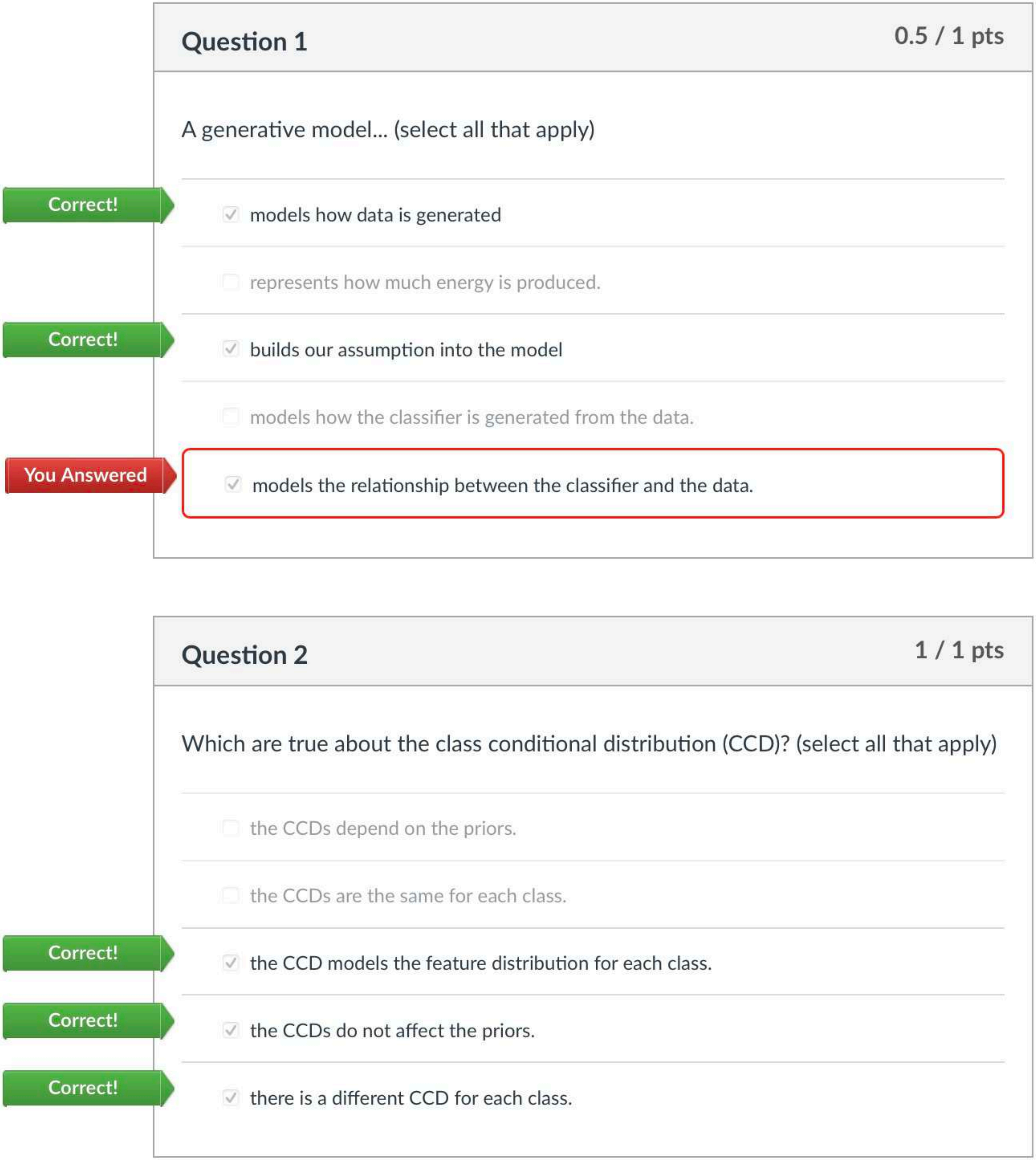


Question 1

0.5 / 1 pts

A generative model... (select all that apply)

Correct! ☑ models how data is generated

□ represents how much energy is produced.

Correct! ☑ builds our assumption into the model

□ models how the classifier is generated from the data.

You Answered ☑ models the relationship between the classifier and the data.

Question 2

1 / 1 pts

Which are true about the class conditional distribution (CCD)? (select all that apply)

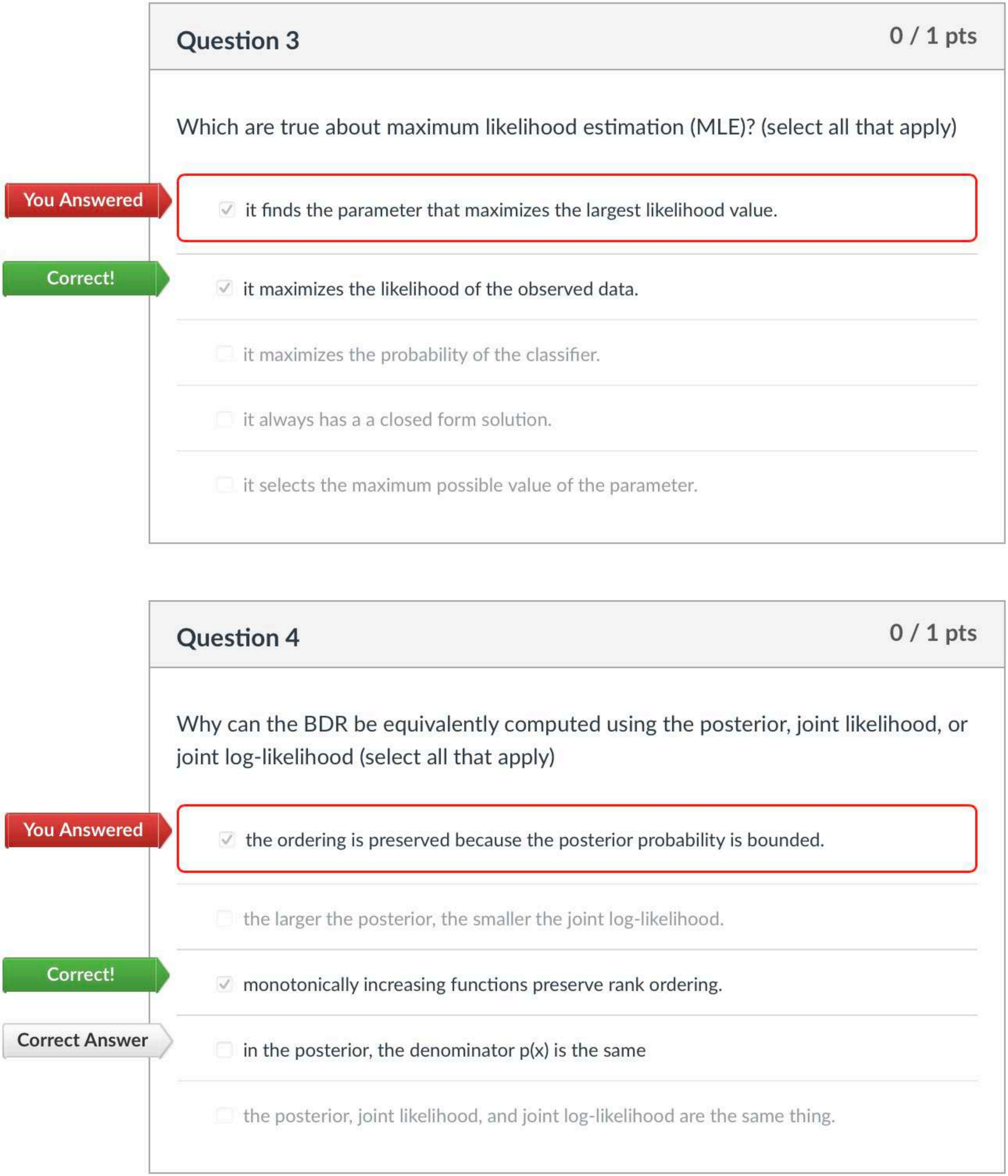
□ the CCDs depend on the priors.

□ the CCDs are the same for each class.

Correct! ☑ the CCD models the feature distribution for each class.

Correct! ☑ the CCDs do not affect the priors.

Correct! ☑ there is a different CCD for each class.



Question 3

0 / 1 pts

Which are true about maximum likelihood estimation (MLE)? (select all that apply)

You Answered — ☑ it finds the parameter that maximizes the largest likelihood value.

Correct! — ☑ it maximizes the likelihood of the observed data.

□ it maximizes the probability of the classifier.

□ it always has a a closed form solution.

□ it selects the maximum possible value of the parameter.

Question 4

0 / 1 pts

Why can the BDR be equivalently computed using the posterior, joint likelihood, or joint log-likelihood (select all that apply)

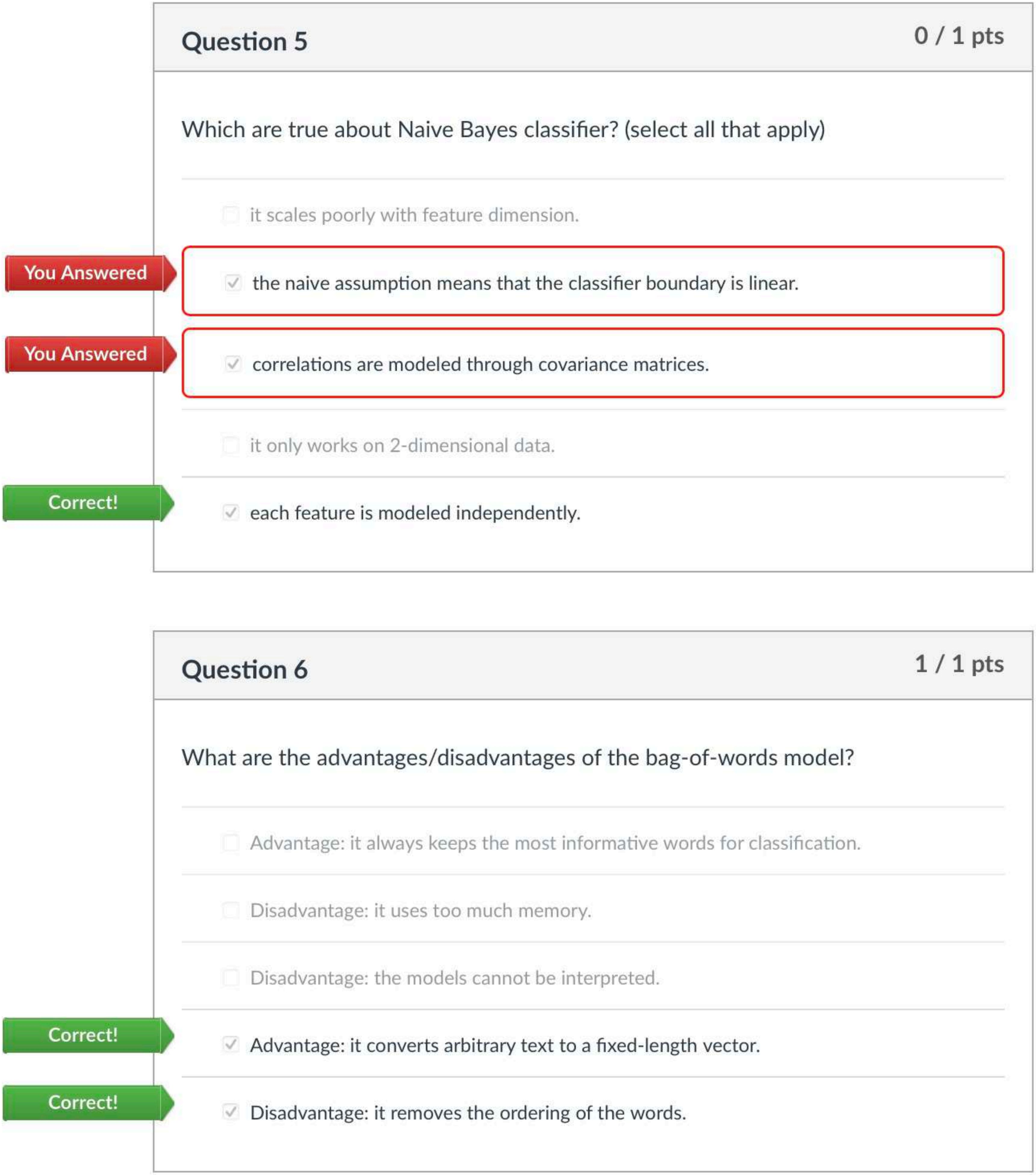
You Answered — ☑ the ordering is preserved because the posterior probability is bounded.

□ the larger the posterior, the smaller the joint log-likelihood.

Correct! — ☑ monotonically increasing functions preserve rank ordering.

Correct Answer — □ in the posterior, the denominator p(x) is the same

□ the posterior, joint likelihood, and joint log-likelihood are the same thing.



Question 5

0 / 1 pts

Which are true about Naive Bayes classifier? (select all that apply)

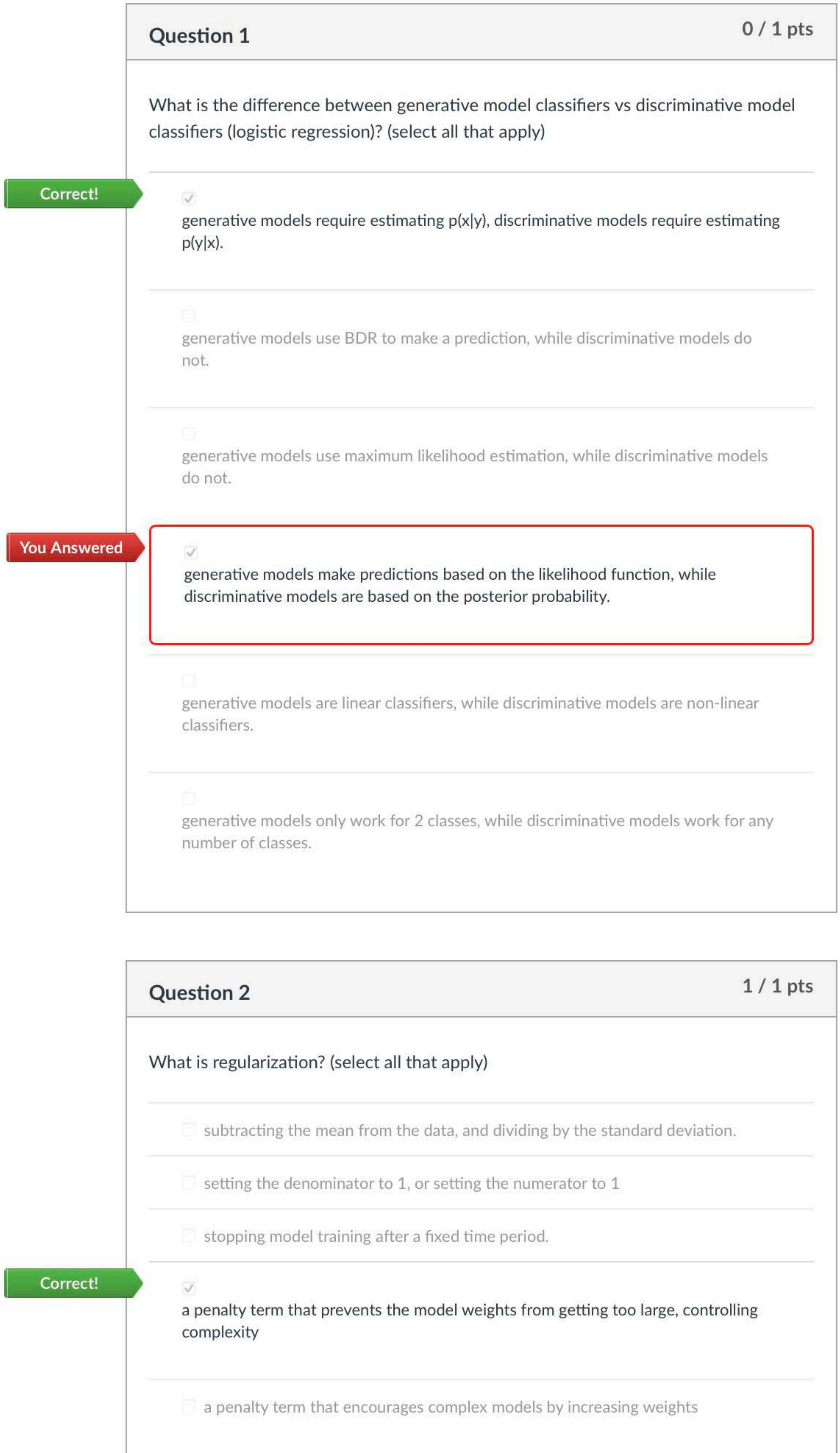
- □ it scales poorly with feature dimension.
- You Answered ☑ the naive assumption means that the classifier boundary is linear.
- You Answered ☑ correlations are modeled through covariance matrices.
- □ it only works on 2-dimensional data.
- Correct! ☑ each feature is modeled independently.

Question 6

1 / 1 pts

What are the advantages/disadvantages of the bag-of-words model?

- □ Advantage: it always keeps the most informative words for classification.
- □ Disadvantage: it uses too much memory.
- □ Disadvantage: the models cannot be interpreted.
- Correct! ☑ Advantage: it converts arbitrary text to a fixed-length vector.
- Correct! ☑ Disadvantage: it removes the ordering of the words.



Question 1

0 / 1 pts

What is the difference between generative model classifiers vs discriminative model classifiers (logistic regression)? (select all that apply)

Correct!

☑ generative models require estimating p(x|y), discriminative models require estimating p(y|x).

☐ generative models use BDR to make a prediction, while discriminative models do not.

☐ generative models use maximum likelihood estimation, while discriminative models do not.

You Answered

☑ generative models make predictions based on the likelihood function, while discriminative models are based on the posterior probability.

☐ generative models are linear classifiers, while discriminative models are non-linear classifiers.

☐ generative models only work for 2 classes, while discriminative models work for any number of classes.

Question 2

1 / 1 pts

What is regularization? (select all that apply)

☐ subtracting the mean from the data, and dividing by the standard deviation.

