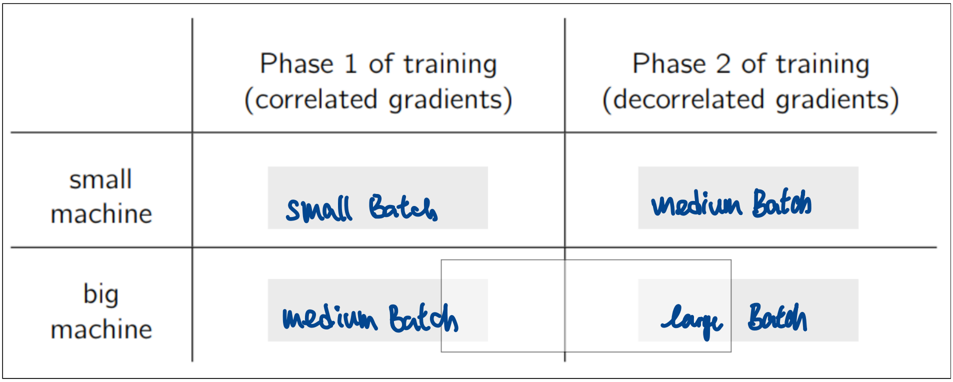
Status	Beendet
Begonnen	Freitag, 8. November 2024, 22:10
Abgeschlossen	Mittwoch, 12. Februar 2025, 14:17
Dauer	95 Tage 16 Stunden
Punkte	0,00/11,00
Bewertung	**0,00** von 10,00 (**0**%)

Frage 1

Nicht beantwortet

Erreichbare Punkte: 1,00

Choosing Batch Size in SGD:



	Phase 1 of training (correlated gradients)	Phase 2 of training (decorrelated gradients)
small machine	small Batch	medium Batch
big machine	medium Batch	large Batch

large batch | medium batch | medium batch

small batch

Die Antwort ist falsch.

Frage 2

Nicht beantwortet

Erreichbare Punkte: 1,00

Prozessing heavy

Data Parallelism divides the training data across the available workers and runs a copy of the model on each worker.

Model Parallelism is to divide our model across different workers, with each worker running the same data on a different segment of the model. Memery heavy

Model parallelism | Data parallelism

Your answer is incorrect.

Die richtige Antwort lautet:
[Data parallelism] divides the training data across the available workers and runs a copy of the model on each worker.

[Model parallelism] is to divide our model across different workers, with each worker running the same data on a different segment of the model.

Frage 3

Nicht beantwortet

Erreichbare Punkte: 1,00

Which of the following statements about Adam is **FALSE**?

Wählen Sie eine Antwort:

[✓ a. Adam guarantees convergence to the global minimum of the loss function.]

○ b. Adam combines the advantages of AdaGrad and RMSProp.

○ c. Adam calculates an exponential moving average of the gradient and the squared gradient.

○ d. The initial value of β_1 and β_2 values close to 1.0 result in a bias of moment estimates towards zero.

$$m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$$

Die Antwort ist falsch.

Die richtige Antwort ist:
Adam guarantees convergence to the global minimum of the loss function.

Frage 4

Nicht beantwortet

Erreichbare Punkte: 1,00

Which of the following statements about conditioning is **FALSE**?

Wählen Sie eine Antwort:

○ a. Data normalization could be used to improve conditioning.

○ b. Well-conditioned functions make it easier to escape local minima.

○ c. A poor conditioning happens when there is a strong divergence of curvature between the different dimensions.

[✓ d. The choice of activation function has no effect on conditioning.]

Your answer is incorrect.

Die richtige Antwort ist:
The choice of activation function has no effect on conditioning.

Frage 5

Nicht beantwortet

Erreichbare Punkte: 1,00

Which of the following statements is **FALSE**?

Wählen Sie eine oder mehrere Antworten:

□ a. A non-convex loss function has only one global minimum and no local minima.

□ b. A convex loss function has both local and global minima.

☒ c. A convex loss function has only one global minimum and no local minima.

☒ d. A non-convex loss function has both local and global minima.

正解답

Your answer is incorrect.

Die richtigen Antworten sind:
A convex loss function has only one global minimum and no local minima.,

A non-convex loss function has both local and global minima.

