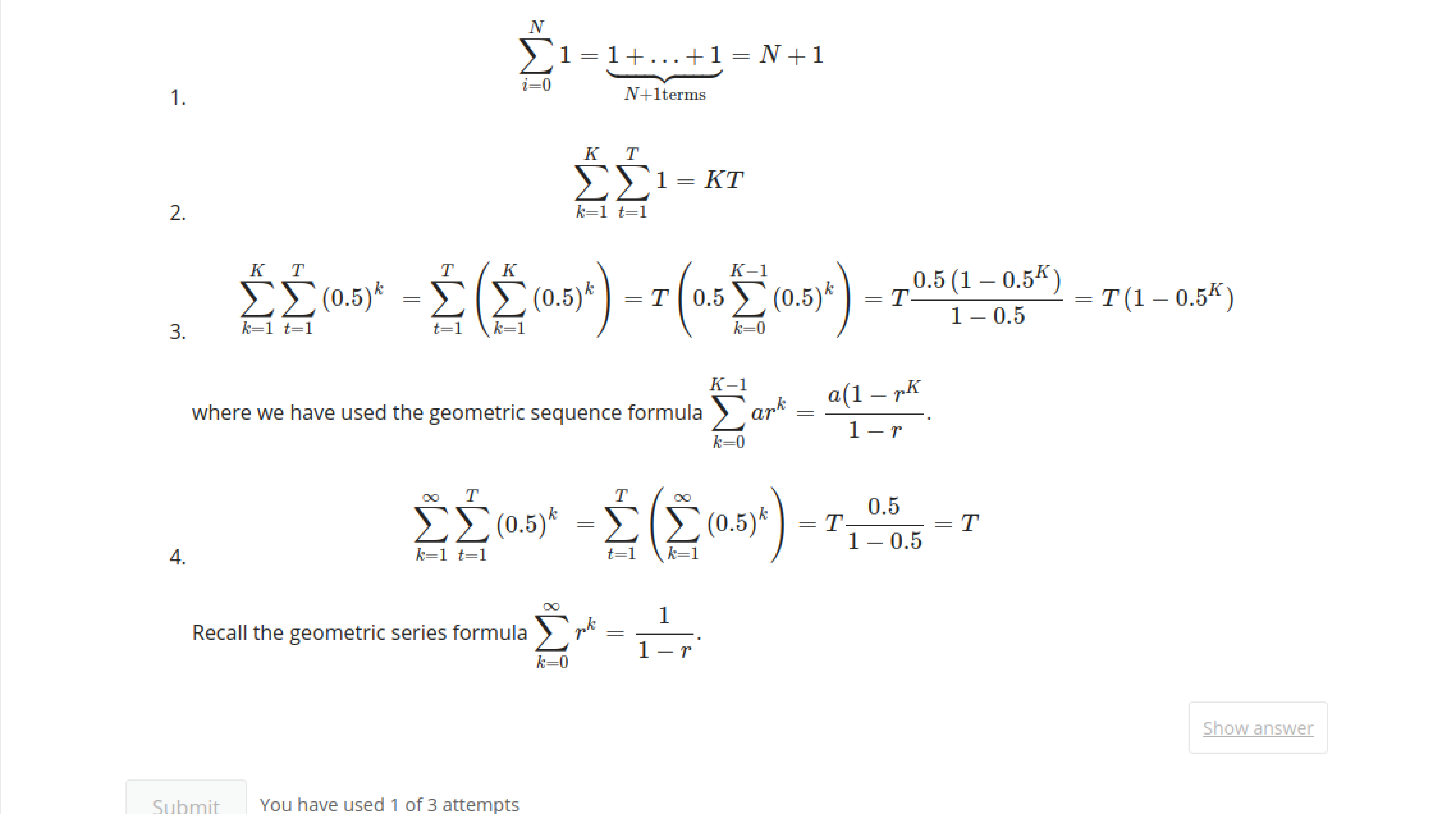
   $T \cdot K$

3. $\sum_{k=1}^{K}\sum_{t=1}^{T} 0.5^k =$ T*(1-0.5^K) ✓ **Answer:** T*(1-0.5^K)

   $T \cdot \left(1 - 0.5^{K}\right)$

4. $\sum_{k=1}^{\infty}\sum_{t=1}^{T} 0.5^k =$ T ✓ **Answer:** T

Solution:



1. $$\sum_{i=0}^{N} 1 = \underbrace{1 + \ldots + 1}_{N+1\text{terms}} = N + 1$$

2. $$\sum_{k=1}^{K} \sum_{t=1}^{T} 1 = KT$$

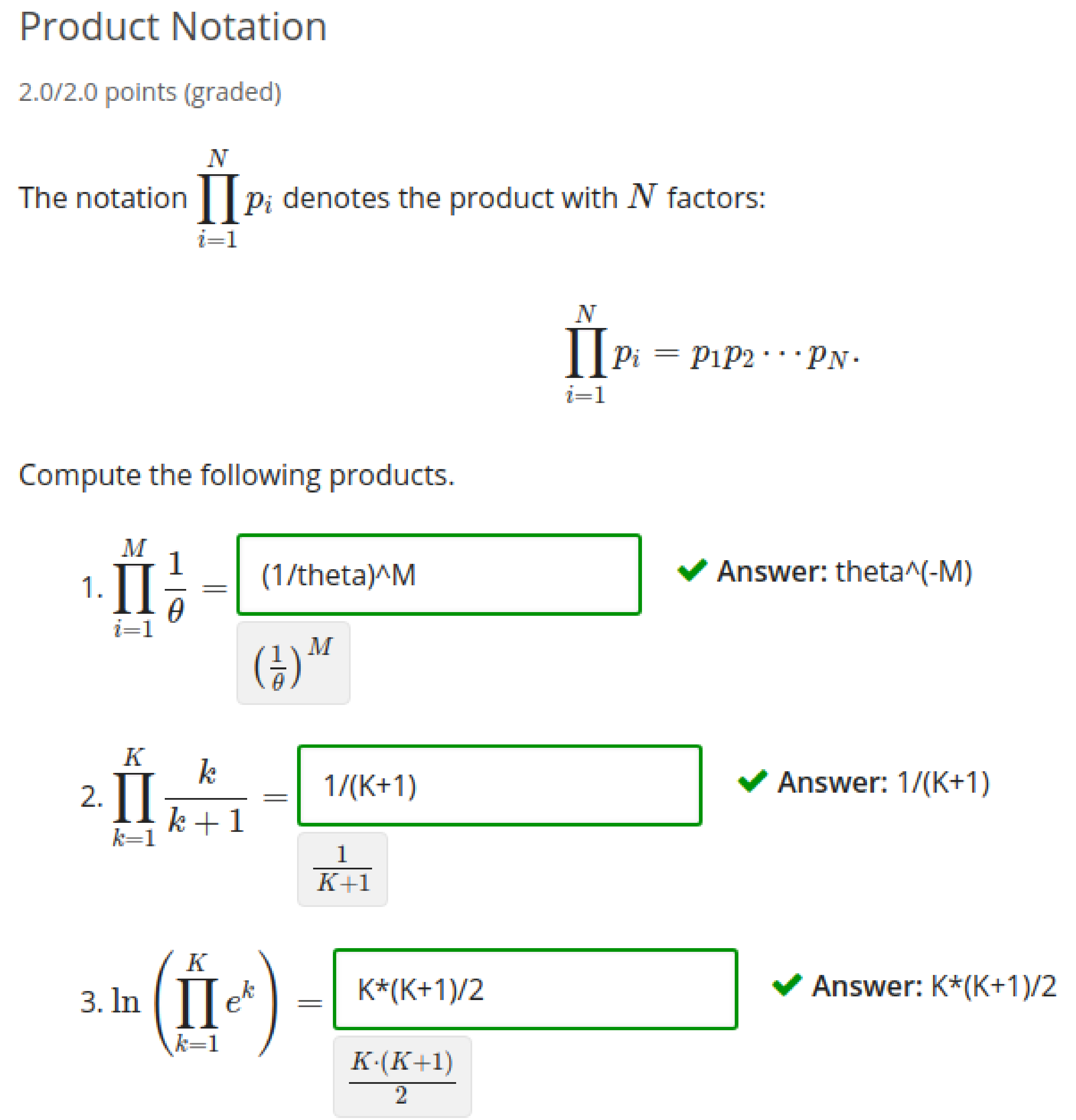
3. $$\sum_{k=1}^{K} \sum_{t=1}^{T} (0.5)^k \ = \sum_{t=1}^{T} \left(\sum_{k=1}^{K} (0.5)^k \right) = T \left(0.5 \sum_{k=0}^{K-1} (0.5)^k \right) = T \frac{0.5\left(1 - 0.5^K\right)}{1 - 0.5} = T\left(1 - 0.5^K\right)$$

where we have used the geometric sequence formula $\sum_{k=0}^{K-1} ar^k = \frac{a(1 - r^K}{1 - r}$.

4. $$\sum_{k=1}^{\infty} \sum_{t=1}^{T} (0.5)^k \ = \sum_{t=1}^{T} \left(\sum_{k=1}^{\infty} (0.5)^k \right) = T \frac{0.5}{1 - 0.5} = T$$

Recall the geometric series formula $\sum_{k=0}^{\infty} r^k = \frac{1}{1 - r}$.

You have used 1 of 3 attempts



Product Notation

2.0/2.0 points (graded)

The notation $\prod_{i=1}^{N} p_i$ denotes the product with N factors:

$$\prod_{i=1}^{N} p_i = p_1 p_2 \cdots p_N.$$

Compute the following products.

1. $\prod_{i=1}^{M} \frac{1}{\theta} =$ (1/theta)^M ✓ **Answer:** theta^(-M)

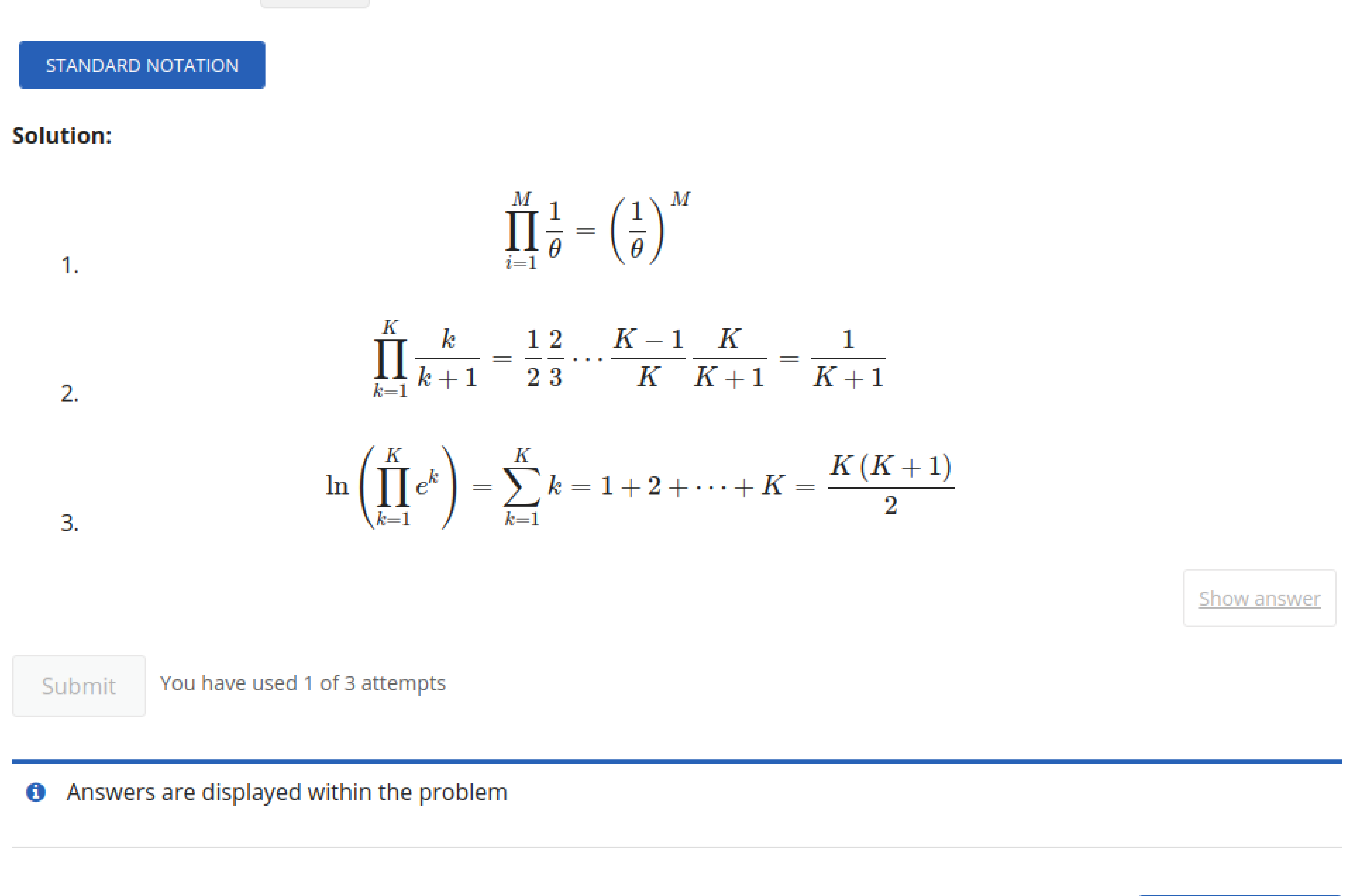
   $\left(\frac{1}{\theta}\right)^M$

2. $\prod_{k=1}^{K} \frac{k}{k+1} =$ 1/(K+1) ✓ **Answer:** 1/(K+1)

   $\frac{1}{K+1}$

3. $\ln\left(\prod_{k=1}^{K} e^k\right) =$ K*(K+1)/2 ✓ **Answer:** K*(K+1)/2

   $\frac{K\cdot(K+1)}{2}$



Solution:

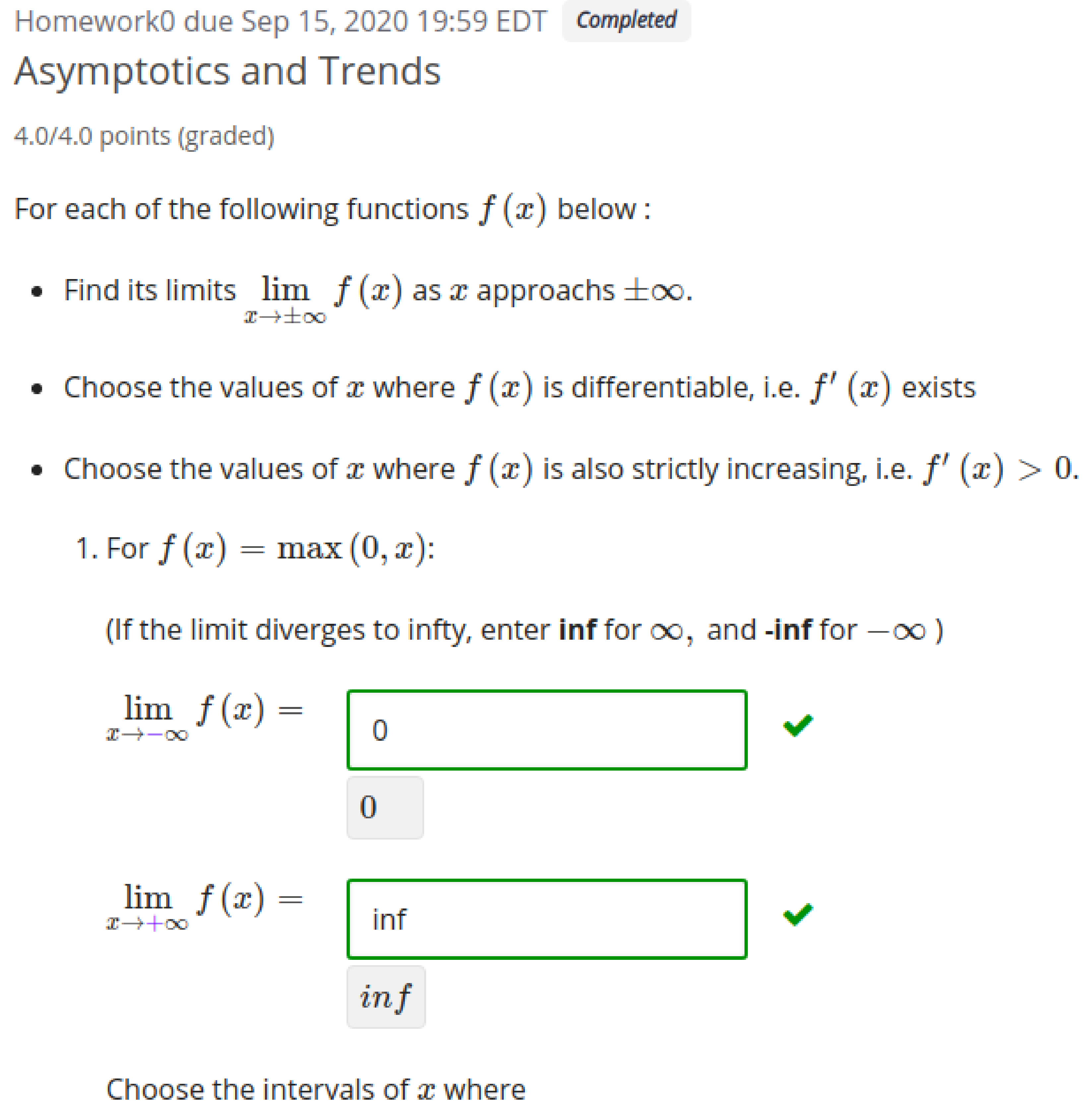
1. $$\prod_{i=1}^{M} \frac{1}{\theta} = \left(\frac{1}{\theta}\right)^{M}$$

2. $$\prod_{k=1}^{K} \frac{k}{k+1} = \frac{1}{2}\frac{2}{3}\cdots\frac{K-1}{K}\frac{K}{K+1} = \frac{1}{K+1}$$

3. $$\ln\left(\prod_{k=1}^{K} e^{k}\right) = \sum_{k=1}^{K} k = 1+2+\cdots+K = \frac{K\left(K+1\right)}{2}$$

You have used 1 of 3 attempts

Answers are displayed within the problem



Homework0 due Sep 15, 2020 19:59 EDT *Completed*

Asymptotics and Trends

4.0/4.0 points (graded)

For each of the following functions $f(x)$ below :

- Find its limits $\lim_{x\to\pm\infty} f(x)$ as x approaches $\pm\infty$.
- Choose the values of x where $f(x)$ is differentiable, i.e. $f'(x)$ exists
- Choose the values of x where $f(x)$ is also strictly increasing, i.e. $f'(x) > 0$.

1. For $f(x) = \max(0, x)$:

(If the limit diverges to infty, enter **inf** for ∞, and **-inf** for $-\infty$)

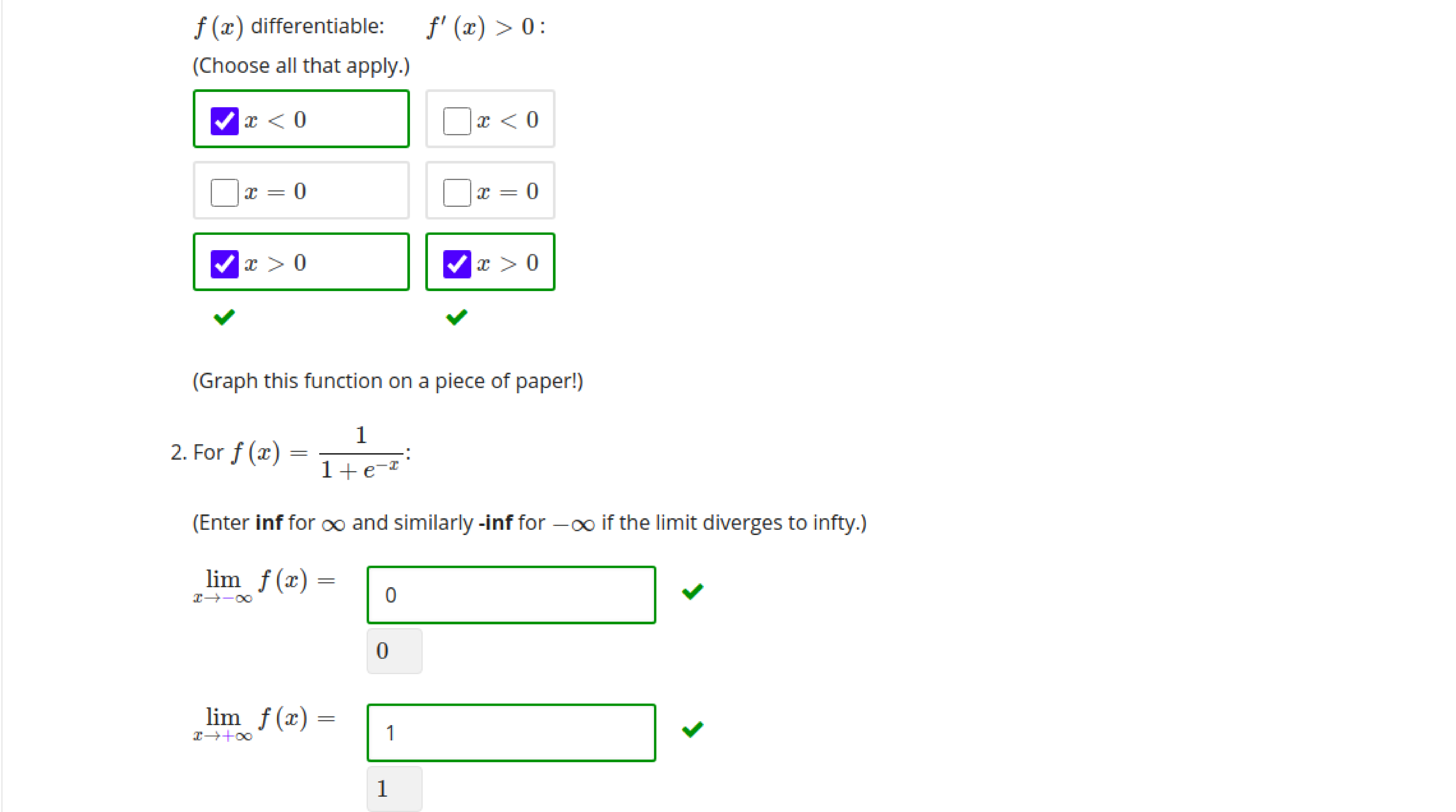
$\lim_{x\to-\infty} f(x) =$ 0 ✓

0

$\lim_{x\to+\infty} f(x) =$ inf ✓

inf

Choose the intervals of x where



$f(x)$ differentiable: $f'(x) > 0$:

(Choose all that apply.)

☑ $x < 0$	□ $x < 0$
□ $x = 0$	□ $x = 0$
☑ $x > 0$	☑ $x > 0$
✓	✓

(Graph this function on a piece of paper!)

2. For $f(x) = \dfrac{1}{1 + e^{-x}}$:

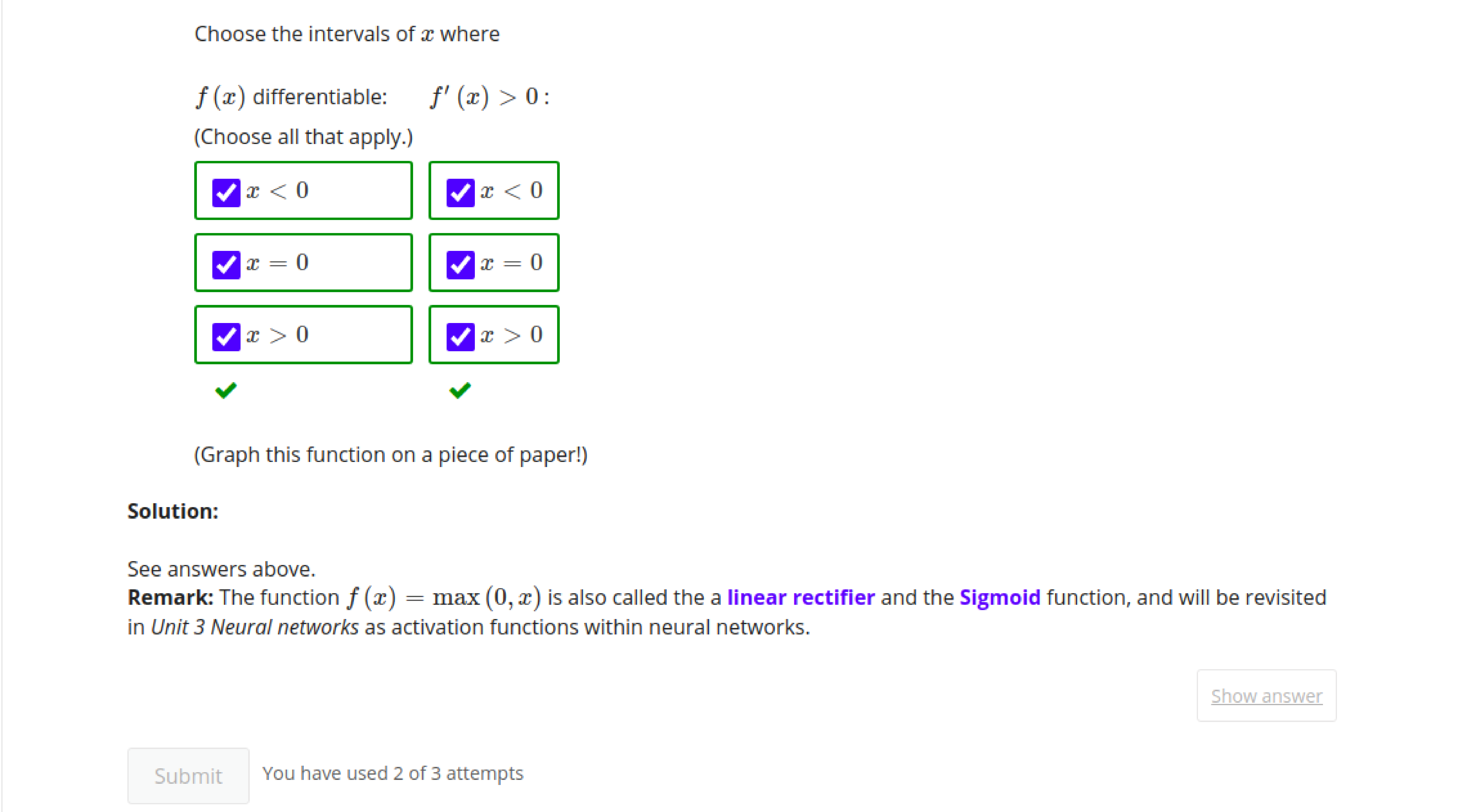
(Enter **inf** for ∞ and similarly **-inf** for $-\infty$ if the limit diverges to infty.)

$\lim\limits_{x \to -\infty} f(x) =$ 0 ✓

0

$\lim\limits_{x \to +\infty} f(x) =$ 1 ✓

1



Choose the intervals of x where

$f(x)$ differentiable: $f'(x) > 0$:

(Choose all that apply.)

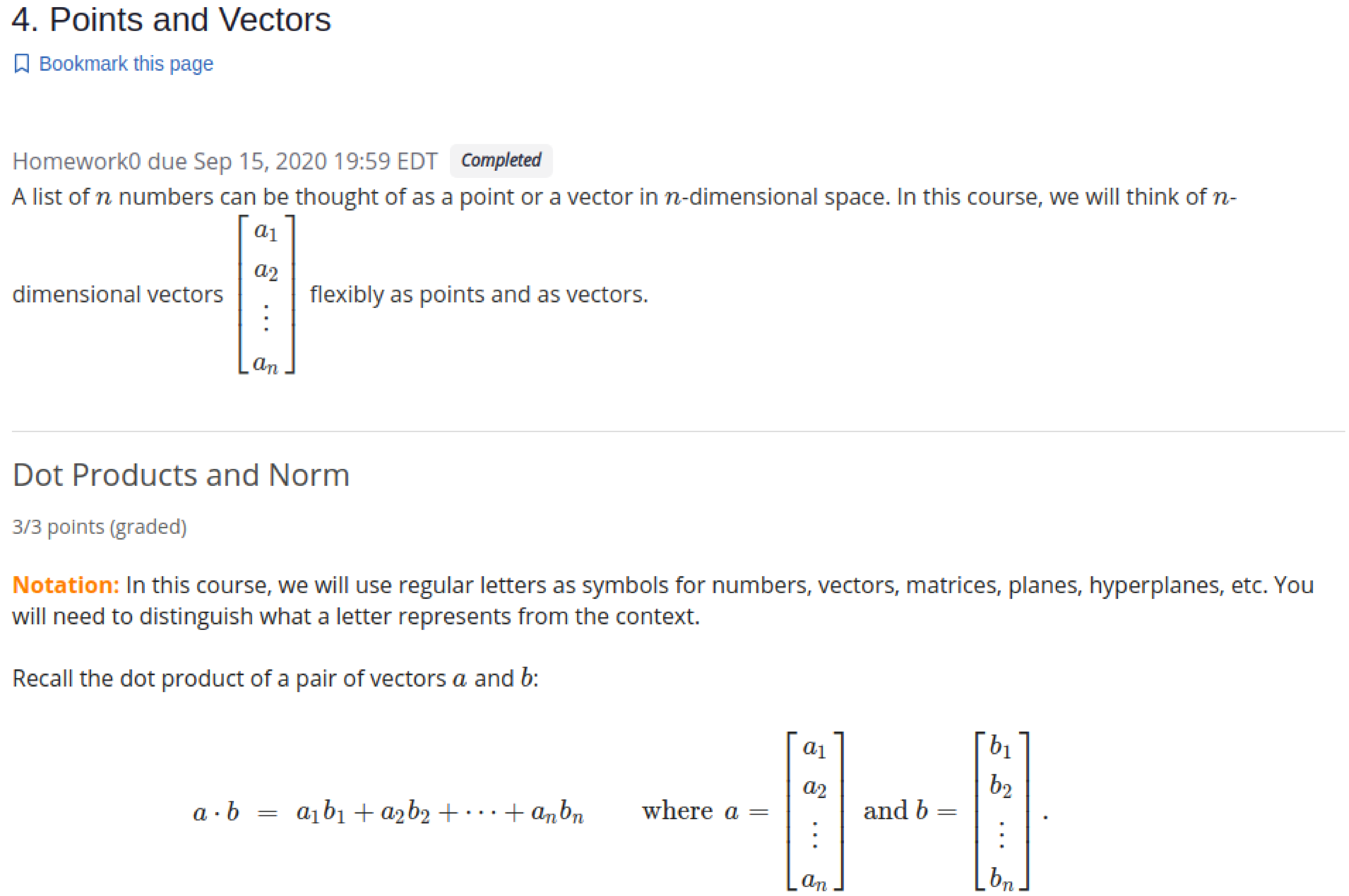
☑ $x < 0$	☑ $x < 0$
☑ $x = 0$	☑ $x = 0$
☑ $x > 0$	☑ $x > 0$
✓	✓

(Graph this function on a piece of paper!)

Solution:

See answers above.
Remark: The function $f(x) = \max(0, x)$ is also called the a **linear rectifier** and the **Sigmoid** function, and will be revisited in *Unit 3 Neural networks* as activation functions within neural networks.

You have used 2 of 3 attempts



4. Points and Vectors

Bookmark this page

Homework0 due Sep 15, 2020 19:59 EDT Completed

A list of n numbers can be thought of as a point or a vector in n-dimensional space. In this course, we will think of n-dimensional vectors $\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$ flexibly as points and as vectors.

Dot Products and Norm

3/3 points (graded)

Notation: In this course, we will use regular letters as symbols for numbers, vectors, matrices, planes, hyperplanes, etc. You will need to distinguish what a letter represents from the context.

Recall the dot product of a pair of vectors a and b:

$$a \cdot b \;=\; a_1 b_1 + a_2 b_2 + \cdots + a_n b_n \qquad \text{where } a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \text{ and } b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

When thinking about a and b as vectors in n-dimensional space, we can also express the dot product as

$$a \cdot b \;=\; \|a\|\|b\| \cos \alpha,$$

where α is the angle formed between the vectors a and b in n-dimensional Euclidean space. Here, $\|a\|$ refers to the length, also known as **norm**, of a:

$$\|a\| = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}.$$

What is the length of the vector $\begin{bmatrix} 0.4 \\ 0.3 \end{bmatrix}$?