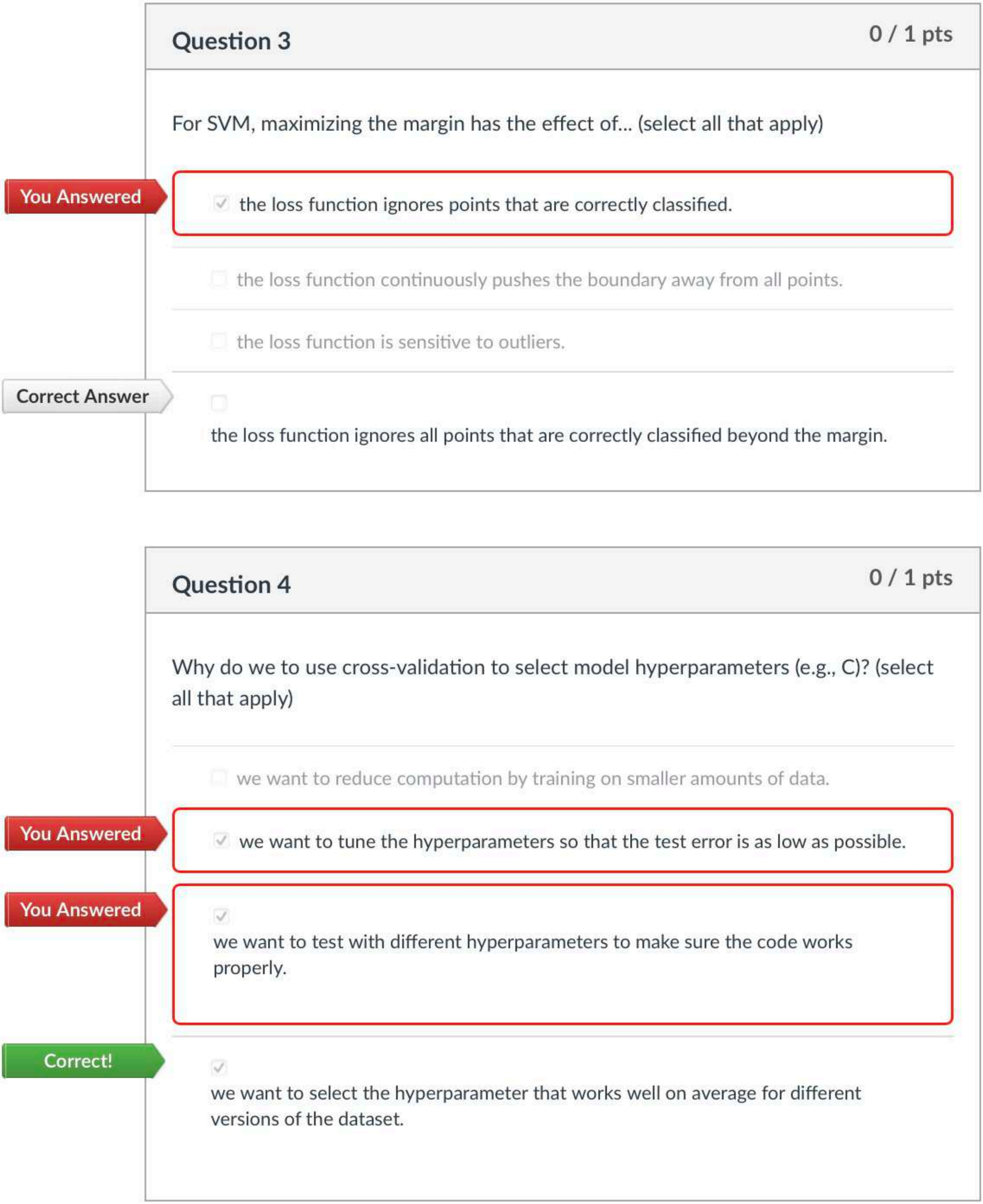
☐ setting the denominator to 1, or setting the numerator to 1

☐ stopping model training after a fixed time period.

Correct!

☑ a penalty term that prevents the model weights from getting too large, controlling complexity

☐ a penalty term that encourages complex models by increasing weights



Question 3

0 / 1 pts

For SVM, maximizing the margin has the effect of... (select all that apply)

You Answered

☑ the loss function ignores points that are correctly classified.

□ the loss function continuously pushes the boundary away from all points.

□ the loss function is sensitive to outliers.

Correct Answer

□ the loss function ignores all points that are correctly classified beyond the margin.

Question 4

0 / 1 pts

Why do we to use cross-validation to select model hyperparameters (e.g., C)? (select all that apply)

□ we want to reduce computation by training on smaller amounts of data.

You Answered

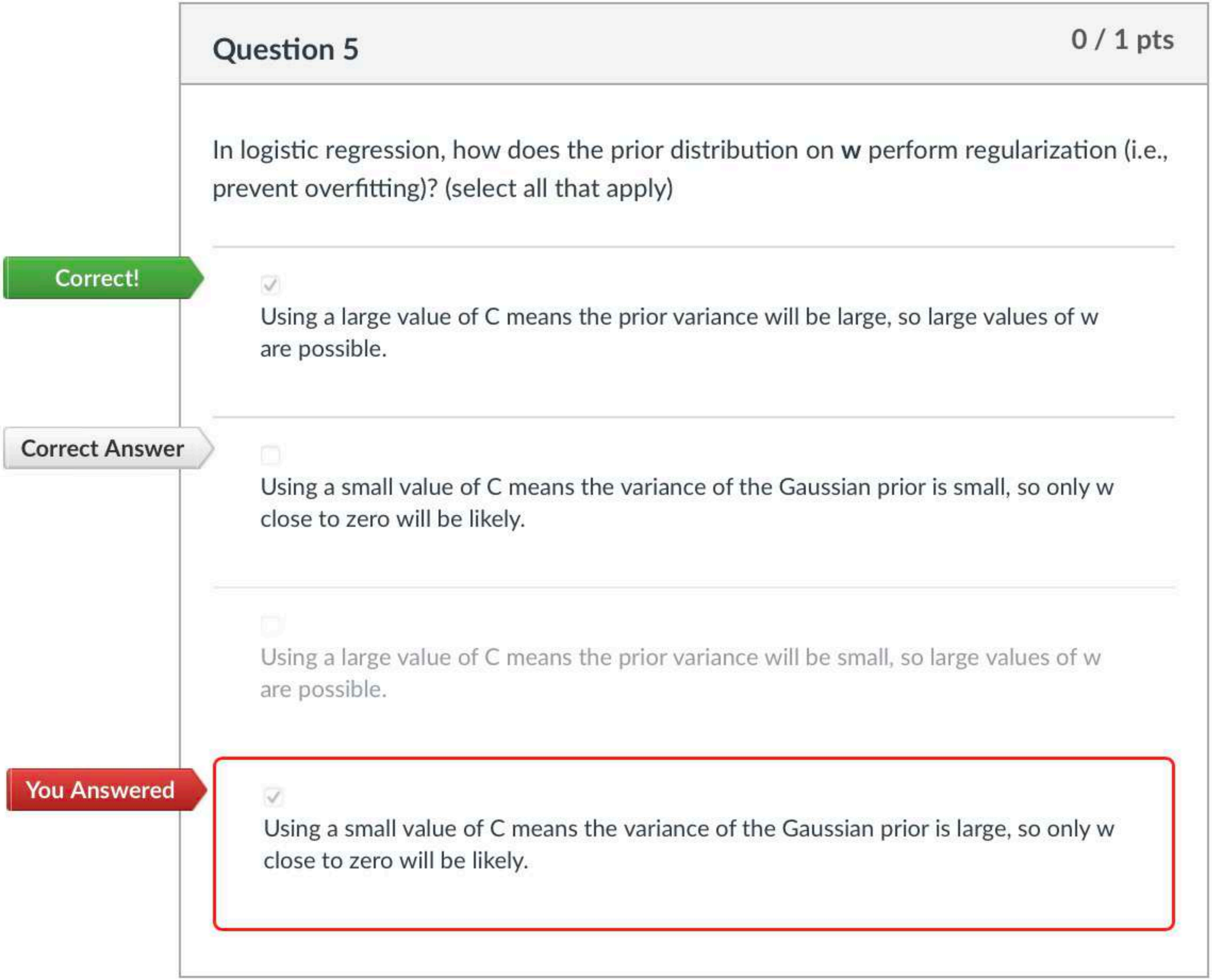
☑ we want to tune the hyperparameters so that the test error is as low as possible.

You Answered

☑ we want to test with different hyperparameters to make sure the code works properly.

Correct!

☑ we want to select the hyperparameter that works well on average for different versions of the dataset.



Question 5

0 / 1 pts

In logistic regression, how does the prior distribution on **w** perform regularization (i.e., prevent overfitting)? (select all that apply)

Correct!

☑ Using a large value of C means the prior variance will be large, so large values of w are possible.

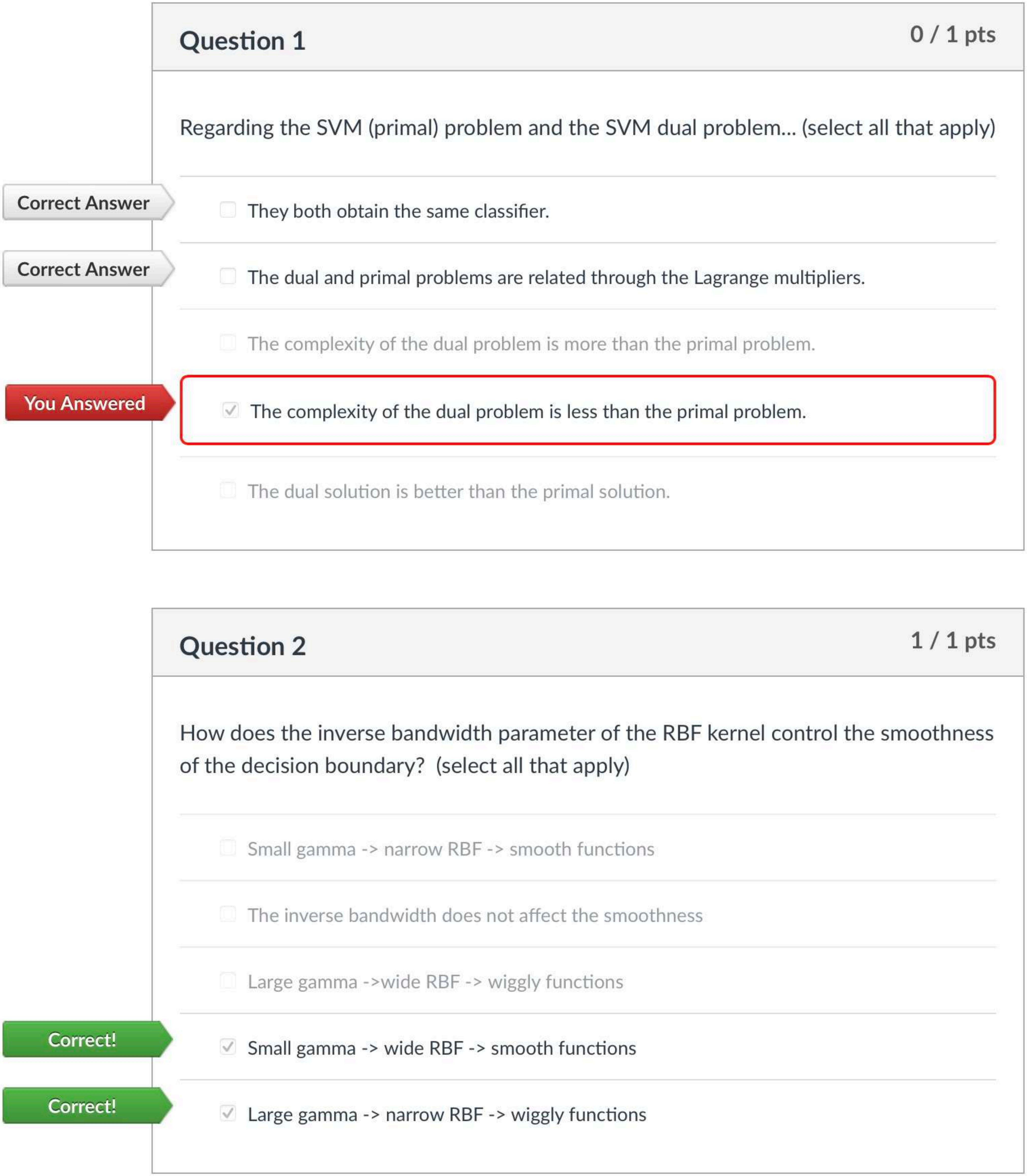
Correct Answer

□ Using a small value of C means the variance of the Gaussian prior is small, so only w close to zero will be likely.

□ Using a large value of C means the prior variance will be small, so large values of w are possible.

You Answered

☑ Using a small value of C means the variance of the Gaussian prior is large, so only w close to zero will be likely.



Question 1

0 / 1 pts

Regarding the SVM (primal) problem and the SVM dual problem... (select all that apply)

Correct Answer ☐ They both obtain the same classifier.

Correct Answer ☐ The dual and primal problems are related through the Lagrange multipliers.

☐ The complexity of the dual problem is more than the primal problem.

You Answered ☑ The complexity of the dual problem is less than the primal problem.

☐ The dual solution is better than the primal solution.

Question 2

1 / 1 pts

How does the inverse bandwidth parameter of the RBF kernel control the smoothness of the decision boundary? (select all that apply)

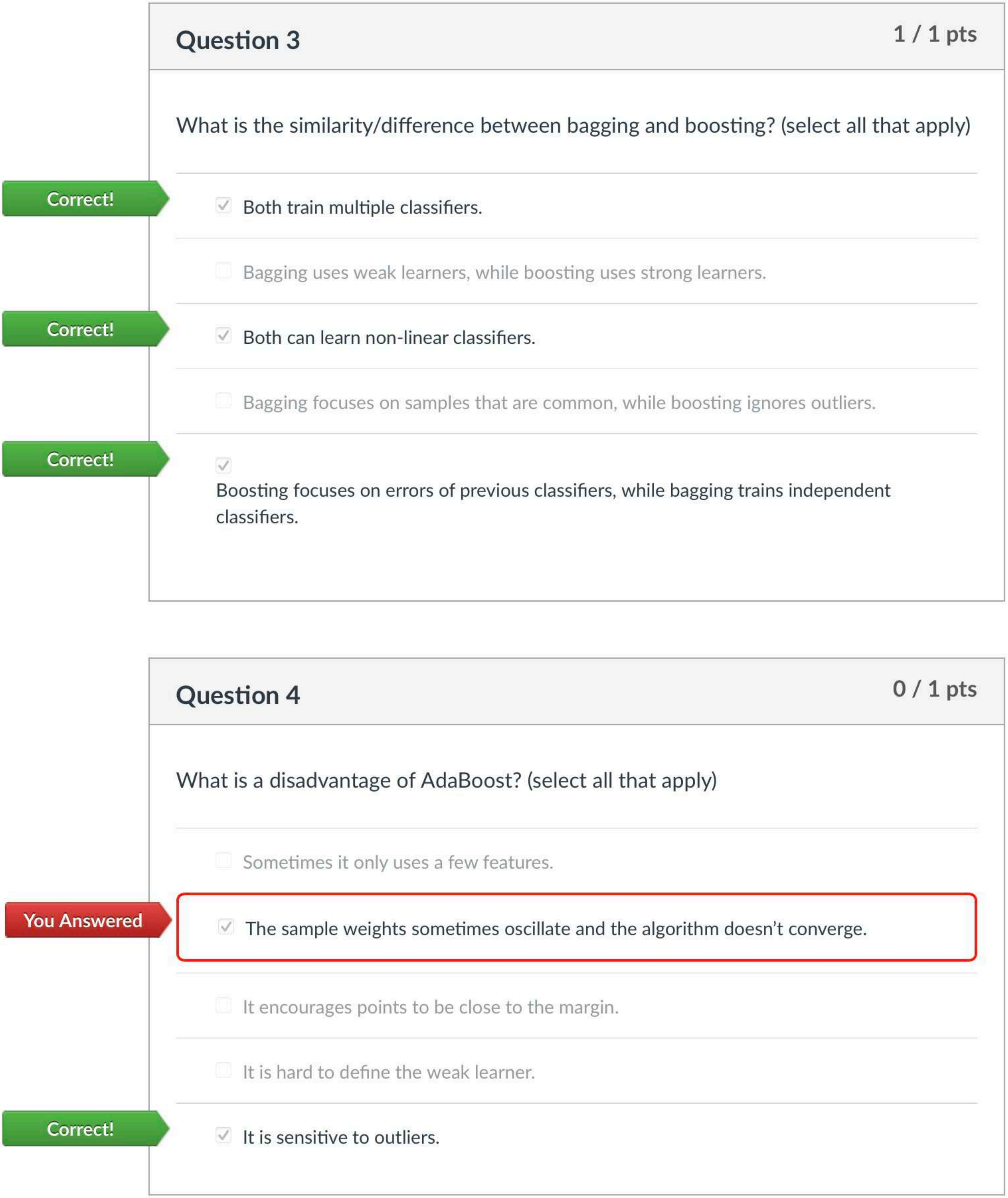
☐ Small gamma -> narrow RBF -> smooth functions

☐ The inverse bandwidth does not affect the smoothness

☐ Large gamma ->wide RBF -> wiggly functions

Correct! ☑ Small gamma -> wide RBF -> smooth functions

Correct! ☑ Large gamma -> narrow RBF -> wiggly functions



Question 3

1 / 1 pts

What is the similarity/difference between bagging and boosting? (select all that apply)

Correct! ☑ Both train multiple classifiers.

□ Bagging uses weak learners, while boosting uses strong learners.

Correct! ☑ Both can learn non-linear classifiers.

□ Bagging focuses on samples that are common, while boosting ignores outliers.

Correct! ☑ Boosting focuses on errors of previous classifiers, while bagging trains independent classifiers.

Question 4

0 / 1 pts

What is a disadvantage of AdaBoost? (select all that apply)

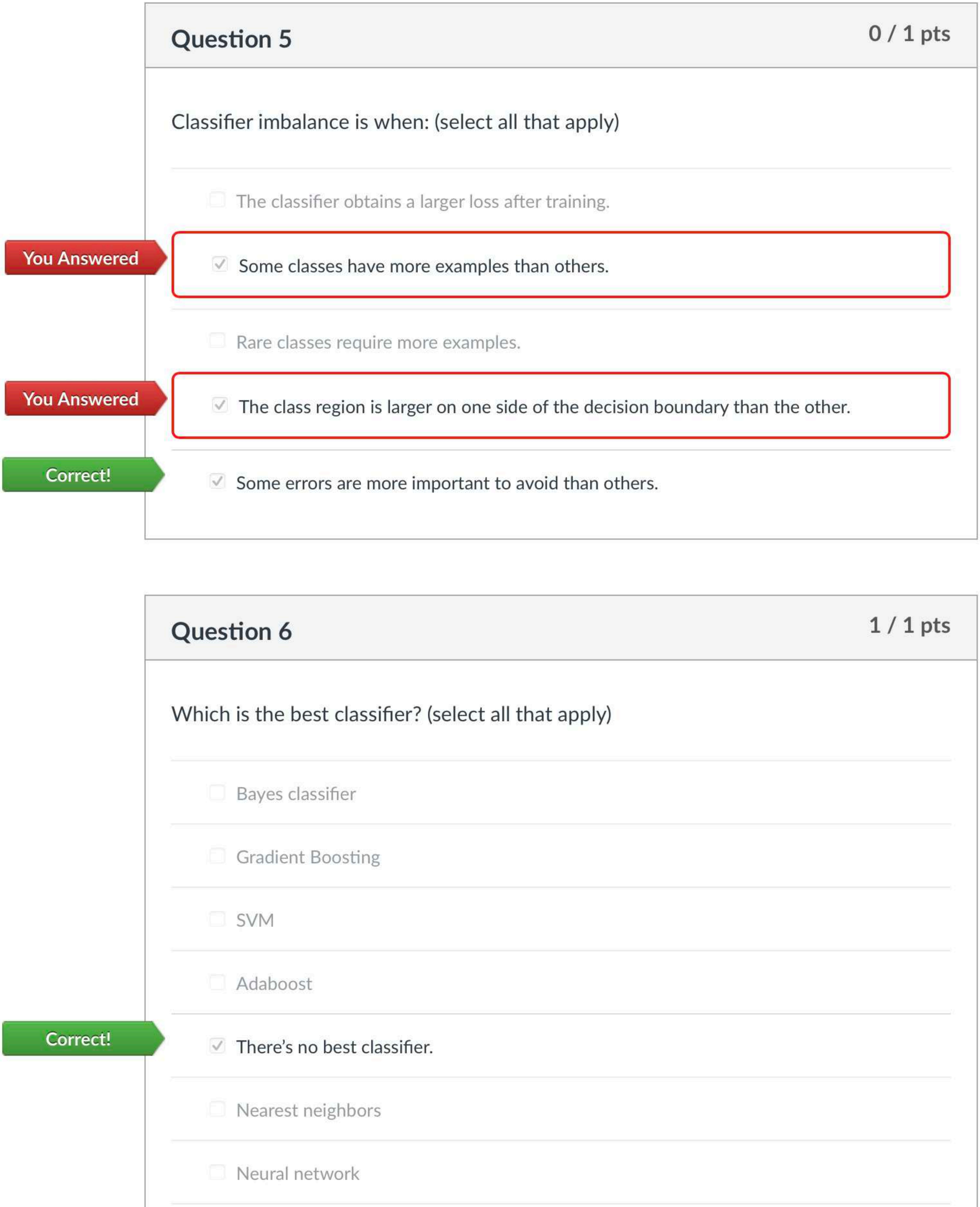
□ Sometimes it only uses a few features.