Frage 6

Nicht beantwortet

Erreichbare Punkte: 1,00

We could use a locally connected neural net, when

☒ a. features in different regions are largely independent of each other.

☒ b. there is a strong correlation between local features.

Your answer is incorrect.

Die richtigen Antworten sind:
there is a strong correlation between local features.,

features in different regions are largely independent of each other.

Frage 7

Nicht beantwortet

Erreichbare Punkte: 1,00

Convolutional layers are technically

✗ a. locally connected layers

○ b. globally connected layers

Your answer is incorrect.

Die richtige Antwort ist:
locally connected layers

Frage 8

Nicht beantwortet

Erreichbare Punkte: 1,00

SGD is guaranteed to converge if the sequence of the learning rates$\gamma^{(t)}$ converges to 0 and the series $\sum_{t=1}^{\infty} \gamma^{(t)}$ diverges.

Then, which of the following learning rates can be used in SGD:

Wählen Sie eine Antwort:

○ a. $\gamma^{(t)} = 1$

✗ b. $\gamma^{(t)} = t^{-1}$

○ c. $\gamma^{(t)} = e^{-t}$

Die Antwort ist falsch.

Die richtige Antwort ist:
$\gamma^{(t)} = t^{-1}$

Frage 9

Nicht beantwortet

Erreichbare Punkte: 1,00

Which of the following statements about gradient descent with momentum is **TRUE**?

Wählen Sie eine oder mehrere Antworten:

☒ a. Momentum takes into account past gradients so as to smooth down gradient measures.

☒ b. The convergence of the gradient descent algorithm can be accelerated by adding Momentum.

□ c. Higher momentum values always lead to better convergence and faster training.

☒ d. Gradient descent algorithm with momentum is likely to skip local minima.

Die Antwort ist falsch.

Die richtigen Antworten sind:
Momentum takes into account past gradients so as to smooth down gradient measures., Gradient descent algorithm with momentum is likely to skip local minima.,

The convergence of the gradient descent algorithm can be accelerated by adding Momentum.

Frage 10

Nicht beantwortet

Erreichbare Punkte: 1,00

Which of the following ideas is helpful to land in some local minima?

Wählen Sie eine oder mehrere Antworten:

☒ a. Do not increase the depth of the neural network beyond necessity.

☒ b. Use a sufficient number of neurons at each layer.

☒ c. Set the learning rate larger.

☒ d. Retrain the network with multiple random initializations.

Your answer is incorrect.

Die richtigen Antworten sind:
Retrain the network with multiple random initializations.,

Set the learning rate larger.,

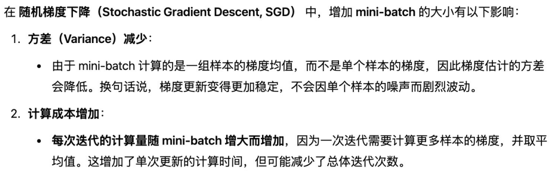
Use a sufficient number of neurons at each layer.,

Do not increase the depth of the neural network beyond necessity.

Frage 11

Nicht beantwortet

Erreichbare Punkte: 1,00



在 **随机梯度下降（Stochastic Gradient Descent, SGD）** 中，增加 **mini-batch** 的大小有以下影响：

1. **方差（Variance）减少**：
   - 由于 mini-batch 计算的是一组样本的梯度均值，而不是单个样本的梯度，因此梯度估计的方差会降低。换句话说，梯度更新变得更加稳定，不会因单个样本的噪声而剧烈波动。
2. **计算成本增加**：
   - **每次迭代的计算量随 mini-batch 增大而增加**，因为一次迭代需要计算更多样本的梯度，并取平均值。这增加了单次更新的计算时间，但可能减少了总体迭代次数。

Taking larger mini-batches in stochastic gradient descent:

Wählen Sie eine Antwort:

○ a. reduces the variance, increases the computational cost of an iteration

○ b. does not change the variance, but improves communication costs

○ c. reduces the variance, at no additional computational expense

○ d. reduces the bias, reduces the variance

Die Antwort ist falsch.

Die richtige Antwort ist:
reduces the variance, increases the computational cost of an iteration