0.5

✓ **Answer:** 0.5

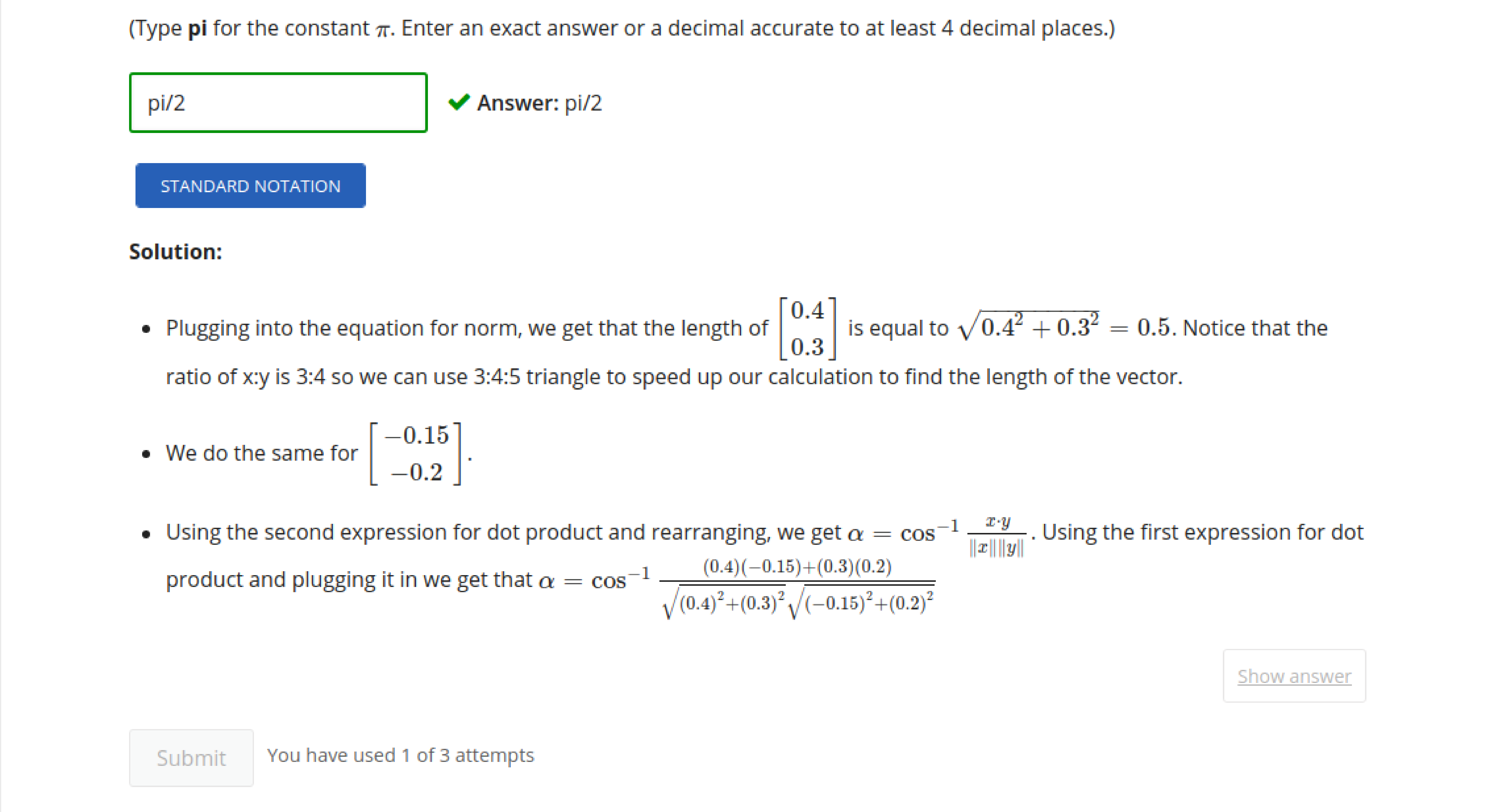
What is the length of the vector $\begin{bmatrix} -0.15 \\ 0.2 \end{bmatrix}$?

0.25

✓ **Answer:** 0.25

What is the angle (in radians) between $\begin{bmatrix} 0.4 \\ 0.3 \end{bmatrix}$ and $\begin{bmatrix} -0.15 \\ 0.2 \end{bmatrix}$? Choose the answer that lies between 0 and π.

(Type **pi** for the constant π. Enter an exact answer or a decimal accurate to at least 4 decimal places.)



(Type **pi** for the constant π. Enter an exact answer or a decimal accurate to at least 4 decimal places.)

pi/2 ✓ **Answer:** pi/2

Solution:

- Plugging into the equation for norm, we get that the length of $\begin{bmatrix} 0.4 \\ 0.3 \end{bmatrix}$ is equal to $\sqrt{0.4^2 + 0.3^2} = 0.5$. Notice that the ratio of x:y is 3:4 so we can use 3:4:5 triangle to speed up our calculation to find the length of the vector.
- We do the same for $\begin{bmatrix} -0.15 \\ -0.2 \end{bmatrix}$.
- Using the second expression for dot product and rearranging, we get $\alpha = \cos^{-1} \frac{x \cdot y}{\|x\| \|y\|}$. Using the first expression for dot product and plugging it in we get that $\alpha = \cos^{-1} \frac{(0.4)(-0.15)+(0.3)(0.2)}{\sqrt{(0.4)^2+(0.3)^2}\sqrt{(-0.15)^2+(0.2)^2}}$

You have used 1 of 3 attempts

Dot Products and Orthogonality

1/1 point (graded)

Given 3-dimensional vectors $x^{(1)} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$ and $x^{(2)} = \begin{bmatrix} a_1 \\ -a_2 \\ a_3 \end{bmatrix}$, when is $x^{(1)}$ orthogonal to $x^{(2)}$, i.e. the angle between them is $\pi/2$?

○ when $2a_1 + 2a_3 = 0$

● when $a_1^2 - a_2^2 + a_3^2 = 0$

○ when $a_1^2 + a_2^2 + a_3^2 = 0$

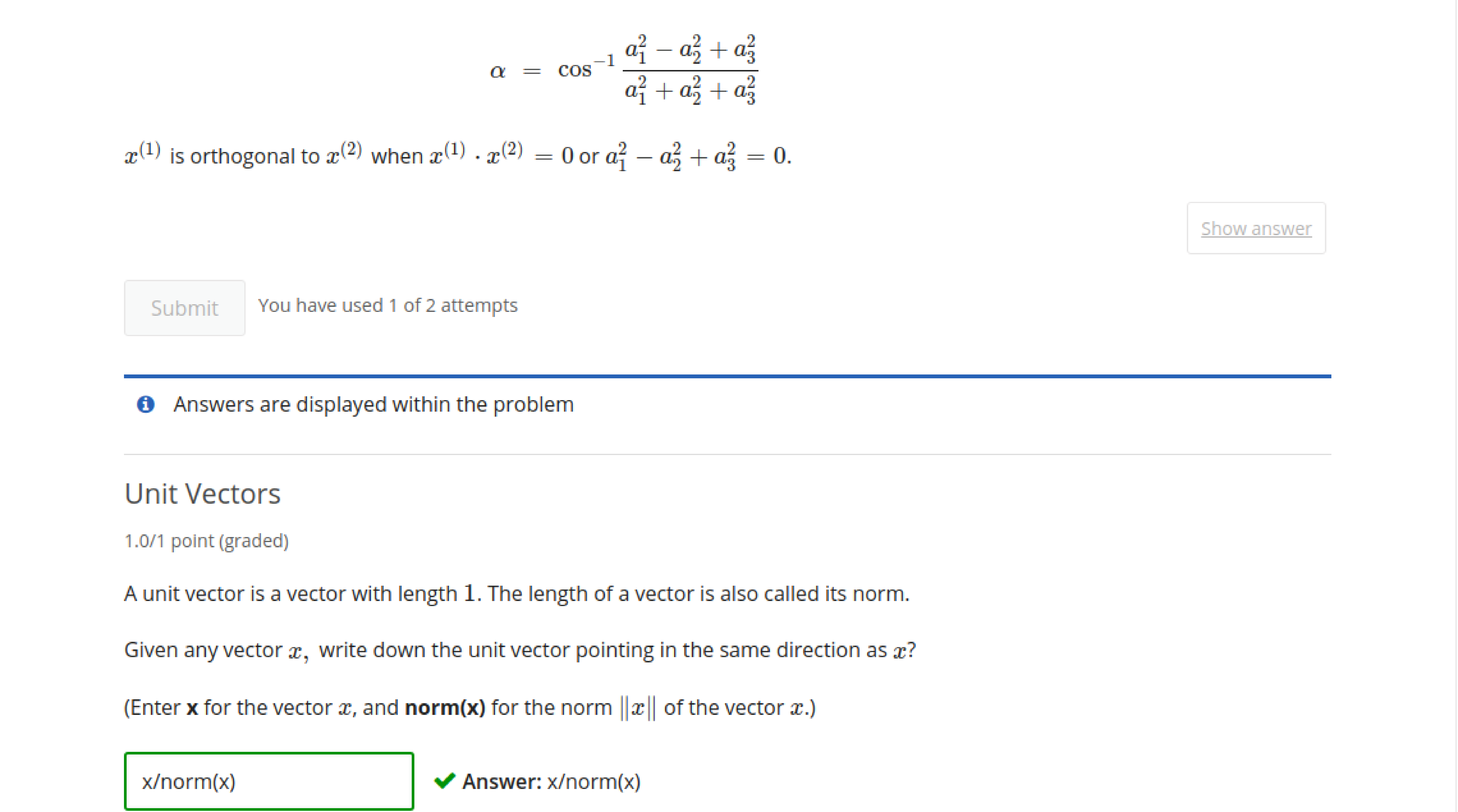
✓

STANDARD NOTATION

Solution:

Based on the previous equations for the dot product, we find that the angle between $x^{(1)}$ and $x^{(2)}$ is:

$$\alpha = \cos^{-1} \frac{x^{(1)} \cdot x^{(2)}}{||x^{(1)}||||x^{(2)}||}$$



$$\alpha = \cos^{-1}\frac{a_1^2 - a_2^2 + a_3^2}{a_1^2 + a_2^2 + a_3^2}$$

$x^{(1)}$ is orthogonal to $x^{(2)}$ when $x^{(1)} \cdot x^{(2)} = 0$ or $a_1^2 - a_2^2 + a_3^2 = 0.$

You have used 1 of 2 attempts

Answers are displayed within the problem

Unit Vectors

1.0/1 point (graded)

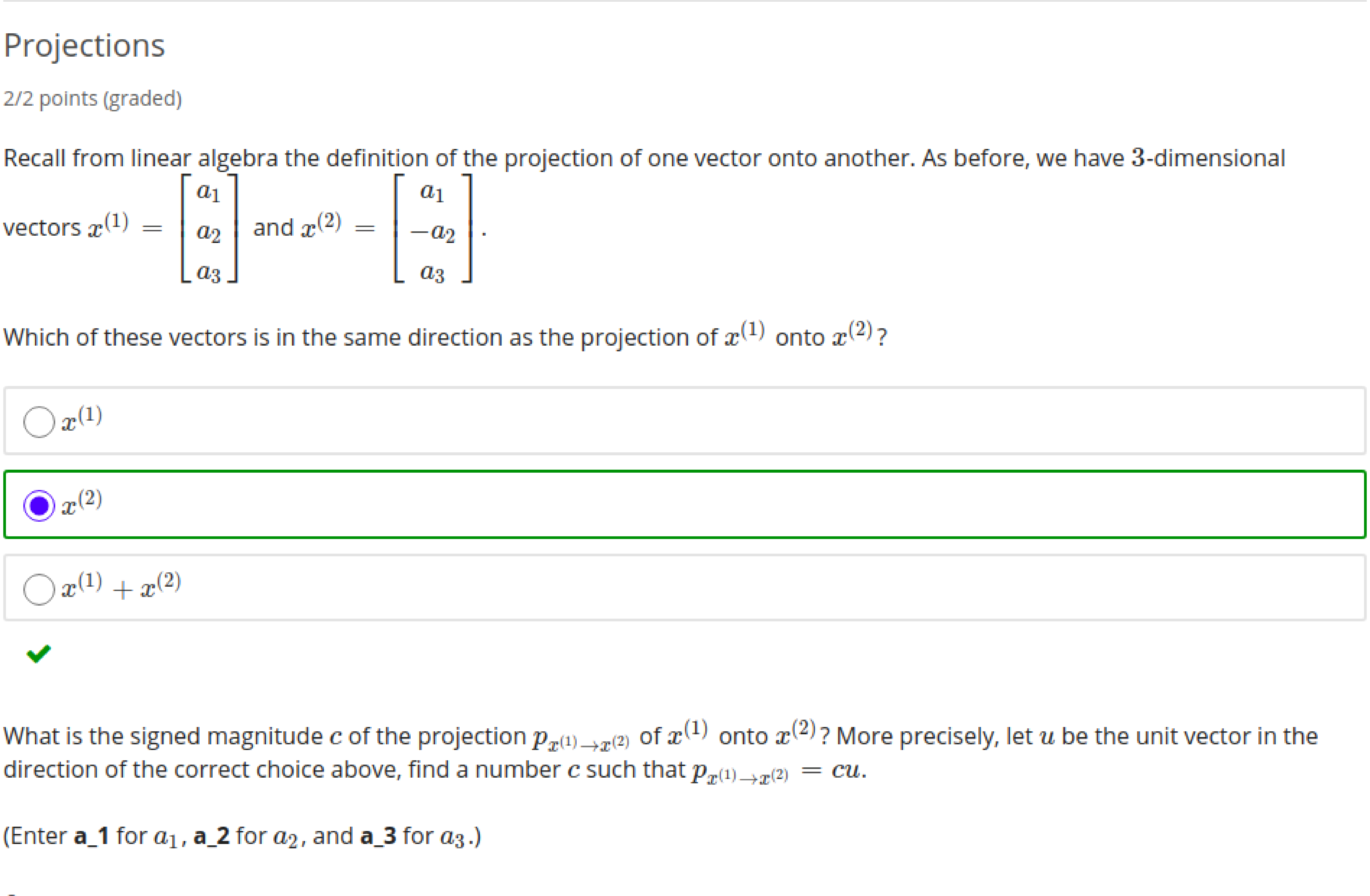
A unit vector is a vector with length 1. The length of a vector is also called its norm.

Given any vector $x,$ write down the unit vector pointing in the same direction as x?

(Enter **x** for the vector x, and **norm(x)** for the norm $||x||$ of the vector x.)

x/norm(x)

✓ **Answer:** x/norm(x)



Projections

2/2 points (graded)

Recall from linear algebra the definition of the projection of one vector onto another. As before, we have 3-dimensional vectors $x^{(1)} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$ and $x^{(2)} = \begin{bmatrix} a_1 \\ -a_2 \\ a_3 \end{bmatrix}$.

Which of these vectors is in the same direction as the projection of $x^{(1)}$ onto $x^{(2)}$?

○ $x^{(1)}$

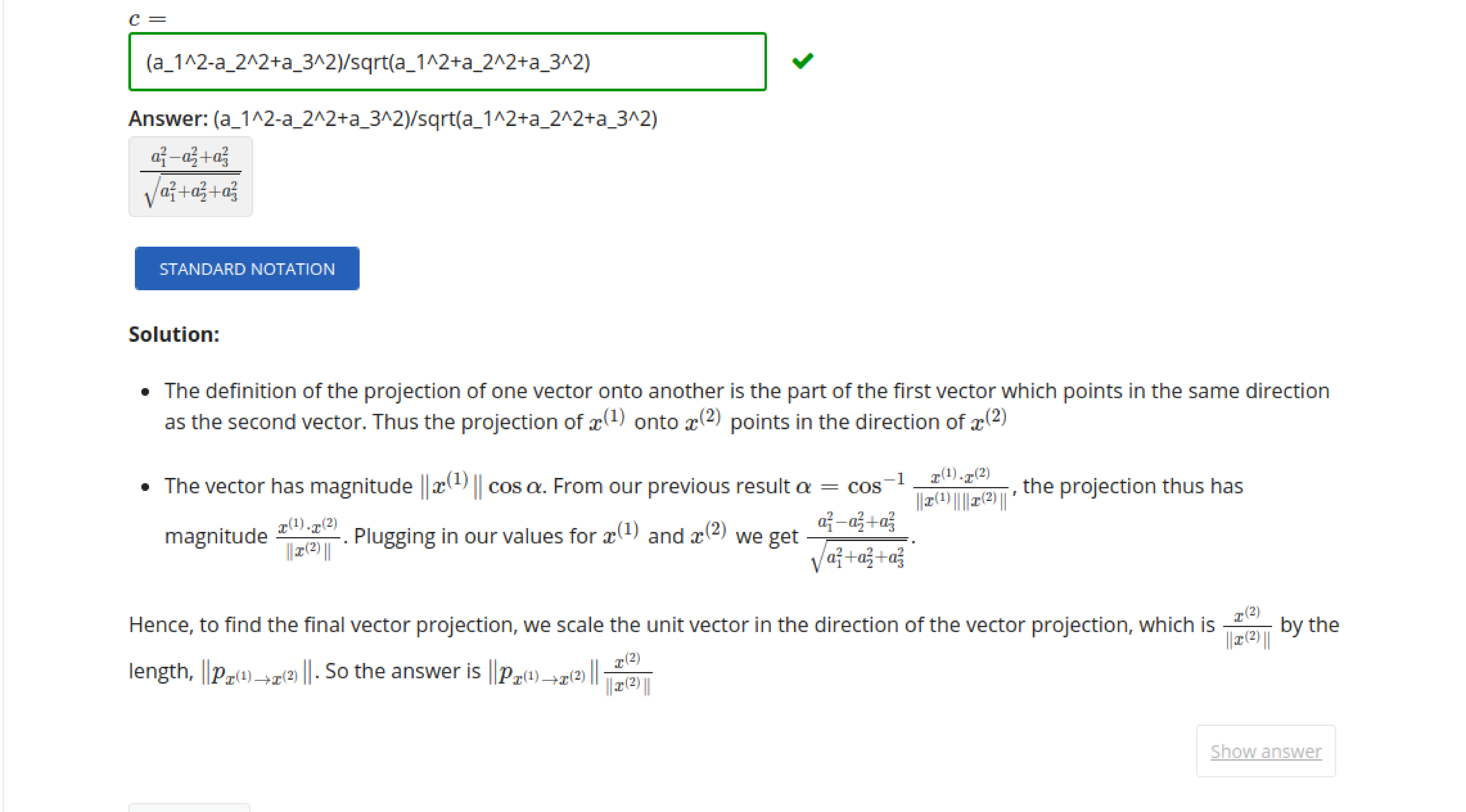
● $x^{(2)}$

○ $x^{(1)} + x^{(2)}$

✓

What is the signed magnitude c of the projection $p_{x^{(1)} \to x^{(2)}}$ of $x^{(1)}$ onto $x^{(2)}$? More precisely, let u be the unit vector in the direction of the correct choice above, find a number c such that $p_{x^{(1)} \to x^{(2)}} = cu$.

(Enter **a_1** for a_1, **a_2** for a_2, and **a_3** for a_3.)



$c =$

(a_1^2-a_2^2+a_3^2)/sqrt(a_1^2+a_2^2+a_3^2) ✓

Answer: (a_1^2-a_2^2+a_3^2)/sqrt(a_1^2+a_2^2+a_3^2)

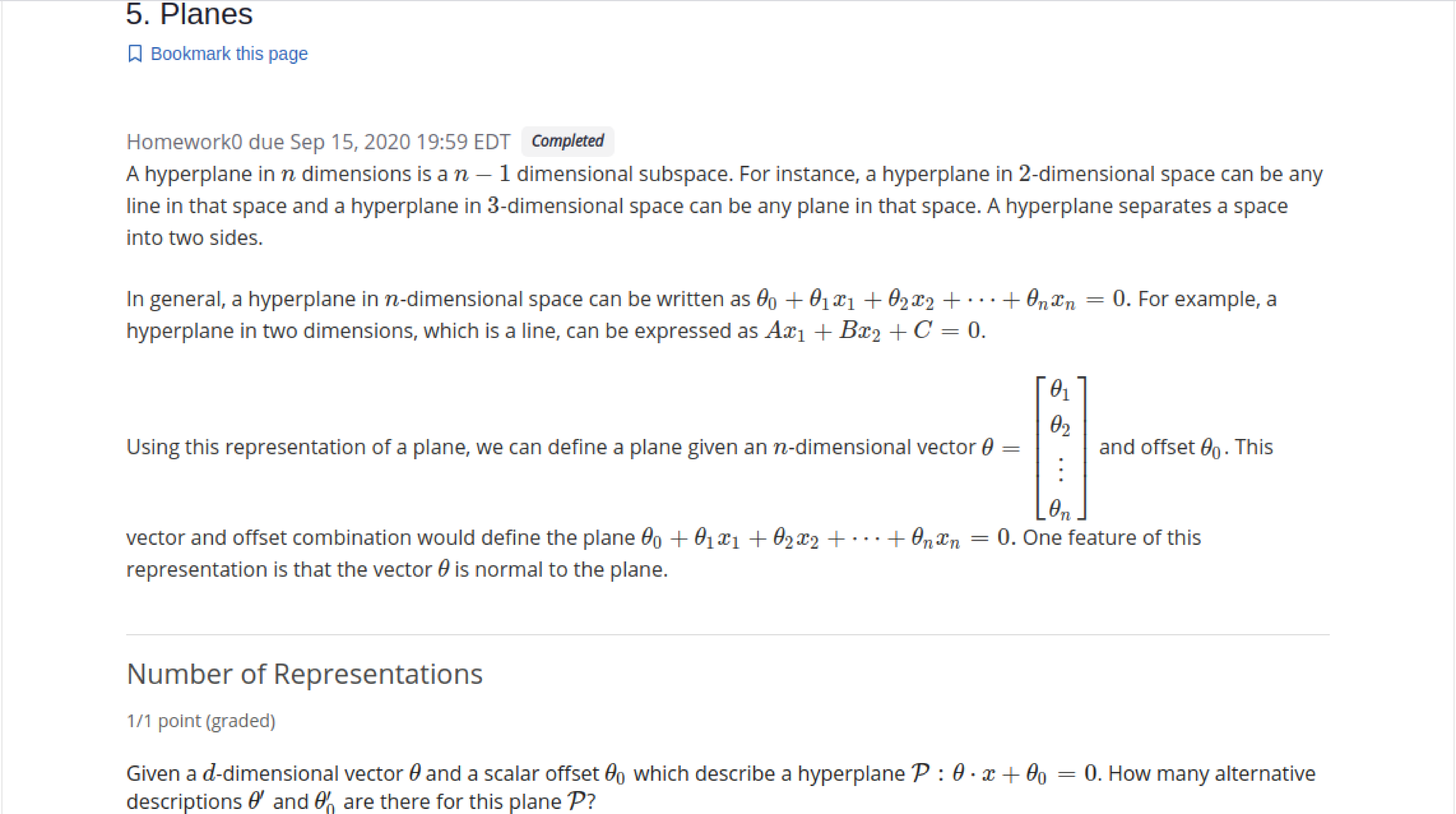
$$\frac{a_1^2-a_2^2+a_3^2}{\sqrt{a_1^2+a_2^2+a_3^2}}$$

Solution:

- The definition of the projection of one vector onto another is the part of the first vector which points in the same direction as the second vector. Thus the projection of $x^{(1)}$ onto $x^{(2)}$ points in the direction of $x^{(2)}$

- The vector has magnitude $\|x^{(1)}\| \cos\alpha$. From our previous result $\alpha = \cos^{-1} \frac{x^{(1)} \cdot x^{(2)}}{\|x^{(1)}\| \|x^{(2)}\|}$, the projection thus has magnitude $\frac{x^{(1)} \cdot x^{(2)}}{\|x^{(2)}\|}$. Plugging in our values for $x^{(1)}$ and $x^{(2)}$ we get $\frac{a_1^2-a_2^2+a_3^2}{\sqrt{a_1^2+a_2^2+a_3^2}}$.

Hence, to find the final vector projection, we scale the unit vector in the direction of the vector projection, which is $\frac{x^{(2)}}{\|x^{(2)}\|}$ by the length, $\|p_{x^{(1)} \to x^{(2)}}\|$. So the answer is $\|p_{x^{(1)} \to x^{(2)}}\| \frac{x^{(2)}}{\|x^{(2)}\|}$



5. Planes

Bookmark this page

Homework0 due Sep 15, 2020 19:59 EDT Completed

A hyperplane in n dimensions is a $n-1$ dimensional subspace. For instance, a hyperplane in 2-dimensional space can be any line in that space and a hyperplane in 3-dimensional space can be any plane in that space. A hyperplane separates a space into two sides.

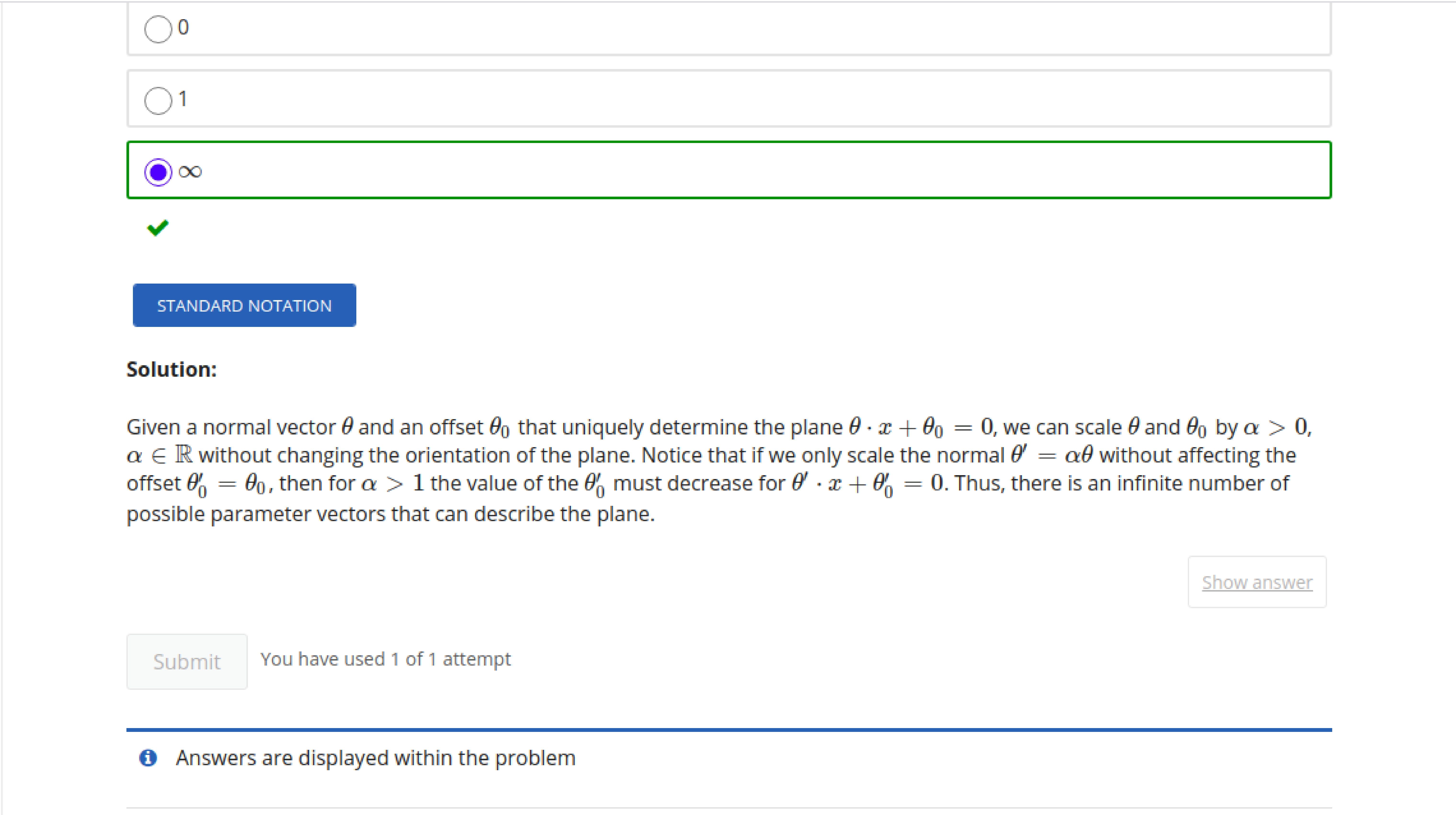
In general, a hyperplane in n-dimensional space can be written as $\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n = 0$. For example, a hyperplane in two dimensions, which is a line, can be expressed as $Ax_1 + Bx_2 + C = 0$.

Using this representation of a plane, we can define a plane given an n-dimensional vector $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix}$ and offset θ_0. This vector and offset combination would define the plane $\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n = 0$. One feature of this representation is that the vector θ is normal to the plane.

Number of Representations

1/1 point (graded)

Given a d-dimensional vector θ and a scalar offset θ_0 which describe a hyperplane $\mathcal{P} : \theta \cdot x + \theta_0 = 0$. How many alternative descriptions θ' and θ'_0 are there for this plane $\mathcal{P}$?



○ 0

○ 1

● ∞

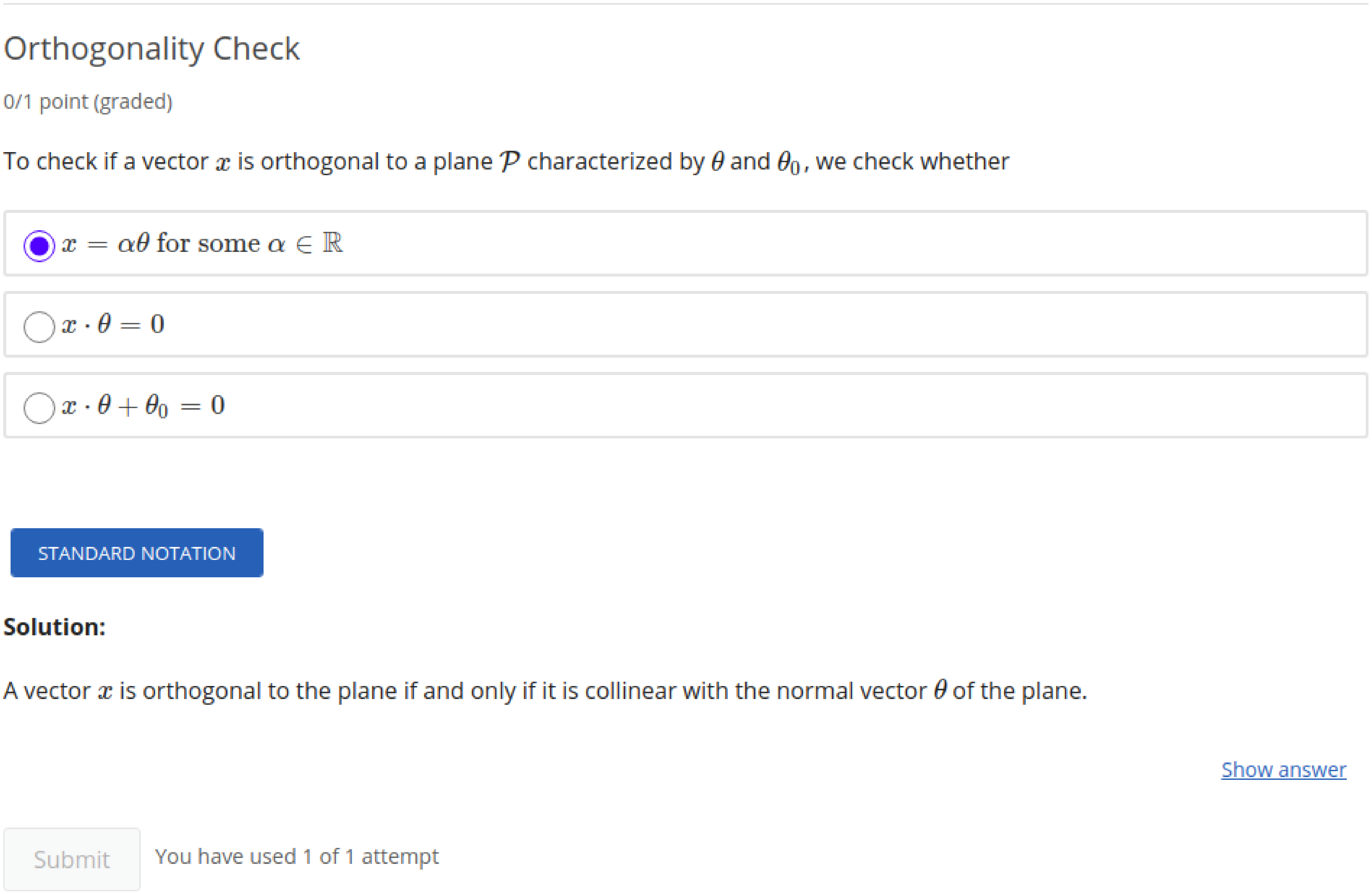
✓

Solution:

Given a normal vector θ and an offset θ_0 that uniquely determine the plane $\theta \cdot x + \theta_0 = 0$, we can scale θ and θ_0 by $\alpha > 0$, $\alpha \in \mathbb{R}$ without changing the orientation of the plane. Notice that if we only scale the normal $\theta' = \alpha\theta$ without affecting the offset $\theta_0' = \theta_0$, then for $\alpha > 1$ the value of the θ_0' must decrease for $\theta' \cdot x + \theta_0' = 0$. Thus, there is an infinite number of possible parameter vectors that can describe the plane.