You Answered ☑ The sample weights sometimes oscillate and the algorithm doesn’t converge.

□ It encourages points to be close to the margin.

□ It is hard to define the weak learner.

Correct! ☑ It is sensitive to outliers.



Question 5

0 / 1 pts

Classifier imbalance is when: (select all that apply)

□ The classifier obtains a larger loss after training.

You Answered ☑ Some classes have more examples than others.

□ Rare classes require more examples.

You Answered ☑ The class region is larger on one side of the decision boundary than the other.

Correct! ☑ Some errors are more important to avoid than others.

Question 6

1 / 1 pts

Which is the best classifier? (select all that apply)

□ Bayes classifier

□ Gradient Boosting

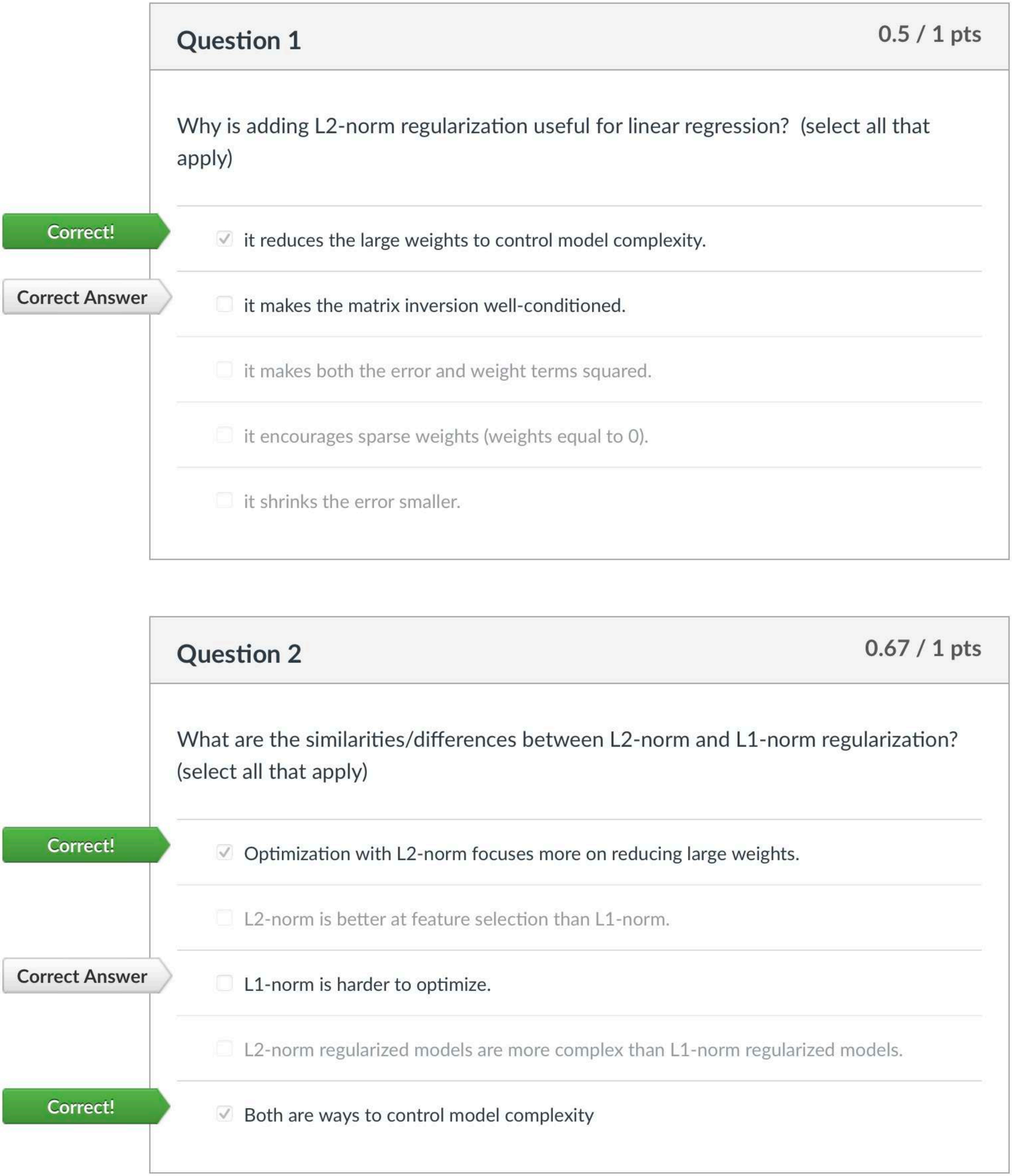
□ SVM

□ Adaboost

Correct! ☑ There's no best classifier.

□ Nearest neighbors

□ Neural network



Question 1

0.5 / 1 pts

Why is adding L2-norm regularization useful for linear regression? (select all that apply)

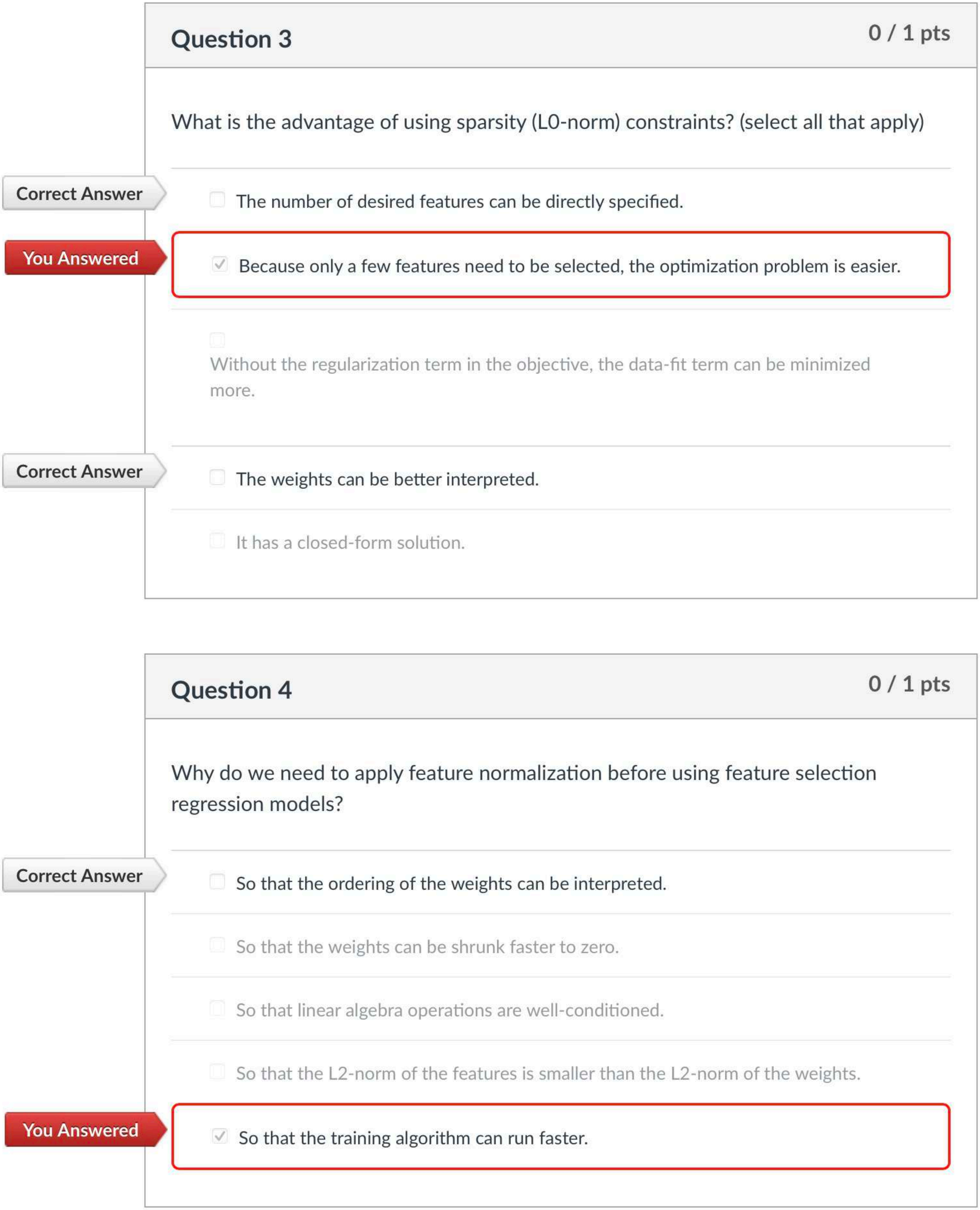
- Correct! ☑ it reduces the large weights to control model complexity.
- Correct Answer □ it makes the matrix inversion well-conditioned.
- □ it makes both the error and weight terms squared.
- □ it encourages sparse weights (weights equal to 0).
- □ it shrinks the error smaller.

Question 2

0.67 / 1 pts

What are the similarities/differences between L2-norm and L1-norm regularization? (select all that apply)

- Correct! ☑ Optimization with L2-norm focuses more on reducing large weights.
- □ L2-norm is better at feature selection than L1-norm.
- Correct Answer □ L1-norm is harder to optimize.
- □ L2-norm regularized models are more complex than L1-norm regularized models.
- Correct! ☑ Both are ways to control model complexity



Question 3

0 / 1 pts

What is the advantage of using sparsity (L0-norm) constraints? (select all that apply)

Correct Answer

□ The number of desired features can be directly specified.

You Answered

☑ Because only a few features need to be selected, the optimization problem is easier.

□ Without the regularization term in the objective, the data-fit term can be minimized more.

Correct Answer

□ The weights can be better interpreted.

□ It has a closed-form solution.

Question 4

0 / 1 pts

Why do we need to apply feature normalization before using feature selection regression models?

Correct Answer

□ So that the ordering of the weights can be interpreted.

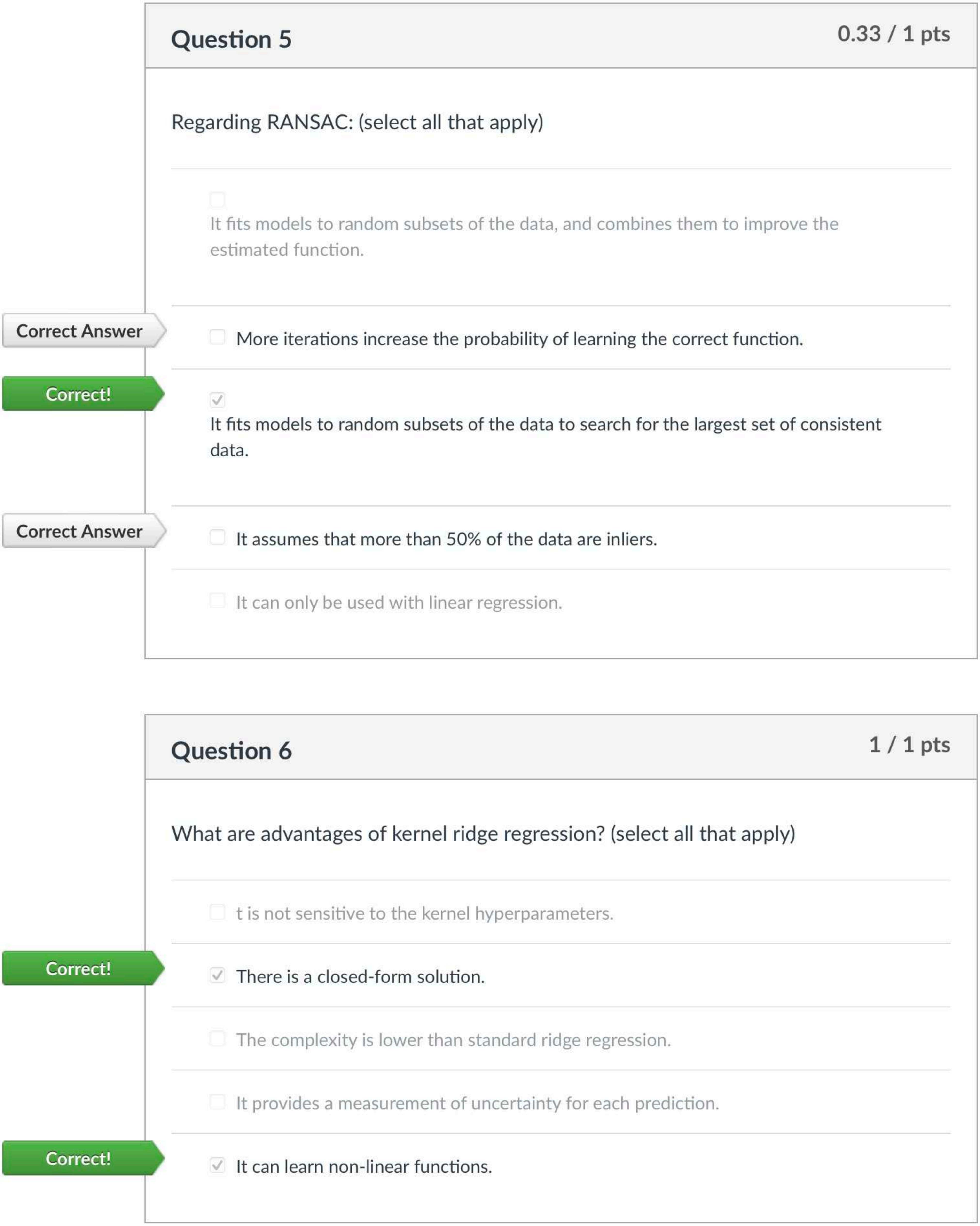
□ So that the weights can be shrunk faster to zero.

□ So that linear algebra operations are well-conditioned.

□ So that the L2-norm of the features is smaller than the L2-norm of the weights.

You Answered

☑ So that the training algorithm can run faster.



Question 5

0.33 / 1 pts

Regarding RANSAC: (select all that apply)

□ It fits models to random subsets of the data, and combines them to improve the estimated function.

Correct Answer □ More iterations increase the probability of learning the correct function.

Correct! ☑ It fits models to random subsets of the data to search for the largest set of consistent data.

Correct Answer □ It assumes that more than 50% of the data are inliers.

□ It can only be used with linear regression.

Question 6

1 / 1 pts

What are advantages of kernel ridge regression? (select all that apply)

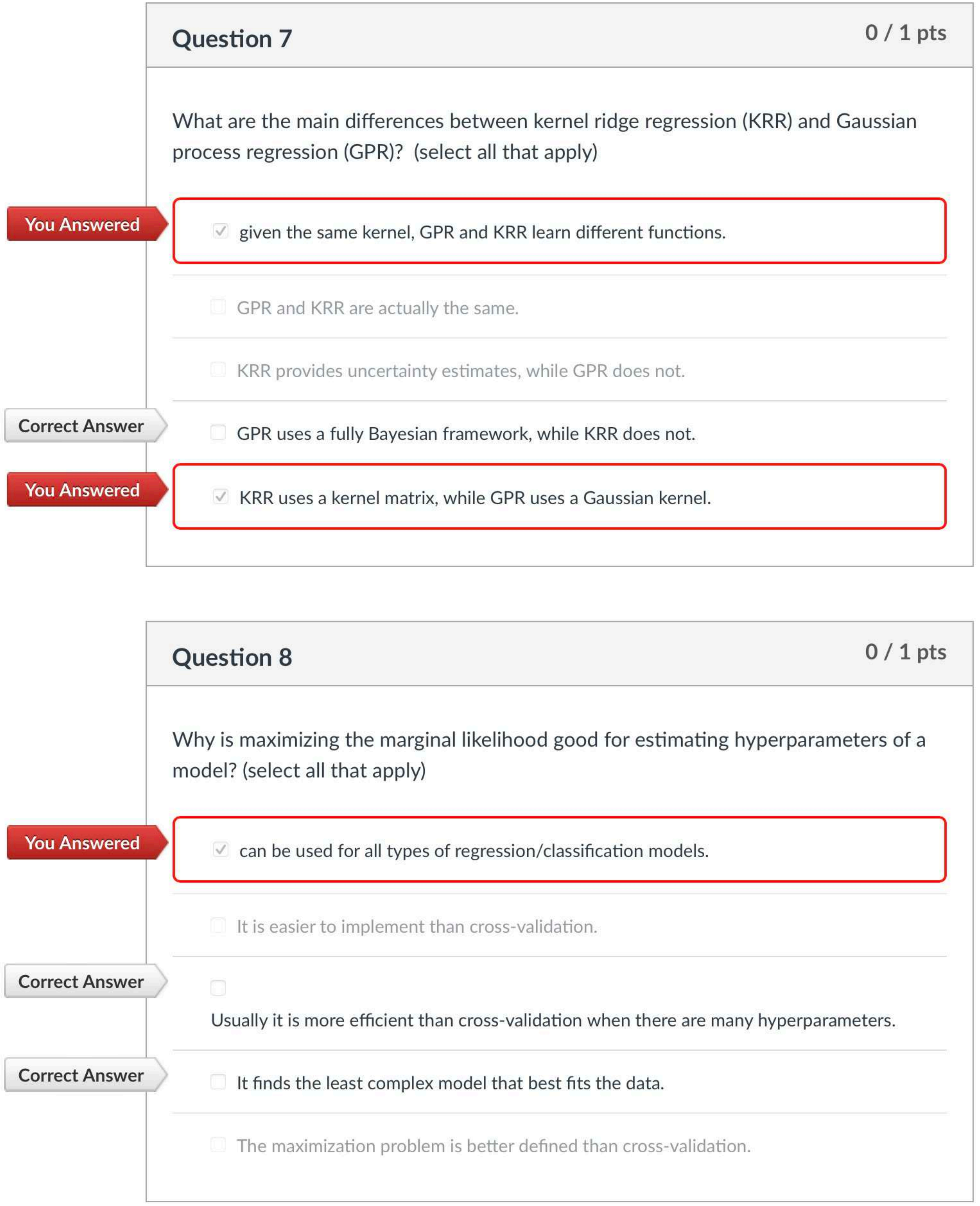
□ t is not sensitive to the kernel hyperparameters.

Correct! ☑ There is a closed-form solution.

□ The complexity is lower than standard ridge regression.

□ It provides a measurement of uncertainty for each prediction.

Correct! ☑ It can learn non-linear functions.



Question 7

0 / 1 pts

What are the main differences between kernel ridge regression (KRR) and Gaussian process regression (GPR)? (select all that apply)

You Answered ☑ given the same kernel, GPR and KRR learn different functions.

☐ GPR and KRR are actually the same.

☐ KRR provides uncertainty estimates, while GPR does not.

Correct Answer ☐ GPR uses a fully Bayesian framework, while KRR does not.

You Answered ☑ KRR uses a kernel matrix, while GPR uses a Gaussian kernel.

Question 8

0 / 1 pts

Why is maximizing the marginal likelihood good for estimating hyperparameters of a model? (select all that apply)

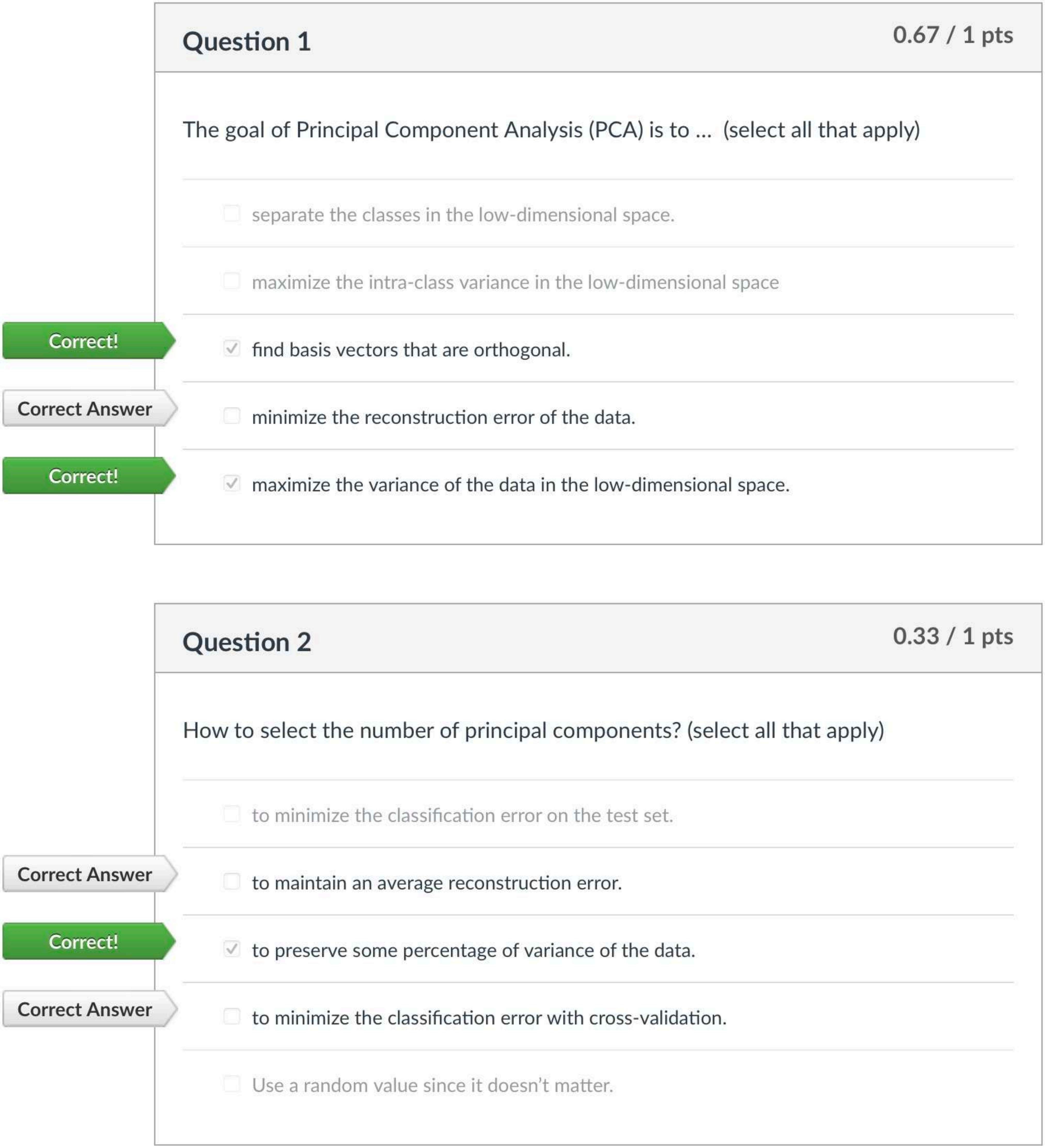
You Answered ☑ can be used for all types of regression/classification models.

☐ It is easier to implement than cross-validation.

Correct Answer ☐ Usually it is more efficient than cross-validation when there are many hyperparameters.

Correct Answer ☐ It finds the least complex model that best fits the data.

☐ The maximization problem is better defined than cross-validation.



Question 1

0.67 / 1 pts

The goal of Principal Component Analysis (PCA) is to ... (select all that apply)

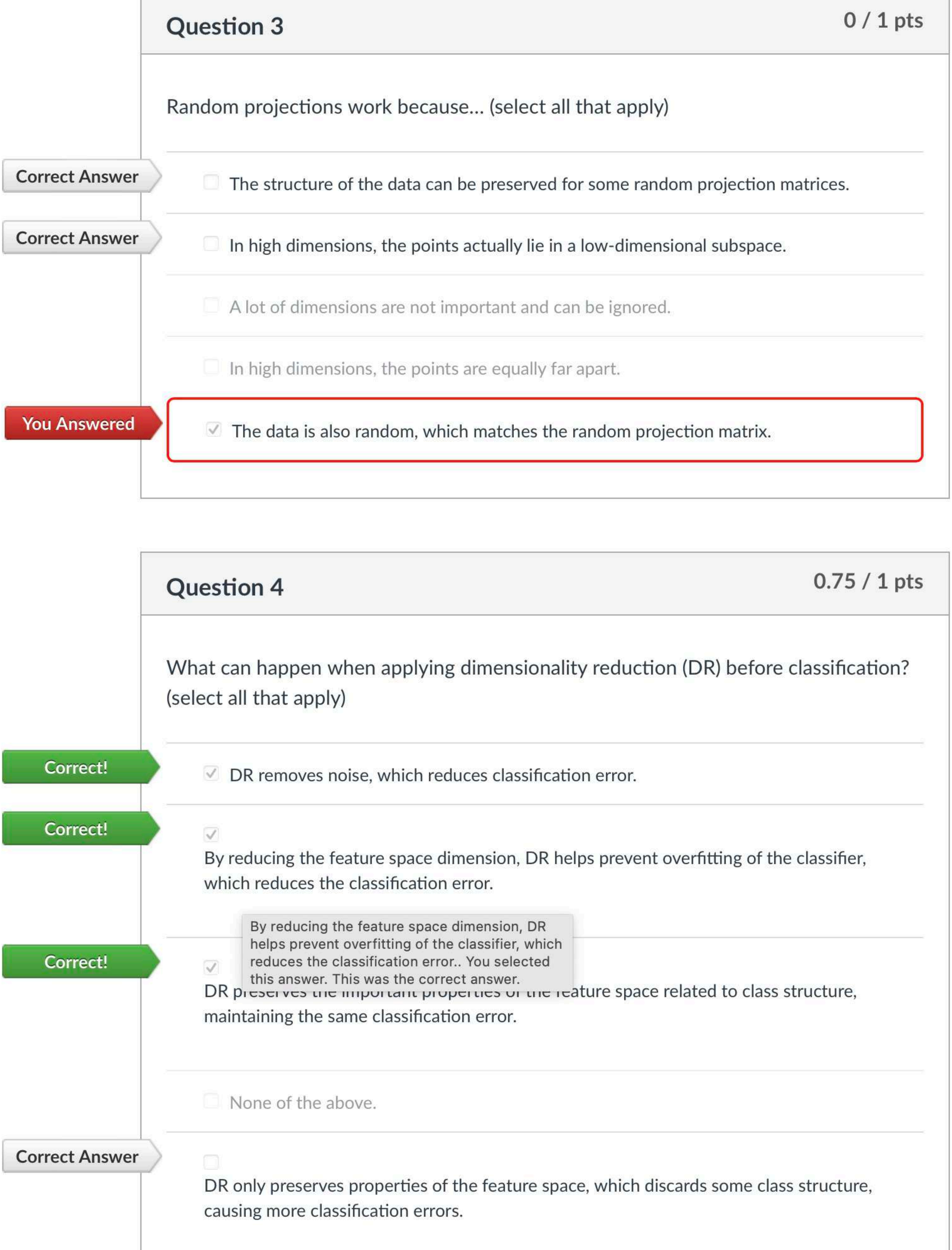
- □ separate the classes in the low-dimensional space.
- □ maximize the intra-class variance in the low-dimensional space
- Correct! ☑ find basis vectors that are orthogonal.
- Correct Answer □ minimize the reconstruction error of the data.
- Correct! ☑ maximize the variance of the data in the low-dimensional space.

Question 2

0.33 / 1 pts

How to select the number of principal components? (select all that apply)

- □ to minimize the classification error on the test set.
- Correct Answer □ to maintain an average reconstruction error.
- Correct! ☑ to preserve some percentage of variance of the data.
- Correct Answer □ to minimize the classification error with cross-validation.
- □ Use a random value since it doesn't matter.



Question 3

0 / 1 pts

Random projections work because... (select all that apply)

Correct Answer □ The structure of the data can be preserved for some random projection matrices.

Correct Answer □ In high dimensions, the points actually lie in a low-dimensional subspace.

□ A lot of dimensions are not important and can be ignored.

□ In high dimensions, the points are equally far apart.

You Answered ☑ The data is also random, which matches the random projection matrix.

Question 4

0.75 / 1 pts

What can happen when applying dimensionality reduction (DR) before classification? (select all that apply)

Correct! ☑ DR removes noise, which reduces classification error.

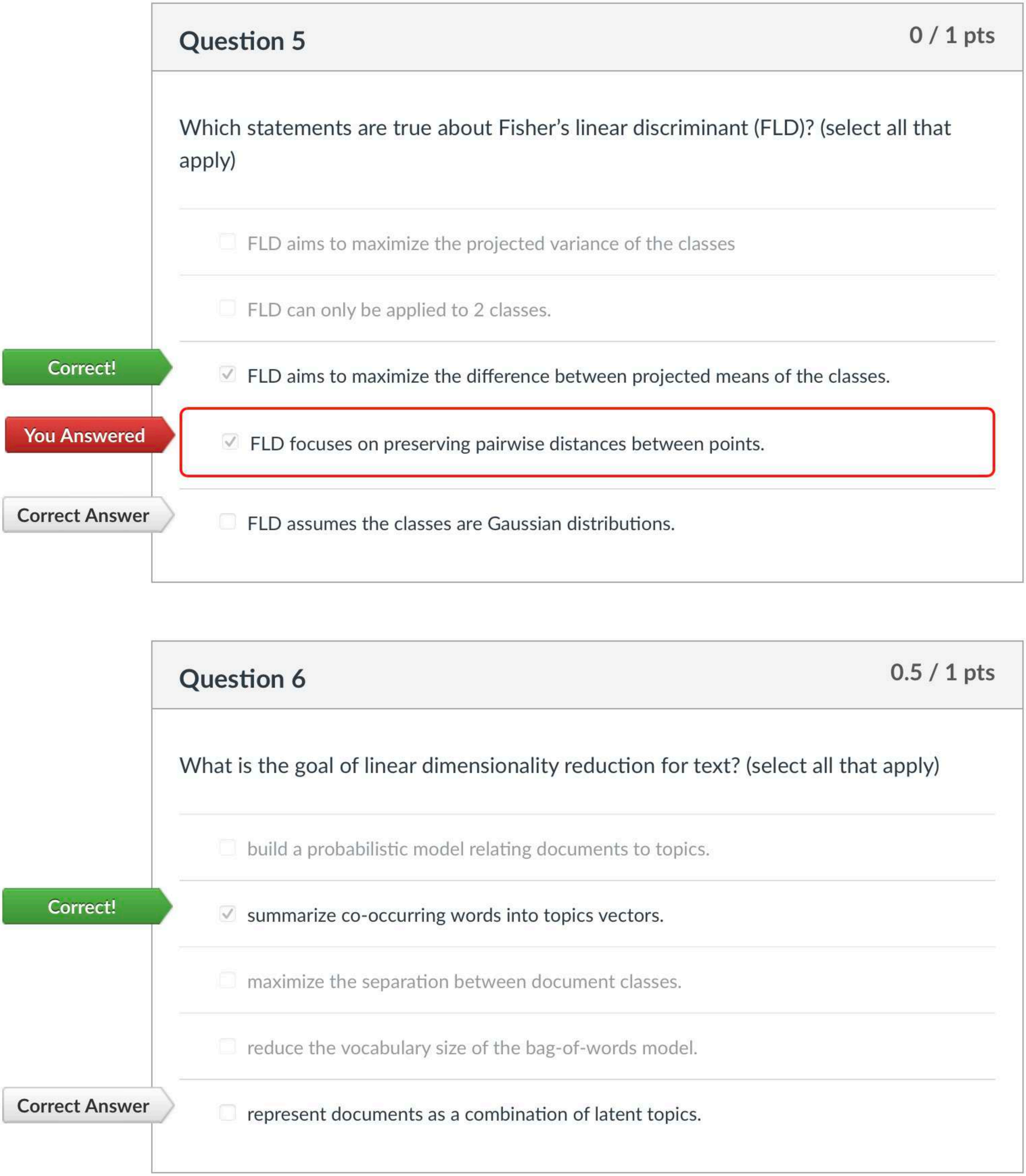
Correct! ☑ By reducing the feature space dimension, DR helps prevent overfitting of the classifier, which reduces the classification error.

By reducing the feature space dimension, DR helps prevent overfitting of the classifier, which reduces the classification error.. You selected this answer. This was the correct answer.

Correct! ☑ DR preserves the important properties of the feature space related to class structure, maintaining the same classification error.

□ None of the above.

Correct Answer □ DR only preserves properties of the feature space, which discards some class structure, causing more classification errors.



Question 5

0 / 1 pts

Which statements are true about Fisher's linear discriminant (FLD)? (select all that apply)

☐ FLD aims to maximize the projected variance of the classes

☐ FLD can only be applied to 2 classes.

Correct! ☑ FLD aims to maximize the difference between projected means of the classes.

You Answered ☑ FLD focuses on preserving pairwise distances between points.

Correct Answer ☐ FLD assumes the classes are Gaussian distributions.

Question 6

0.5 / 1 pts

What is the goal of linear dimensionality reduction for text? (select all that apply)

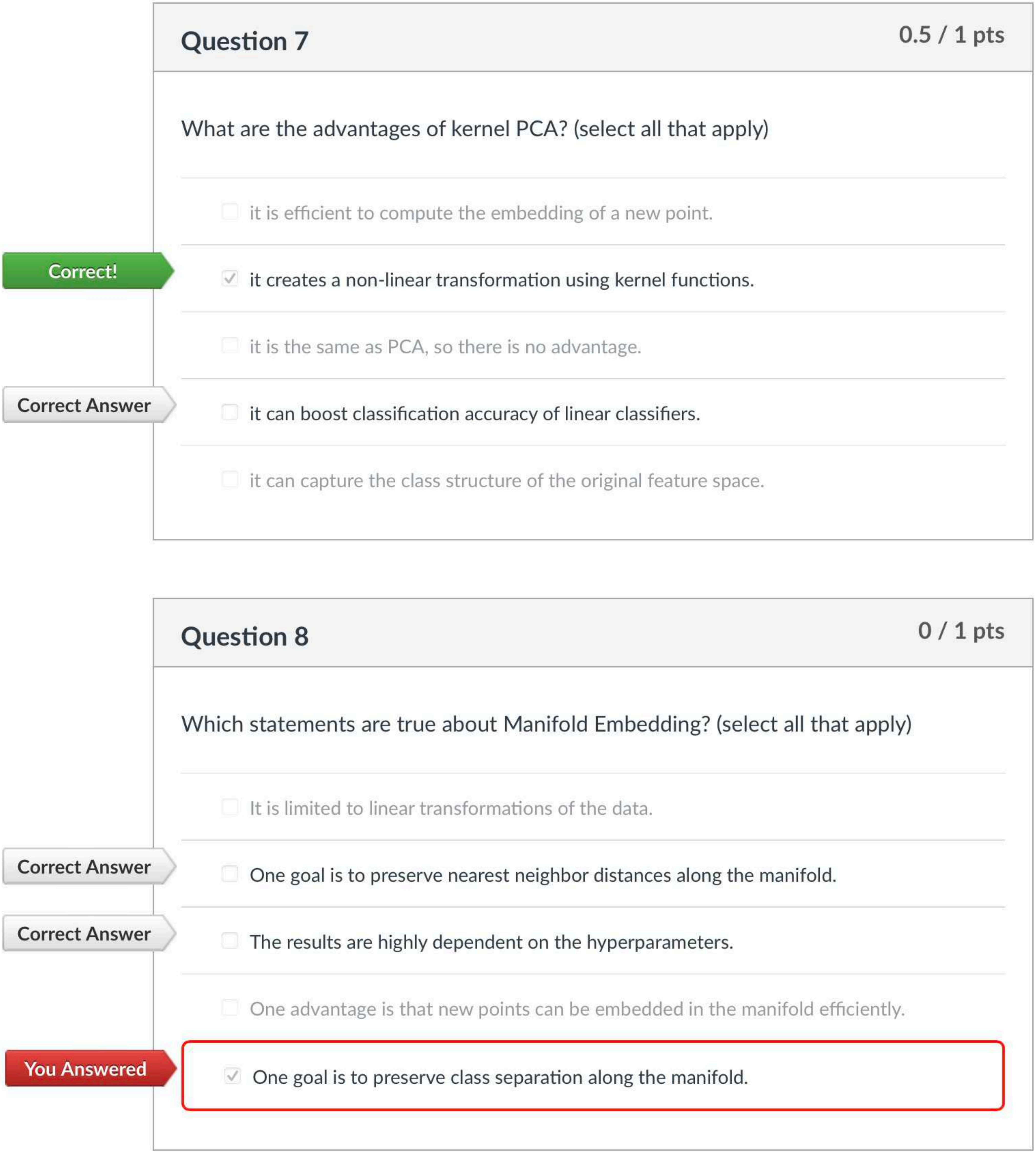
☐ build a probabilistic model relating documents to topics.

Correct! ☑ summarize co-occurring words into topics vectors.

☐ maximize the separation between document classes.

☐ reduce the vocabulary size of the bag-of-words model.

Correct Answer ☐ represent documents as a combination of latent topics.



Question 7

0.5 / 1 pts

What are the advantages of kernel PCA? (select all that apply)

□ it is efficient to compute the embedding of a new point.

Correct!