You have used 1 of 1 attempt

Answers are displayed within the problem



Orthogonality Check

0/1 point (graded)

To check if a vector x is orthogonal to a plane $\mathcal{P}$ characterized by θ and θ_0, we check whether

● $x = \alpha\theta \text{ for some } \alpha \in \mathbb{R}$

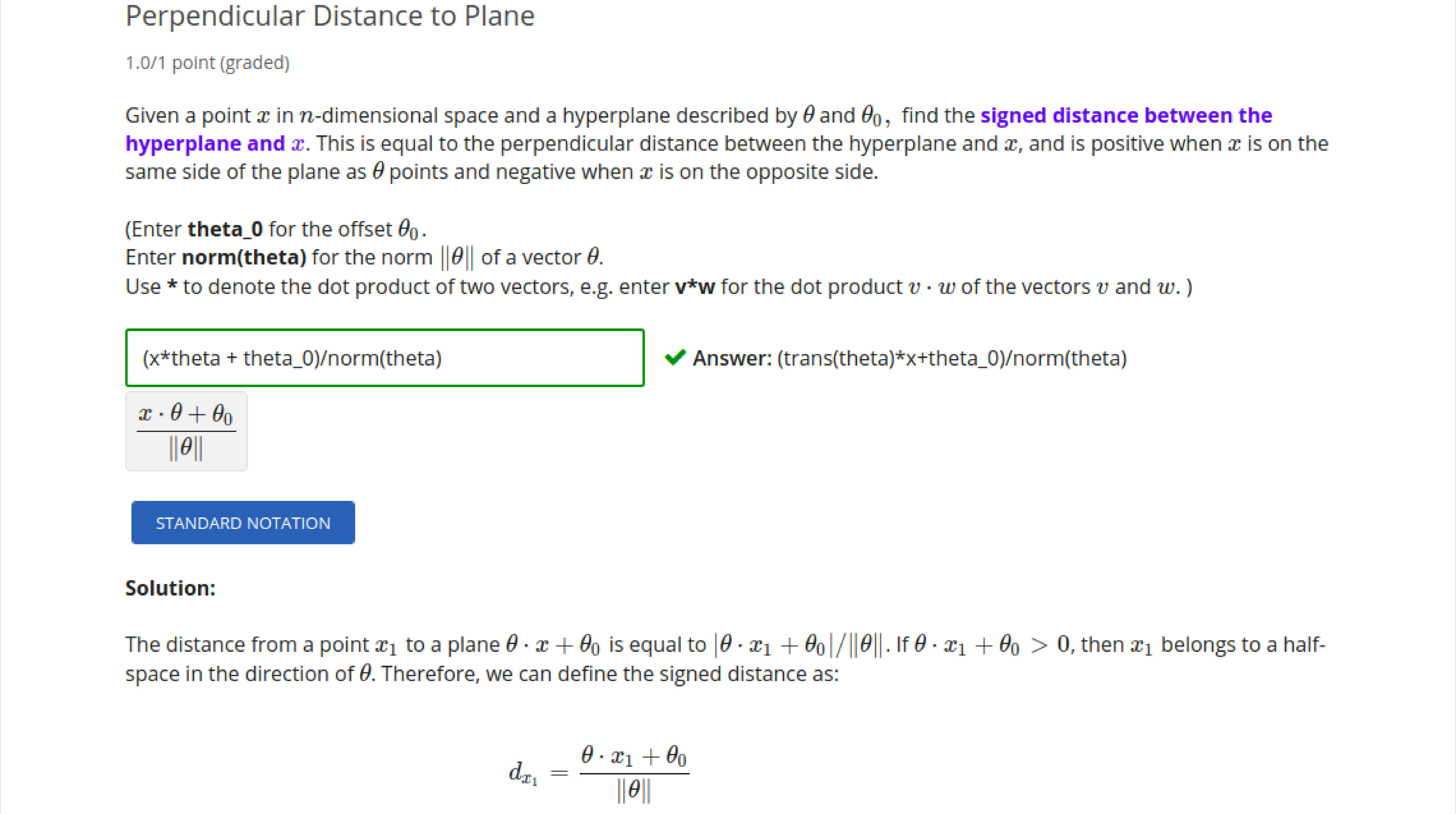
○ $x \cdot \theta = 0$

○ $x \cdot \theta + \theta_0 = 0$

Solution:

A vector x is orthogonal to the plane if and only if it is collinear with the normal vector θ of the plane.

You have used 1 of 1 attempt



Perpendicular Distance to Plane

1.0/1 point (graded)

Given a point x in n-dimensional space and a hyperplane described by θ and θ_0, find the **signed distance between the hyperplane and x**. This is equal to the perpendicular distance between the hyperplane and x, and is positive when x is on the same side of the plane as θ points and negative when x is on the opposite side.

(Enter **theta_0** for the offset θ_0.
Enter **norm(theta)** for the norm $||\theta||$ of a vector θ.
Use ***** to denote the dot product of two vectors, e.g. enter **v*w** for the dot product $v \cdot w$ of the vectors v and w.)

(x*theta + theta_0)/norm(theta) ✓ **Answer:** (trans(theta)*x+theta_0)/norm(theta)

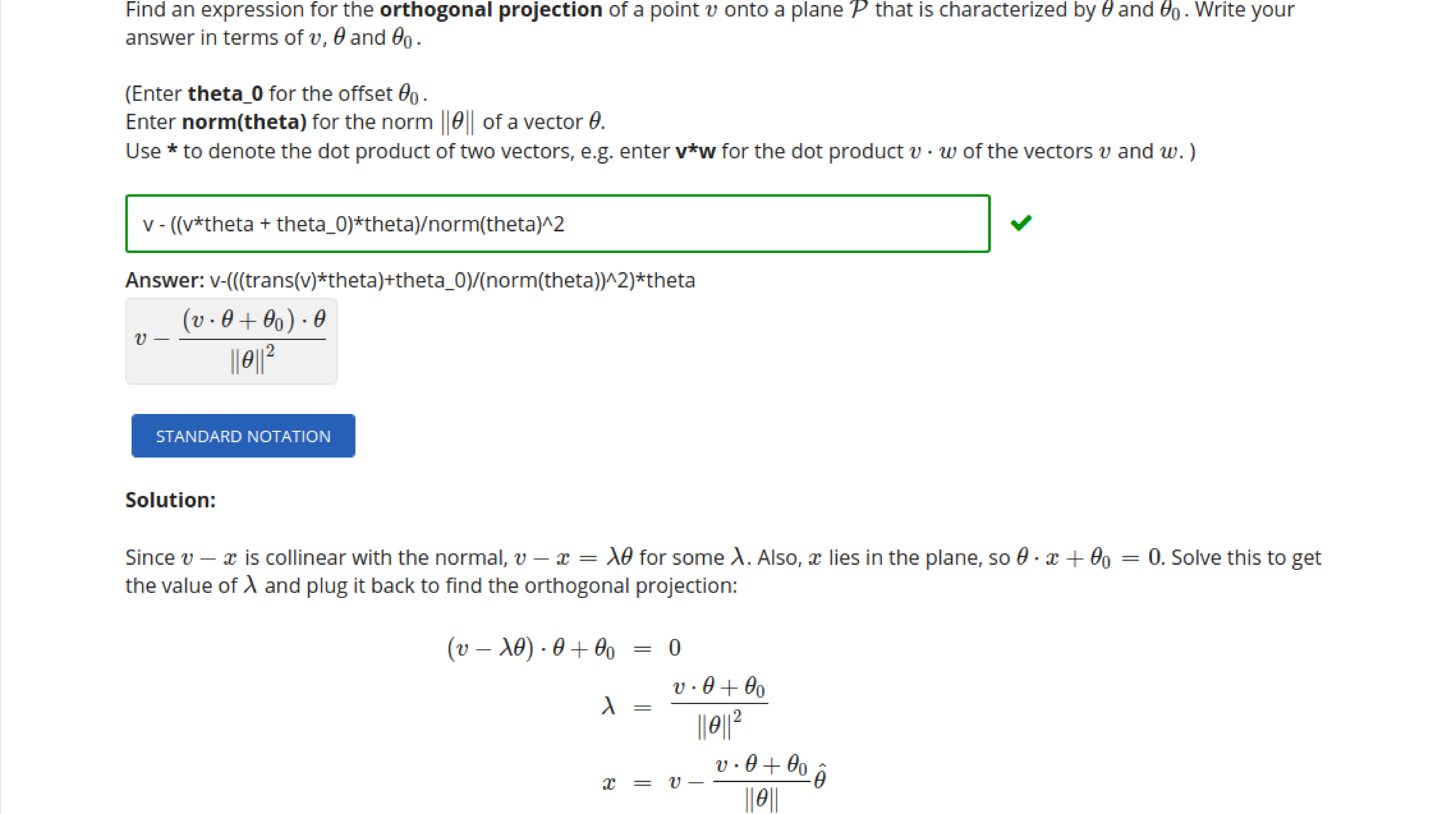
$$\frac{x \cdot \theta + \theta_0}{||\theta||}$$

STANDARD NOTATION

Solution:

The distance from a point x_1 to a plane $\theta \cdot x + \theta_0$ is equal to $|\theta \cdot x_1 + \theta_0|/||\theta||$. If $\theta \cdot x_1 + \theta_0 > 0$, then x_1 belongs to a half-space in the direction of θ. Therefore, we can define the signed distance as:

$$d_{x_1} = \frac{\theta \cdot x_1 + \theta_0}{||\theta||}$$



Find an expression for the **orthogonal projection** of a point v onto a plane $\mathcal{P}$ that is characterized by θ and θ_0. Write your answer in terms of v, θ and θ_0.

(Enter **theta_0** for the offset θ_0.
Enter **norm(theta)** for the norm $||\theta||$ of a vector θ.
Use * to denote the dot product of two vectors, e.g. enter **v*w** for the dot product $v \cdot w$ of the vectors v and w.)

```
v - ((v*theta + theta_0)*theta)/norm(theta)^2
```
✓

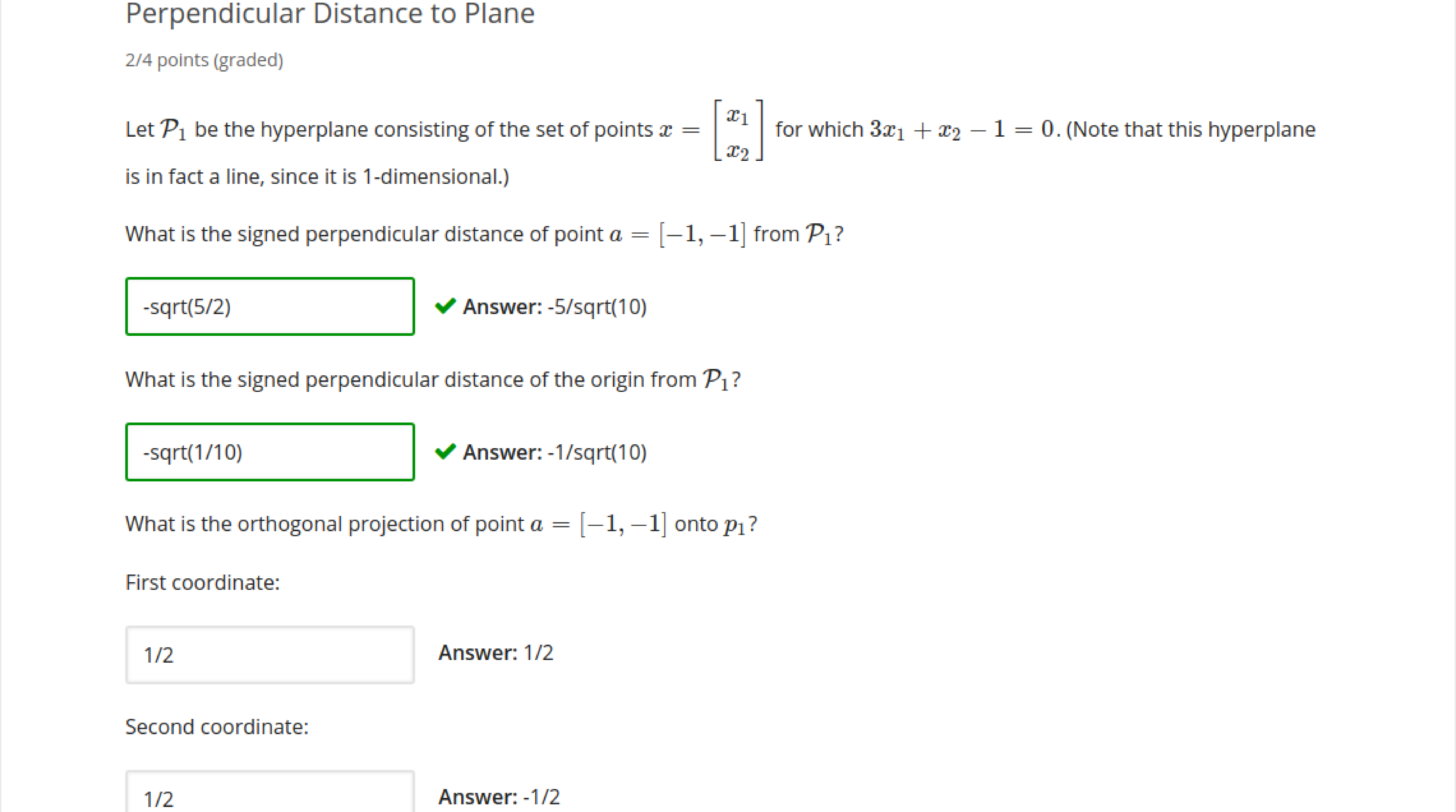
Answer: v-(((trans(v)*theta)+theta_0)/(norm(theta))^2)*theta

$$v - \frac{(v \cdot \theta + \theta_0) \cdot \theta}{||\theta||^2}$$

Solution:

Since $v - x$ is collinear with the normal, $v - x = \lambda\theta$ for some λ. Also, x lies in the plane, so $\theta \cdot x + \theta_0 = 0$. Solve this to get the value of λ and plug it back to find the orthogonal projection:

$$\begin{aligned} (v - \lambda\theta) \cdot \theta + \theta_0 &= 0 \\ \lambda &= \frac{v \cdot \theta + \theta_0}{||\theta||^2} \\ x &= v - \frac{v \cdot \theta + \theta_0}{||\theta||}\hat{\theta} \end{aligned}$$



Perpendicular Distance to Plane

2/4 points (graded)

Let $\mathcal{P}_1$ be the hyperplane consisting of the set of points $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ for which $3x_1 + x_2 - 1 = 0$. (Note that this hyperplane is in fact a line, since it is 1-dimensional.)

What is the signed perpendicular distance of point $a = [-1, -1]$ from $\mathcal{P}_1$?

-sqrt(5/2)

✓ **Answer:** -5/sqrt(10)

What is the signed perpendicular distance of the origin from $\mathcal{P}_1$?

-sqrt(1/10)

✓ **Answer:** -1/sqrt(10)

What is the orthogonal projection of point $a = [-1, -1]$ onto p_1?

First coordinate:

1/2

Answer: 1/2

Second coordinate:

1/2

Answer: -1/2

Solution:

1. For $a = [-1, -1]^T$ the signed distance is:

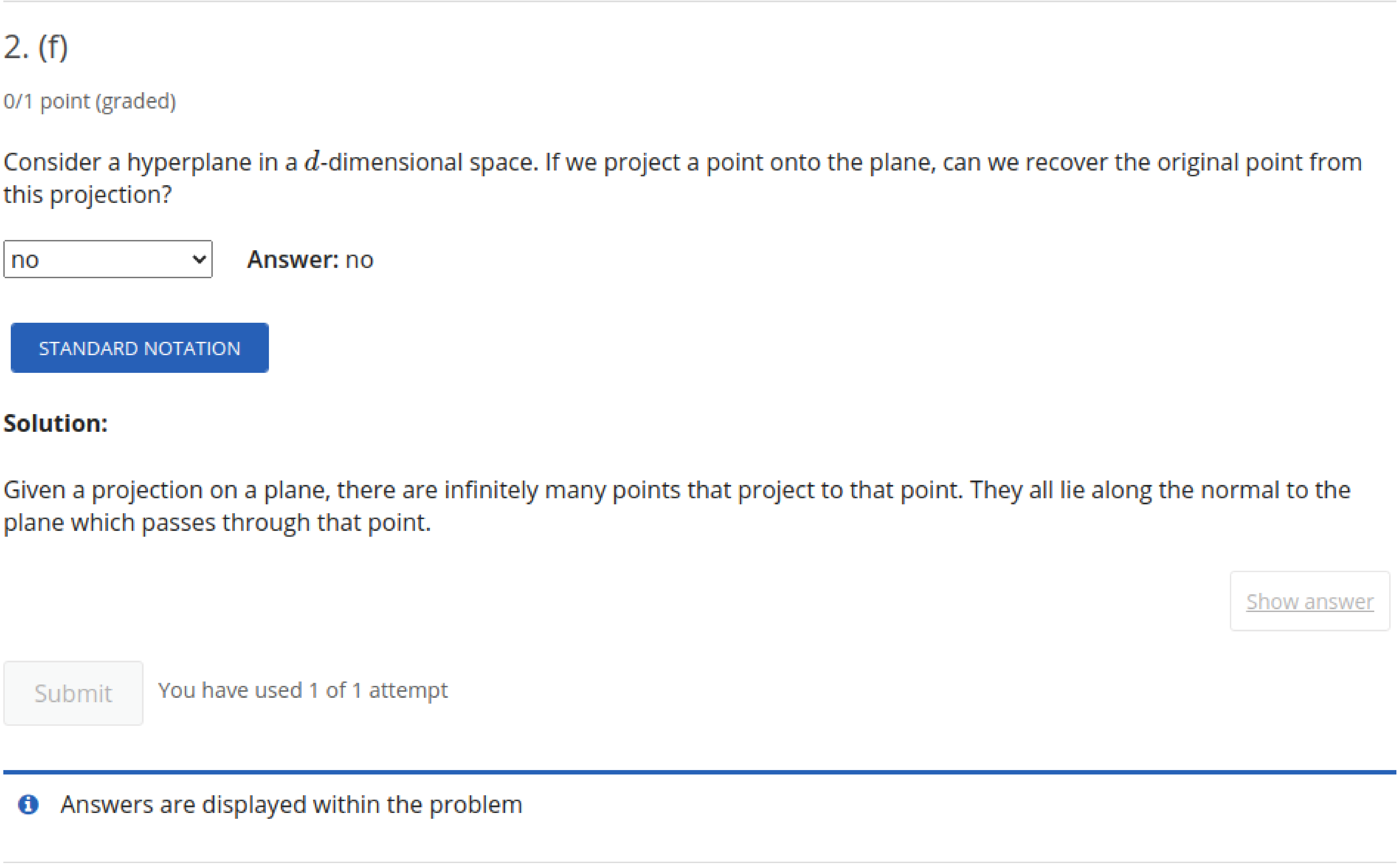
$$\frac{\theta \cdot a + \theta_0}{\|\theta\|} = \frac{(3)\,(-1) + (1)\,(-1) - 1}{\sqrt{(3)^2 + (1)^2}} = -\frac{5}{\sqrt{10}}$$

2. For $a = [0, 0]^T$ the signed distance is:

$$\frac{\theta \cdot 0 + \theta_0}{\|\theta\|} = \frac{-1}{\sqrt{(3)^2 + (1)^2}} = -\frac{1}{\sqrt{10}}$$

3. For $a = [-1, -1]^T$ the orthogonal projection is:

$$\begin{aligned} x &= v - \frac{v \cdot \theta + \theta_0}{\|\theta\|}\hat{\theta} \\ &= [-1, -1]^T - \frac{[-1, -1]^T \cdot [3, 1]^T + (-1)}{\sqrt{(3)^2 + (1)^2}}[3/\sqrt{10}, 1/\sqrt{10}]^T \\ &= [1/2, -1/2]^T \end{aligned}$$



2. (f)

0/1 point (graded)

Consider a hyperplane in a d-dimensional space. If we project a point onto the plane, can we recover the original point from this projection?

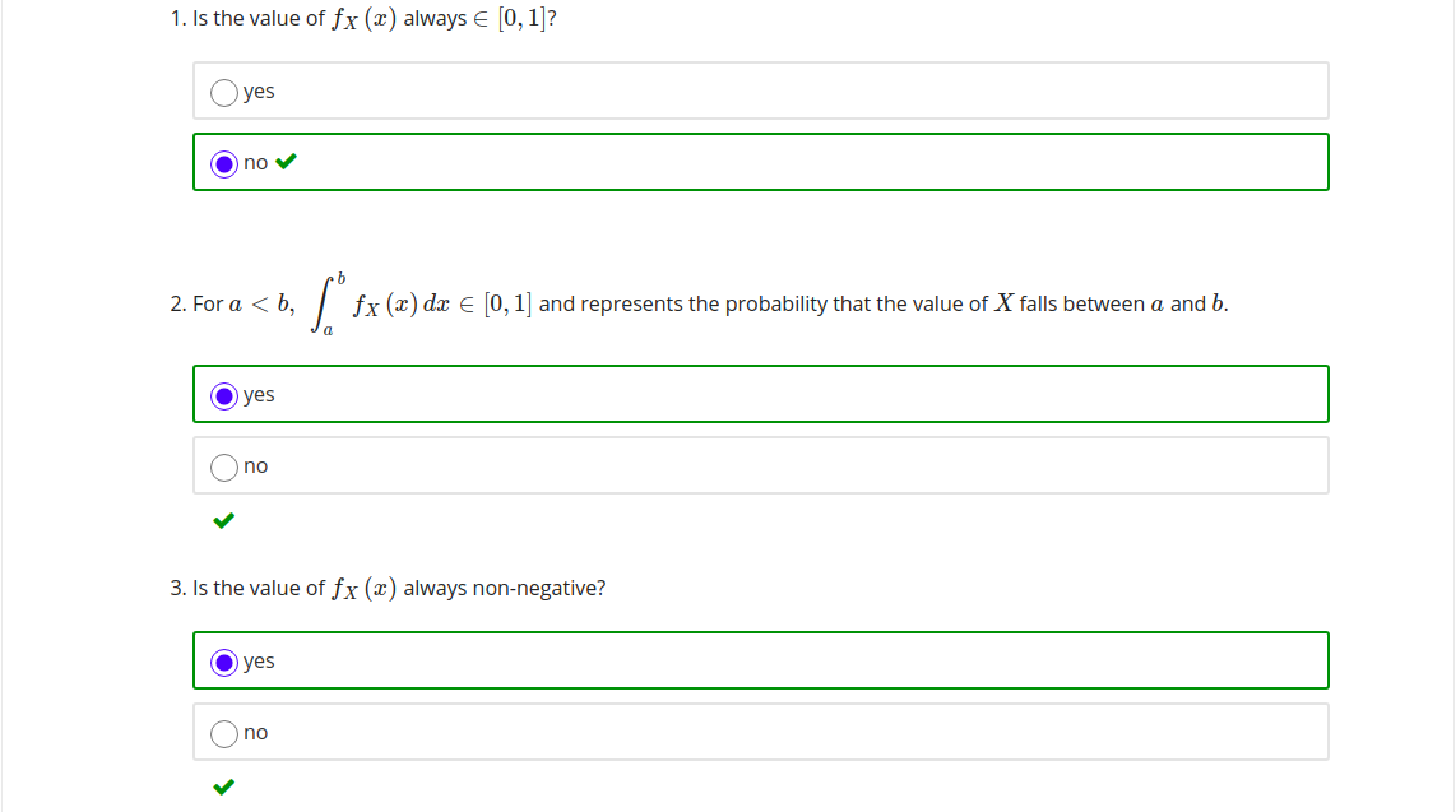
no **Answer:** no

Solution:

Given a projection on a plane, there are infinitely many points that project to that point. They all lie along the normal to the plane which passes through that point.

You have used 1 of 1 attempt

Answers are displayed within the problem



1. Is the value of $f_X(x)$ always $\in [0,1]$?

○ yes

● no ✓

2. For $a < b$, $\int_a^b f_X(x)\,dx \in [0,1]$ and represents the probability that the value of X falls between a and b.

● yes

○ no

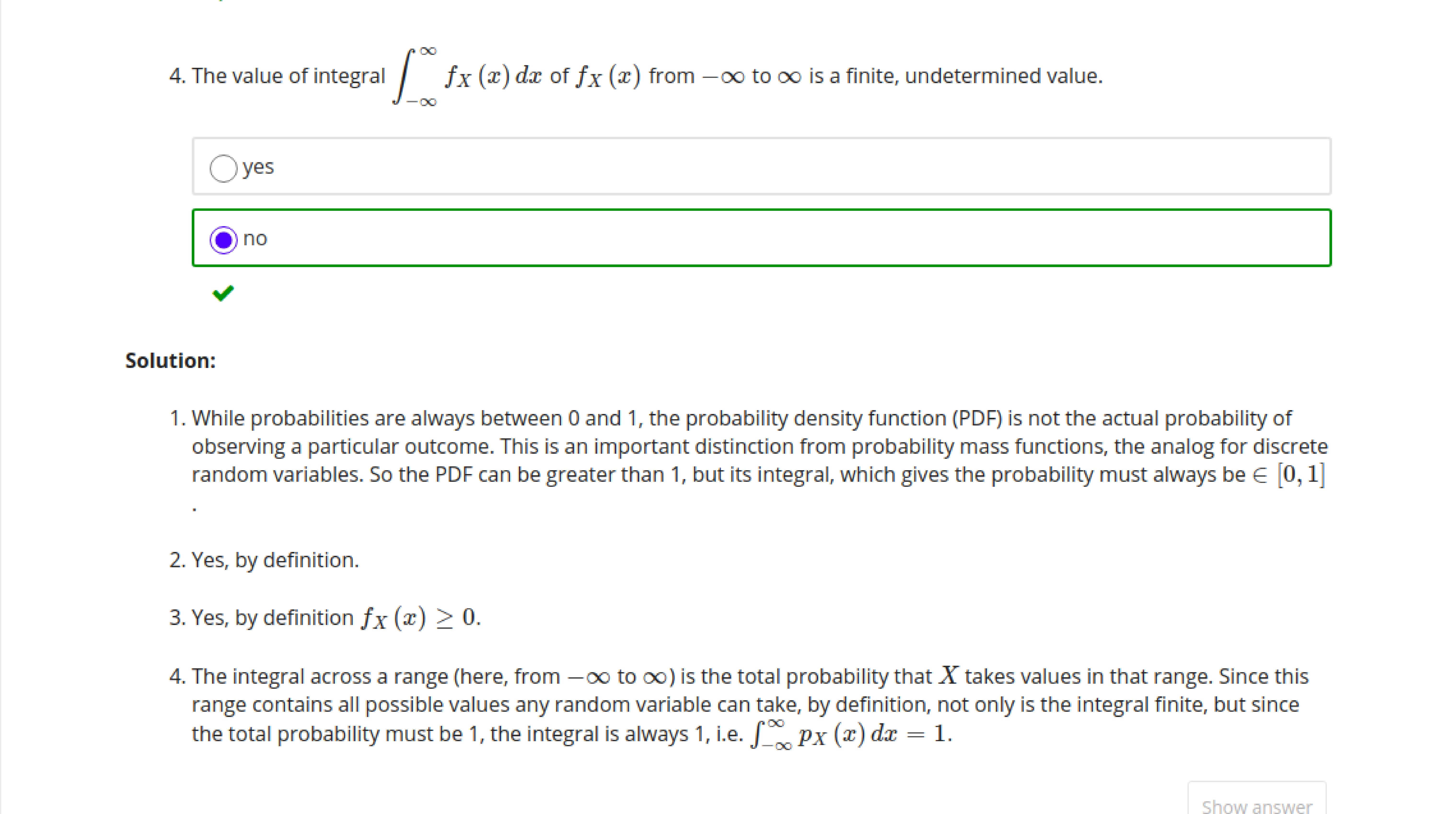
✓

3. Is the value of $f_X(x)$ always non-negative?

● yes

○ no

✓



4. The value of integral $\int_{-\infty}^{\infty} f_X\left(x\right)dx$ of $f_X\left(x\right)$ from $-\infty$ to ∞ is a finite, undetermined value.

○ yes

● no

✓

Solution:

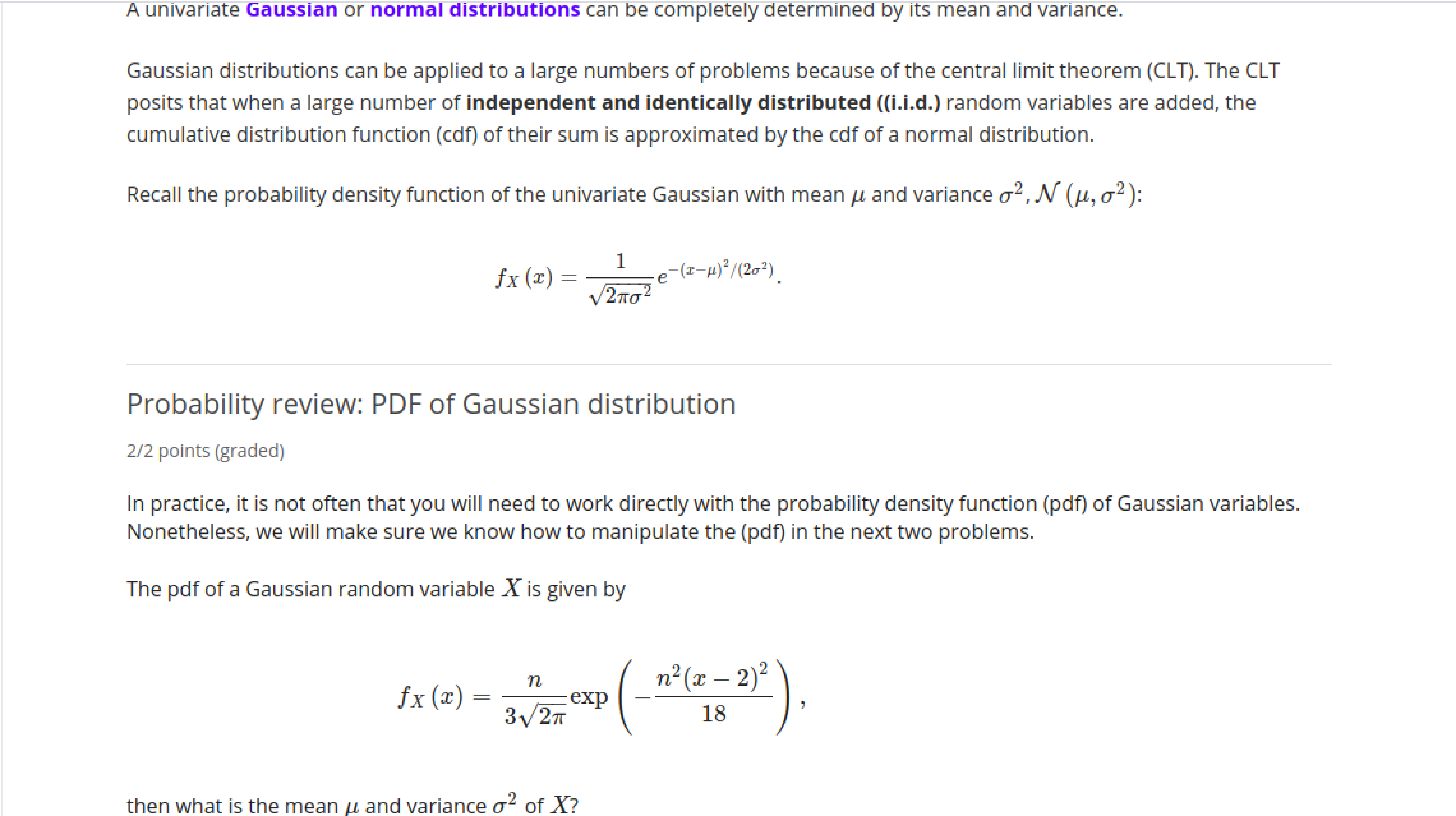
1. While probabilities are always between 0 and 1, the probability density function (PDF) is not the actual probability of observing a particular outcome. This is an important distinction from probability mass functions, the analog for discrete random variables. So the PDF can be greater than 1, but its integral, which gives the probability must always be $\in \left[0,1\right]$.

2. Yes, by definition.

3. Yes, by definition $f_X\left(x\right) \geq 0$.

4. The integral across a range (here, from $-\infty$ to ∞) is the total probability that X takes values in that range. Since this range contains all possible values any random variable can take, by definition, not only is the integral finite, but since the total probability must be 1, the integral is always 1, i.e. $\int_{-\infty}^{\infty} p_X\left(x\right)dx = 1$.

Show answer



A univariate **Gaussian** or **normal distributions** can be completely determined by its mean and variance.

Gaussian distributions can be applied to a large numbers of problems because of the central limit theorem (CLT). The CLT posits that when a large number of **independent and identically distributed ((i.i.d.)** random variables are added, the cumulative distribution function (cdf) of their sum is approximated by the cdf of a normal distribution.

Recall the probability density function of the univariate Gaussian with mean μ and variance $\sigma^2, \mathcal{N}\left(\mu, \sigma^2\right)$:

$$f_X\left(x\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}.$$

Probability review: PDF of Gaussian distribution

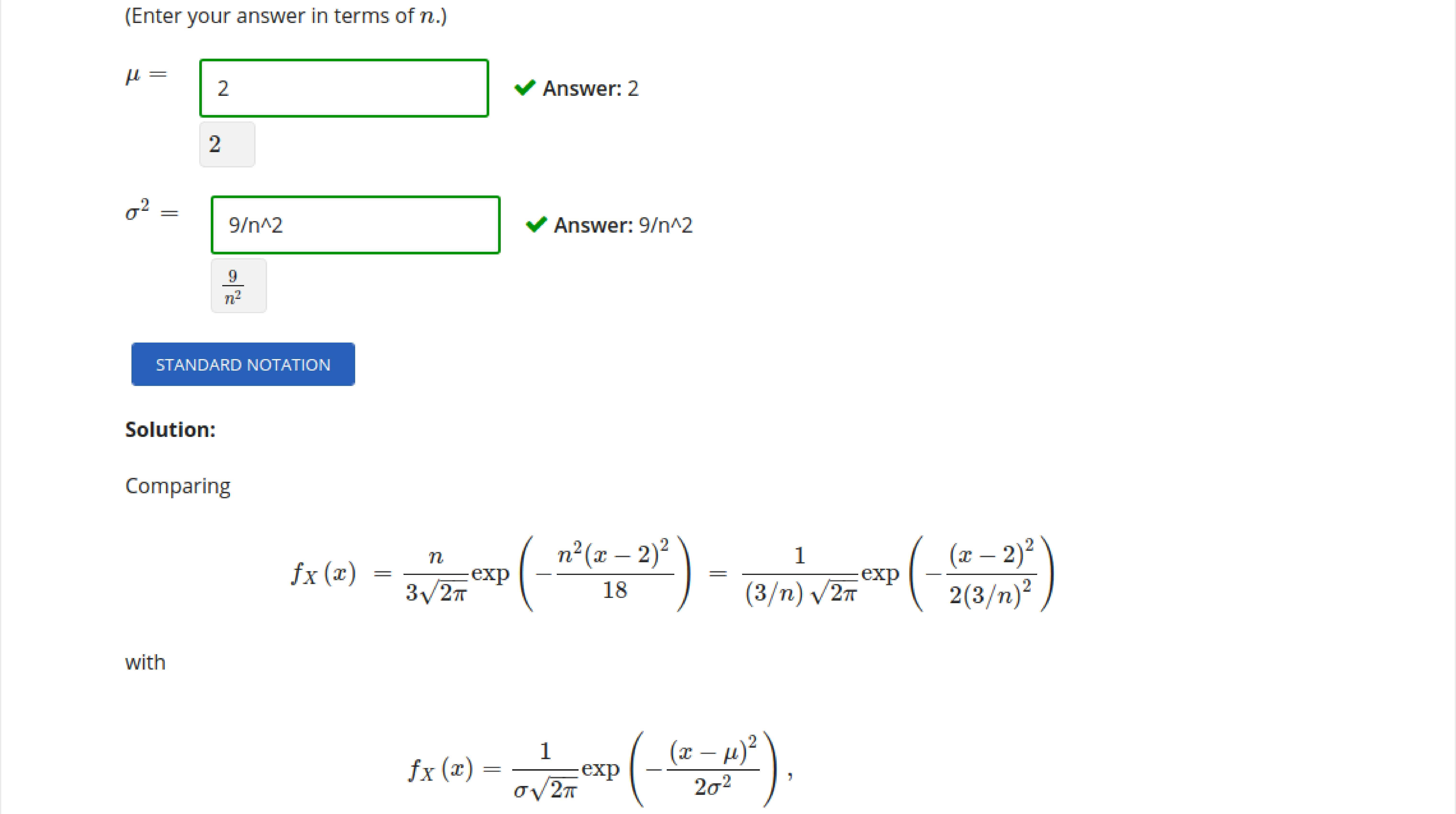
2/2 points (graded)

In practice, it is not often that you will need to work directly with the probability density function (pdf) of Gaussian variables. Nonetheless, we will make sure we know how to manipulate the (pdf) in the next two problems.