☑ it creates a non-linear transformation using kernel functions.

□ it is the same as PCA, so there is no advantage.

Correct Answer

□ it can boost classification accuracy of linear classifiers.

□ it can capture the class structure of the original feature space.

Question 8

0 / 1 pts

Which statements are true about Manifold Embedding? (select all that apply)

□ It is limited to linear transformations of the data.

Correct Answer

□ One goal is to preserve nearest neighbor distances along the manifold.

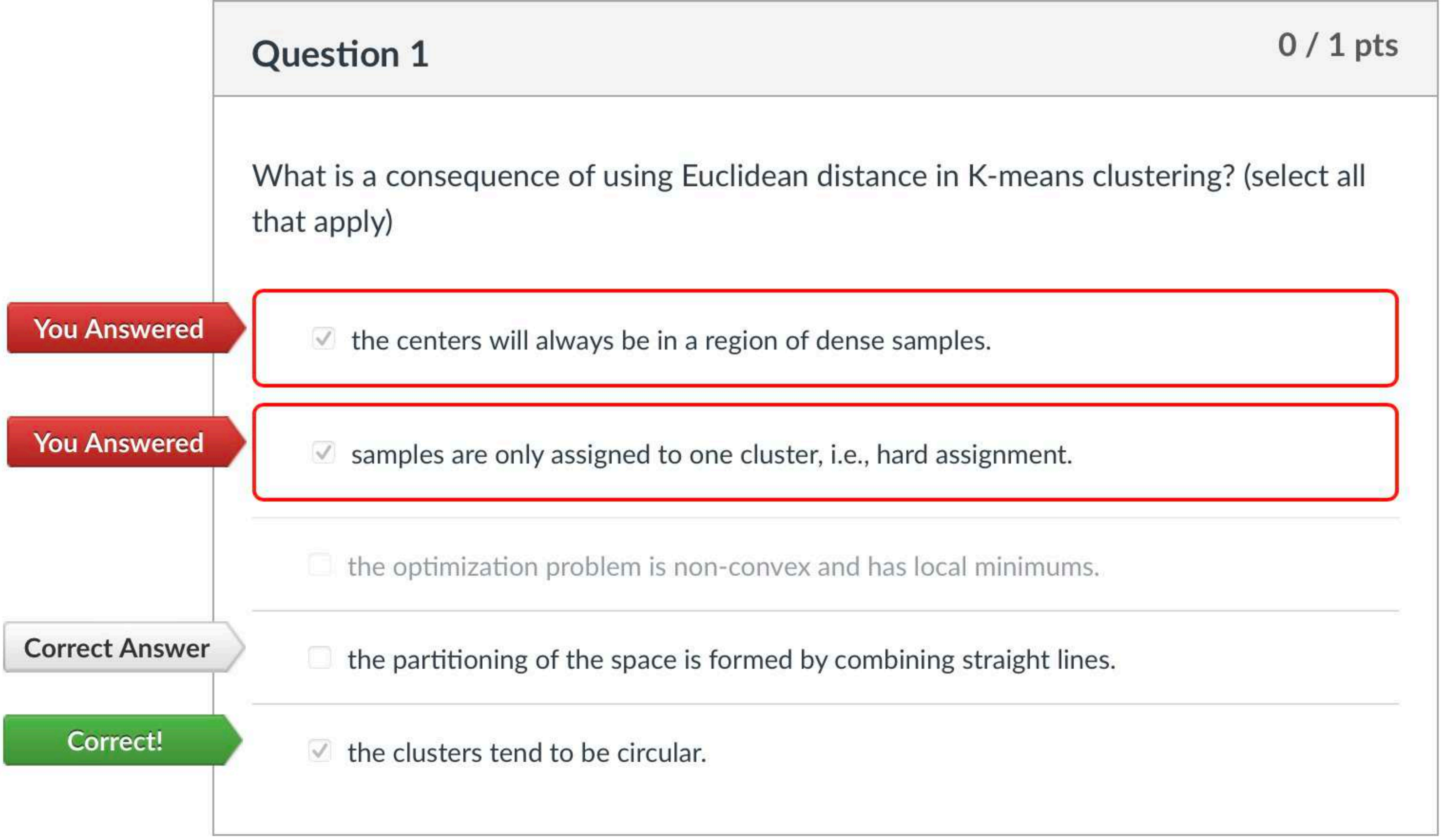
Correct Answer

□ The results are highly dependent on the hyperparameters.

□ One advantage is that new points can be embedded in the manifold efficiently.

You Answered

☑ One goal is to preserve class separation along the manifold.



Question 1

0 / 1 pts

What is a consequence of using Euclidean distance in K-means clustering? (select all that apply)

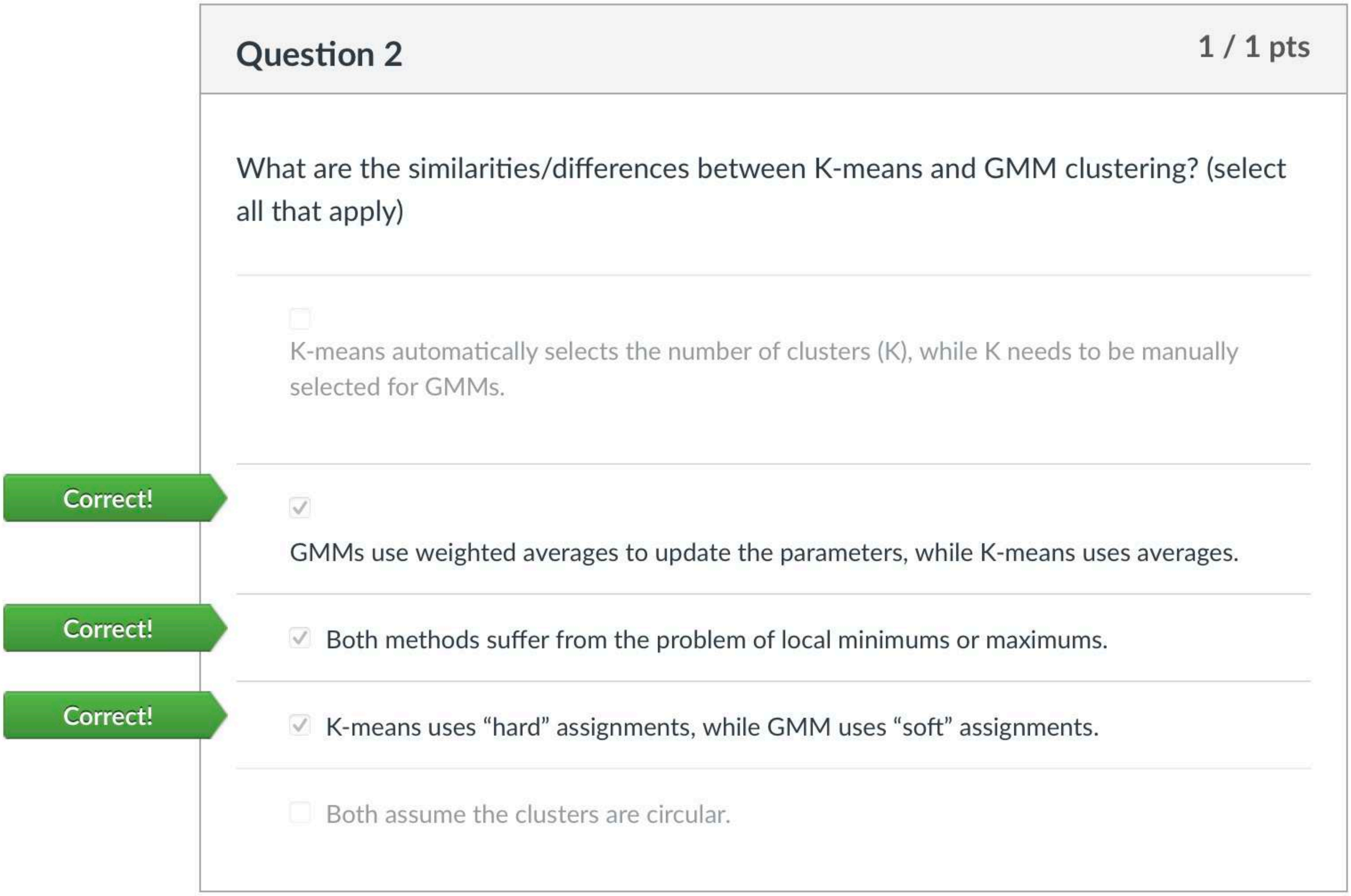
You Answered ☑ the centers will always be in a region of dense samples.

You Answered ☑ samples are only assigned to one cluster, i.e., hard assignment.

□ the optimization problem is non-convex and has local minimums.

Correct Answer □ the partitioning of the space is formed by combining straight lines.

Correct! ☑ the clusters tend to be circular.



Question 2

1 / 1 pts

What are the similarities/differences between K-means and GMM clustering? (select all that apply)

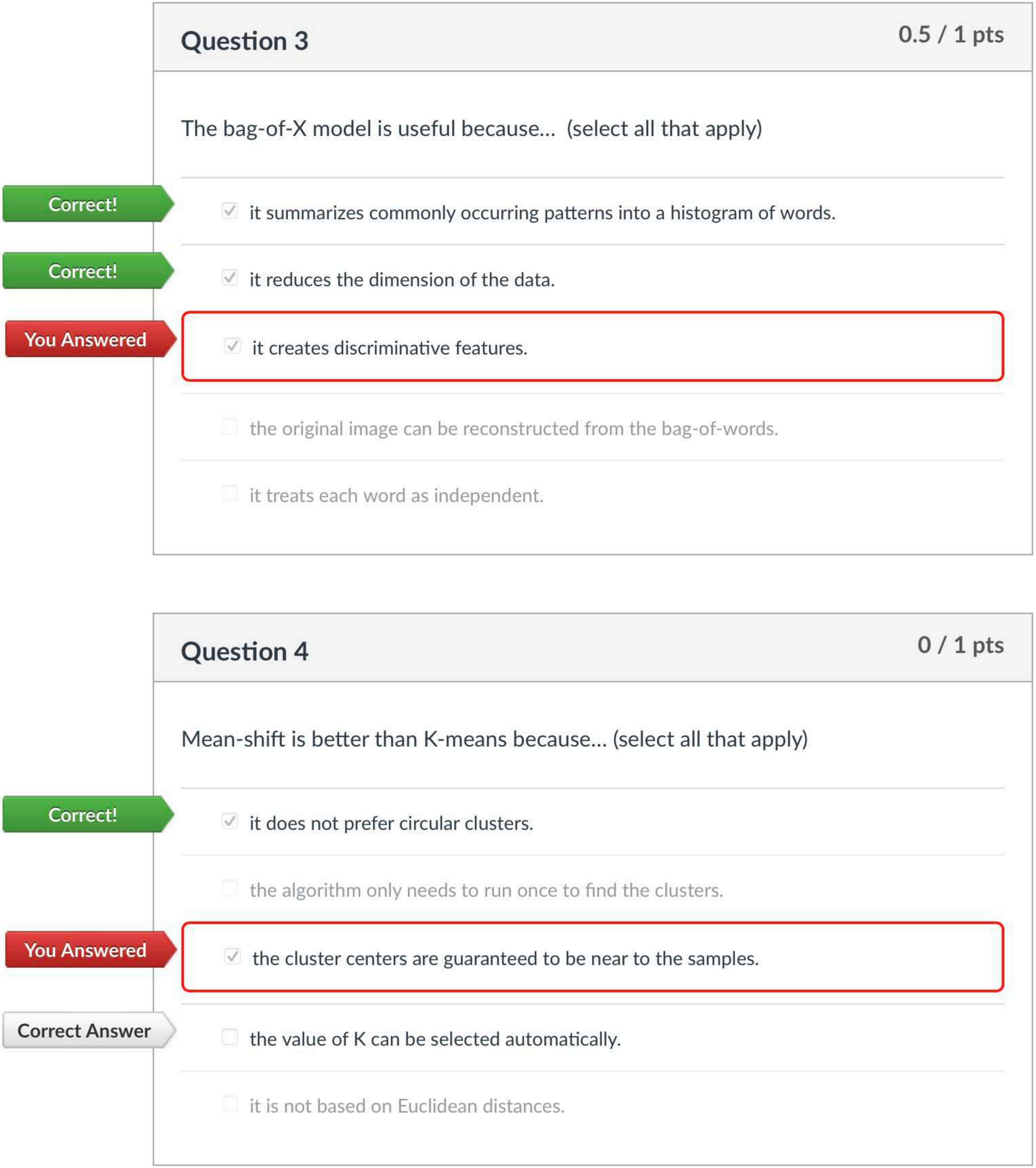
□ K-means automatically selects the number of clusters (K), while K needs to be manually selected for GMMs.

Correct! ☑ GMMs use weighted averages to update the parameters, while K-means uses averages.

Correct! ☑ Both methods suffer from the problem of local minimums or maximums.

Correct! ☑ K-means uses “hard” assignments, while GMM uses “soft” assignments.

□ Both assume the clusters are circular.



Question 3

0.5 / 1 pts

The bag-of-X model is useful because... (select all that apply)

Correct! ☑ it summarizes commonly occurring patterns into a histogram of words.

Correct! ☑ it reduces the dimension of the data.

You Answered ☑ it creates discriminative features.

☐ the original image can be reconstructed from the bag-of-words.

☐ it treats each word as independent.

Question 4

0 / 1 pts

Mean-shift is better than K-means because... (select all that apply)

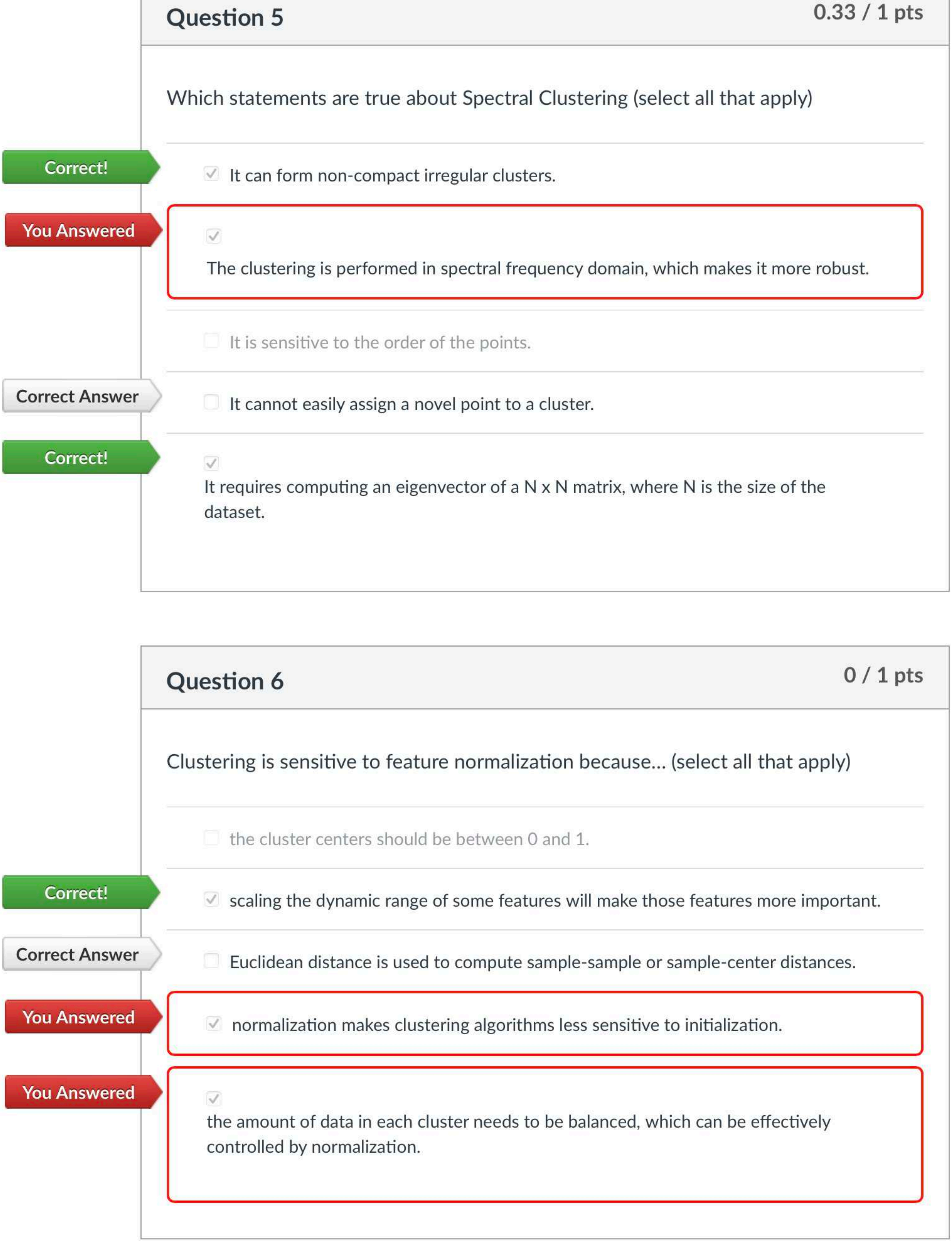
Correct! ☑ it does not prefer circular clusters.

☐ the algorithm only needs to run once to find the clusters.

You Answered ☑ the cluster centers are guaranteed to be near to the samples.

Correct Answer ☐ the value of K can be selected automatically.

☐ it is not based on Euclidean distances.



Question 5

0.33 / 1 pts

Which statements are true about Spectral Clustering (select all that apply)

Correct! ☑ It can form non-compact irregular clusters.

You Answered ☑ The clustering is performed in spectral frequency domain, which makes it more robust.

□ It is sensitive to the order of the points.

Correct Answer □ It cannot easily assign a novel point to a cluster.

Correct! ☑ It requires computing an eigenvector of a N x N matrix, where N is the size of the dataset.