The pdf of a Gaussian random variable X is given by

$$f_X\left(x\right) = \frac{n}{3\sqrt{2\pi}} \exp\left(-\frac{n^2(x-2)^2}{18}\right),$$

then what is the mean μ and variance σ^2 of X?



(Enter your answer in terms of n.)

$\mu =$ 2

2

✓ **Answer:** 2

$\sigma^2 =$ 9/n^2

$\frac{9}{n^2}$

✓ **Answer:** 9/n^2

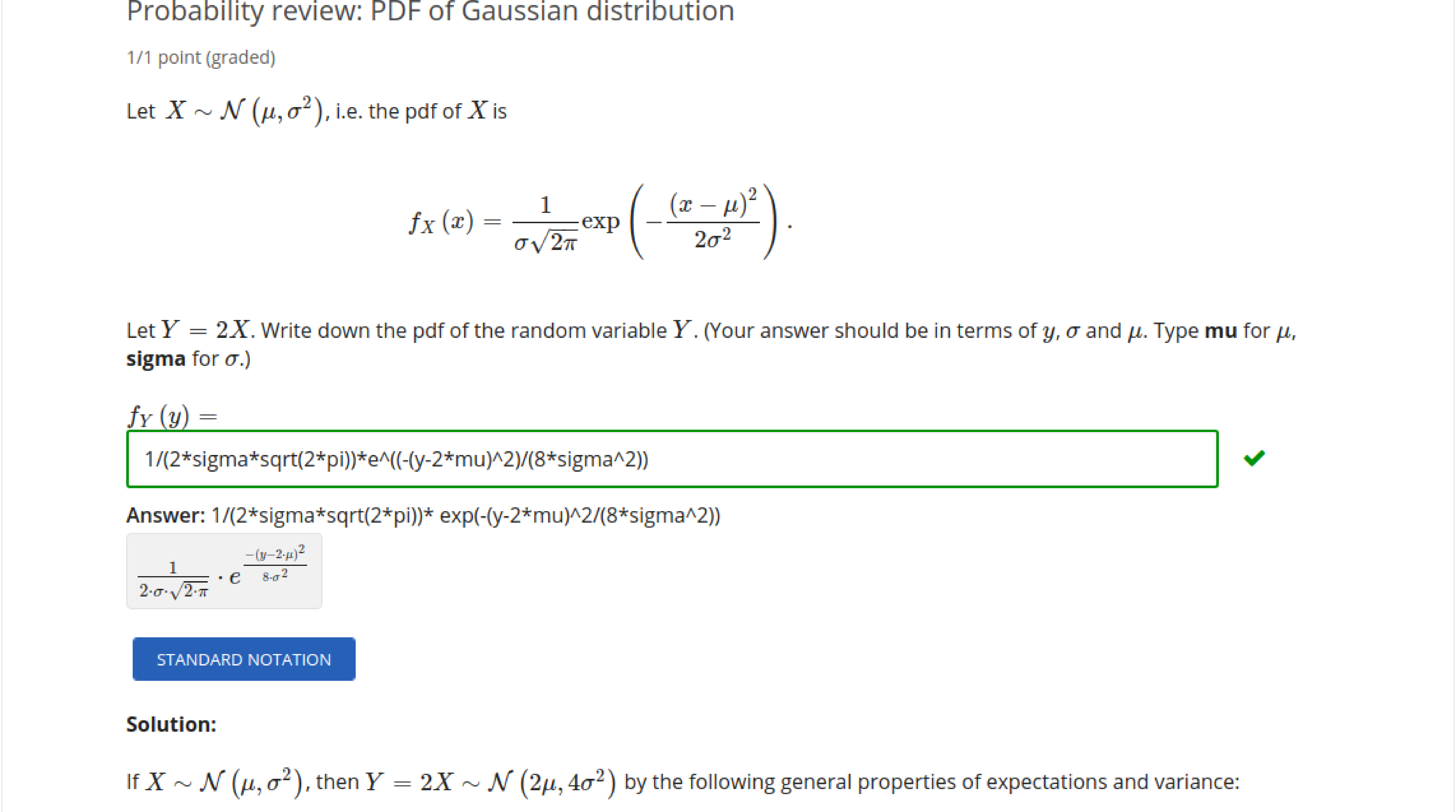
Solution:

Comparing

$$f_X(x) \;=\; \frac{n}{3\sqrt{2\pi}}\exp\left(-\frac{n^2(x-2)^2}{18}\right) \;=\; \frac{1}{(3/n)\sqrt{2\pi}}\exp\left(-\frac{(x-2)^2}{2(3/n)^2}\right)$$

with

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



Probability review: PDF of Gaussian distribution

1/1 point (graded)

Let $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, i.e. the pdf of X is

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

Let $Y = 2X$. Write down the pdf of the random variable Y. (Your answer should be in terms of y, σ and μ. Type **mu** for μ, **sigma** for σ.)

$f_Y\left(y\right) =$

1/(2*sigma*sqrt(2*pi))*e^((-(y-2*mu)^2)/(8*sigma^2)) ✓

Answer: 1/(2*sigma*sqrt(2*pi))* exp(-(y-2*mu)^2/(8*sigma^2))

$$\frac{1}{2\cdot\sigma\cdot\sqrt{2\cdot\pi}}\cdot e^{\frac{-(y-2\cdot\mu)^2}{8\cdot\sigma^2}}$$

STANDARD NOTATION

Solution:

If $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, then $Y = 2X \sim \mathcal{N}\left(2\mu, 4\sigma^2\right)$ by the following general properties of expectations and variance:

$$
\begin{aligned}
\mathbf{E}\left[2X\right] &= 2\mathbf{E}\left[X\right] \\
\mathsf{Var}\left[2X\right] &= 2^2\mathsf{Var}\left[X\right] = 4\mathsf{Var}\left[X\right].
\end{aligned}
$$

Therefore,

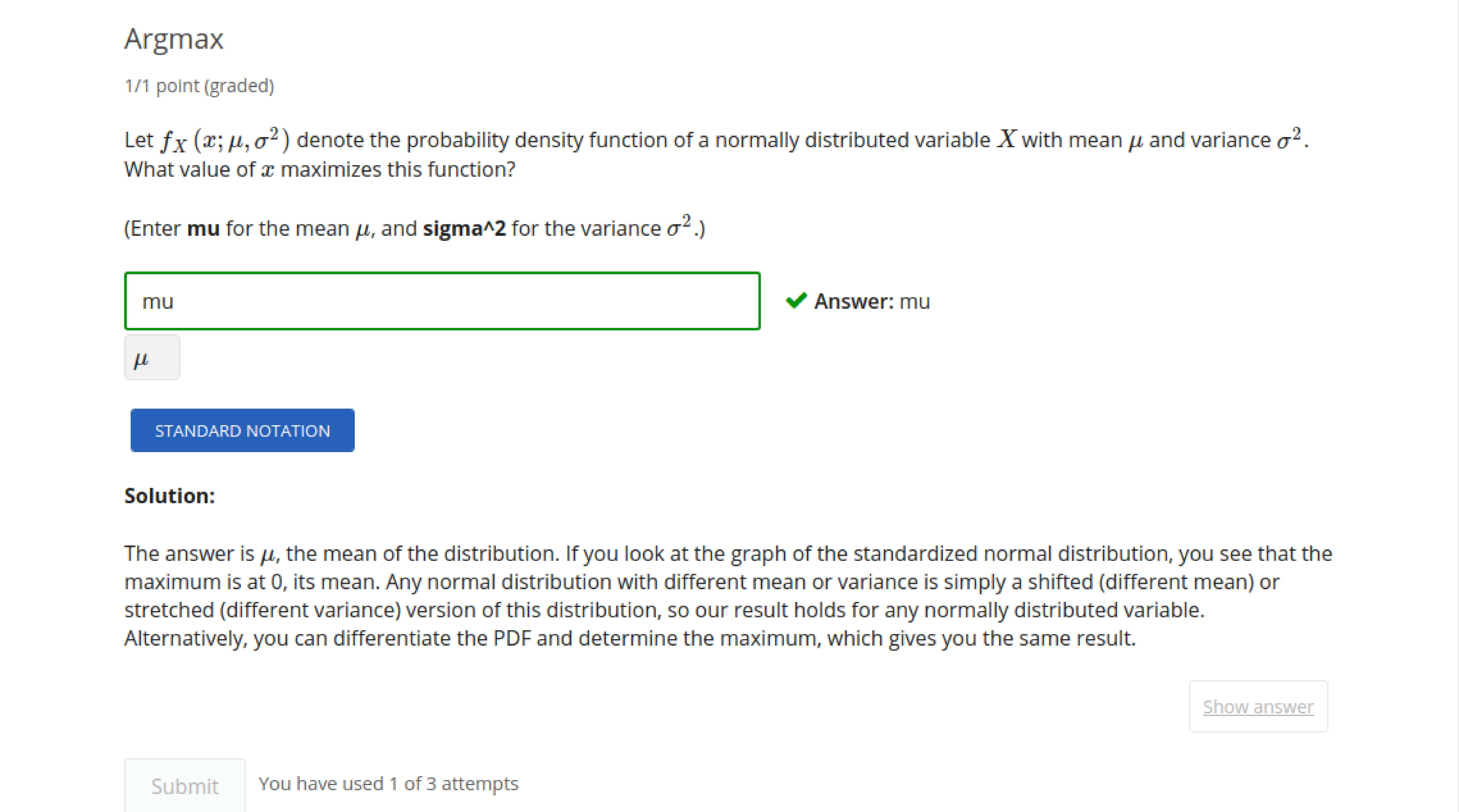
$$
f_Y\left(y\right) = \frac{1}{2\sigma\sqrt{2\pi}}\exp\left(-\frac{\left(y-2\mu\right)^2}{2\left(4\sigma^2\right)}\right).
$$

Alternate solution: In general, for any continuous random variables X and any continuous monotonous (i.e. always increasing or always decreasing) function g, such that $Y = g\left(X\right)$, the pdf of Y is given by:

$$
f_Y\left(y\right) \;=\; \frac{f_X\left(x\right)}{\left|g'\left(x\right)\right|} \qquad \text{where } x = g^{-1}\left(y\right).
$$

In this problem, $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, $Y = g\left(X\right) = 2X$, and $g'\left(x\right) = 2$. Therefore:

$$
\begin{aligned}
f_Y\left(y\right) &= \frac{f_X\left(\frac{y}{2}\right)}{\left|g'\left(\frac{y}{2}\right)\right|} \\
&= \frac{1}{g'\left(y/2\right)\sigma\sqrt{2\pi}}\exp\left(-\frac{\left(y/2-\mu\right)^2}{2\sigma^2}\right) \\
&= \frac{1}{2\sigma\sqrt{2\pi}}\exp\left(-\frac{\left(\left(y-2\mu\right)/2\right)^2}{2\sigma^2}\right)
\end{aligned}
$$



Argmax

1/1 point (graded)

Let $f_X\left(x; \mu, \sigma^2\right)$ denote the probability density function of a normally distributed variable X with mean μ and variance σ^2. What value of x maximizes this function?

(Enter **mu** for the mean μ, and **sigma^2** for the variance σ^2.)

mu

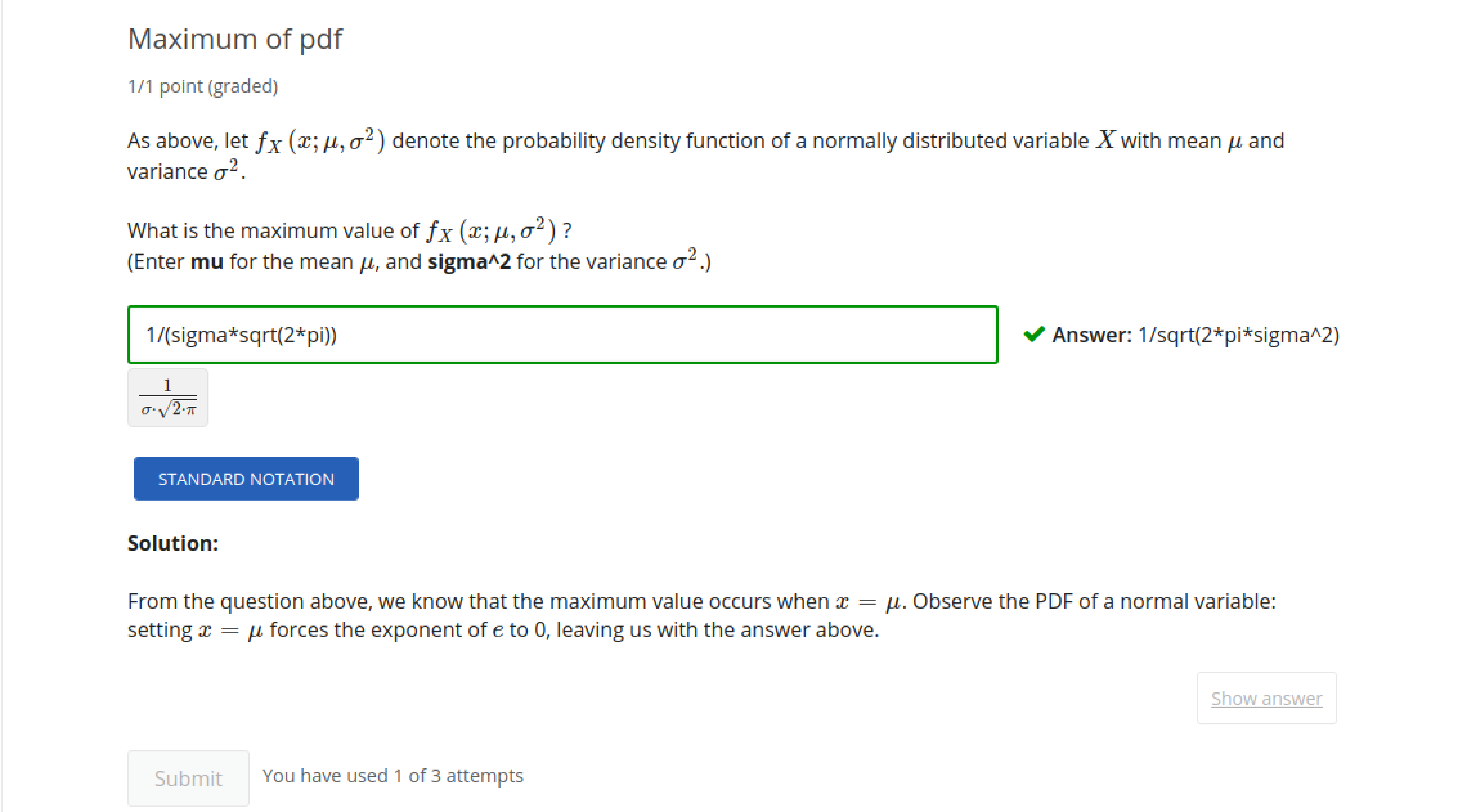
✓ **Answer:** mu

μ

Solution:

The answer is μ, the mean of the distribution. If you look at the graph of the standardized normal distribution, you see that the maximum is at 0, its mean. Any normal distribution with different mean or variance is simply a shifted (different mean) or stretched (different variance) version of this distribution, so our result holds for any normally distributed variable. Alternatively, you can differentiate the PDF and determine the maximum, which gives you the same result.

You have used 1 of 3 attempts



Maximum of pdf

1/1 point (graded)

As above, let $f_X\left(x;\mu,\sigma^2\right)$ denote the probability density function of a normally distributed variable X with mean μ and variance σ^2.

What is the maximum value of $f_X\left(x;\mu,\sigma^2\right)$?
(Enter **mu** for the mean μ, and **sigma^2** for the variance σ^2.)

1/(sigma*sqrt(2*pi))

✓ **Answer:** 1/sqrt(2*pi*sigma^2)

$$\frac{1}{\sigma\cdot\sqrt{2\cdot\pi}}$$

Solution:

From the question above, we know that the maximum value occurs when $x=\mu$. Observe the PDF of a normal variable: setting $x=\mu$ forces the exponent of e to 0, leaving us with the answer above.

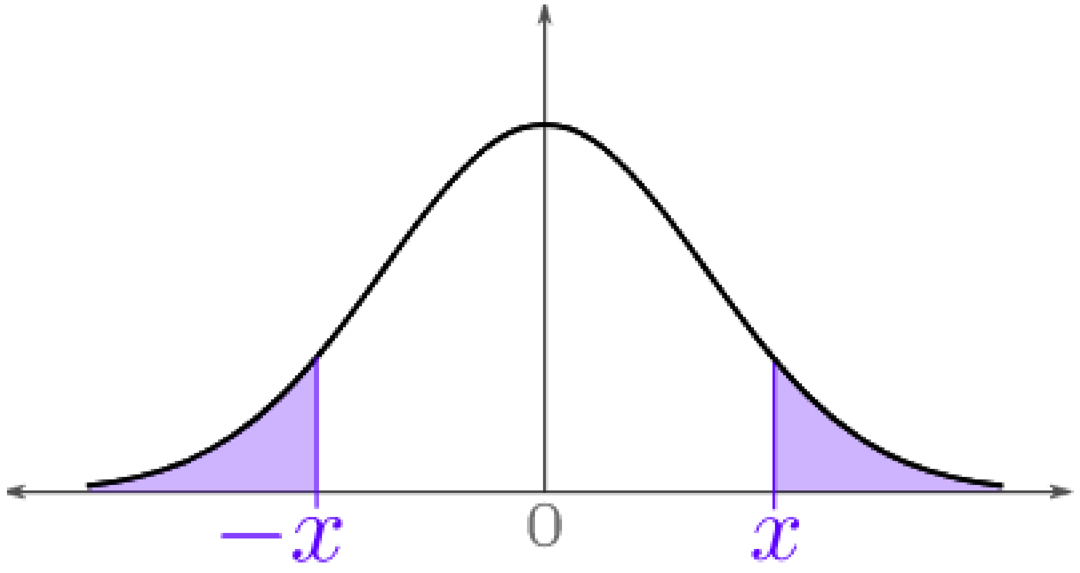
You have used 1 of 3 attempts

Quantiles

0/1 point (graded)

The **quantile** of order $1-\alpha$ of a variable X, denoted by q_α (specific to a particular X), is the number such that $\mathbf{P}\left(X \le q_\alpha\right) = 1-\alpha$.

Graphed below is the pdf of the normal distribution with generic/unknown (but fixed) variance σ^2. If the total area of the two shaded regions is 0.03, then what is x?
(Choose all that apply.)



The total area of the two shaded regions is 0.03.

□ $\mathbf{P}\left(|X| \le 0.03\right)$

☐ $\mathbf{P}\left(|X| \leq 0.015\right)$

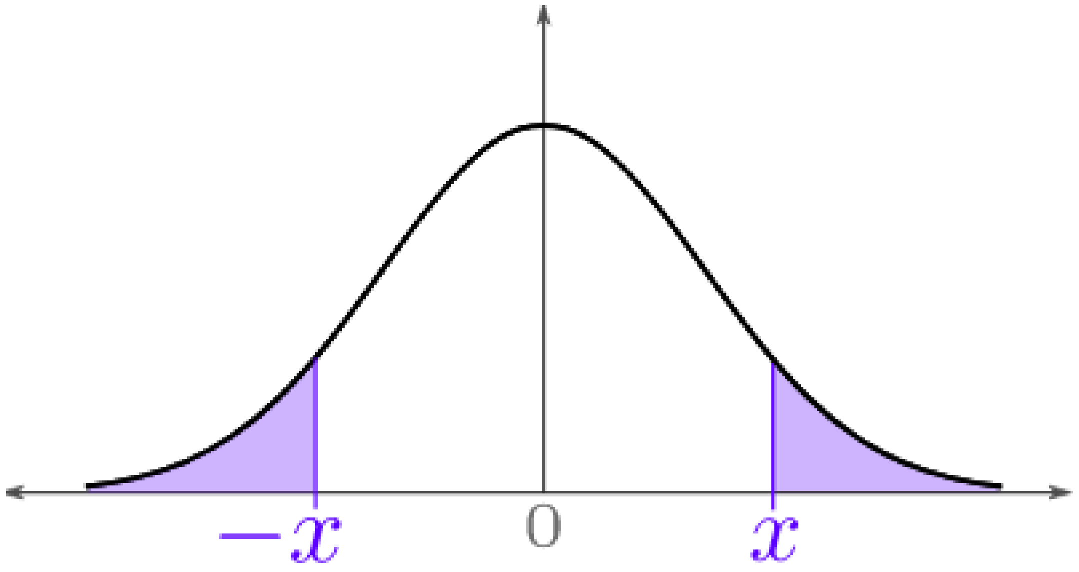
☐ 0.97

☑ 0.985

☐ $q_{0.03}$

☑ $q_{0.015}$

Solution:



The total area of the two shaded regions equals $\mathbf{P}\left(|X| \geq x\right) = 0.03$. By symmetry, the probability in the positive tail is $\mathbf{P}\left(X \geq x\right) = 0.015$; hence $x = q_{\alpha}$ with $\alpha = 0.015$.

For the wrong choices:

- The first pair of choices mixed up the values of probability with the value of the variable.
- The third and fourth choices “0.97" and “0.985" are meant to play the role resembling $1-\alpha$ in this example, but these are wrong for the same reasons as the first pair of choices. In any case, to give a particular numerical value of x, the answer must depend on σ.
- The fifth choice would have been correct again if the area of one of the tails is 0.03.

Show answer

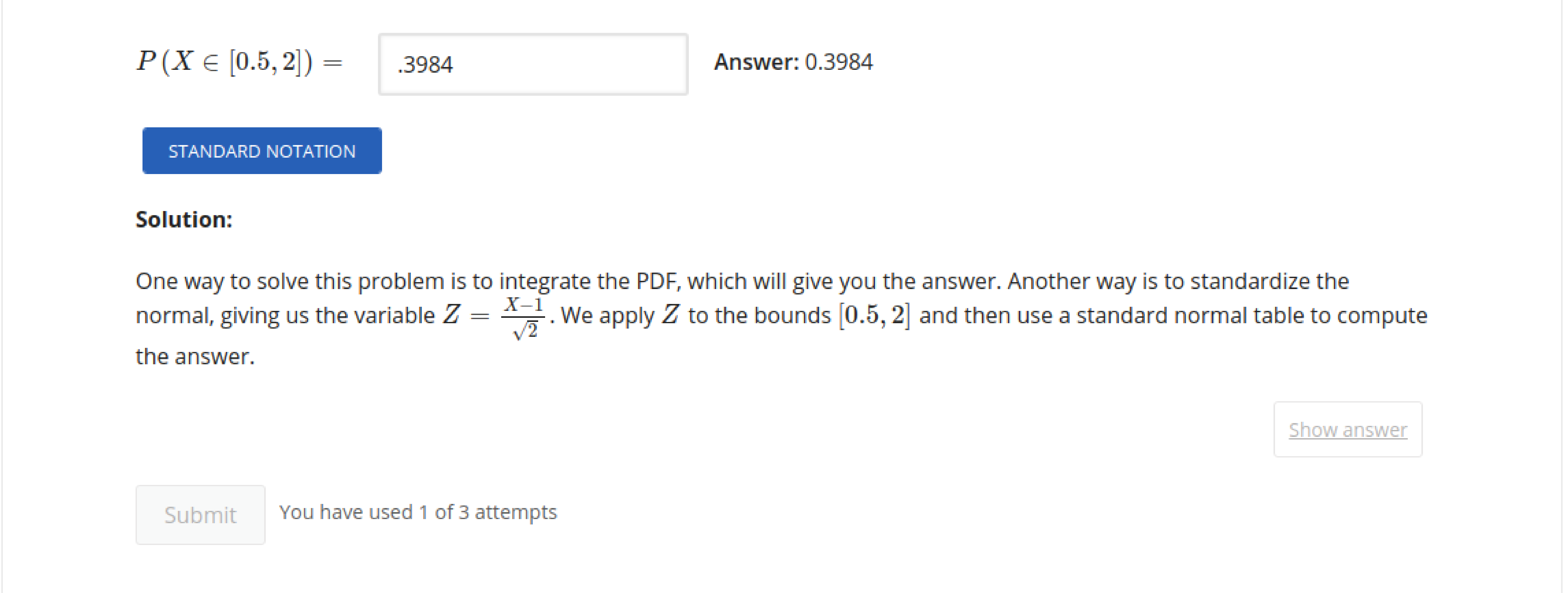
Submit You have used 2 of 2 attempts

Probability

0/1 point (graded)

Let $X \sim \mathcal{N}\left(1,2\right)$, i.e., the random variable X is normally distributed with mean 1 and variance 2. What is the probability that $X \in [0.5, 2]$?

(Enter your answer accurate to at least 4 decimal places.)

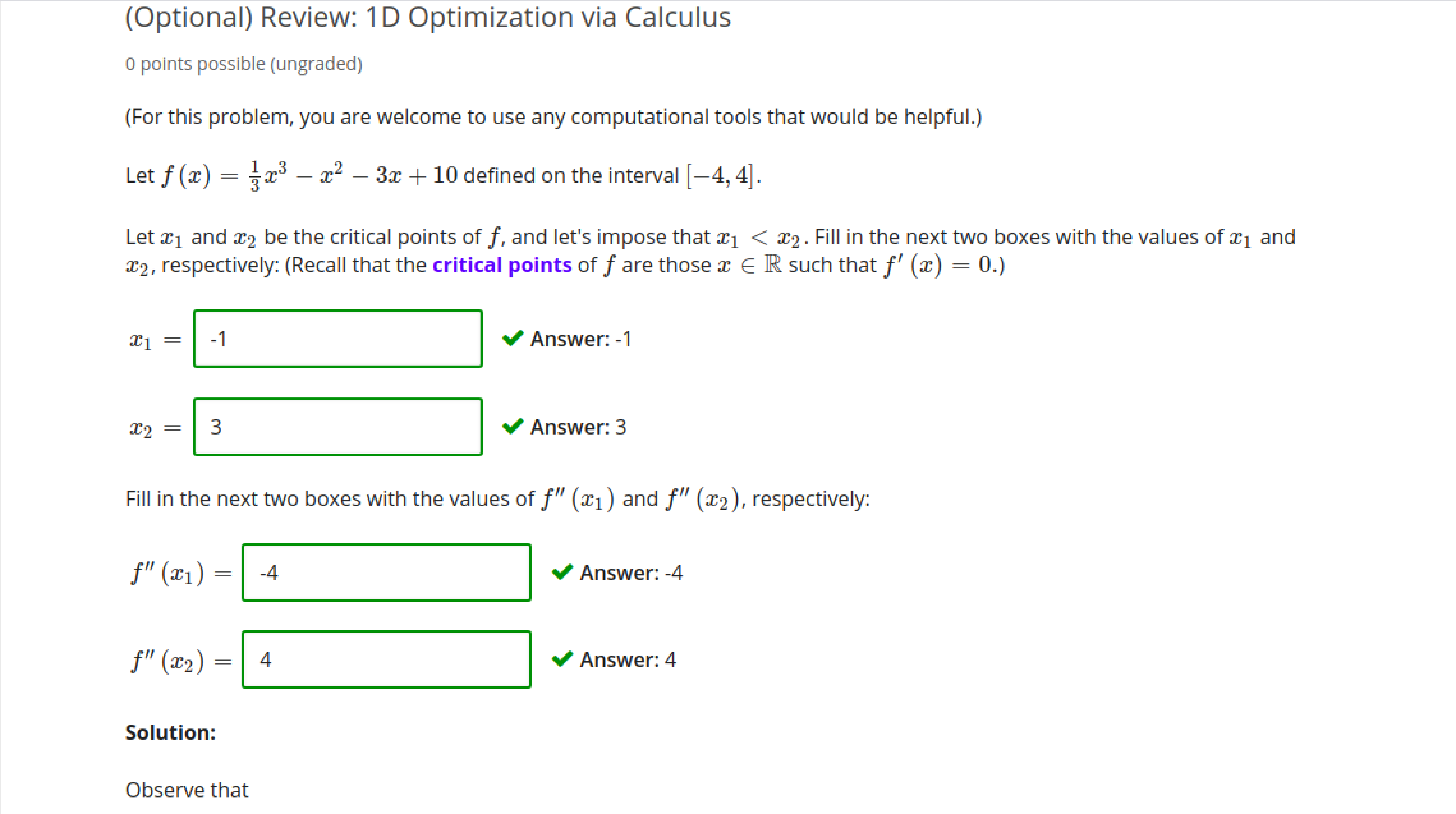


$P\left(X \in [0.5, 2]\right) =$ .3984 **Answer:** 0.3984

Solution:

One way to solve this problem is to integrate the PDF, which will give you the answer. Another way is to standardize the normal, giving us the variable $Z = \frac{X-1}{\sqrt{2}}$. We apply Z to the bounds $[0.5, 2]$ and then use a standard normal table to compute the answer.

You have used 1 of 3 attempts



(Optional) Review: 1D Optimization via Calculus

0 points possible (ungraded)

(For this problem, you are welcome to use any computational tools that would be helpful.)

Let $f(x) = \frac{1}{3}x^3 - x^2 - 3x + 10$ defined on the interval $[-4, 4]$.

Let x_1 and x_2 be the critical points of f, and let's impose that $x_1 < x_2$. Fill in the next two boxes with the values of x_1 and x_2, respectively: (Recall that the **critical points** of f are those $x \in \mathbb{R}$ such that $f'(x) = 0$.)