Question 6

0 / 1 pts

Clustering is sensitive to feature normalization because... (select all that apply)

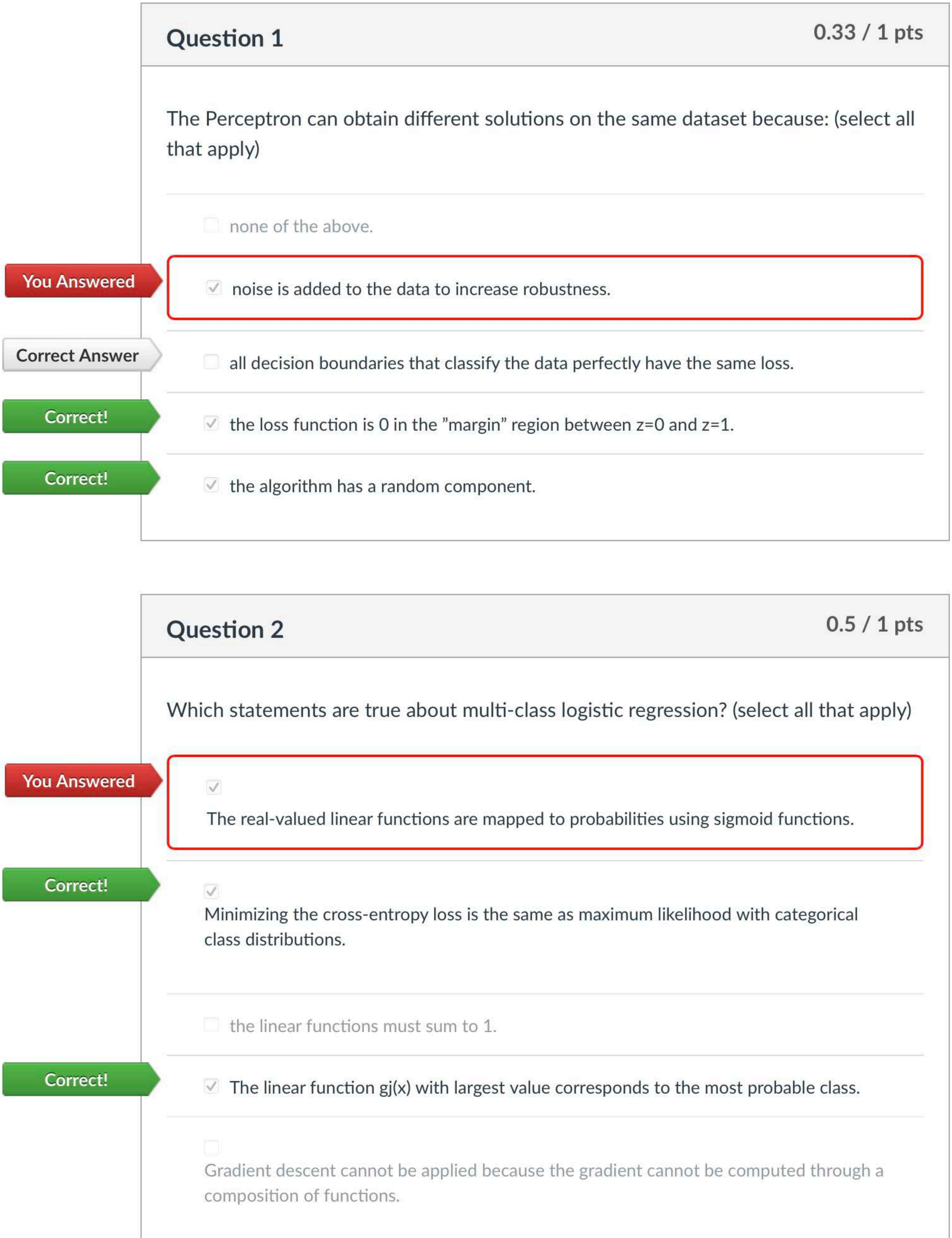
□ the cluster centers should be between 0 and 1.

Correct! ☑ scaling the dynamic range of some features will make those features more important.

Correct Answer □ Euclidean distance is used to compute sample-sample or sample-center distances.

You Answered ☑ normalization makes clustering algorithms less sensitive to initialization.

You Answered ☑ the amount of data in each cluster needs to be balanced, which can be effectively controlled by normalization.



Question 1

0.33 / 1 pts

The Perceptron can obtain different solutions on the same dataset because: (select all that apply)

□ none of the above.

You Answered

☑ noise is added to the data to increase robustness.

Correct Answer

□ all decision boundaries that classify the data perfectly have the same loss.

Correct!

☑ the loss function is 0 in the ”margin” region between z=0 and z=1.

Correct!

☑ the algorithm has a random component.

Question 2

0.5 / 1 pts

Which statements are true about multi-class logistic regression? (select all that apply)

You Answered

☑ The real-valued linear functions are mapped to probabilities using sigmoid functions.

Correct!

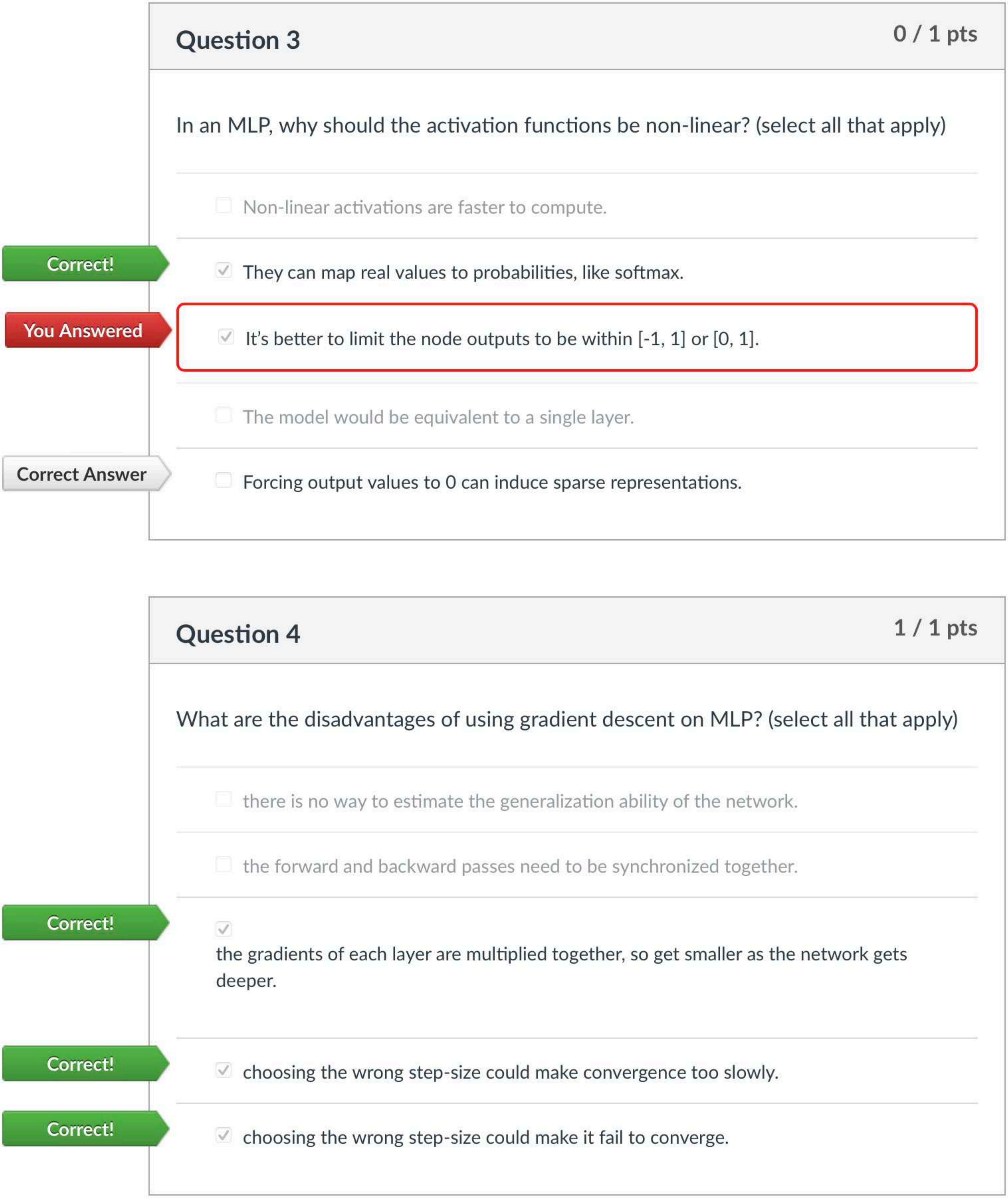
☑ Minimizing the cross-entropy loss is the same as maximum likelihood with categorical class distributions.

□ the linear functions must sum to 1.

Correct!

☑ The linear function gj(x) with largest value corresponds to the most probable class.

□ Gradient descent cannot be applied because the gradient cannot be computed through a composition of functions.



Question 3

0 / 1 pts

In an MLP, why should the activation functions be non-linear? (select all that apply)

□ Non-linear activations are faster to compute.

Correct!

☑ They can map real values to probabilities, like softmax.

You Answered

☑ It’s better to limit the node outputs to be within [-1, 1] or [0, 1].

□ The model would be equivalent to a single layer.

Correct Answer

□ Forcing output values to 0 can induce sparse representations.

Question 4

1 / 1 pts

What are the disadvantages of using gradient descent on MLP? (select all that apply)

□ there is no way to estimate the generalization ability of the network.

□ the forward and backward passes need to be synchronized together.

Correct!

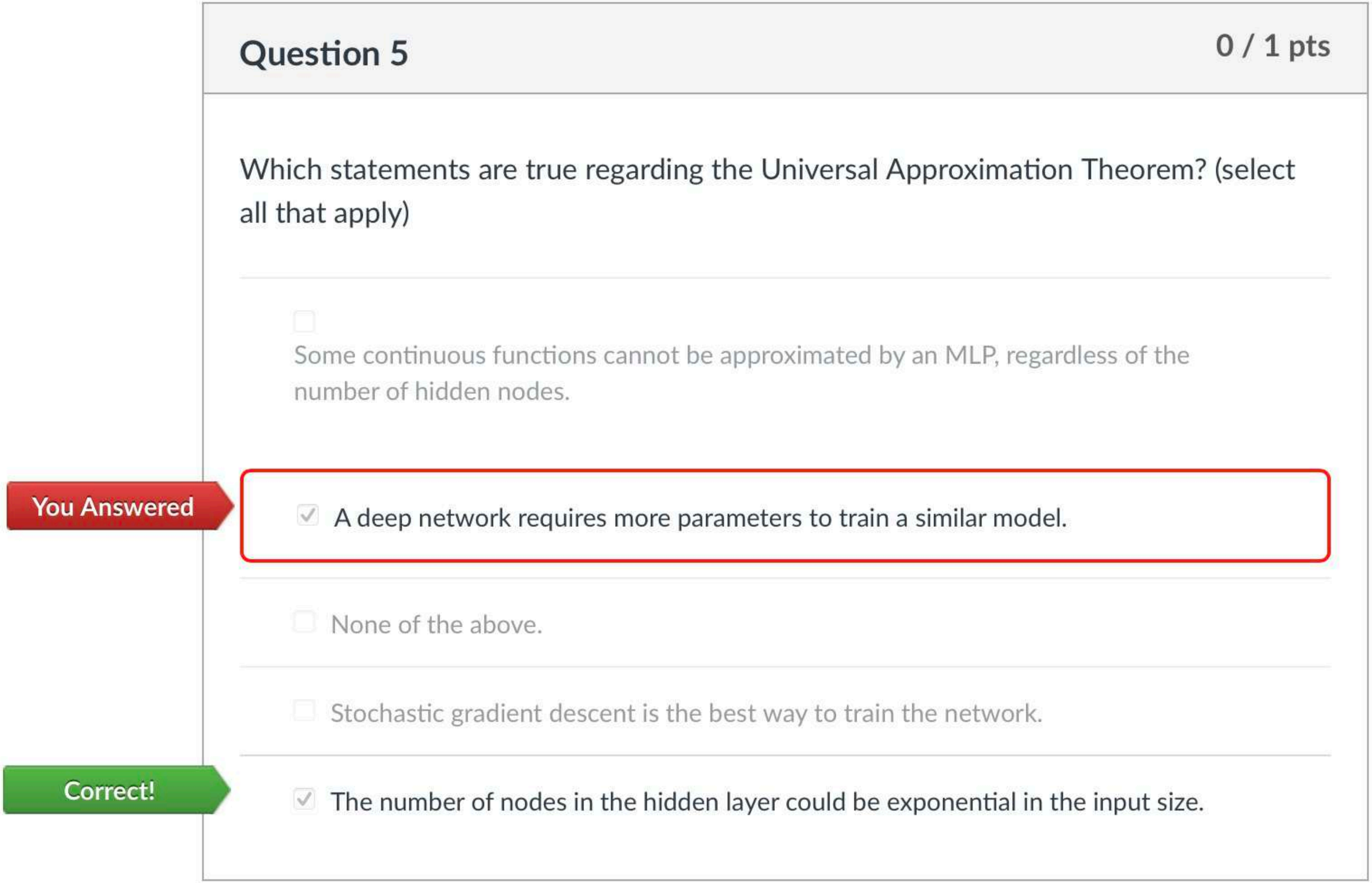
☑ the gradients of each layer are multiplied together, so get smaller as the network gets deeper.

Correct!

☑ choosing the wrong step-size could make convergence too slowly.

Correct!

☑ choosing the wrong step-size could make it fail to converge.



Question 5

0 / 1 pts

Which statements are true regarding the Universal Approximation Theorem? (select all that apply)

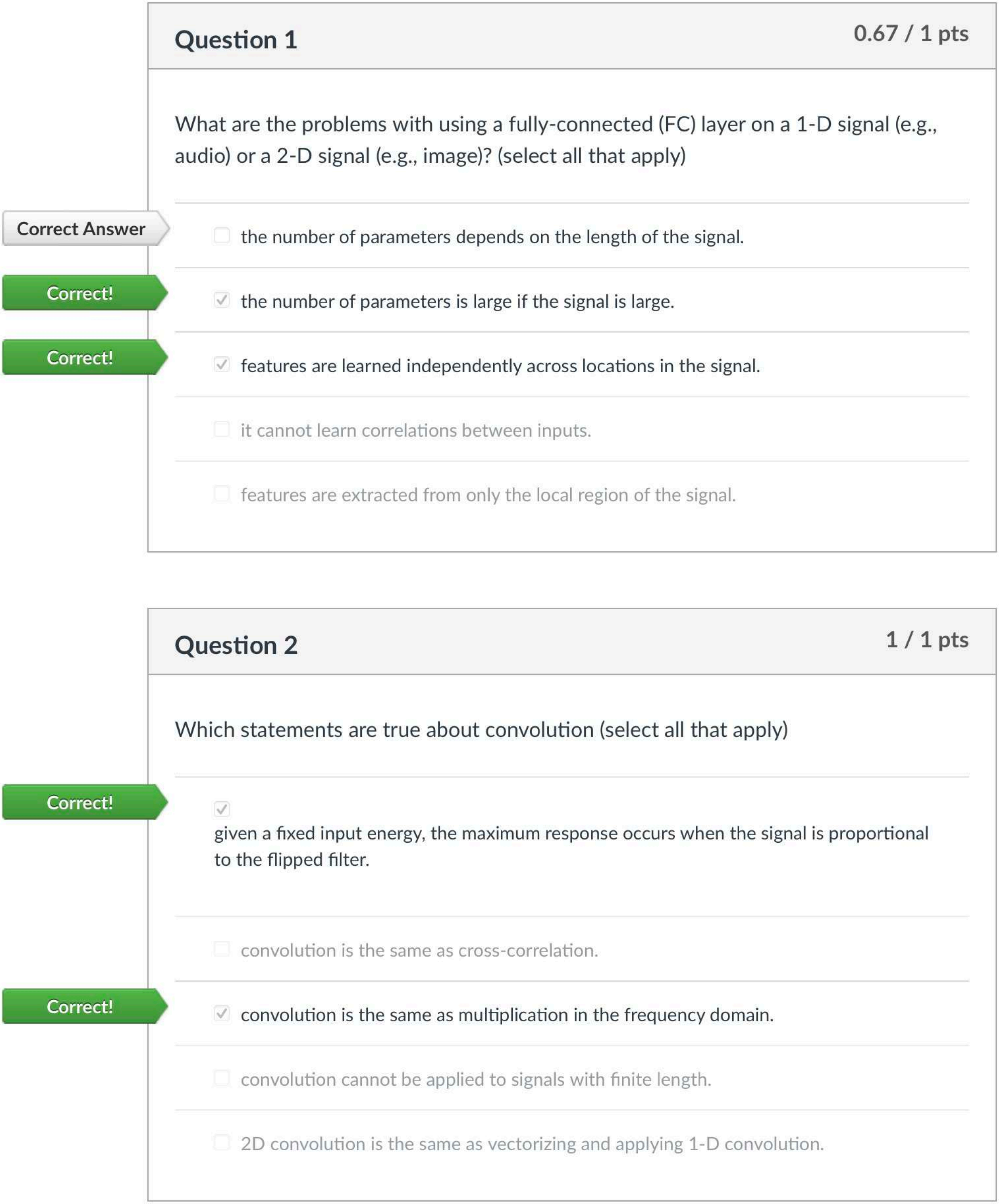
□ Some continuous functions cannot be approximated by an MLP, regardless of the number of hidden nodes.

You Answered ☑ A deep network requires more parameters to train a similar model.

□ None of the above.

□ Stochastic gradient descent is the best way to train the network.

Correct! ☑ The number of nodes in the hidden layer could be exponential in the input size.



Question 1

0.67 / 1 pts

What are the problems with using a fully-connected (FC) layer on a 1-D signal (e.g., audio) or a 2-D signal (e.g., image)? (select all that apply)

Correct Answer □ the number of parameters depends on the length of the signal.

Correct! ☑ the number of parameters is large if the signal is large.

Correct! ☑ features are learned independently across locations in the signal.

□ it cannot learn correlations between inputs.

□ features are extracted from only the local region of the signal.

Question 2

1 / 1 pts

Which statements are true about convolution (select all that apply)

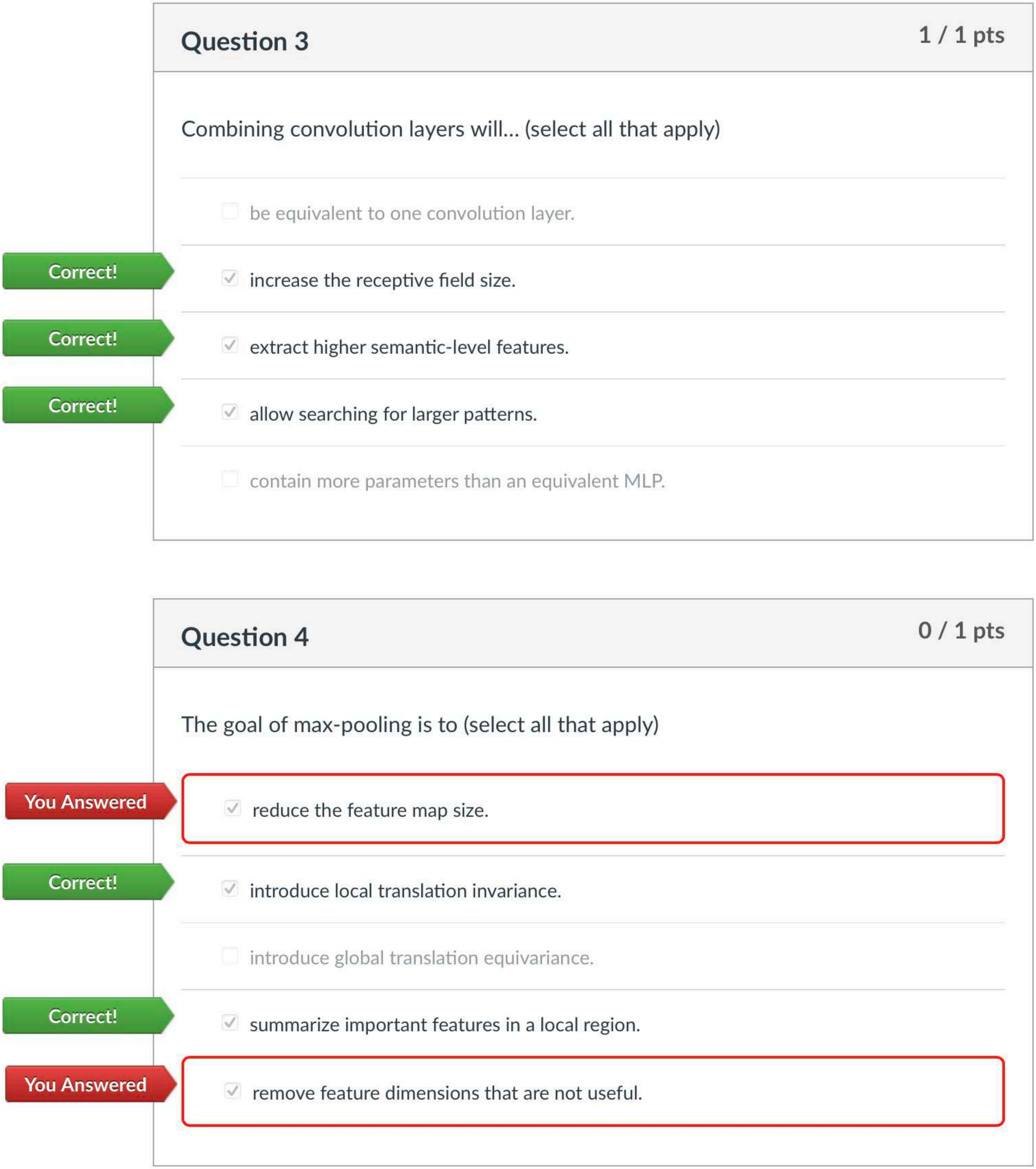
Correct! ☑ given a fixed input energy, the maximum response occurs when the signal is proportional to the flipped filter.