$x_1 =$ -1 ✓ **Answer:** -1

$x_2 =$ 3 ✓ **Answer:** 3

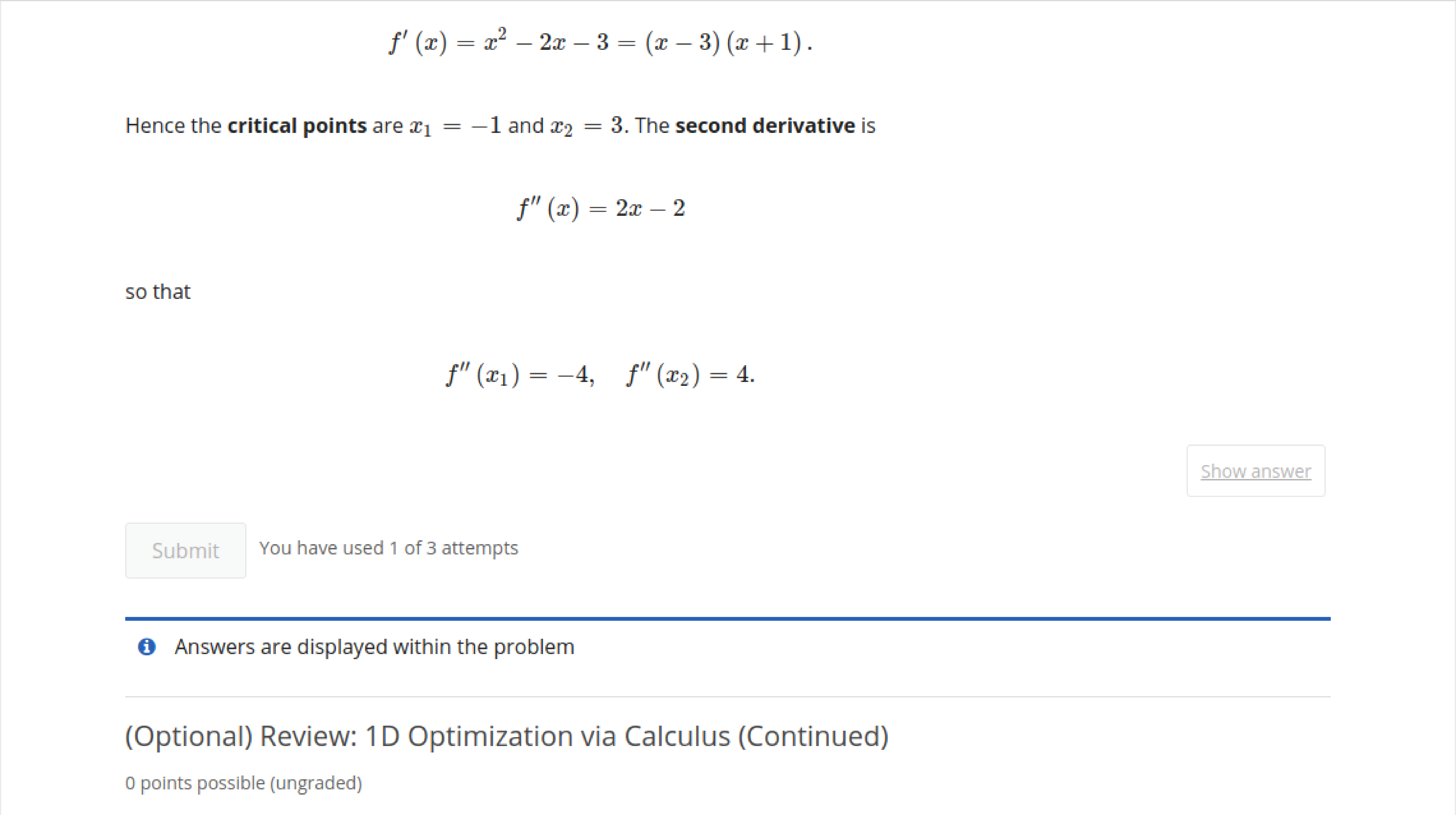
Fill in the next two boxes with the values of $f''(x_1)$ and $f''(x_2)$, respectively:

$f''(x_1) =$ -4 ✓ **Answer:** -4

$f''(x_2) =$ 4 ✓ **Answer:** 4

Solution:

Observe that



$$f'(x) = x^2 - 2x - 3 = (x-3)(x+1)\,.$$

Hence the **critical points** are $x_1 = -1$ and $x_2 = 3$. The **second derivative** is

$$f''(x) = 2x - 2$$

so that

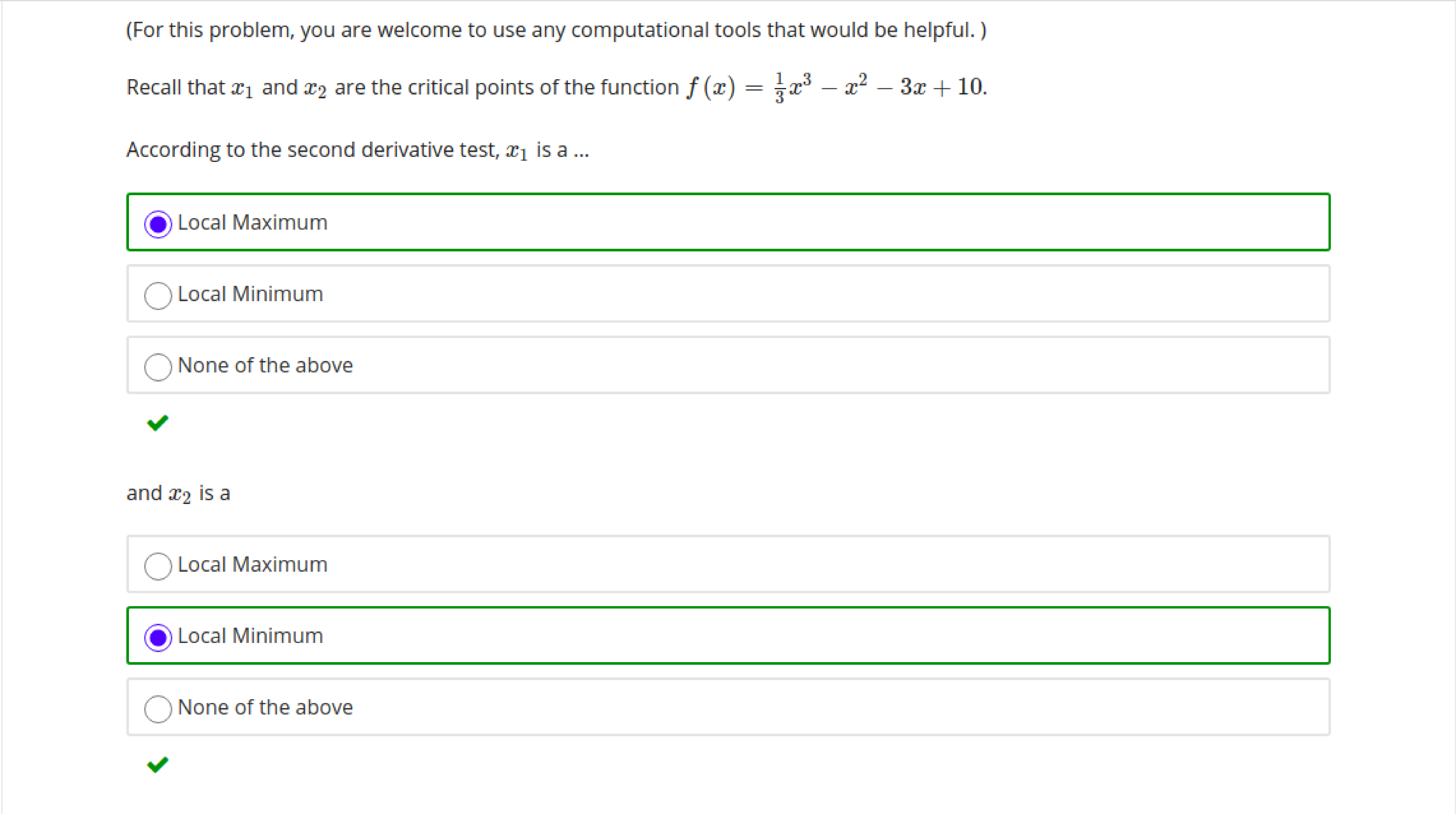
$$f''(x_1) = -4, \quad f''(x_2) = 4.$$

You have used 1 of 3 attempts

Answers are displayed within the problem

(Optional) Review: 1D Optimization via Calculus (Continued)

0 points possible (ungraded)



(For this problem, you are welcome to use any computational tools that would be helpful.)

Recall that x_1 and x_2 are the critical points of the function $f\left(x\right) = \frac{1}{3}x^3 - x^2 - 3x + 10.$

According to the second derivative test, x_1 is a ...

● Local Maximum

○ Local Minimum

○ None of the above

✓

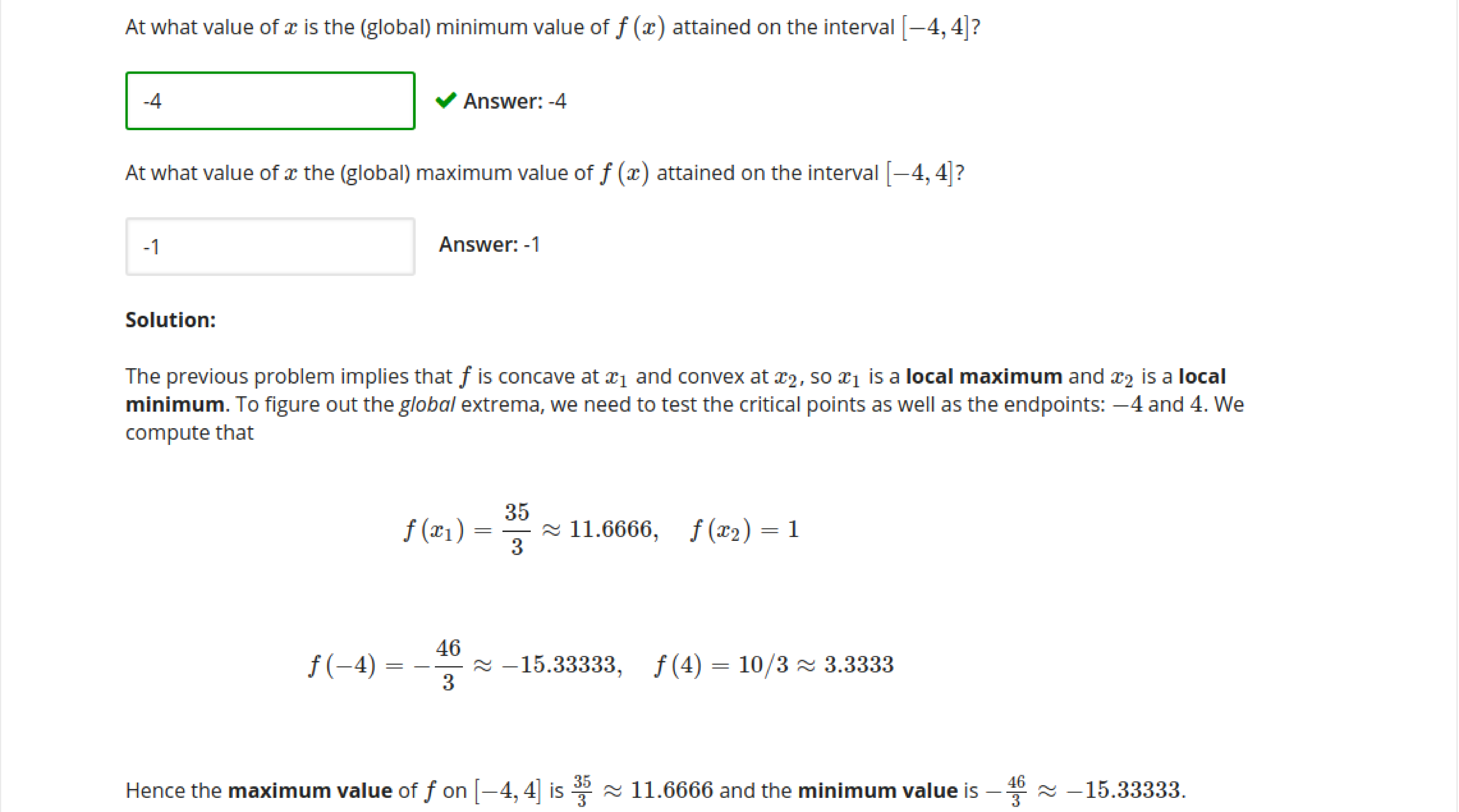
and x_2 is a

○ Local Maximum

● Local Minimum

○ None of the above

✓



At what value of x is the (global) minimum value of $f(x)$ attained on the interval $[-4, 4]$?

-4

✓ **Answer:** -4

At what value of x the (global) maximum value of $f(x)$ attained on the interval $[-4, 4]$?

-1

Answer: -1

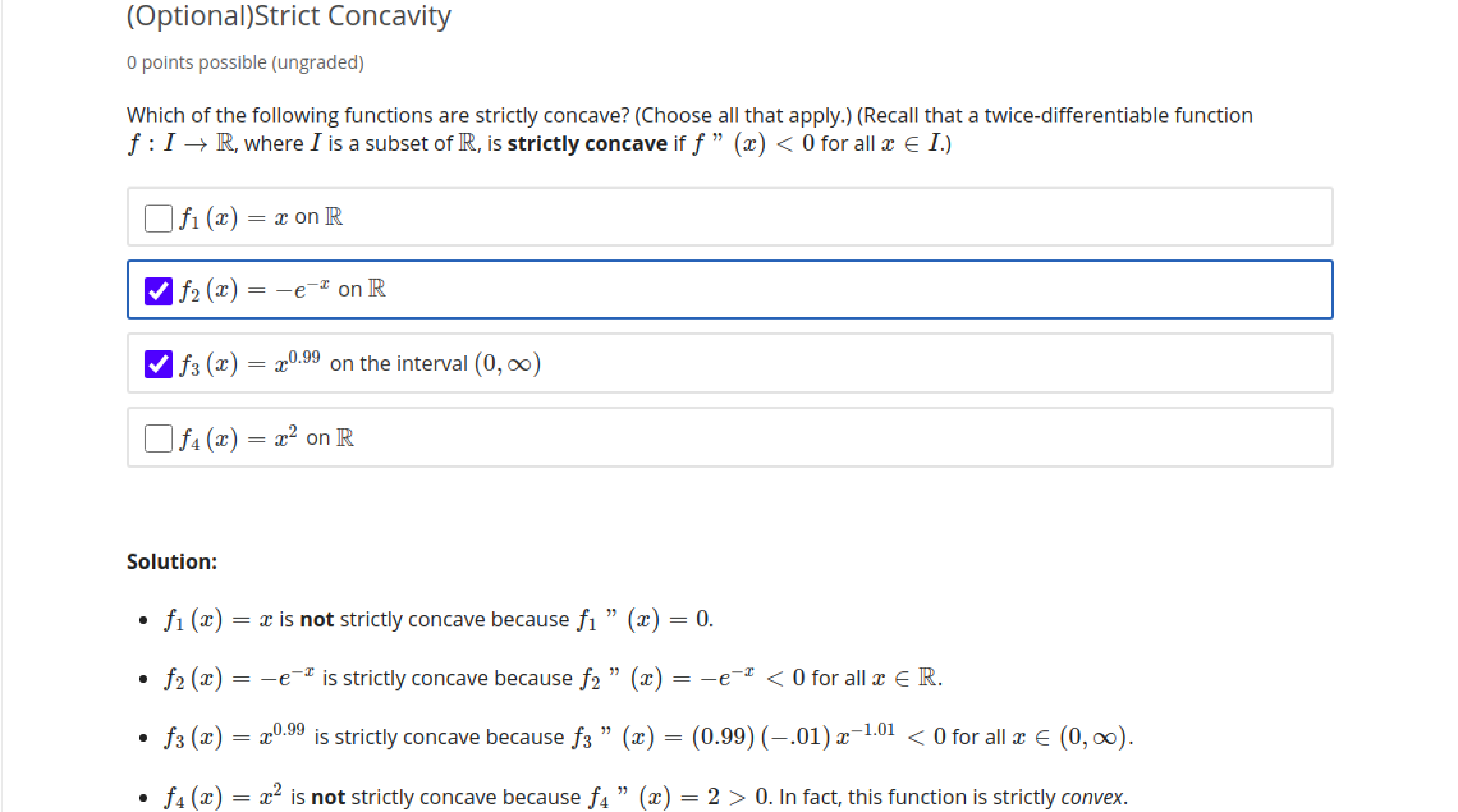
Solution:

The previous problem implies that f is concave at x_1 and convex at x_2, so x_1 is a **local maximum** and x_2 is a **local minimum**. To figure out the *global* extrema, we need to test the critical points as well as the endpoints: -4 and 4. We compute that

$$f(x_1) = \frac{35}{3} \approx 11.6666, \quad f(x_2) = 1$$

$$f(-4) = -\frac{46}{3} \approx -15.33333, \quad f(4) = 10/3 \approx 3.3333$$

Hence the **maximum value** of f on $[-4, 4]$ is $\frac{35}{3} \approx 11.6666$ and the **minimum value** is $-\frac{46}{3} \approx -15.33333$.



(Optional)Strict Concavity

0 points possible (ungraded)

Which of the following functions are strictly concave? (Choose all that apply.) (Recall that a twice-differentiable function $f : I \to \mathbb{R}$, where I is a subset of $\mathbb{R}$, is **strictly concave** if $f\text{ ”}\left(x\right) < 0$ for all $x \in I$.)

□ $f_1\left(x\right) = x$ on $\mathbb{R}$

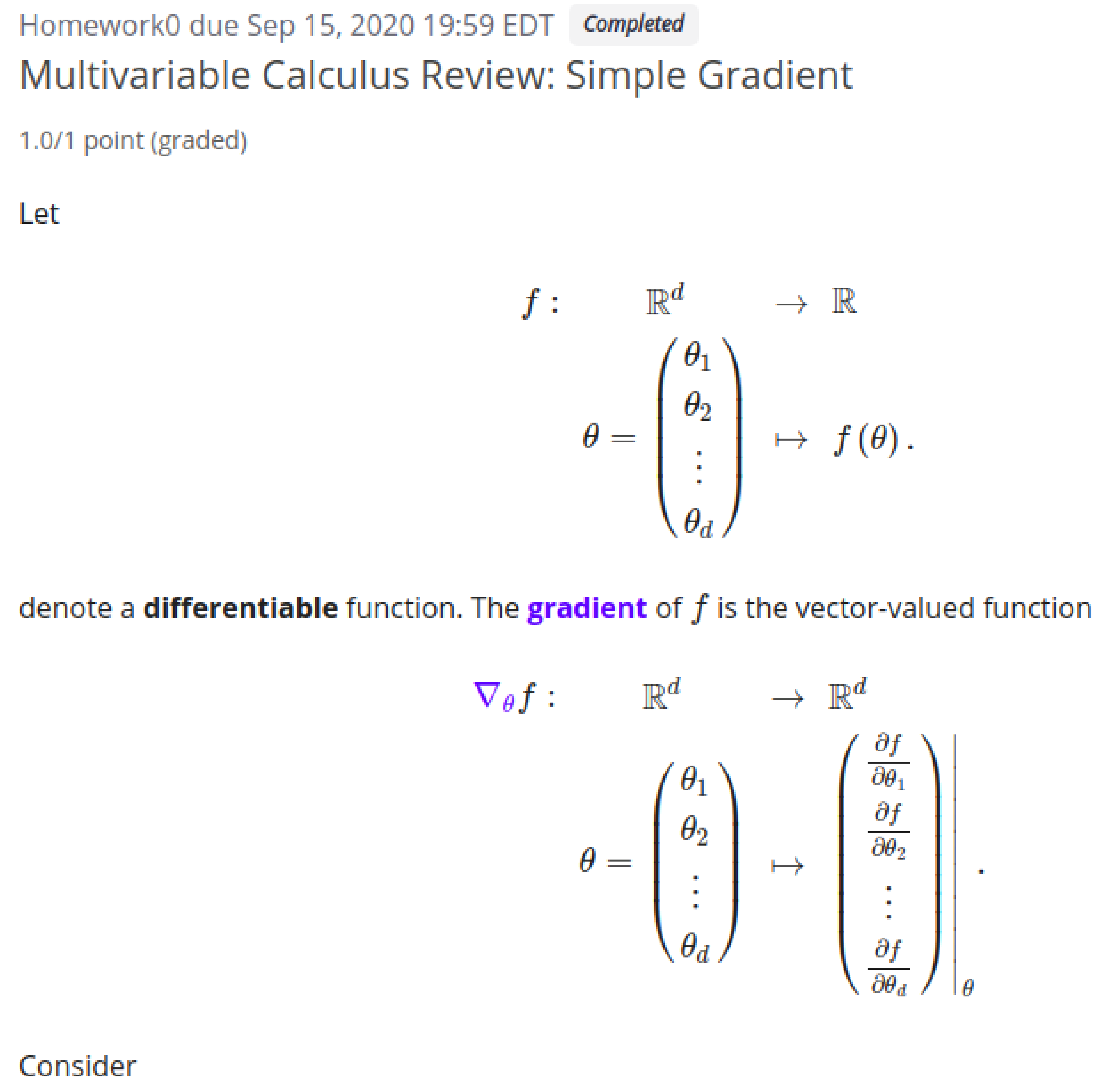
☑ $f_2\left(x\right) = -e^{-x}$ on $\mathbb{R}$

☑ $f_3\left(x\right) = x^{0.99}$ on the interval $\left(0, \infty\right)$

□ $f_4\left(x\right) = x^2$ on $\mathbb{R}$

Solution:

- $f_1\left(x\right) = x$ is **not** strictly concave because $f_1\text{ ”}\left(x\right) = 0$.
- $f_2\left(x\right) = -e^{-x}$ is strictly concave because $f_2\text{ ”}\left(x\right) = -e^{-x} < 0$ for all $x \in \mathbb{R}$.
- $f_3\left(x\right) = x^{0.99}$ is strictly concave because $f_3\text{ ”}\left(x\right) = \left(0.99\right)\left(-.01\right)x^{-1.01} < 0$ for all $x \in \left(0, \infty\right)$.
- $f_4\left(x\right) = x^2$ is **not** strictly concave because $f_4\text{ ”}\left(x\right) = 2 > 0$. In fact, this function is strictly *convex*.



Homework0 due Sep 15, 2020 19:59 EDT Completed

Multivariable Calculus Review: Simple Gradient

1.0/1 point (graded)

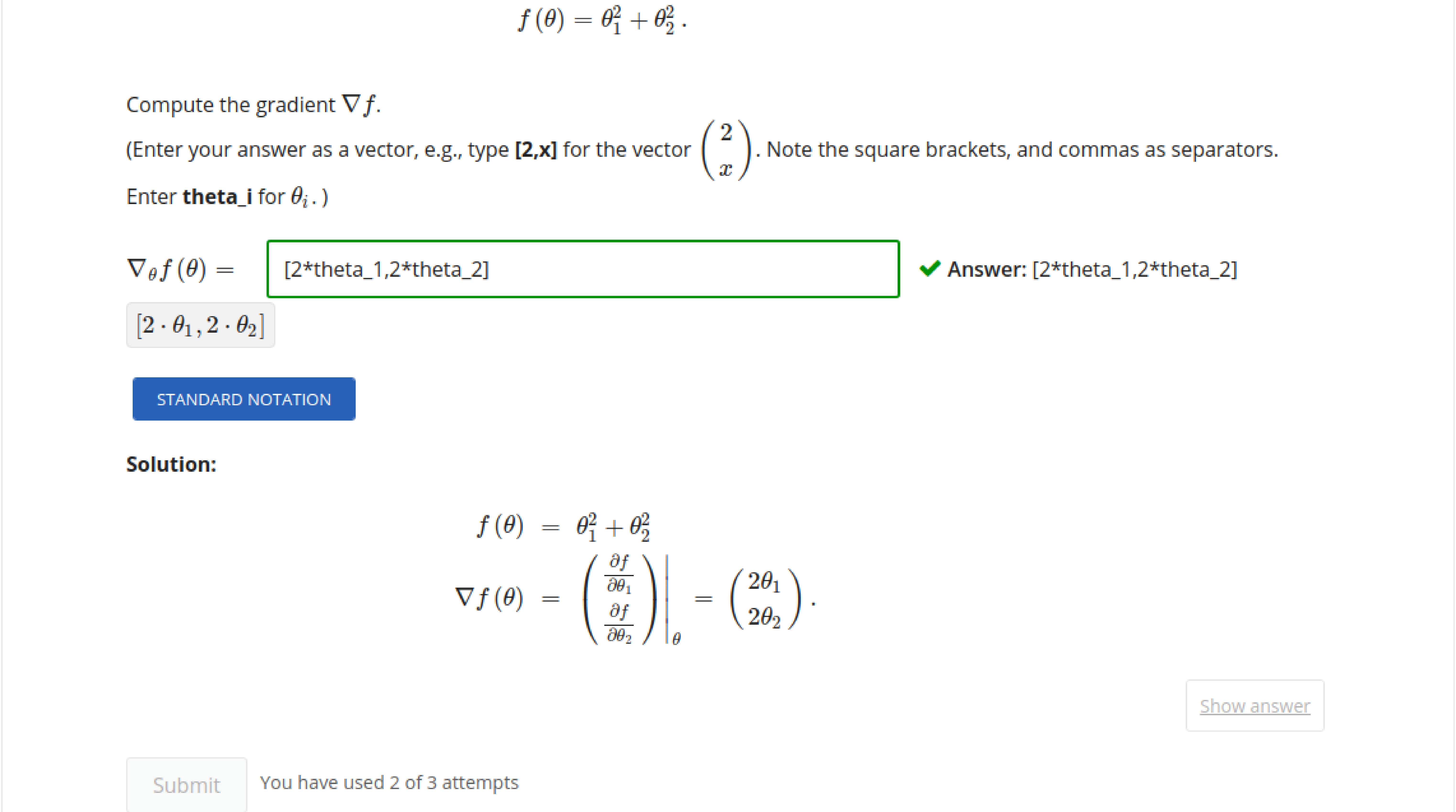
Let

$$\begin{aligned} f: \quad & \mathbb{R}^d && \to \mathbb{R} \\ \theta = & \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{pmatrix} && \mapsto f(\theta). \end{aligned}$$

denote a **differentiable** function. The **gradient** of f is the vector-valued function

$$\begin{aligned} \nabla_\theta f: \quad & \mathbb{R}^d && \to \mathbb{R}^d \\ \theta = & \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{pmatrix} && \mapsto \left.\begin{pmatrix} \frac{\partial f}{\partial \theta_1} \\ \frac{\partial f}{\partial \theta_2} \\ \vdots \\ \frac{\partial f}{\partial \theta_d} \end{pmatrix}\right|_\theta . \end{aligned}$$

Consider



$$f(\theta) = \theta_1^2 + \theta_2^2 \,.$$

Compute the gradient ∇f.

(Enter your answer as a vector, e.g., type **[2,x]** for the vector $\begin{pmatrix} 2 \\ x \end{pmatrix}$. Note the square brackets, and commas as separators.

Enter **theta_i** for θ_i.)

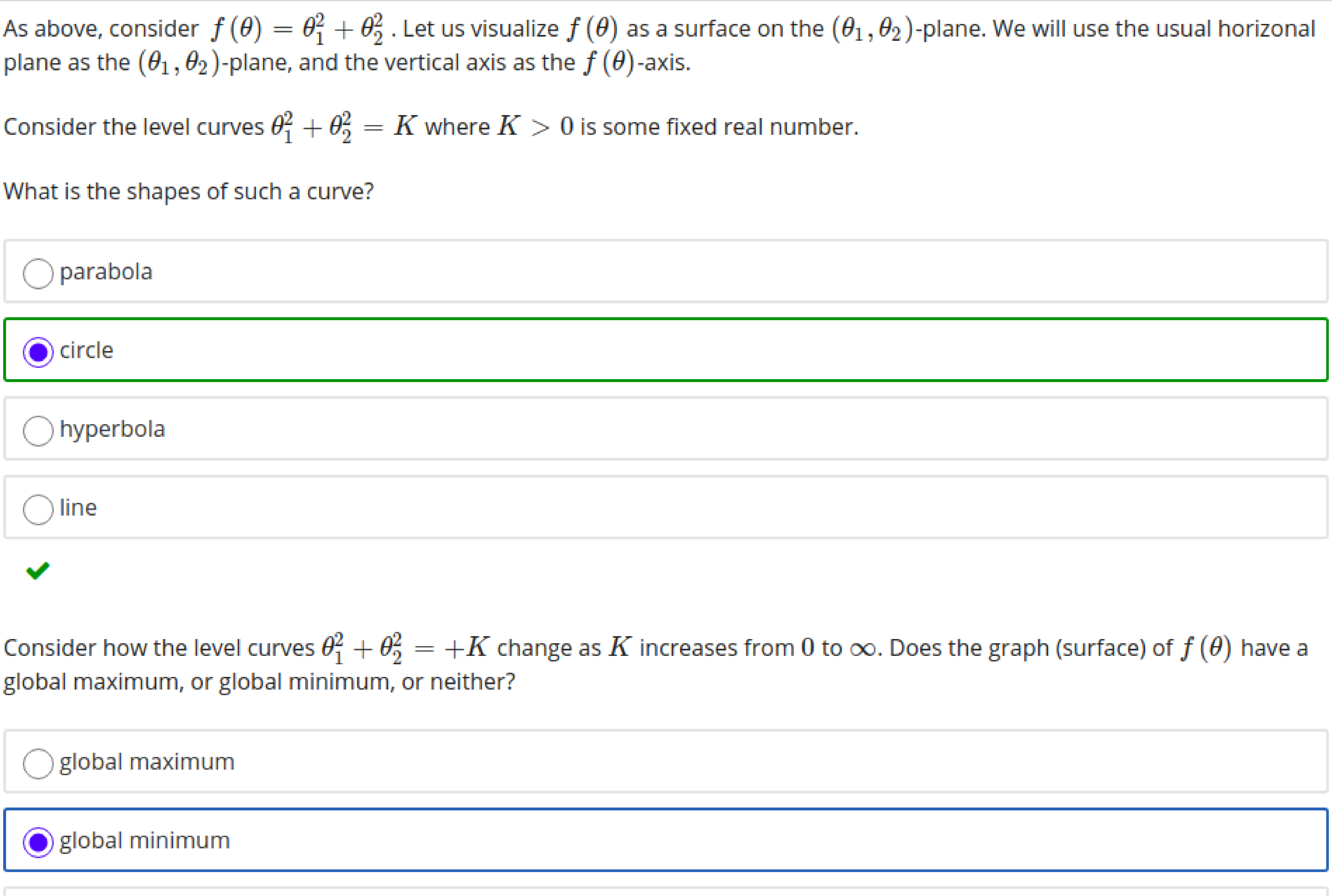
$\nabla_\theta f(\theta) =$ [2*theta_1,2*theta_2] ✓ **Answer:** [2*theta_1,2*theta_2]

$[2 \cdot \theta_1 , 2 \cdot \theta_2]$

Solution:

$$\begin{aligned} f(\theta) &= \theta_1^2 + \theta_2^2 \\ \nabla f(\theta) &= \left. \begin{pmatrix} \frac{\partial f}{\partial \theta_1} \\ \frac{\partial f}{\partial \theta_2} \end{pmatrix} \right|_\theta = \begin{pmatrix} 2\theta_1 \\ 2\theta_2 \end{pmatrix} . \end{aligned}$$

You have used 2 of 3 attempts



As above, consider $f(\theta) = \theta_1^2 + \theta_2^2$. Let us visualize $f(\theta)$ as a surface on the (θ_1, θ_2)-plane. We will use the usual horizonal plane as the (θ_1, θ_2)-plane, and the vertical axis as the $f(\theta)$-axis.

Consider the level curves $\theta_1^2 + \theta_2^2 = K$ where $K > 0$ is some fixed real number.

What is the shapes of such a curve?

○ parabola

● circle

○ hyperbola

○ line

✓

Consider how the level curves $\theta_1^2 + \theta_2^2 = +K$ change as K increases from 0 to ∞. Does the graph (surface) of $f(\theta)$ have a global maximum, or global minimum, or neither?

○ global maximum

● global minimum

At each point $\theta = (\theta_1, \theta_2)$ in the (θ_1, θ_2)-plane, $f(\theta)$ decreases in the direction of...

○ $\nabla_\theta f(\theta)$

● $-\nabla_\theta f(\theta)$

✓

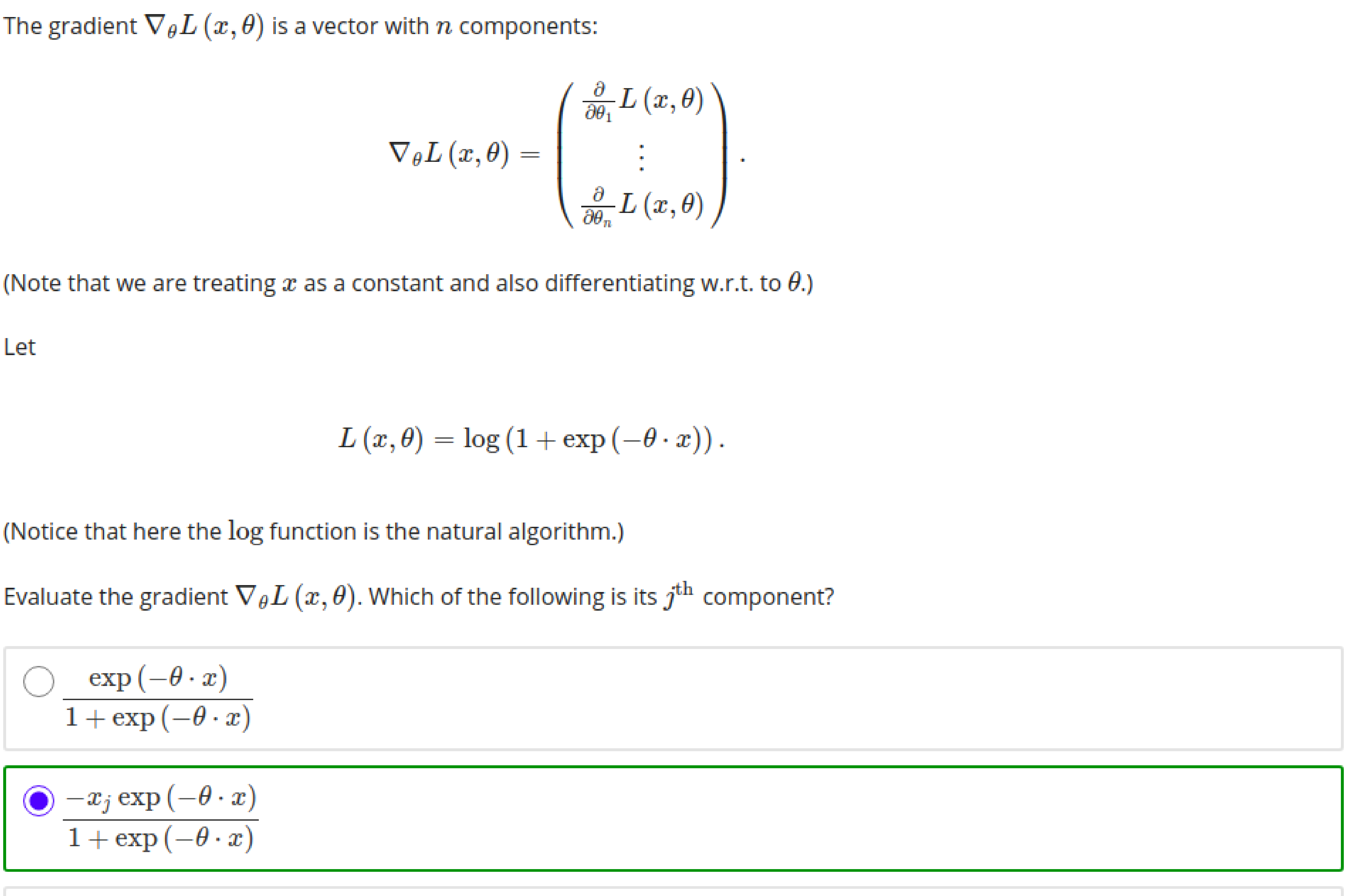
Solution:

The graph of $f(\theta)$ is a paraboloid that opens downwards. Its global maximum is at $\theta = (0, 0)$. We see that $\nabla_\theta f(\theta) = (2\theta_1, 2\theta_2)^T$, and hence $-\nabla_\theta f(\theta)$ points towards the origin at all points θ.

Show answer

Submit You have used 2 of 2 attempts

Gradient ascent/descent methods are typical tools for maximizing/minimizing functions. Consider the function $L(x, \theta)$ where $\theta = [\theta_1, \theta_2, \ldots, \theta_n]^T$ and $x = [x_1, x_2, \ldots, x_n]^T$. Our goal is to select θ such to maximize/minimize the value of L while keeping x fixed.



The gradient $\nabla_\theta L\left(x,\theta\right)$ is a vector with n components:

$$\nabla_\theta L\left(x,\theta\right) = \begin{pmatrix} \frac{\partial}{\partial \theta_1} L\left(x,\theta\right) \\ \vdots \\ \frac{\partial}{\partial \theta_n} L\left(x,\theta\right) \end{pmatrix}.$$

(Note that we are treating x as a constant and also differentiating w.r.t. to θ.)

Let

$$L\left(x,\theta\right) = \log\left(1 + \exp\left(-\theta \cdot x\right)\right).$$

(Notice that here the $\log$ function is the natural algorithm.)

Evaluate the gradient $\nabla_\theta L\left(x,\theta\right)$. Which of the following is its j^{th} component?

○ $\dfrac{\exp\left(-\theta \cdot x\right)}{1 + \exp\left(-\theta \cdot x\right)}$

● $\dfrac{-x_j \exp\left(-\theta \cdot x\right)}{1 + \exp\left(-\theta \cdot x\right)}$

○ $\dfrac{-x_j}{1+\exp(-\theta\cdot x)}$

✓

Solution:

The derivative of $\log(x) = \frac{1}{x}$ and the derivative of $e^{cx} = ce^{cx}$. Applying these rules with the chain rule gives the correct answer.

Show answer

Submit You have used 1 of 1 attempt

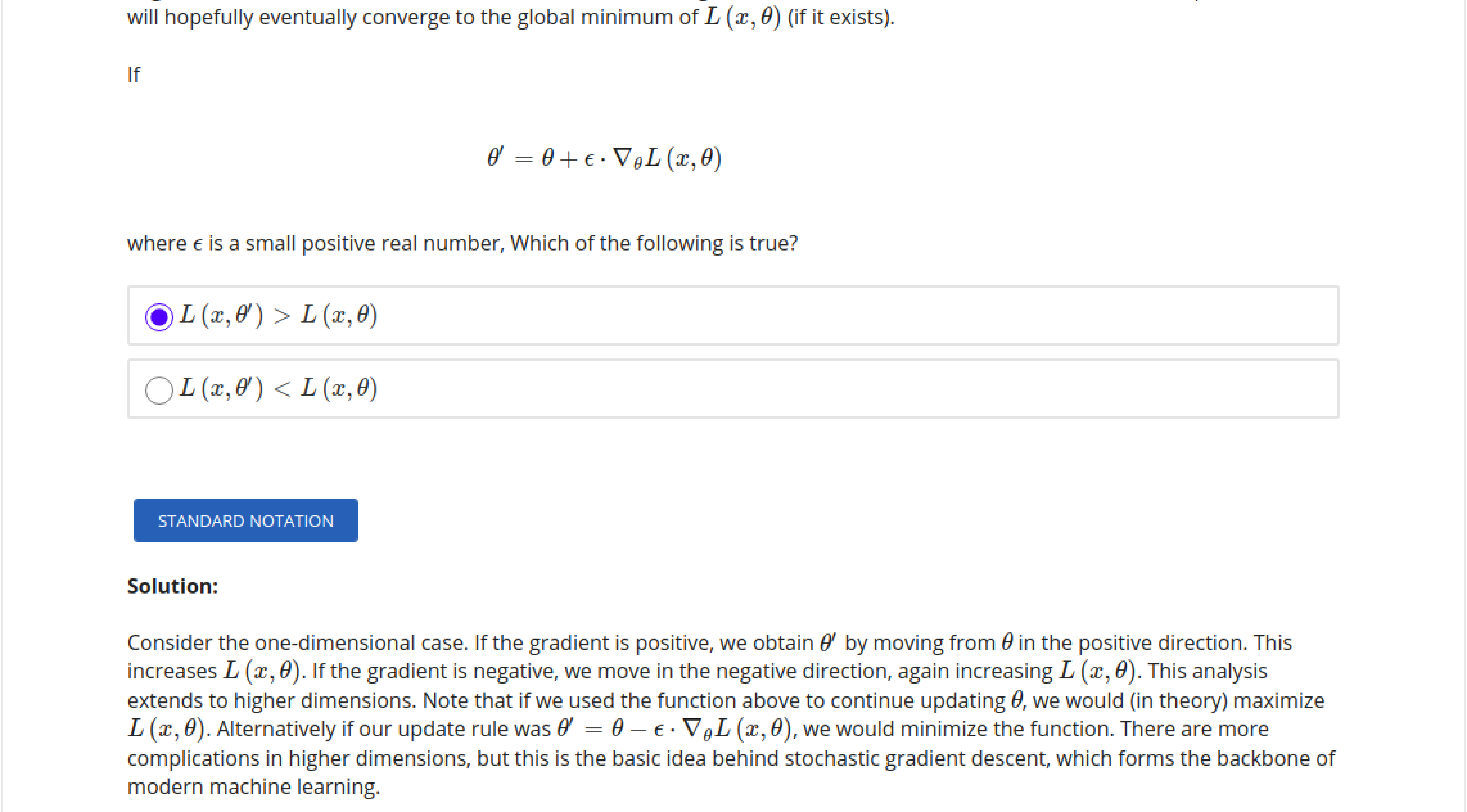
ℹ Answers are displayed within the problem

Gradient Ascent or Descent

0/1 point (graded)

The direction of the derivative of a function gives us the direction of the largest change in the function as the independent variables vary.

In gradient ascent/descent methods, we make an educated guess about the next values of θ, with consecutive updates that



will hopefully eventually converge to the global minimum of $L\left(x,\theta\right)$ (if it exists).

If

$$\theta' = \theta + \epsilon \cdot \nabla_{\theta} L\left(x,\theta\right)$$

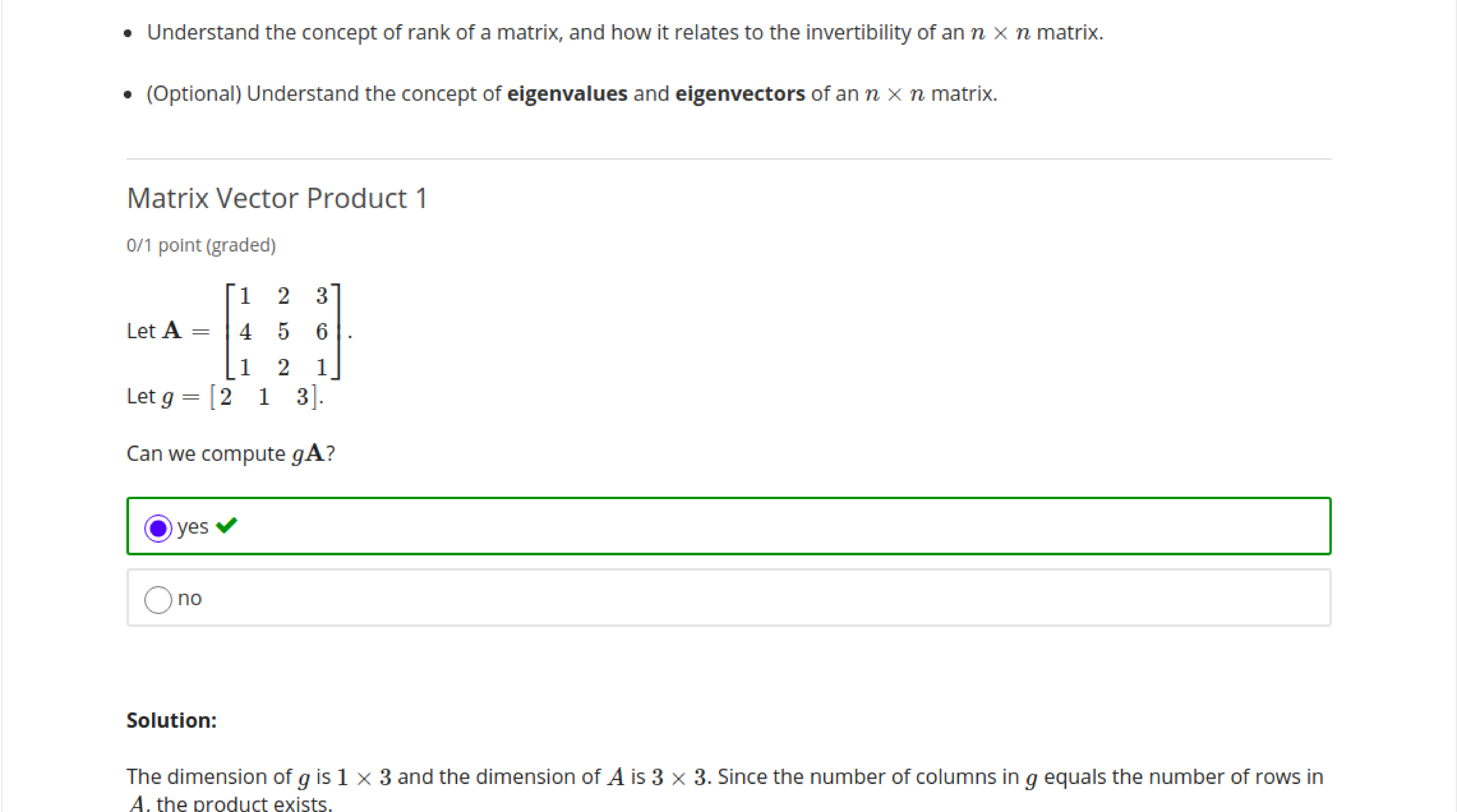
where ϵ is a small positive real number, Which of the following is true?

● $L\left(x,\theta'\right) > L\left(x,\theta\right)$

○ $L\left(x,\theta'\right) < L\left(x,\theta\right)$

Solution:

Consider the one-dimensional case. If the gradient is positive, we obtain θ' by moving from θ in the positive direction. This increases $L\left(x,\theta\right)$. If the gradient is negative, we move in the negative direction, again increasing $L\left(x,\theta\right)$. This analysis extends to higher dimensions. Note that if we used the function above to continue updating θ, we would (in theory) maximize $L\left(x,\theta\right)$. Alternatively if our update rule was $\theta' = \theta - \epsilon \cdot \nabla_{\theta} L\left(x,\theta\right)$, we would minimize the function. There are more complications in higher dimensions, but this is the basic idea behind stochastic gradient descent, which forms the backbone of modern machine learning.



- Understand the concept of rank of a matrix, and how it relates to the invertibility of an $n \times n$ matrix.

- (Optional) Understand the concept of **eigenvalues** and **eigenvectors** of an $n \times n$ matrix.

Matrix Vector Product 1

0/1 point (graded)

Let $\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 1 & 2 & 1 \end{bmatrix}$.

Let $g = \begin{bmatrix} 2 & 1 & 3 \end{bmatrix}$.

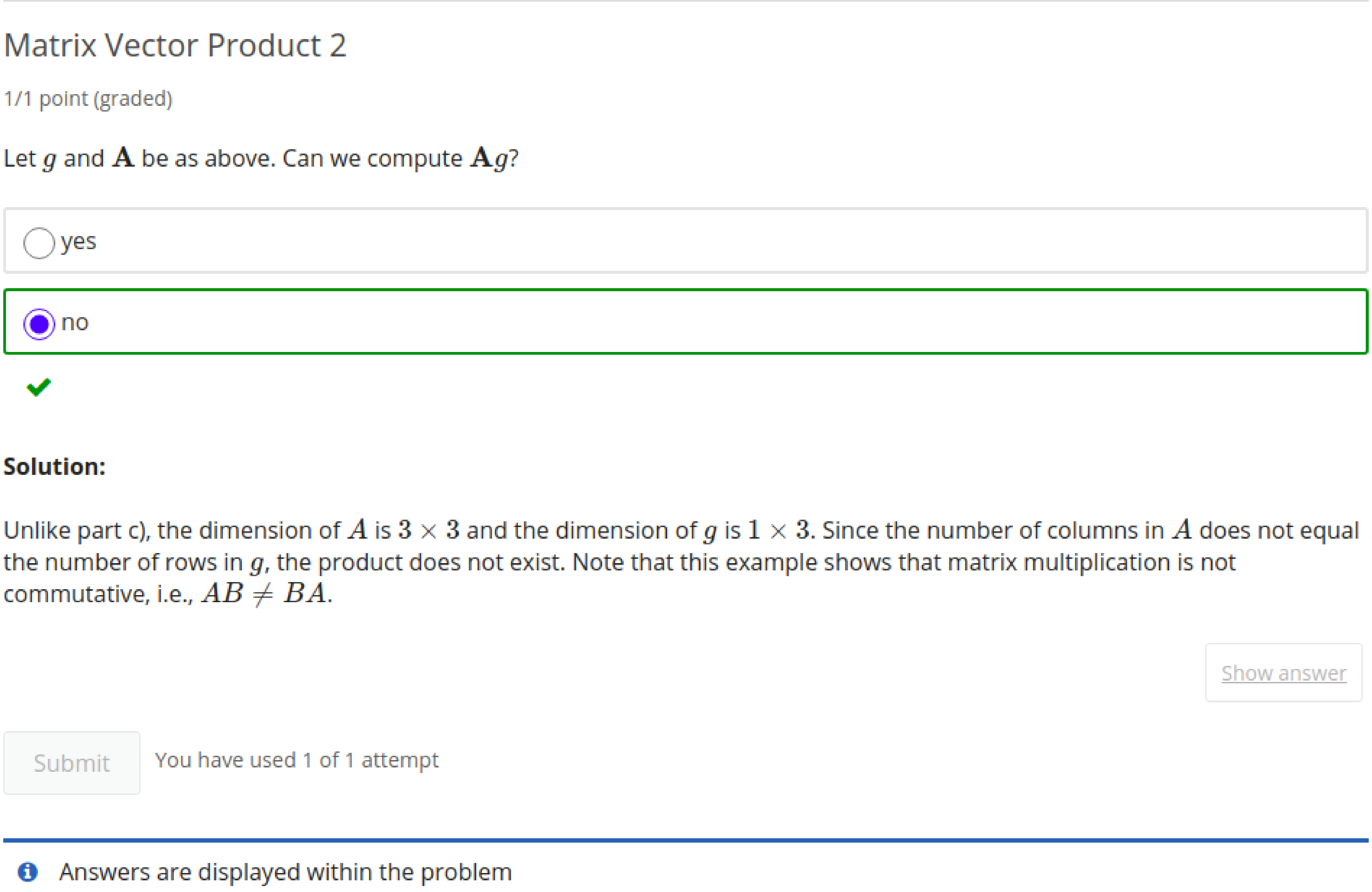
Can we compute $g\mathbf{A}$?

● yes ✓

○ no

Solution:

The dimension of g is 1×3 and the dimension of A is 3×3. Since the number of columns in g equals the number of rows in A, the product exists.



Matrix Vector Product 2

1/1 point (graded)

Let g and $\mathbf{A}$ be as above. Can we compute $\mathbf{A}g$?

○ yes

● no

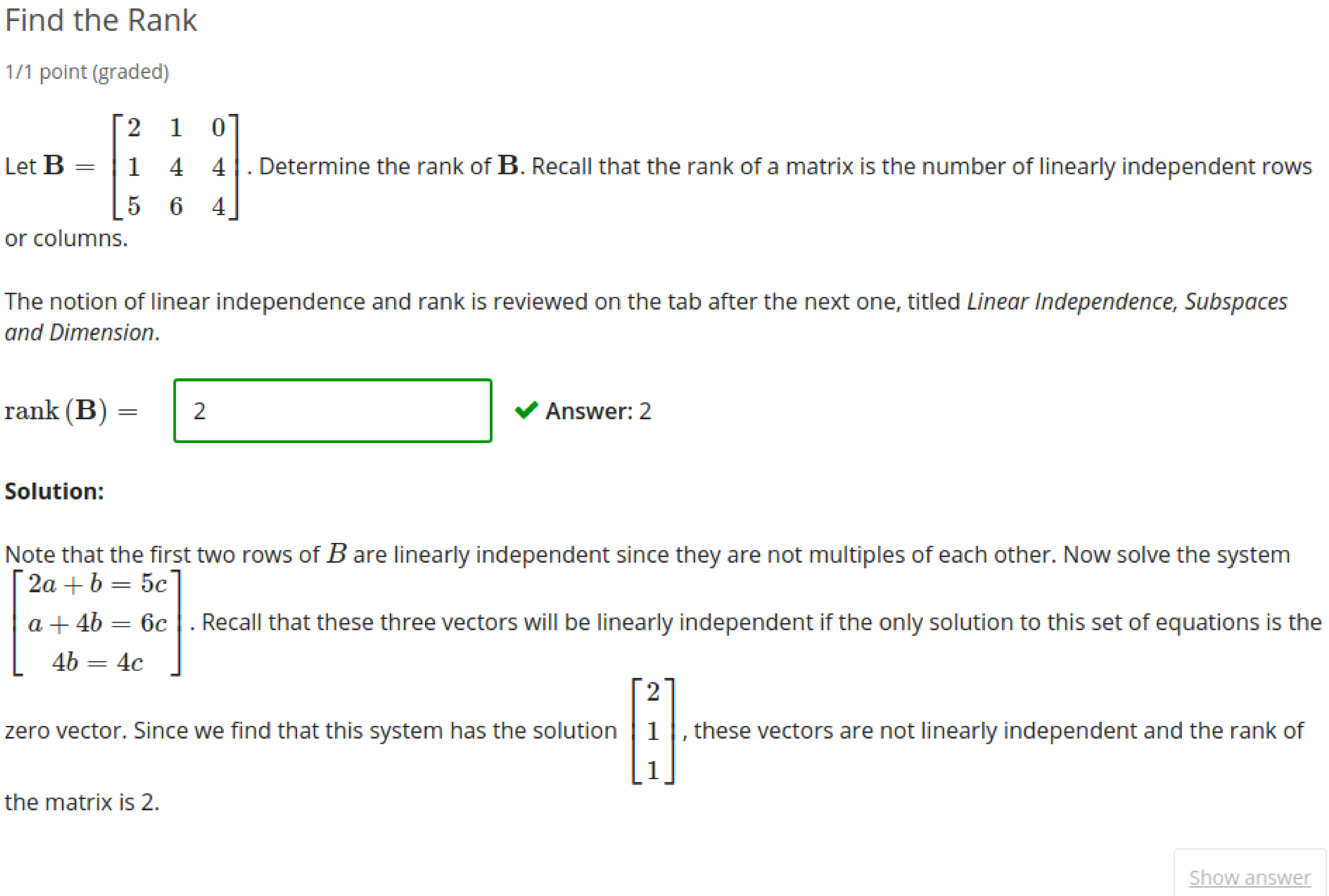
✓

Solution:

Unlike part c), the dimension of A is 3×3 and the dimension of g is 1×3. Since the number of columns in A does not equal the number of rows in g, the product does not exist. Note that this example shows that matrix multiplication is not commutative, i.e., $AB \neq BA$.

You have used 1 of 1 attempt

Answers are displayed within the problem



Find the Rank

1/1 point (graded)

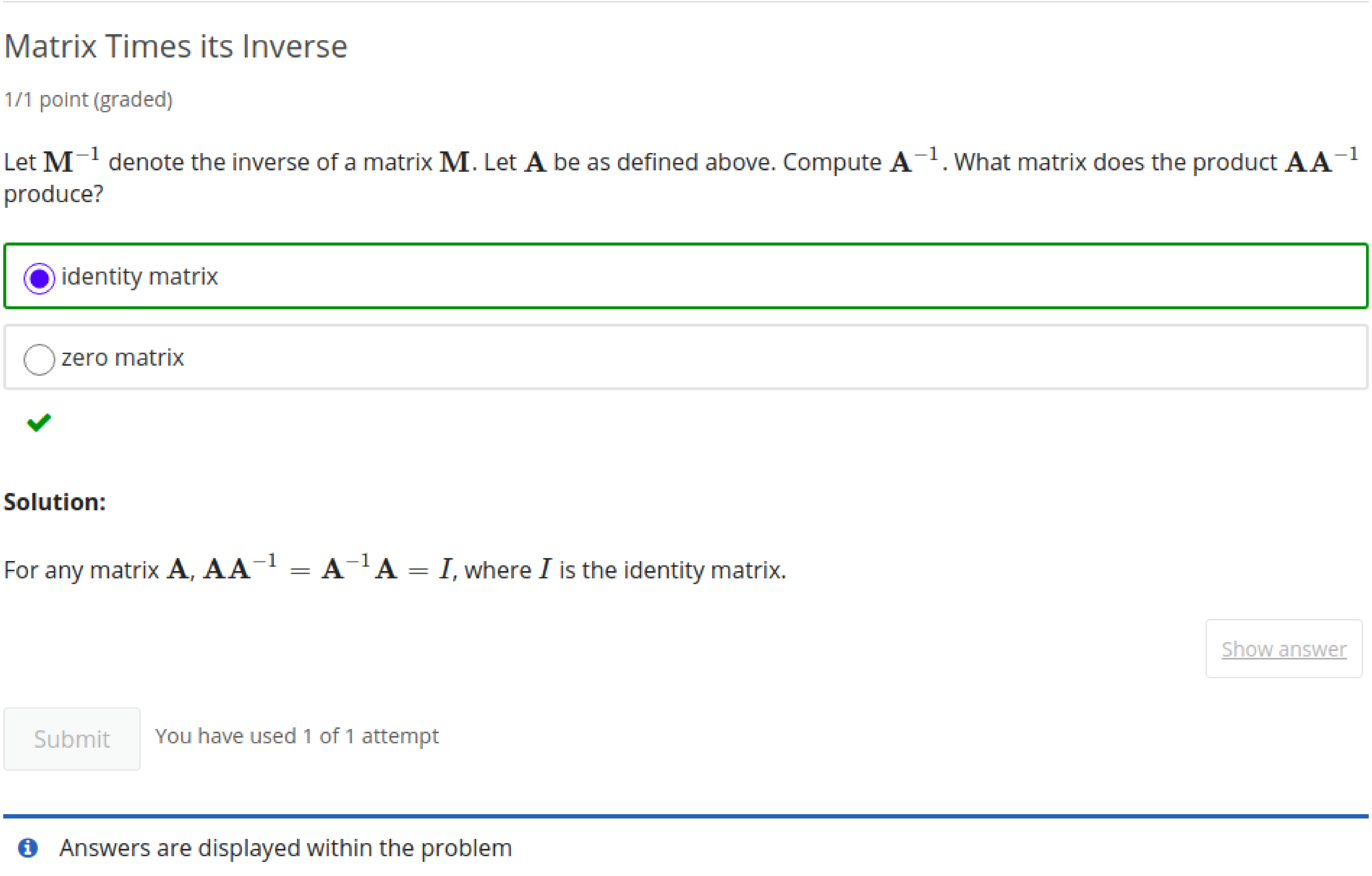
Let $\mathbf{B} = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 4 & 4 \\ 5 & 6 & 4 \end{bmatrix}$. Determine the rank of $\mathbf{B}$. Recall that the rank of a matrix is the number of linearly independent rows or columns.

The notion of linear independence and rank is reviewed on the tab after the next one, titled *Linear Independence, Subspaces and Dimension*.

$\operatorname{rank}(\mathbf{B}) =$ 2 ✓ **Answer:** 2

Solution:

Note that the first two rows of B are linearly independent since they are not multiples of each other. Now solve the system $\begin{bmatrix} 2a + b = 5c \\ a + 4b = 6c \\ 4b = 4c \end{bmatrix}$. Recall that these three vectors will be linearly independent if the only solution to this set of equations is the zero vector. Since we find that this system has the solution $\begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$, these vectors are not linearly independent and the rank of the matrix is 2.



Matrix Times its Inverse

1/1 point (graded)

Let $\mathbf{M}^{-1}$ denote the inverse of a matrix $\mathbf{M}$. Let $\mathbf{A}$ be as defined above. Compute $\mathbf{A}^{-1}$. What matrix does the product $\mathbf{A}\mathbf{A}^{-1}$ produce?

● identity matrix

○ zero matrix

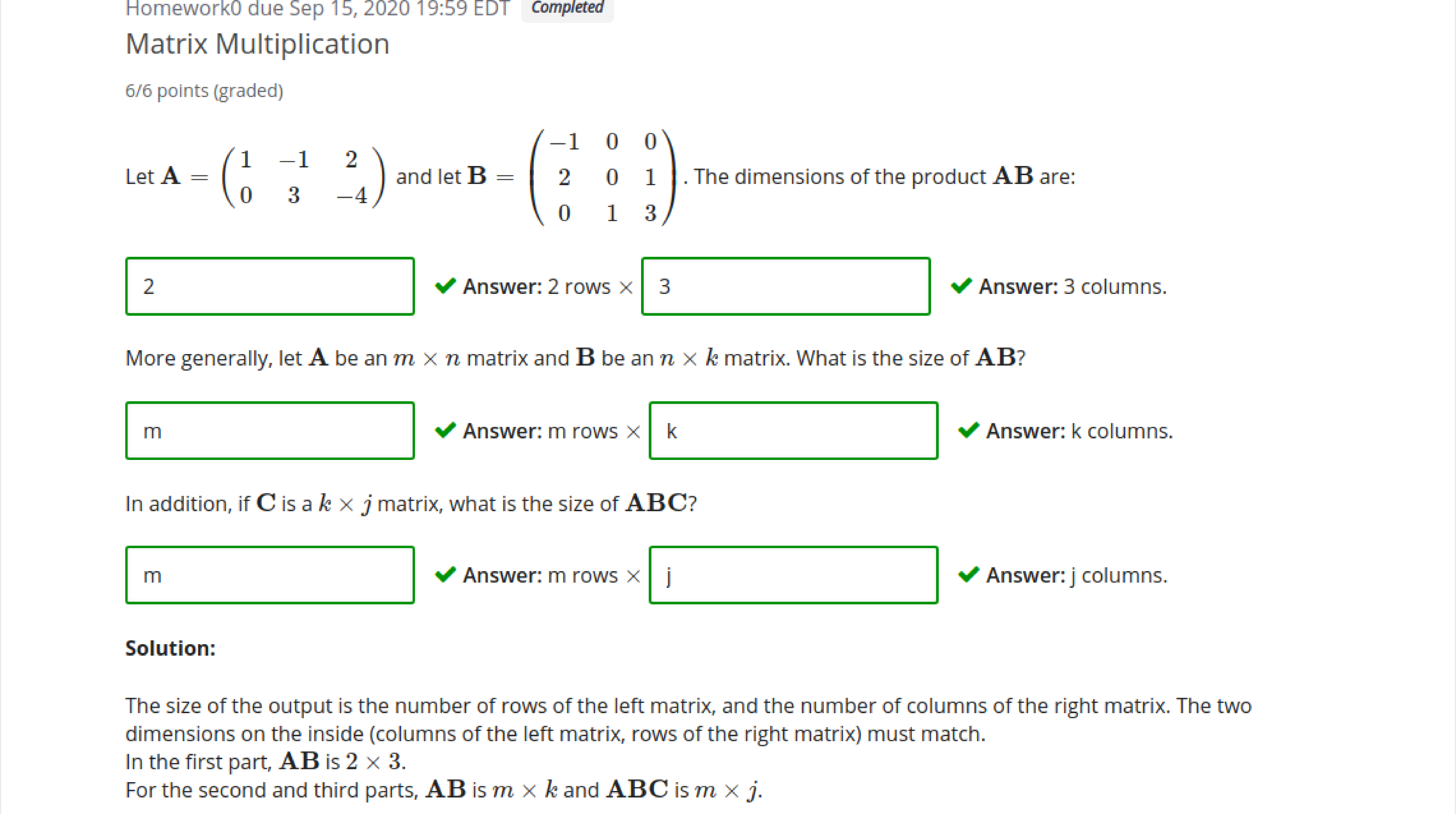
✓

Solution:

For any matrix $\mathbf{A}$, $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = I$, where I is the identity matrix.

You have used 1 of 1 attempt

Answers are displayed within the problem



Homework0 due Sep 15, 2020 19:59 EDT *Completed*

Matrix Multiplication

6/6 points (graded)

Let $\mathbf{A} = \begin{pmatrix} 1 & -1 & 2 \\ 0 & 3 & -4 \end{pmatrix}$ and let $\mathbf{B} = \begin{pmatrix} -1 & 0 & 0 \\ 2 & 0 & 1 \\ 0 & 1 & 3 \end{pmatrix}$. The dimensions of the product $\mathbf{AB}$ are:

2 ✓ **Answer:** 2 rows × 3 ✓ **Answer:** 3 columns.

More generally, let $\mathbf{A}$ be an $m \times n$ matrix and $\mathbf{B}$ be an $n \times k$ matrix. What is the size of $\mathbf{AB}$?

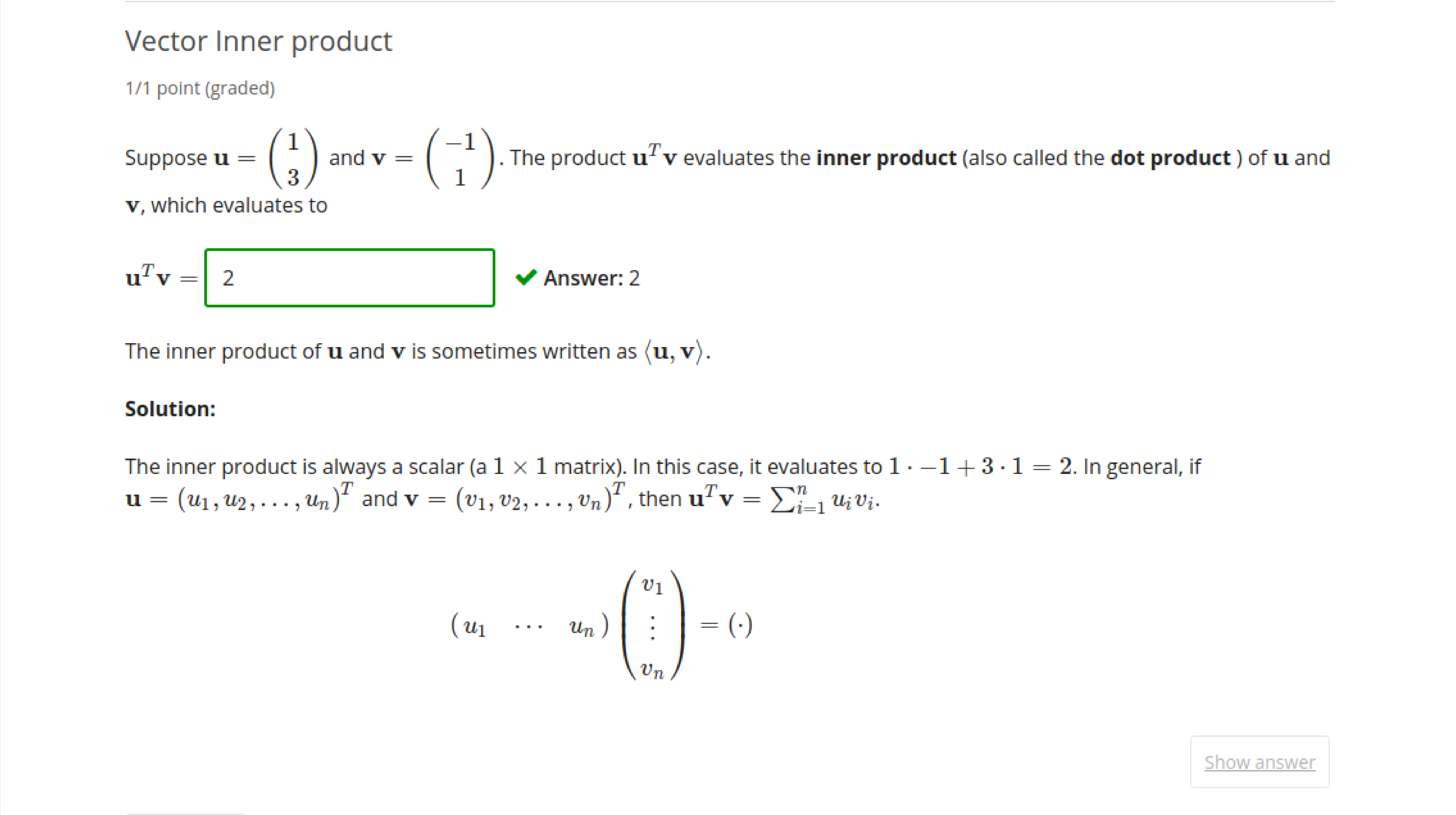
m ✓ **Answer:** m rows × k ✓ **Answer:** k columns.

In addition, if $\mathbf{C}$ is a $k \times j$ matrix, what is the size of $\mathbf{ABC}$?

m ✓ **Answer:** m rows × j ✓ **Answer:** j columns.

Solution:

The size of the output is the number of rows of the left matrix, and the number of columns of the right matrix. The two dimensions on the inside (columns of the left matrix, rows of the right matrix) must match.
In the first part, $\mathbf{AB}$ is 2×3.
For the second and third parts, $\mathbf{AB}$ is $m \times k$ and $\mathbf{ABC}$ is $m \times j$.