□ convolution is the same as cross-correlation.

Correct! ☑ convolution is the same as multiplication in the frequency domain.

□ convolution cannot be applied to signals with finite length.

□ 2D convolution is the same as vectorizing and applying 1-D convolution.



Question 3

1 / 1 pts

Combining convolution layers will... (select all that apply)

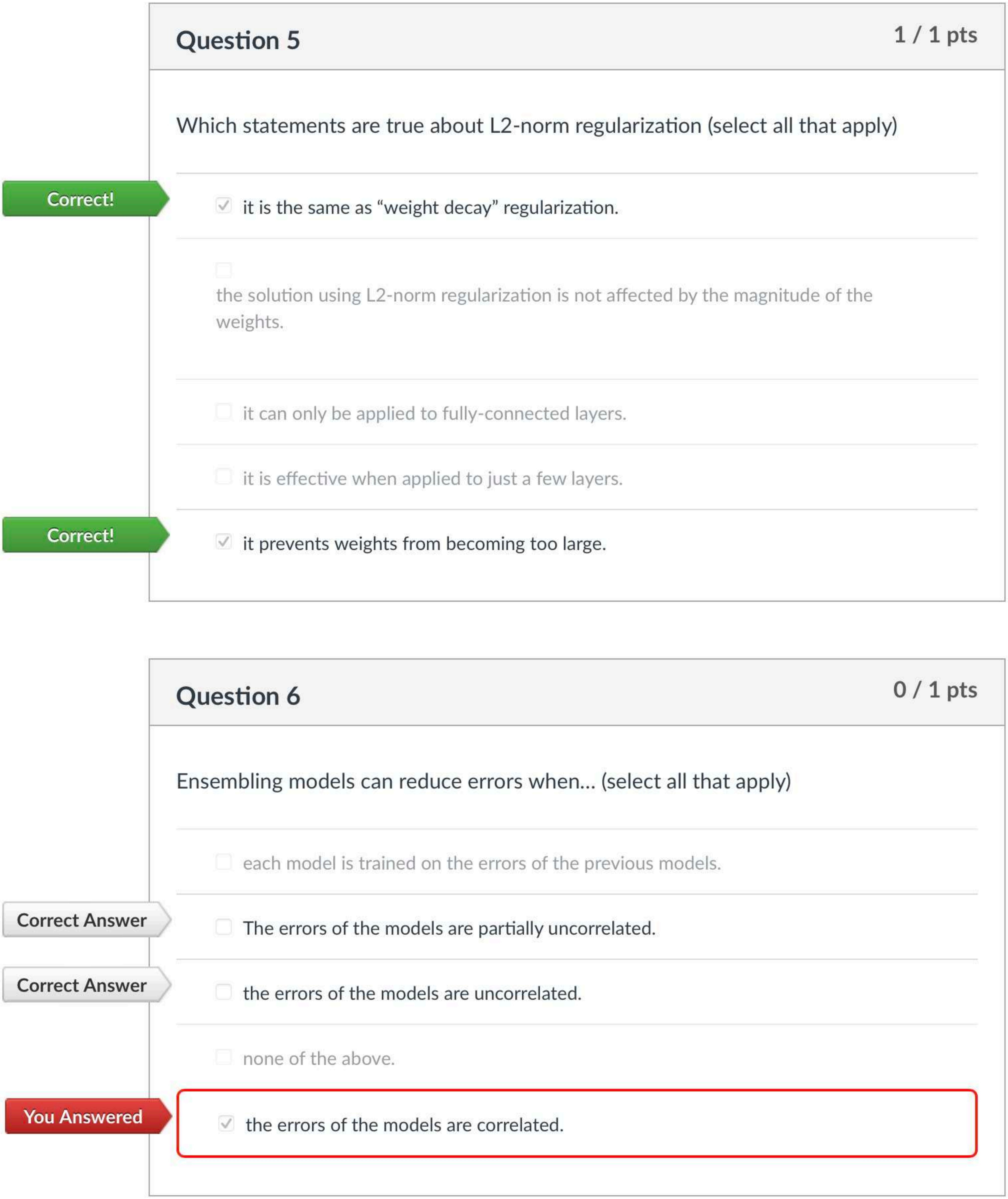
- □ be equivalent to one convolution layer.
- Correct! ☑ increase the receptive field size.
- Correct! ☑ extract higher semantic-level features.
- Correct! ☑ allow searching for larger patterns.
- □ contain more parameters than an equivalent MLP.

Question 4

0 / 1 pts

The goal of max-pooling is to (select all that apply)

- You Answered ☑ reduce the feature map size.
- Correct! ☑ introduce local translation invariance.
- □ introduce global translation equivariance.
- Correct! ☑ summarize important features in a local region.
- You Answered ☑ remove feature dimensions that are not useful.



Question 5

1 / 1 pts

Which statements are true about L2-norm regularization (select all that apply)

Correct! ☑ it is the same as “weight decay” regularization.

□ the solution using L2-norm regularization is not affected by the magnitude of the weights.

□ it can only be applied to fully-connected layers.

□ it is effective when applied to just a few layers.

Correct! ☑ it prevents weights from becoming too large.

Question 6

0 / 1 pts

Ensembling models can reduce errors when... (select all that apply)

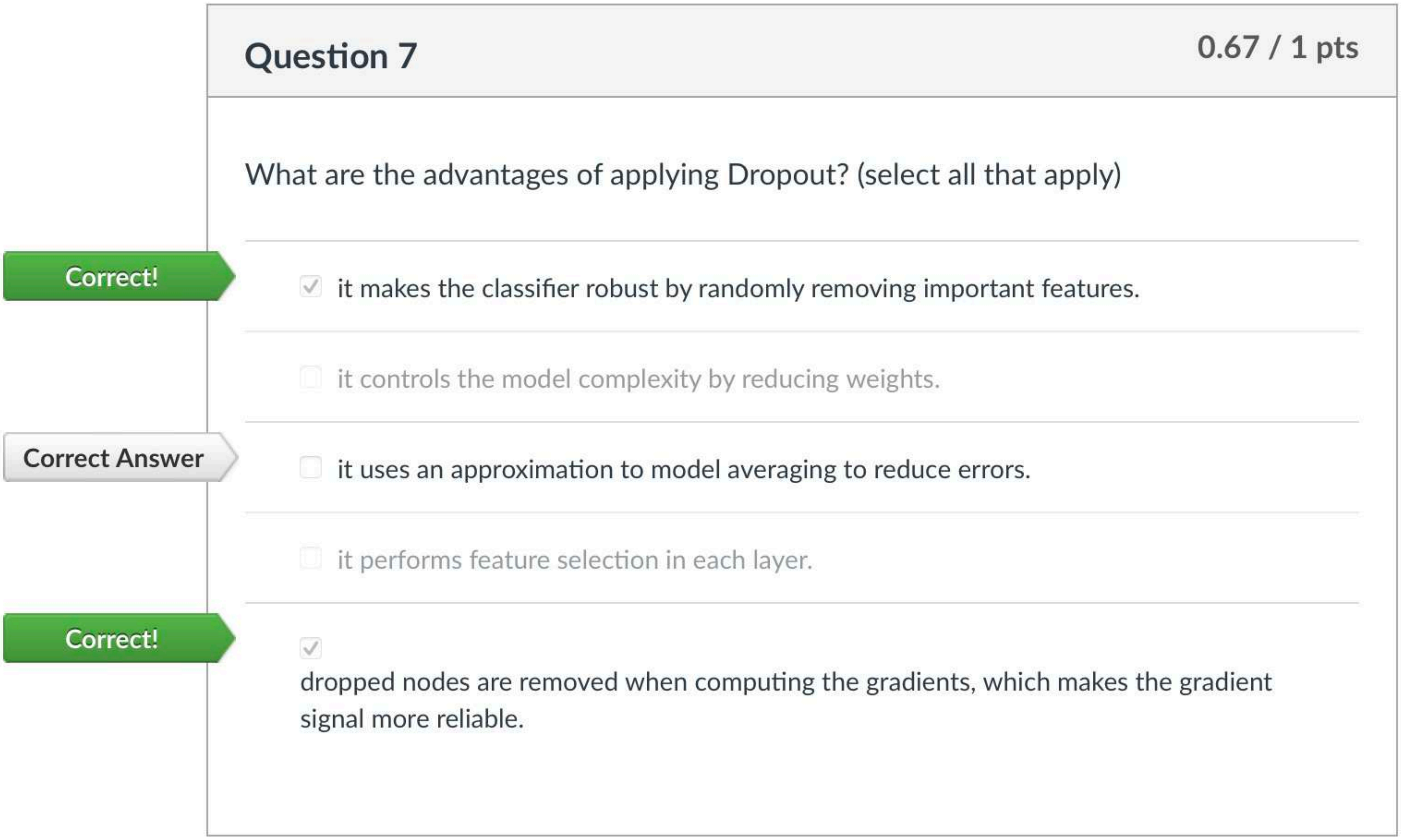
□ each model is trained on the errors of the previous models.

Correct Answer □ The errors of the models are partially uncorrelated.

Correct Answer □ the errors of the models are uncorrelated.

□ none of the above.

You Answered ☑ the errors of the models are correlated.



Question 7

0.67 / 1 pts

What are the advantages of applying Dropout? (select all that apply)

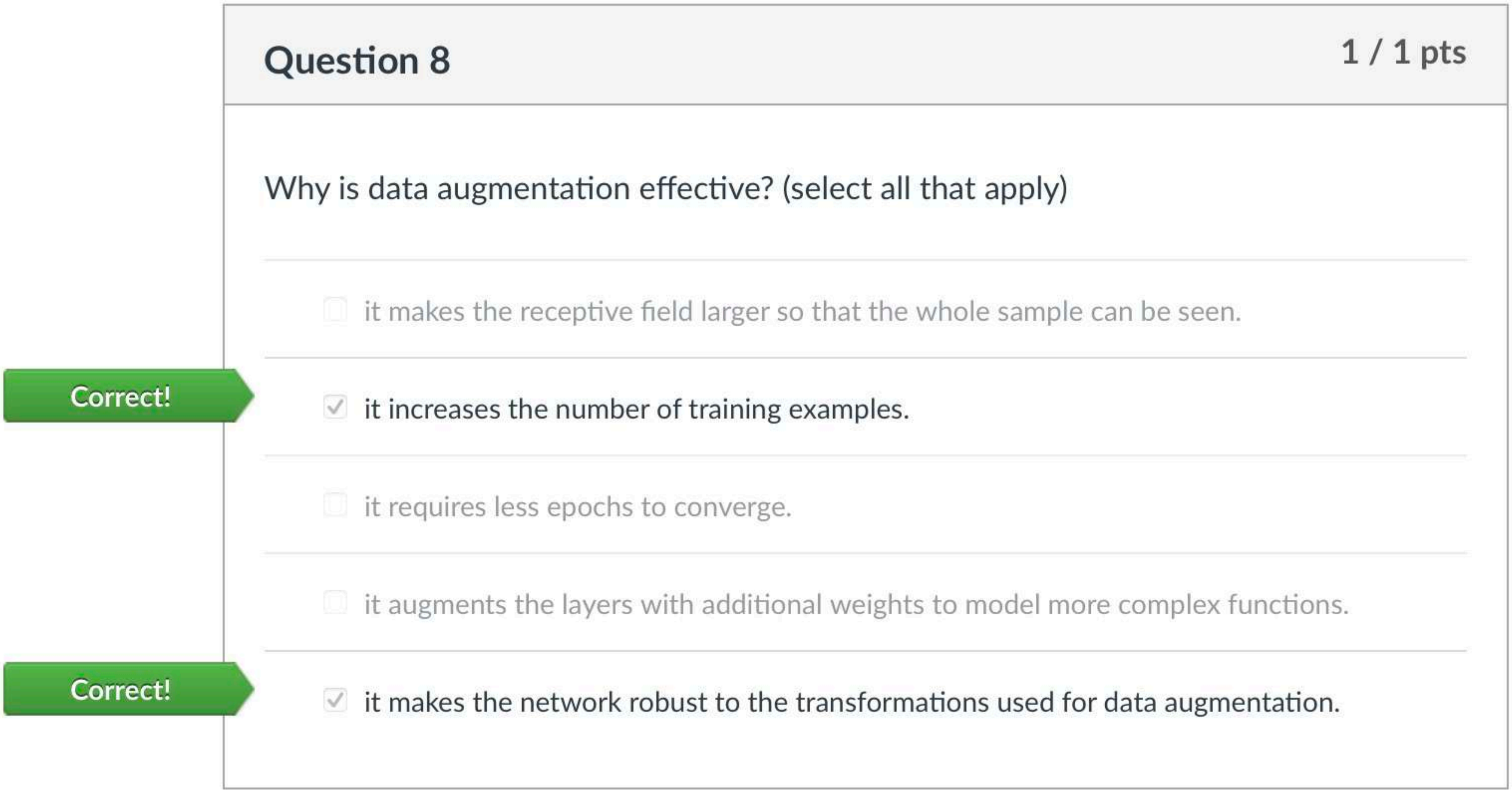
Correct! ☑ it makes the classifier robust by randomly removing important features.

□ it controls the model complexity by reducing weights.

Correct Answer □ it uses an approximation to model averaging to reduce errors.

□ it performs feature selection in each layer.

Correct! ☑ dropped nodes are removed when computing the gradients, which makes the gradient signal more reliable.



Question 8

1 / 1 pts

Why is data augmentation effective? (select all that apply)

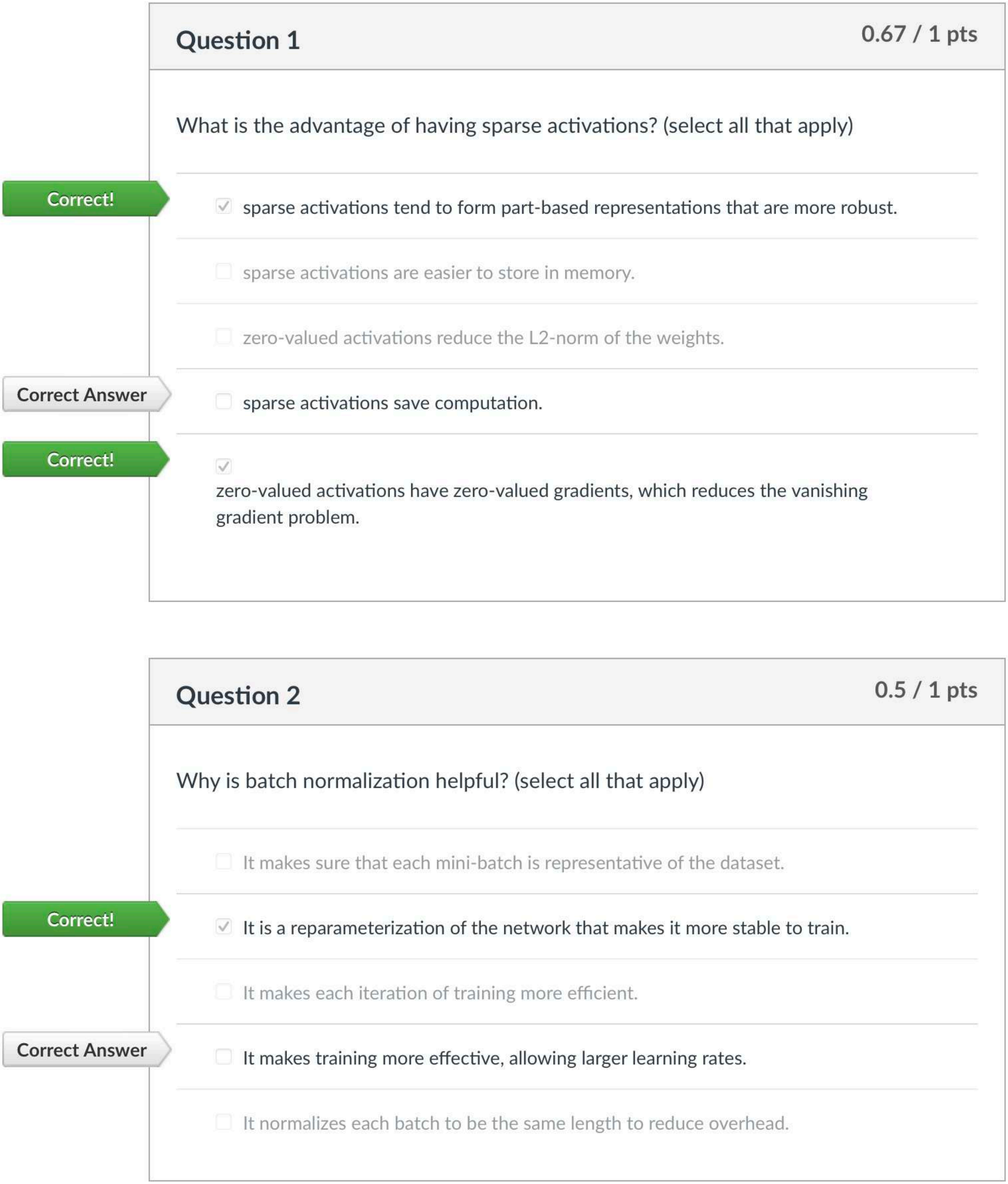
□ it makes the receptive field larger so that the whole sample can be seen.

Correct! ☑ it increases the number of training examples.

□ it requires less epochs to converge.

□ it augments the layers with additional weights to model more complex functions.

Correct! ☑ it makes the network robust to the transformations used for data augmentation.



Question 1

0.67 / 1 pts

What is the advantage of having sparse activations? (select all that apply)

Correct! ☑ sparse activations tend to form part-based representations that are more robust.

□ sparse activations are easier to store in memory.