Vector Inner product

1/1 point (graded)

Suppose $\mathbf{u} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$. The product $\mathbf{u}^T\mathbf{v}$ evaluates the **inner product** (also called the **dot product**) of $\mathbf{u}$ and $\mathbf{v}$, which evaluates to

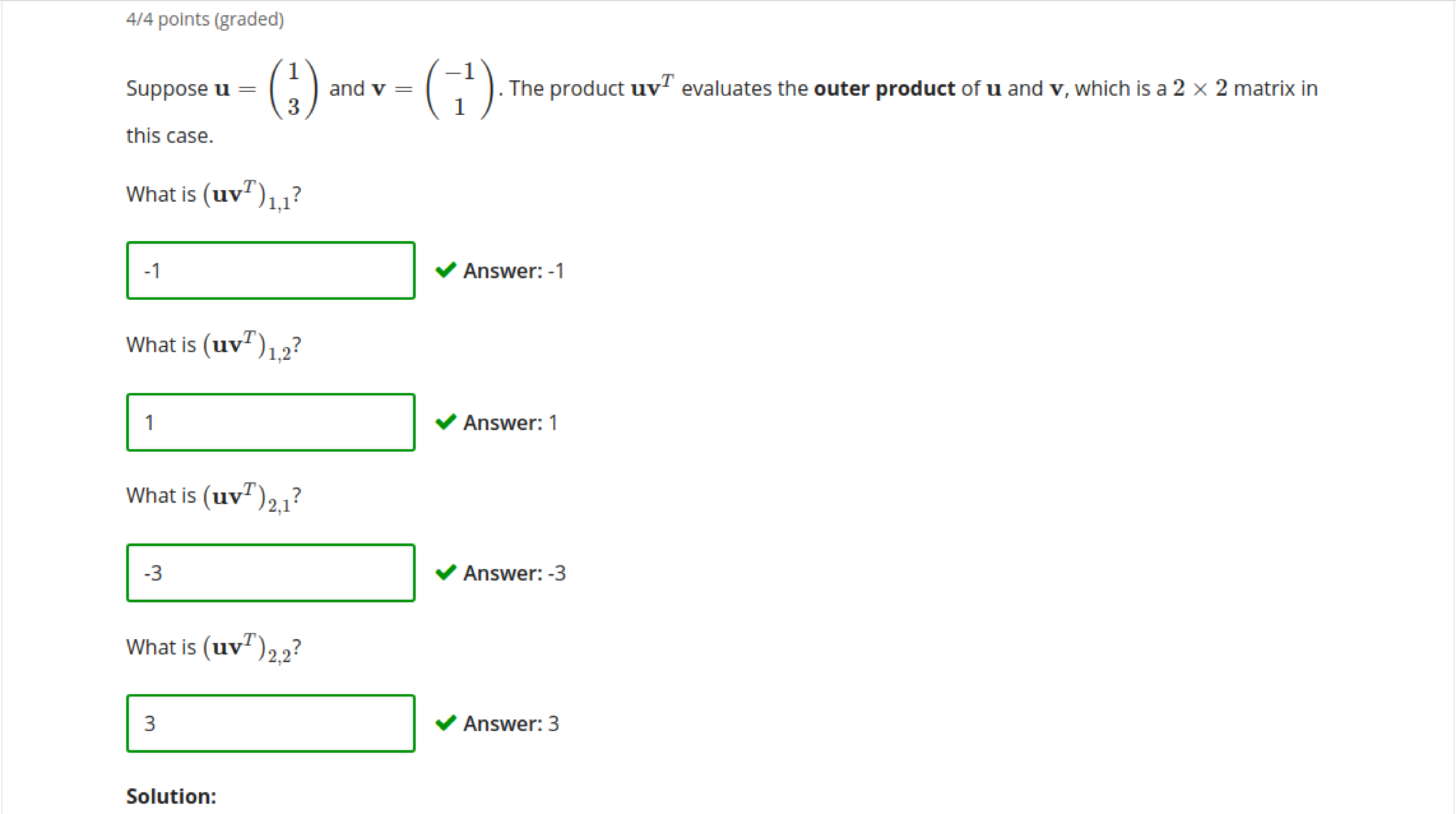
$\mathbf{u}^T\mathbf{v} =$ 2 ✓ **Answer:** 2

The inner product of $\mathbf{u}$ and $\mathbf{v}$ is sometimes written as $\langle \mathbf{u}, \mathbf{v} \rangle$.

Solution:

The inner product is always a scalar (a 1×1 matrix). In this case, it evaluates to $1 \cdot -1 + 3 \cdot 1 = 2$. In general, if $\mathbf{u} = (u_1, u_2, \ldots, u_n)^T$ and $\mathbf{v} = (v_1, v_2, \ldots, v_n)^T$, then $\mathbf{u}^T\mathbf{v} = \sum_{i=1}^{n} u_i v_i$.

$$\begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = (\cdot)$$



4/4 points (graded)

Suppose $\mathbf{u} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$. The product $\mathbf{u}\mathbf{v}^T$ evaluates the **outer product** of $\mathbf{u}$ and $\mathbf{v}$, which is a 2×2 matrix in this case.

What is $(\mathbf{u}\mathbf{v}^T)_{1,1}$?

-1

✓ **Answer:** -1

What is $(\mathbf{u}\mathbf{v}^T)_{1,2}$?

1

✓ **Answer:** 1

What is $(\mathbf{u}\mathbf{v}^T)_{2,1}$?

-3

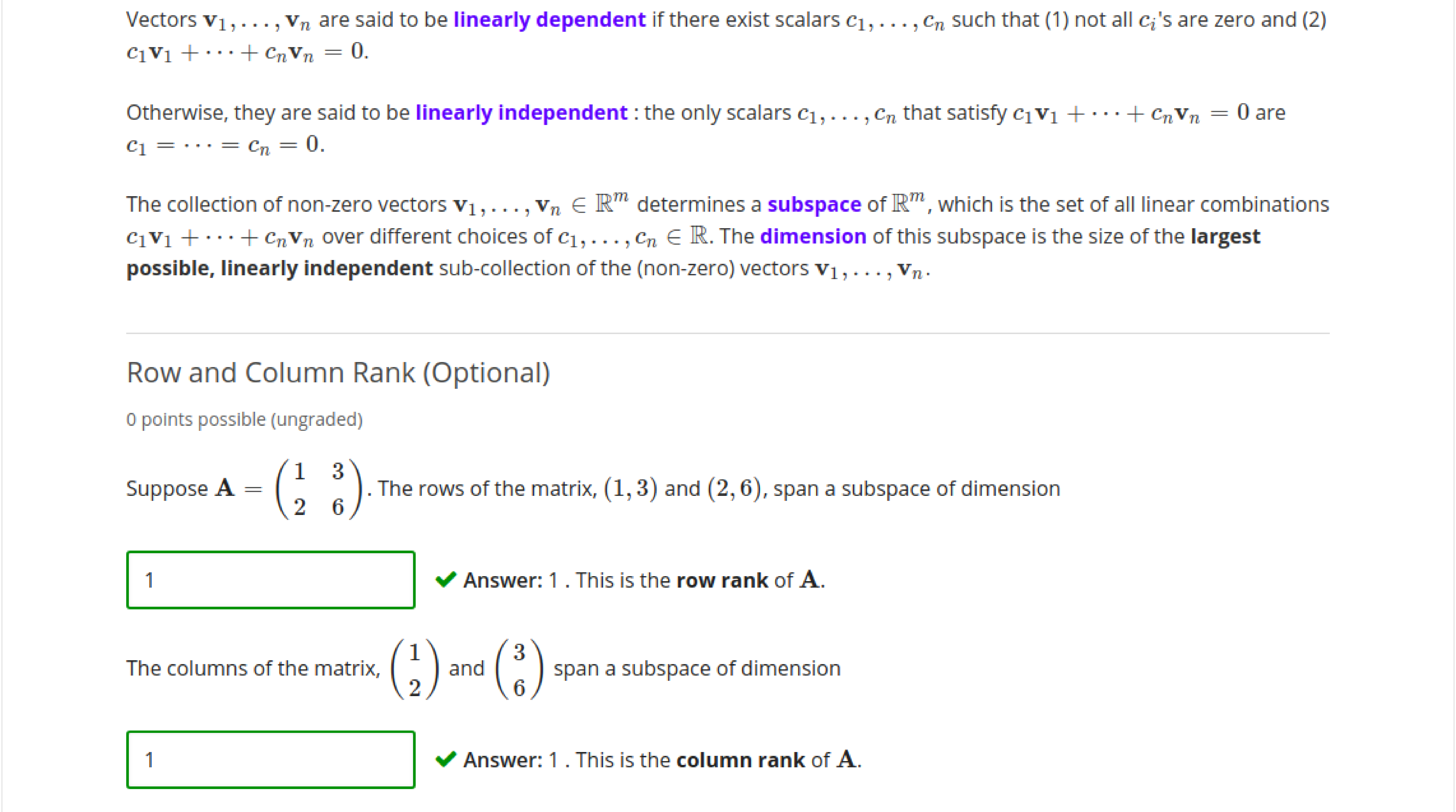
✓ **Answer:** -3

What is $(\mathbf{u}\mathbf{v}^T)_{2,2}$?

3

✓ **Answer:** 3

Solution:



Vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are said to be **linearly dependent** if there exist scalars $c_1, \ldots, c_n$ such that (1) not all c_i's are zero and (2) $c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n = 0$.

Otherwise, they are said to be **linearly independent** : the only scalars $c_1, \ldots, c_n$ that satisfy $c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n = 0$ are $c_1 = \cdots = c_n = 0$.

The collection of non-zero vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^m$ determines a **subspace** of $\mathbb{R}^m$, which is the set of all linear combinations $c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$ over different choices of $c_1, \ldots, c_n \in \mathbb{R}$. The **dimension** of this subspace is the size of the **largest possible, linearly independent** sub-collection of the (non-zero) vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$.

Row and Column Rank (Optional)

0 points possible (ungraded)

Suppose $\mathbf{A} = \begin{pmatrix} 1 & 3 \\ 2 & 6 \end{pmatrix}$. The rows of the matrix, $(1, 3)$ and $(2, 6)$, span a subspace of dimension

1

✓ **Answer:** 1 . This is the **row rank** of $\mathbf{A}$.

The columns of the matrix, $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and $\begin{pmatrix} 3 \\ 6 \end{pmatrix}$ span a subspace of dimension

1

✓ **Answer:** 1 . This is the **column rank** of $\mathbf{A}$.

We will be using these ideas when studying **Linear Regression**, where we will work with larger, possibly rectangular matrices.

Solution:

In both cases, the two vectors are linearly dependent.

$$2 \cdot (1,3) - (2,6) = (0,0)$$

$$3 \begin{pmatrix} 1 \\ 2 \end{pmatrix} - \begin{pmatrix} 3 \\ 6 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

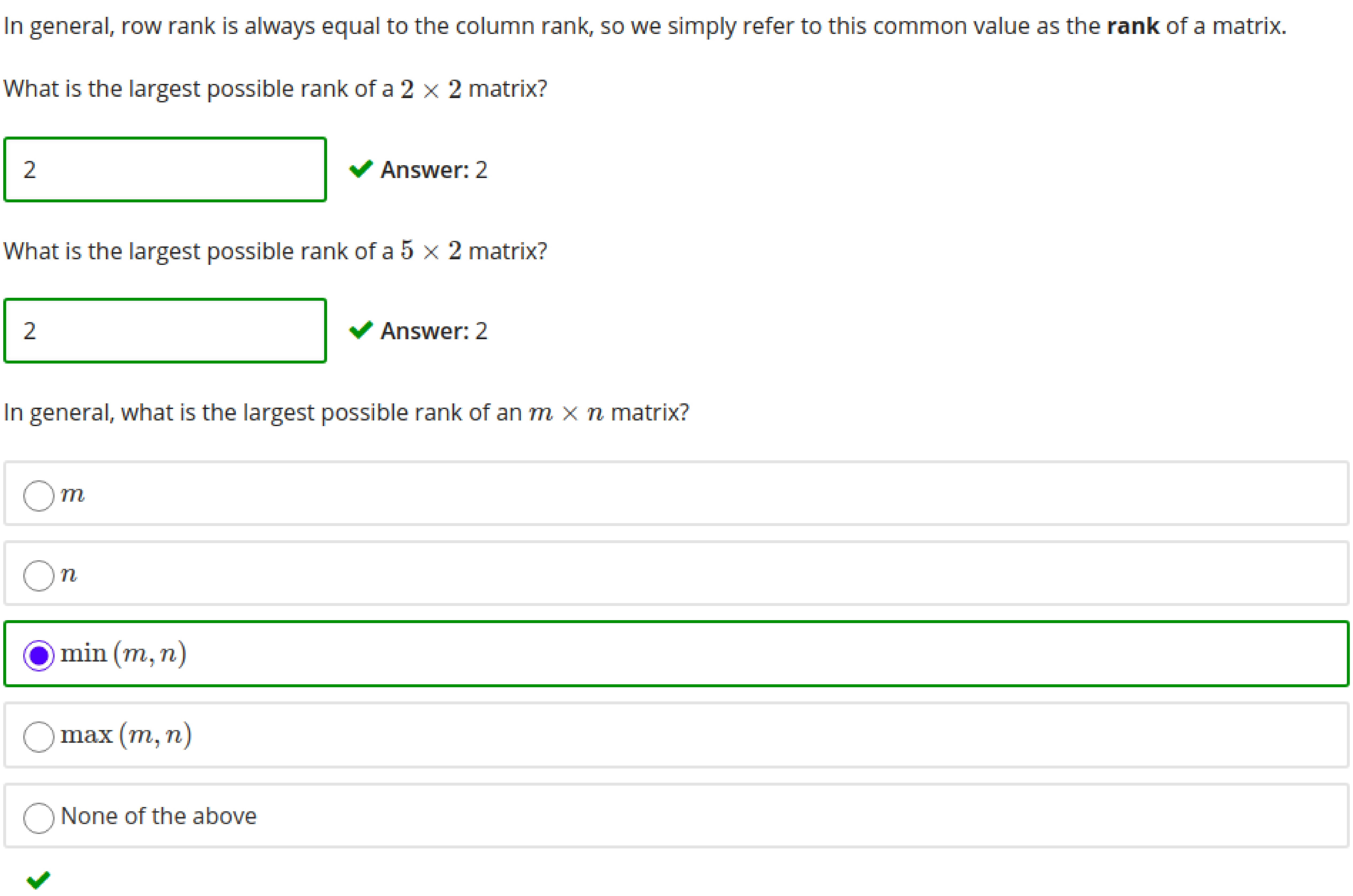
Show answer

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

Rank of a matrix (Optional)

0 points possible (ungraded)



In general, row rank is always equal to the column rank, so we simply refer to this common value as the **rank** of a matrix.

What is the largest possible rank of a 2×2 matrix?

2

✓ **Answer:** 2

What is the largest possible rank of a 5×2 matrix?

2

✓ **Answer:** 2

In general, what is the largest possible rank of an $m \times n$ matrix?

○ m

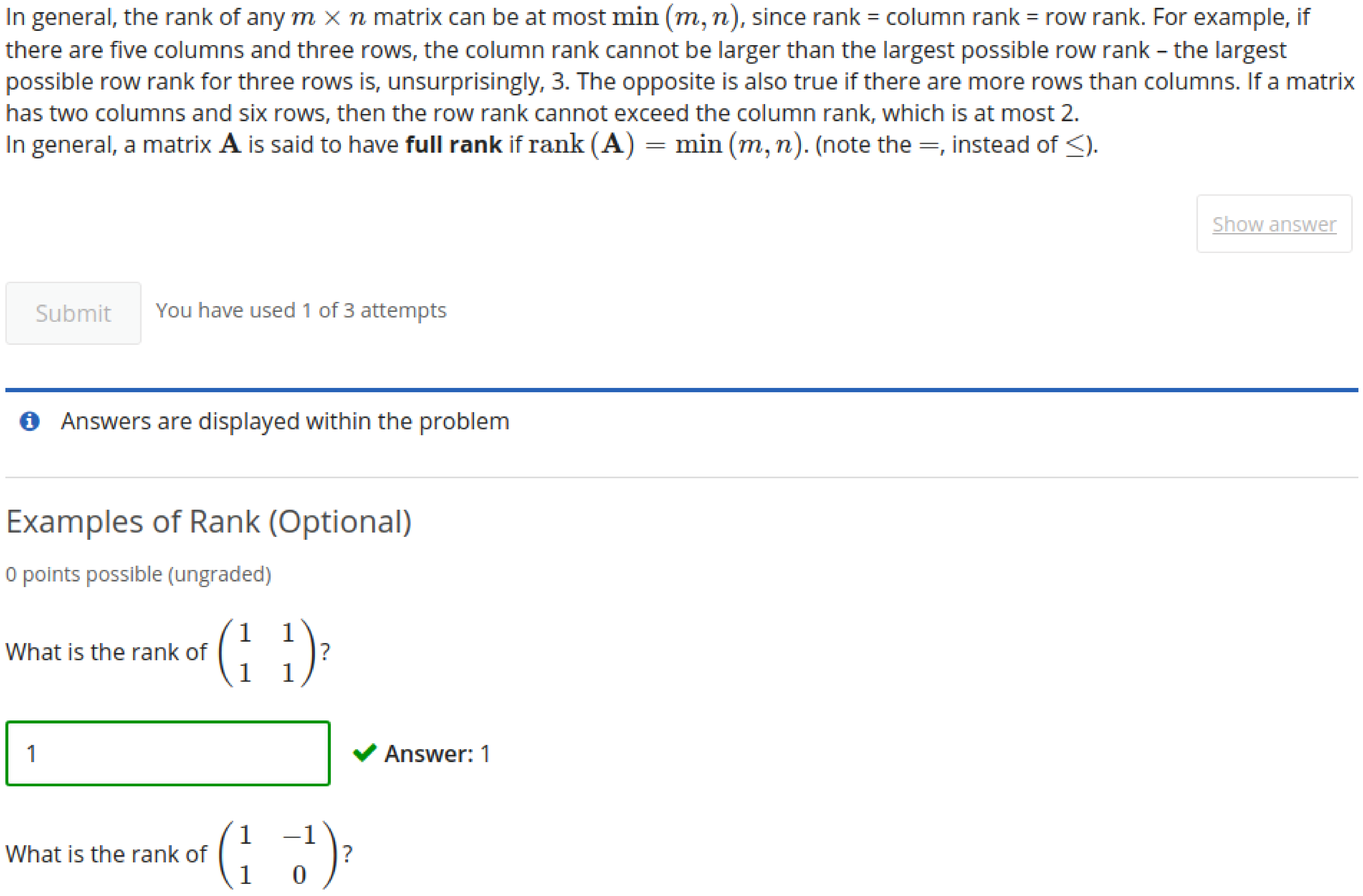
○ n

● $\min(m, n)$

○ $\max(m, n)$

○ None of the above

✓



In general, the rank of any $m \times n$ matrix can be at most $\min(m, n)$, since rank = column rank = row rank. For example, if there are five columns and three rows, the column rank cannot be larger than the largest possible row rank – the largest possible row rank for three rows is, unsurprisingly, 3. The opposite is also true if there are more rows than columns. If a matrix has two columns and six rows, then the row rank cannot exceed the column rank, which is at most 2.
In general, a matrix $\mathbf{A}$ is said to have **full rank** if $\text{rank}(\mathbf{A}) = \min(m, n)$. (note the $=$, instead of $\leq$).

You have used 1 of 3 attempts

Answers are displayed within the problem

Examples of Rank (Optional)

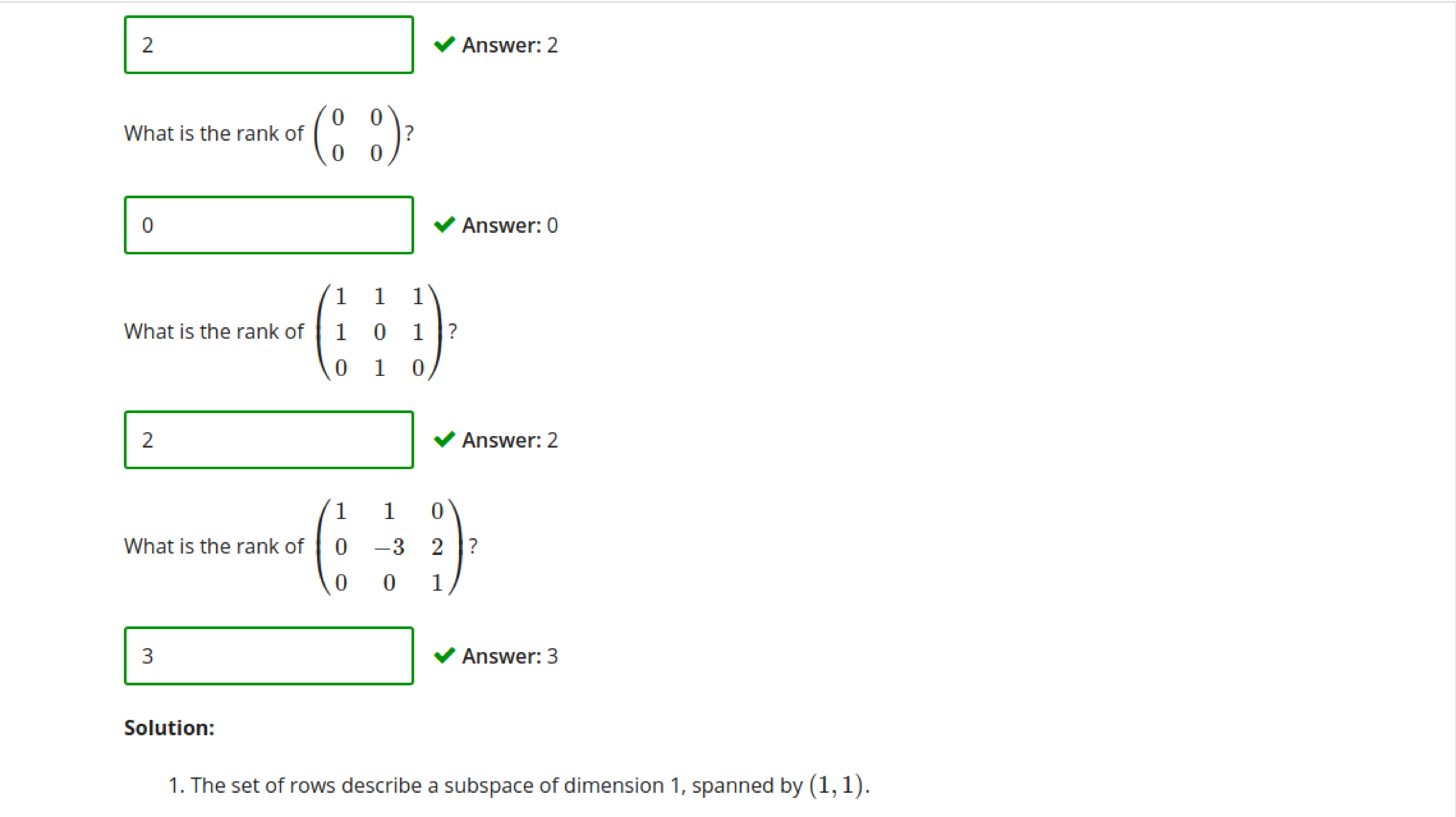
0 points possible (ungraded)

What is the rank of $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$?

1

✓ **Answer:** 1

What is the rank of $\begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}$?



2

✓ **Answer:** 2

What is the rank of $\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$?

0

✓ **Answer:** 0

What is the rank of $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$?

2

✓ **Answer:** 2

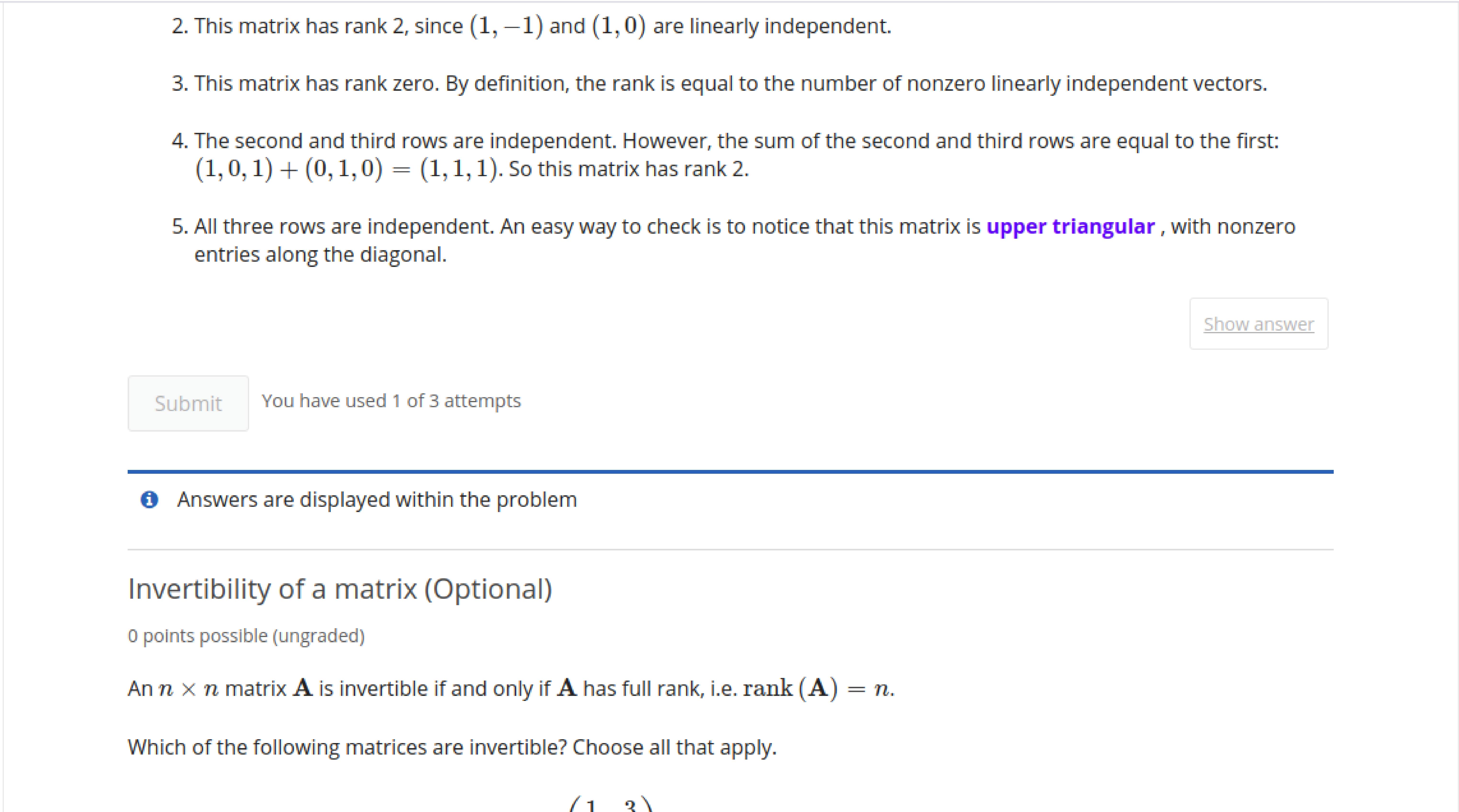
What is the rank of $\begin{pmatrix} 1 & 1 & 0 \\ 0 & -3 & 2 \\ 0 & 0 & 1 \end{pmatrix}$?

3

✓ **Answer:** 3

Solution:

1. The set of rows describe a subspace of dimension 1, spanned by $(1, 1)$.



2. This matrix has rank 2, since $(1, -1)$ and $(1, 0)$ are linearly independent.

3. This matrix has rank zero. By definition, the rank is equal to the number of nonzero linearly independent vectors.

4. The second and third rows are independent. However, the sum of the second and third rows are equal to the first: $(1, 0, 1) + (0, 1, 0) = (1, 1, 1)$. So this matrix has rank 2.

5. All three rows are independent. An easy way to check is to notice that this matrix is **upper triangular** , with nonzero entries along the diagonal.

You have used 1 of 3 attempts

ℹ Answers are displayed within the problem

Invertibility of a matrix (Optional)

0 points possible (ungraded)

An $n \times n$ matrix $\mathbf{A}$ is invertible if and only if $\mathbf{A}$ has full rank, i.e. $\operatorname{rank}(\mathbf{A}) = n$.

Which of the following matrices are invertible? Choose all that apply.

$$\begin{pmatrix} 1 & 3 \end{pmatrix}$$

$$\mathbf{A} = \begin{pmatrix} 1 & 3 \\ 2 & 6 \end{pmatrix}$$

$$\mathbf{B} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\mathbf{D} = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$$

□ **A**

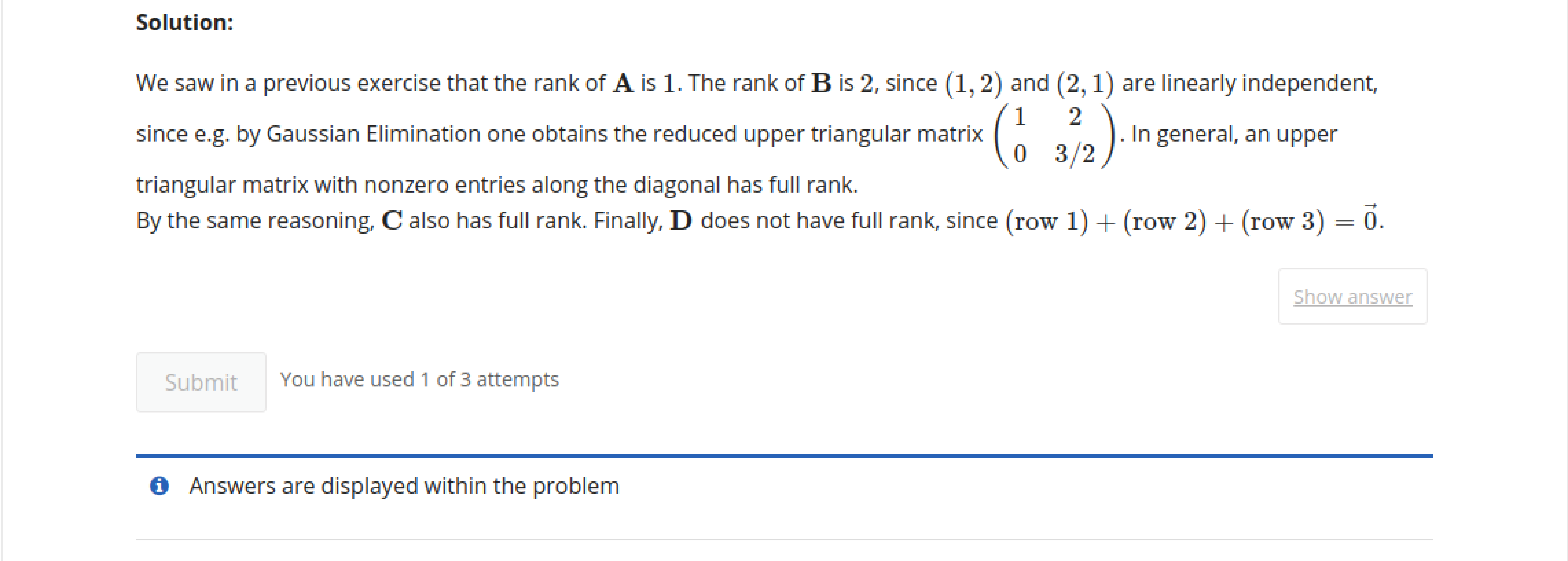
☑ **B**

☑ **C**

□ **D**

✓

Solution:



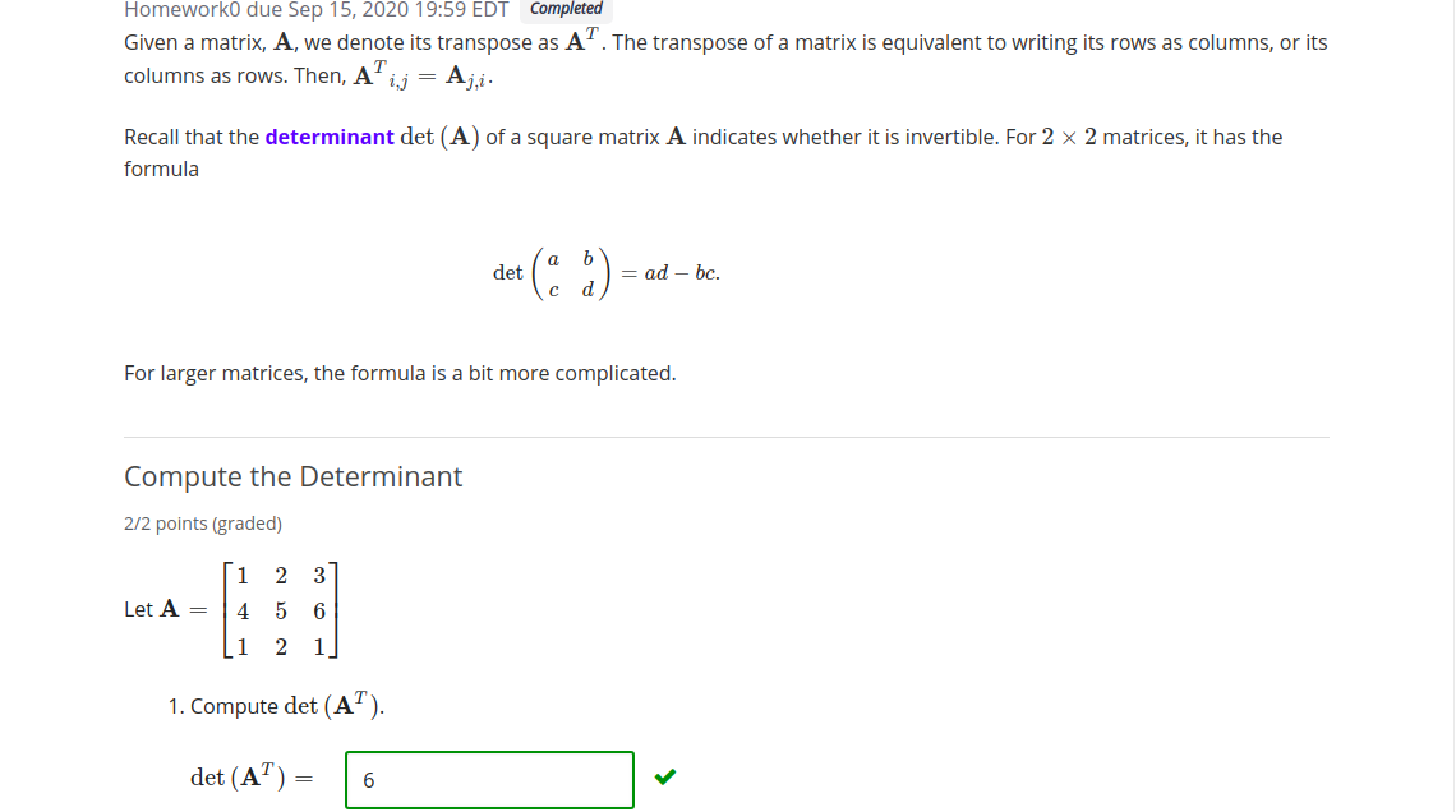
Solution:

We saw in a previous exercise that the rank of $\mathbf{A}$ is 1. The rank of $\mathbf{B}$ is 2, since $(1,2)$ and $(2,1)$ are linearly independent, since e.g. by Gaussian Elimination one obtains the reduced upper triangular matrix $\begin{pmatrix} 1 & 2 \\ 0 & 3/2 \end{pmatrix}$. In general, an upper triangular matrix with nonzero entries along the diagonal has full rank.

By the same reasoning, $\mathbf{C}$ also has full rank. Finally, $\mathbf{D}$ does not have full rank, since $(\text{row } 1) + (\text{row } 2) + (\text{row } 3) = \vec{0}$.

You have used 1 of 3 attempts

Answers are displayed within the problem



Homework0 due Sep 15, 2020 19:59 EDT *Completed*

Given a matrix, $\mathbf{A}$, we denote its transpose as $\mathbf{A}^T$. The transpose of a matrix is equivalent to writing its rows as columns, or its columns as rows. Then, ${\mathbf{A}^T}_{i,j} = \mathbf{A}_{j,i}$.

Recall that the **determinant** $\det(\mathbf{A})$ of a square matrix $\mathbf{A}$ indicates whether it is invertible. For 2×2 matrices, it has the formula

$$\det\begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc.$$

For larger matrices, the formula is a bit more complicated.

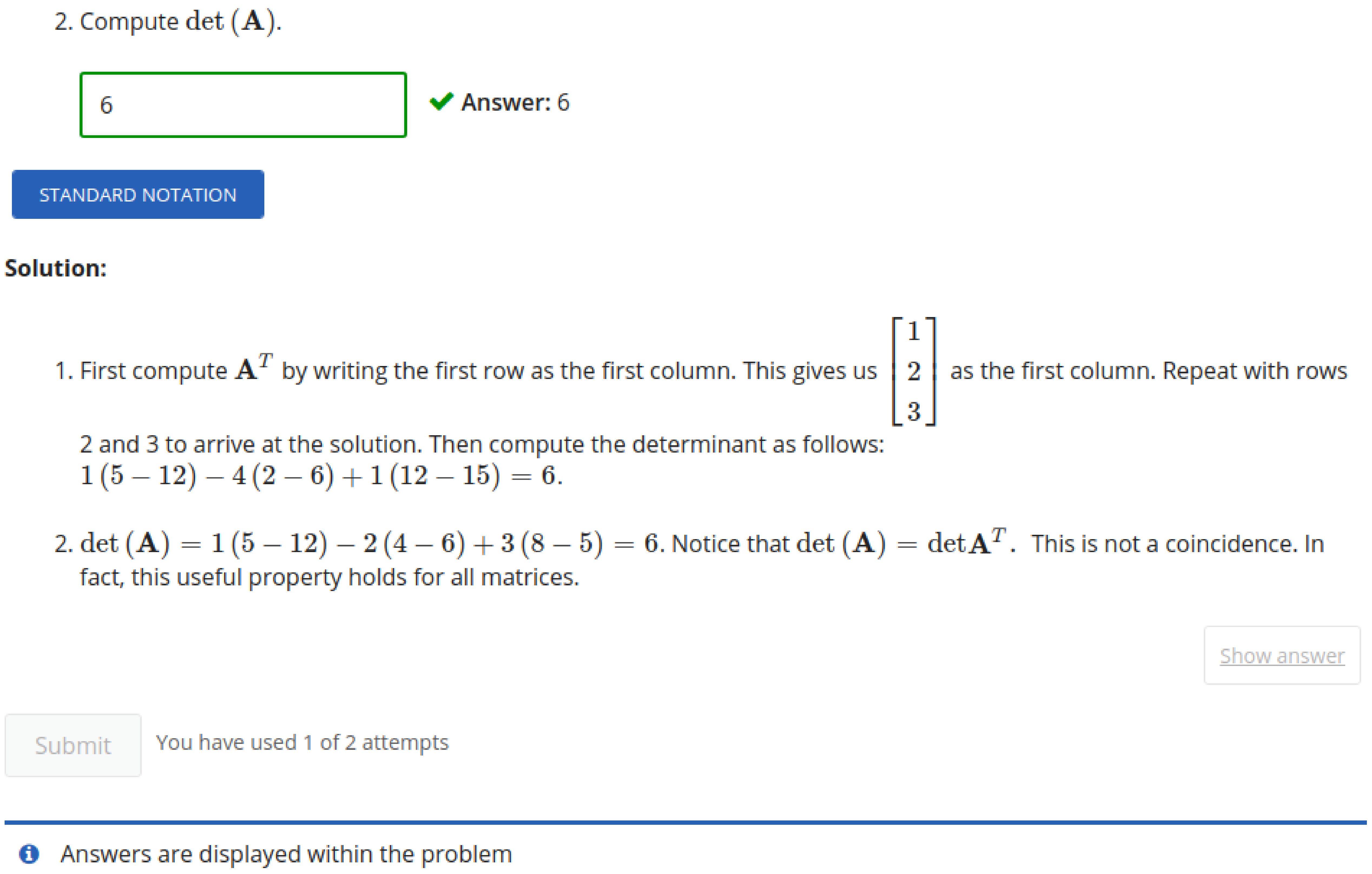
Compute the Determinant

2/2 points (graded)

Let $\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 1 & 2 & 1 \end{bmatrix}$

1. Compute $\det(\mathbf{A}^T)$.

$\det(\mathbf{A}^T) =$ 6 ✓



2. Compute $\det(\mathbf{A})$.

6

✓ **Answer:** 6

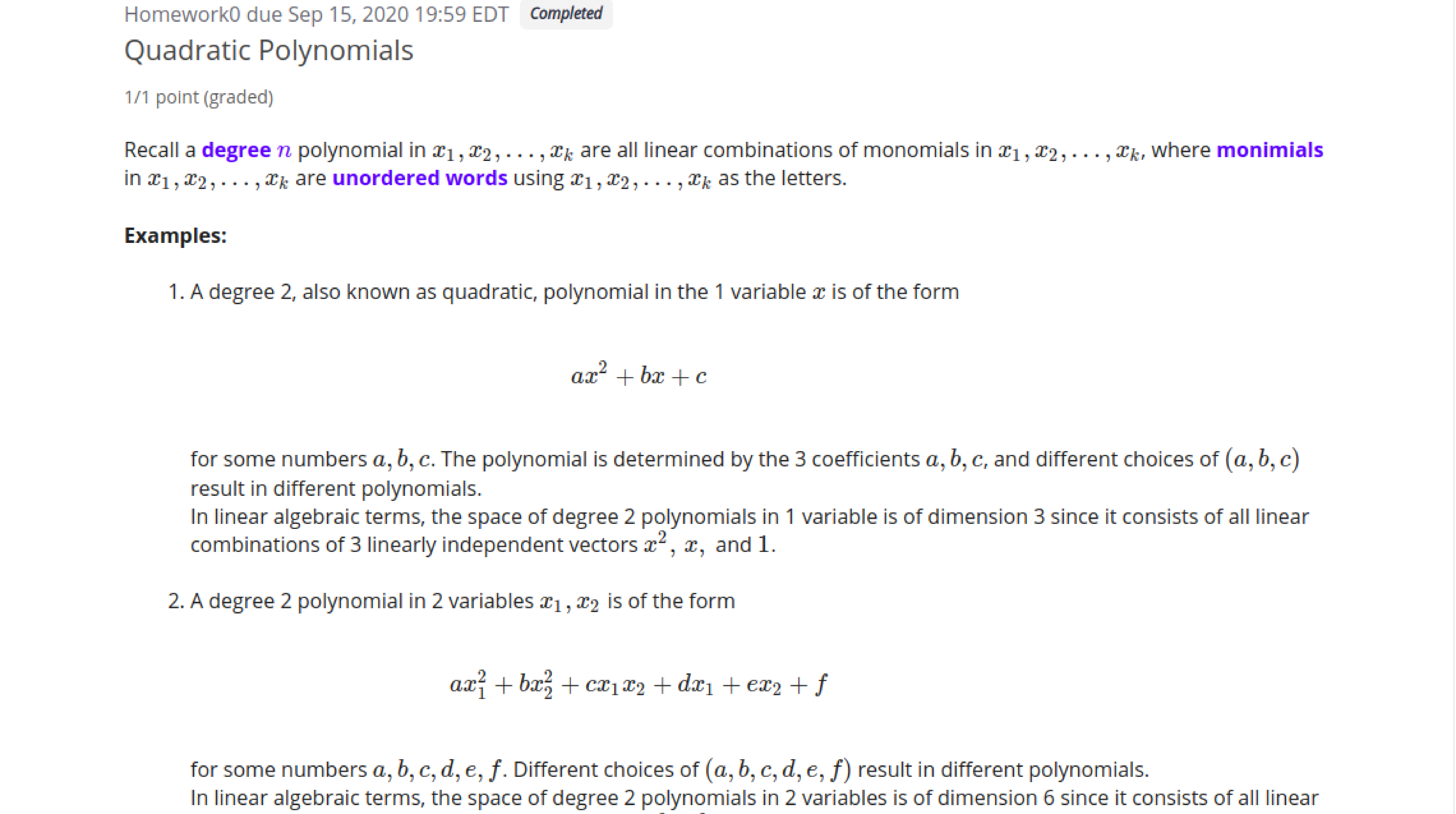
Solution:

1. First compute $\mathbf{A}^T$ by writing the first row as the first column. This gives us $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ as the first column. Repeat with rows 2 and 3 to arrive at the solution. Then compute the determinant as follows: $1\,(5-12)-4\,(2-6)+1\,(12-15)=6.$

2. $\det(\mathbf{A})=1\,(5-12)-2\,(4-6)+3\,(8-5)=6$. Notice that $\det(\mathbf{A})=\det\mathbf{A}^T.$ This is not a coincidence. In fact, this useful property holds for all matrices.

You have used 1 of 2 attempts

Answers are displayed within the problem



Homework0 due Sep 15, 2020 19:59 EDT Completed

Quadratic Polynomials

1/1 point (graded)

Recall a **degree n** polynomial in $x_1, x_2, \dots, x_k$ are all linear combinations of monomials in $x_1, x_2, \dots, x_k$, where **monimials** in $x_1, x_2, \dots, x_k$ are **unordered words** using $x_1, x_2, \dots, x_k$ as the letters.

Examples:

1. A degree 2, also known as quadratic, polynomial in the 1 variable x is of the form

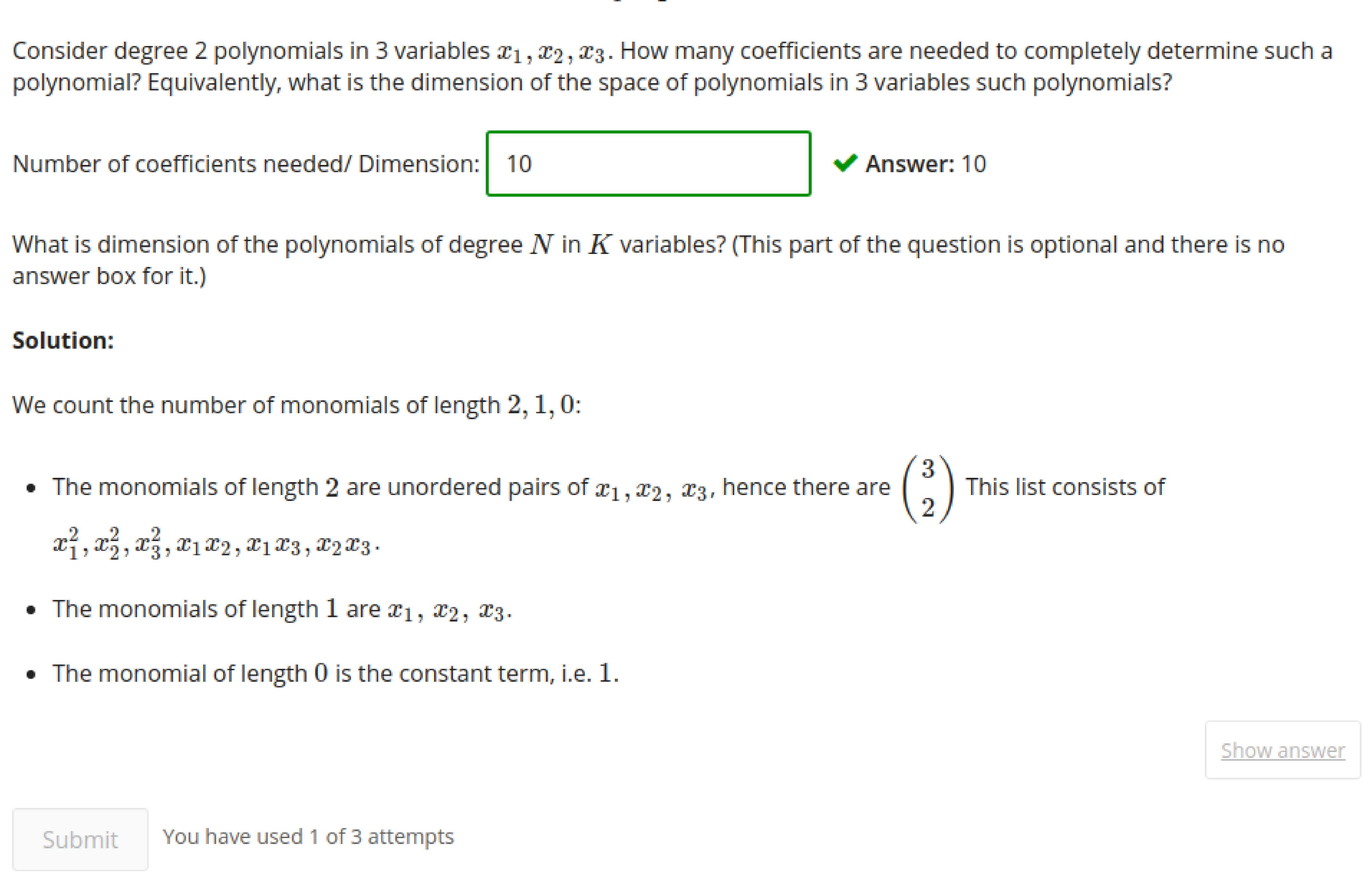
$$ax^2 + bx + c$$

for some numbers a, b, c. The polynomial is determined by the 3 coefficients a, b, c, and different choices of (a, b, c) result in different polynomials.
In linear algebraic terms, the space of degree 2 polynomials in 1 variable is of dimension 3 since it consists of all linear combinations of 3 linearly independent vectors $x^2,\ x,$ and 1.

2. A degree 2 polynomial in 2 variables x_1, x_2 is of the form

$$ax_1^2 + bx_2^2 + cx_1x_2 + dx_1 + ex_2 + f$$

for some numbers a, b, c, d, e, f. Different choices of (a, b, c, d, e, f) result in different polynomials.
In linear algebraic terms, the space of degree 2 polynomials in 2 variables is of dimension 6 since it consists of all linear



Consider degree 2 polynomials in 3 variables x_1, x_2, x_3. How many coefficients are needed to completely determine such a polynomial? Equivalently, what is the dimension of the space of polynomials in 3 variables such polynomials?

Number of coefficients needed/ Dimension: 10 ✓ **Answer:** 10

What is dimension of the polynomials of degree N in K variables? (This part of the question is optional and there is no answer box for it.)

Solution:

We count the number of monomials of length $2, 1, 0$:

- The monomials of length 2 are unordered pairs of $x_1, x_2,\ x_3$, hence there are $\binom{3}{2}$ This list consists of $x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3$.
- The monomials of length 1 are $x_1,\ x_2,\ x_3$.
- The monomial of length 0 is the constant term, i.e. 1.

You have used 1 of 3 attempts

0 points possible (ungraded)

Let $\mathbf{A} = \begin{pmatrix} 3 & 0 \\ \frac{1}{2} & 2 \end{pmatrix}$, $\mathbf{v} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ and $\mathbf{w} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

$\mathbf{Av} = \lambda_1 \mathbf{v}$, where $\lambda_1 =$

Answer: 3 .

$\mathbf{Aw} = \lambda_2 \mathbf{w}$, where $\lambda_2 =$

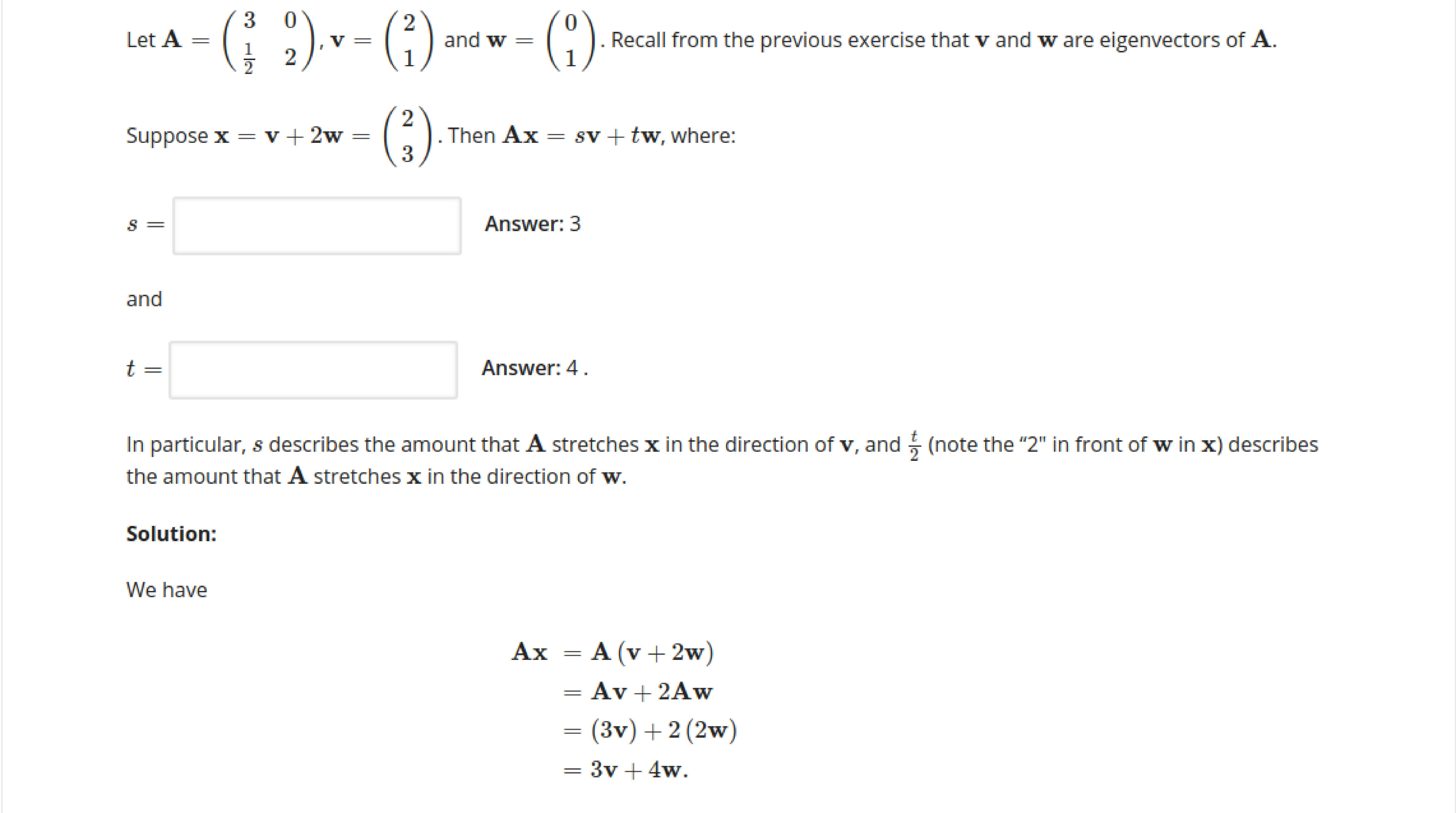
Answer: 2 .

Therefore, $\mathbf{v}$ is an eigenvector of $\mathbf{A}$ with eigenvalue λ_1, and $\mathbf{w}$ is an eigenvector of $\mathbf{A}$ with eigenvalue λ_2.

Solution:

$$\mathbf{Av} = \begin{pmatrix} 3 & 0 \\ \frac{1}{2} & 2 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 6 \\ 3 \end{pmatrix} \implies \lambda_1 = 3$$

$$\mathbf{Aw} = \begin{pmatrix} 3 & 0 \\ \frac{1}{2} & 2 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix} \implies \lambda_2 = 2$$



Let $\mathbf{A} = \begin{pmatrix} 3 & 0 \\ \frac{1}{2} & 2 \end{pmatrix}$, $\mathbf{v} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ and $\mathbf{w} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Recall from the previous exercise that $\mathbf{v}$ and $\mathbf{w}$ are eigenvectors of $\mathbf{A}$.

Suppose $\mathbf{x} = \mathbf{v} + 2\mathbf{w} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$. Then $\mathbf{Ax} = s\mathbf{v} + t\mathbf{w}$, where:

$s =$ **Answer:** 3

and

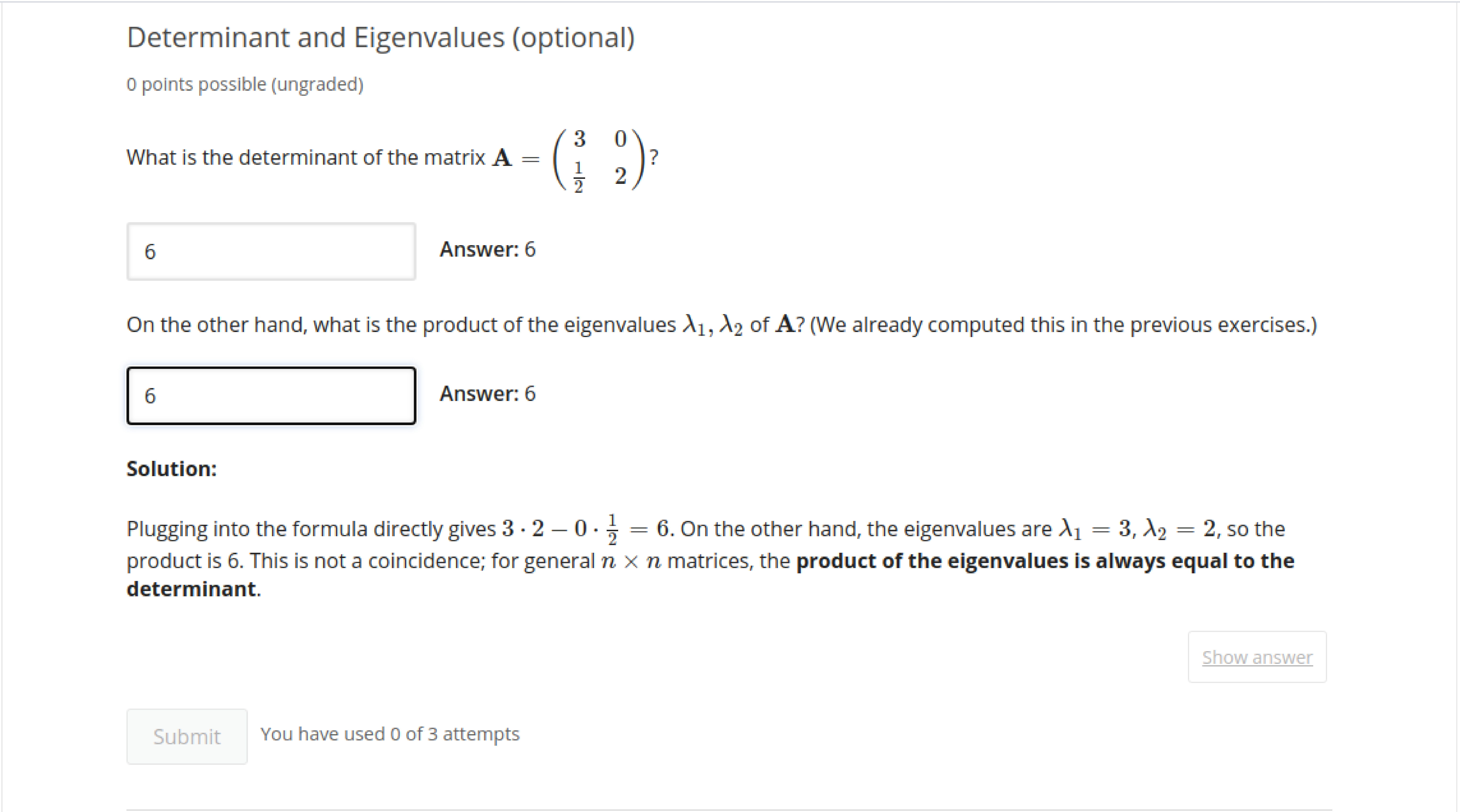
$t =$ **Answer:** 4 .

In particular, s describes the amount that $\mathbf{A}$ stretches $\mathbf{x}$ in the direction of $\mathbf{v}$, and $\frac{t}{2}$ (note the “2" in front of $\mathbf{w}$ in $\mathbf{x}$) describes the amount that $\mathbf{A}$ stretches $\mathbf{x}$ in the direction of $\mathbf{w}$.

Solution:

We have

$$\begin{aligned} \mathbf{Ax} &= \mathbf{A}\left(\mathbf{v} + 2\mathbf{w}\right) \\ &= \mathbf{Av} + 2\mathbf{Aw} \\ &= \left(3\mathbf{v}\right) + 2\left(2\mathbf{w}\right) \\ &= 3\mathbf{v} + 4\mathbf{w}. \end{aligned}$$



Determinant and Eigenvalues (optional)

0 points possible (ungraded)

What is the determinant of the matrix $\mathbf{A} = \begin{pmatrix} 3 & 0 \\ \frac{1}{2} & 2 \end{pmatrix}$?

6

Answer: 6

On the other hand, what is the product of the eigenvalues λ_1, λ_2 of $\mathbf{A}$? (We already computed this in the previous exercises.)

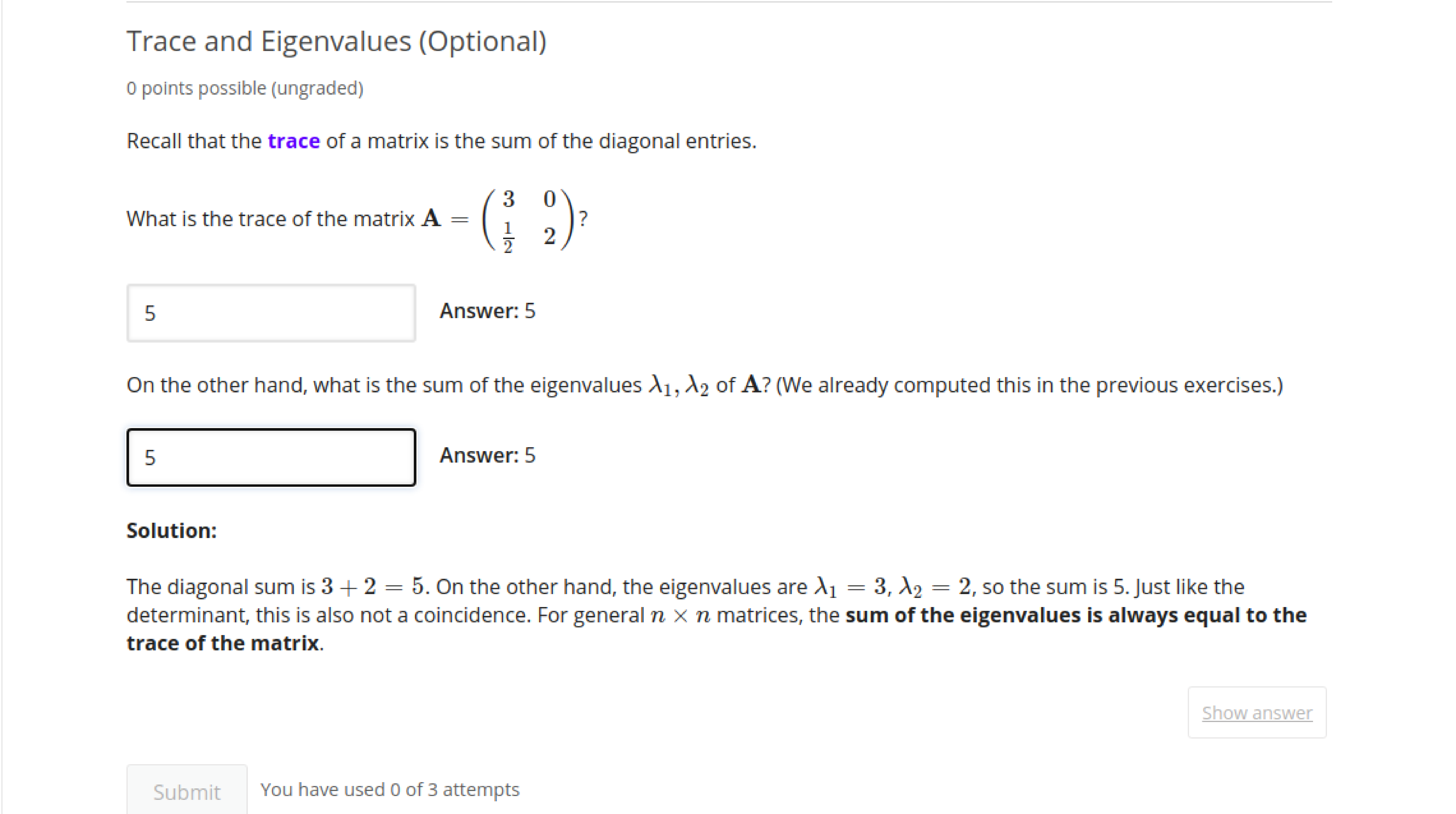
6

Answer: 6

Solution:

Plugging into the formula directly gives $3 \cdot 2 - 0 \cdot \frac{1}{2} = 6$. On the other hand, the eigenvalues are $\lambda_1 = 3$, $\lambda_2 = 2$, so the product is 6. This is not a coincidence; for general $n \times n$ matrices, the **product of the eigenvalues is always equal to the determinant**.

You have used 0 of 3 attempts



Trace and Eigenvalues (Optional)

0 points possible (ungraded)

Recall that the **trace** of a matrix is the sum of the diagonal entries.

What is the trace of the matrix $\mathbf{A} = \begin{pmatrix} 3 & 0 \\ \frac{1}{2} & 2 \end{pmatrix}$?

5

Answer: 5

On the other hand, what is the sum of the eigenvalues λ_1, λ_2 of $\mathbf{A}$? (We already computed this in the previous exercises.)

5

Answer: 5

Solution:

The diagonal sum is $3 + 2 = 5$. On the other hand, the eigenvalues are $\lambda_1 = 3$, $\lambda_2 = 2$, so the sum is 5. Just like the determinant, this is also not a coincidence. For general $n \times n$ matrices, the **sum of the eigenvalues is always equal to the trace of the matrix**.

You have used 0 of 3 attempts

If a (nonzero) vector is in the nullspace of a square matrix $\mathbf{A}$, is it an eigenvector of $\mathbf{A}$?

yes ⌄ **Answer:** yes

Which of the following are equivalent to the statement that 0 is an eigenvalue for a given square matrix $\mathbf{A}$? (Choose all that apply.)

- □ There exists a nonzero solution to $\mathbf{A}\mathbf{v} = \mathbf{0}$. ✓
- □ $\det(\mathbf{A}) = 0$ ✓
- □ $\det(\mathbf{A}) \neq 0$
- □ $\mathrm{NS}(\mathbf{A}) = \mathbf{0}$
- □ $\mathrm{NS}(\mathbf{A}) \neq \mathbf{0}$ ✓

Solution:

- If a vector $\mathbf{v}$ is in the nullspace of $\mathbf{A}$, then $\mathbf{A}\mathbf{v} = \mathbf{0} = (0)\,\mathbf{v}$. So it is an eigenvector of $\mathbf{A}$ associated to the eigenvalue 0.
- If 0 is an eigenvalue for a matrix $\mathbf{A}$, then by definition, there exists a nonzero solution to $\mathbf{A}\mathbf{v} = \mathbf{0}$; that is, $\mathrm{NS}(\mathbf{A}) \neq \mathbf{0}$ and this only happens if and only if $\det(\mathbf{A}) = 0$