□ zero-valued activations reduce the L2-norm of the weights.

Correct Answer □ sparse activations save computation.

Correct! ☑ zero-valued activations have zero-valued gradients, which reduces the vanishing gradient problem.

Question 2

0.5 / 1 pts

Why is batch normalization helpful? (select all that apply)

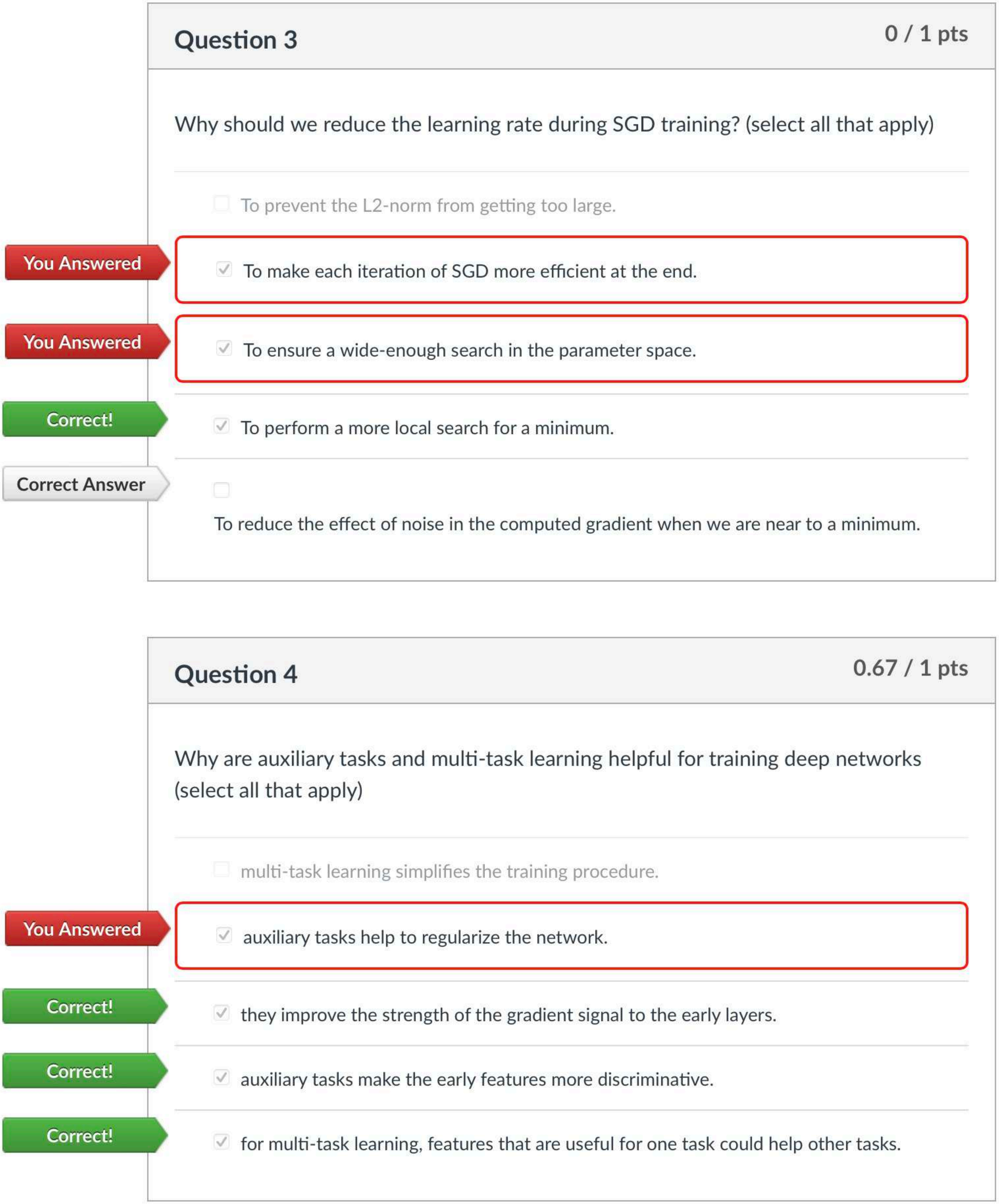
□ It makes sure that each mini-batch is representative of the dataset.

Correct! ☑ It is a reparameterization of the network that makes it more stable to train.

□ It makes each iteration of training more efficient.

Correct Answer □ It makes training more effective, allowing larger learning rates.

□ It normalizes each batch to be the same length to reduce overhead.



Question 3

0 / 1 pts

Why should we reduce the learning rate during SGD training? (select all that apply)

□ To prevent the L2-norm from getting too large.

You Answered ☑ To make each iteration of SGD more efficient at the end.

You Answered ☑ To ensure a wide-enough search in the parameter space.

Correct! ☑ To perform a more local search for a minimum.

Correct Answer □ To reduce the effect of noise in the computed gradient when we are near to a minimum.

Question 4

0.67 / 1 pts

Why are auxiliary tasks and multi-task learning helpful for training deep networks (select all that apply)

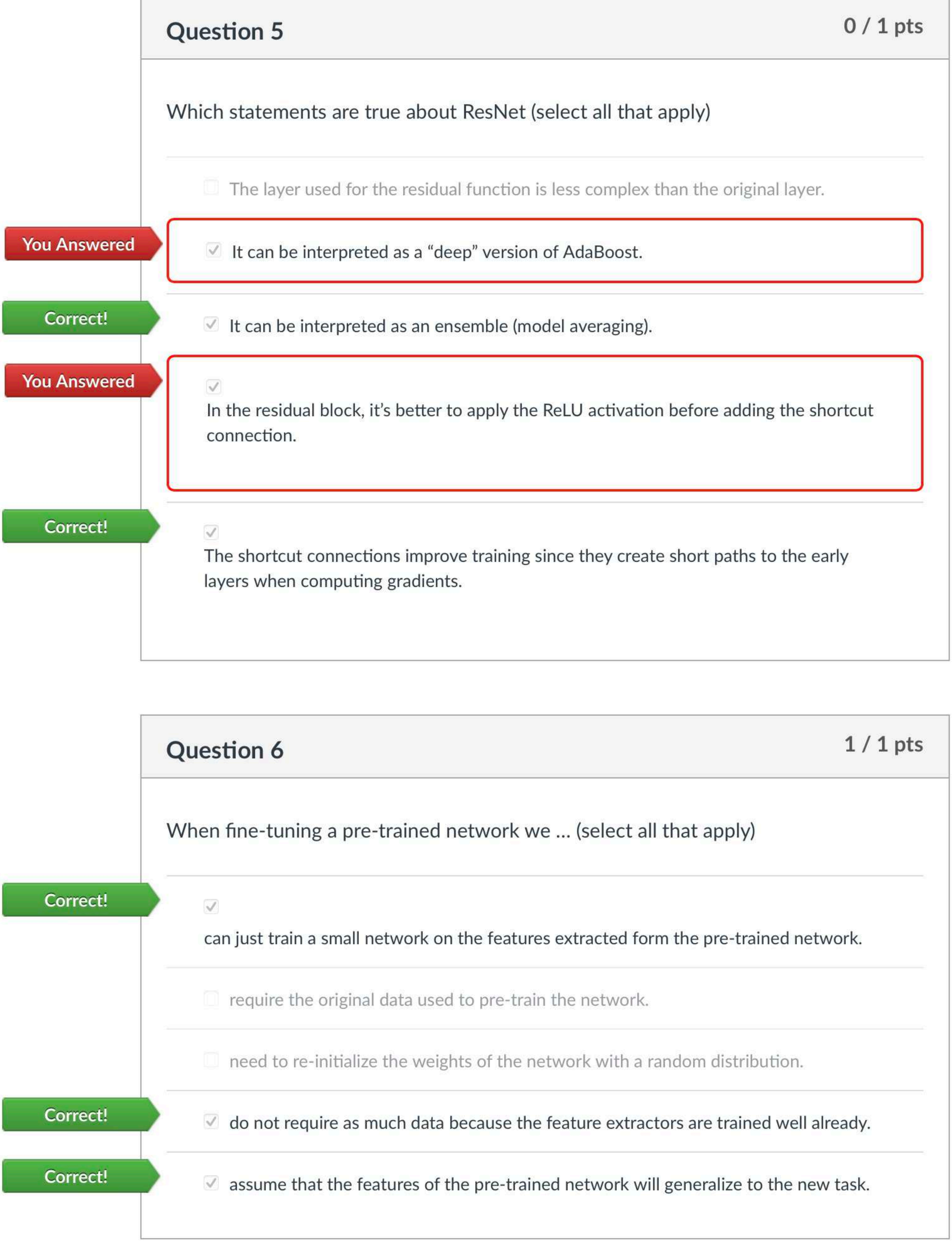
□ multi-task learning simplifies the training procedure.

You Answered ☑ auxiliary tasks help to regularize the network.

Correct! ☑ they improve the strength of the gradient signal to the early layers.

Correct! ☑ auxiliary tasks make the early features more discriminative.

Correct! ☑ for multi-task learning, features that are useful for one task could help other tasks.



Question 5

0 / 1 pts

Which statements are true about ResNet (select all that apply)

□ The layer used for the residual function is less complex than the original layer.

You Answered ☑ It can be interpreted as a “deep” version of AdaBoost.

Correct! ☑ It can be interpreted as an ensemble (model averaging).

You Answered ☑ In the residual block, it’s better to apply the ReLU activation before adding the shortcut connection.

Correct! ☑ The shortcut connections improve training since they create short paths to the early layers when computing gradients.

Question 6

1 / 1 pts

When fine-tuning a pre-trained network we ... (select all that apply)

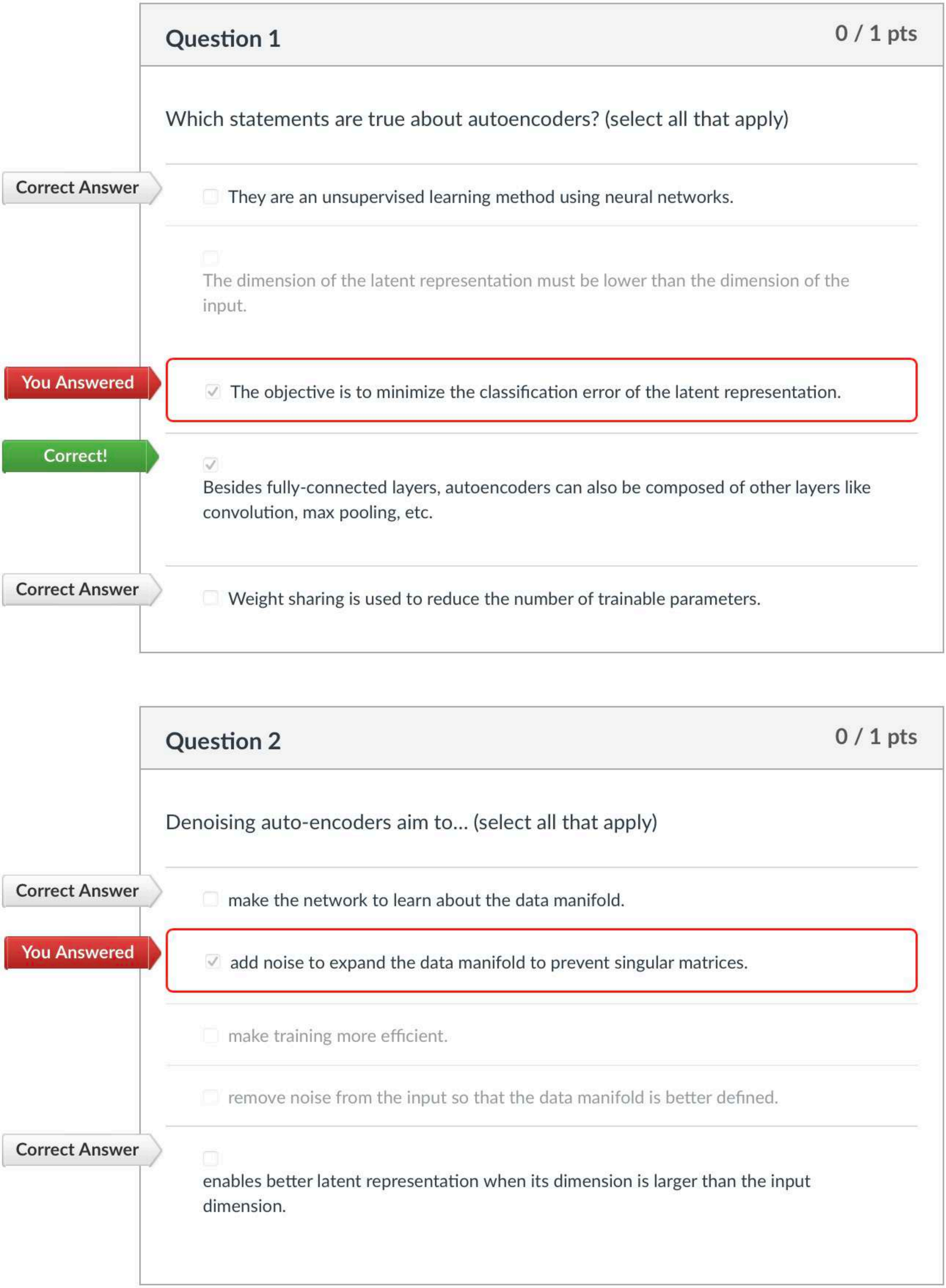
Correct! ☑ can just train a small network on the features extracted form the pre-trained network.

□ require the original data used to pre-train the network.

□ need to re-initialize the weights of the network with a random distribution.

Correct! ☑ do not require as much data because the feature extractors are trained well already.

Correct! ☑ assume that the features of the pre-trained network will generalize to the new task.



Question 1

0 / 1 pts

Which statements are true about autoencoders? (select all that apply)

Correct Answer

□ They are an unsupervised learning method using neural networks.

□ The dimension of the latent representation must be lower than the dimension of the input.

You Answered

☑ The objective is to minimize the classification error of the latent representation.

Correct!

☑ Besides fully-connected layers, autoencoders can also be composed of other layers like convolution, max pooling, etc.

Correct Answer

□ Weight sharing is used to reduce the number of trainable parameters.

Question 2

0 / 1 pts

Denoising auto-encoders aim to... (select all that apply)

Correct Answer

□ make the network to learn about the data manifold.

You Answered

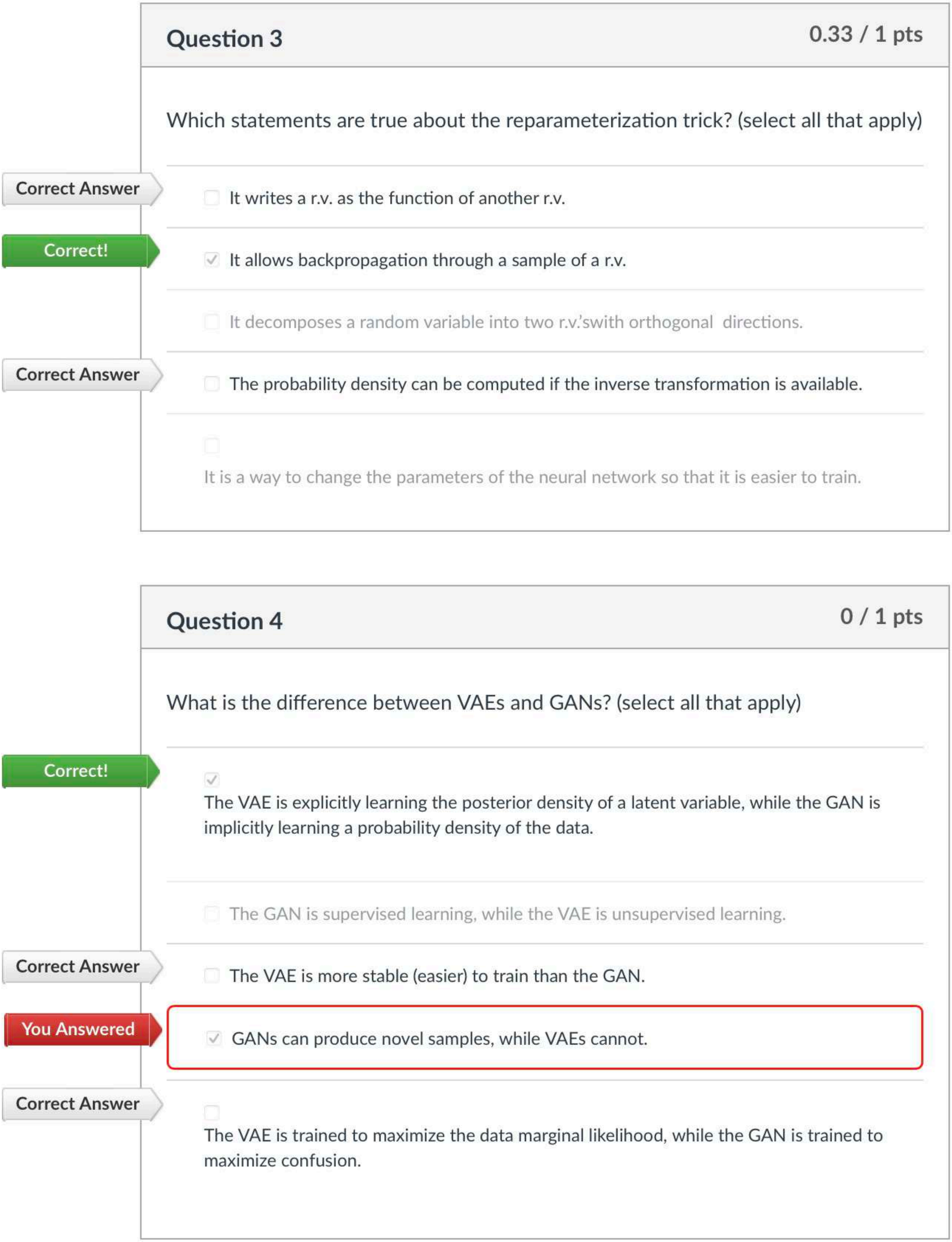
☑ add noise to expand the data manifold to prevent singular matrices.

□ make training more efficient.

□ remove noise from the input so that the data manifold is better defined.

Correct Answer

□ enables better latent representation when its dimension is larger than the input dimension.



Question 3

0.33 / 1 pts

Which statements are true about the reparameterization trick? (select all that apply)

Correct Answer — □ It writes a r.v. as the function of another r.v.

Correct! — ☑ It allows backpropagation through a sample of a r.v.

□ It decomposes a random variable into two r.v.'swith orthogonal directions.

Correct Answer — □ The probability density can be computed if the inverse transformation is available.

□ It is a way to change the parameters of the neural network so that it is easier to train.

Question 4

0 / 1 pts

What is the difference between VAEs and GANs? (select all that apply)

Correct! — ☑ The VAE is explicitly learning the posterior density of a latent variable, while the GAN is implicitly learning a probability density of the data.

□ The GAN is supervised learning, while the VAE is unsupervised learning.

Correct Answer — □ The VAE is more stable (easier) to train than the GAN.

You Answered — ☑ GANs can produce novel samples, while VAEs cannot.

Correct Answer — □ The VAE is trained to maximize the data marginal likelihood, while the GAN is trained to maximize confusion.