Quiz 4 - Results

Attempt 1 of 1

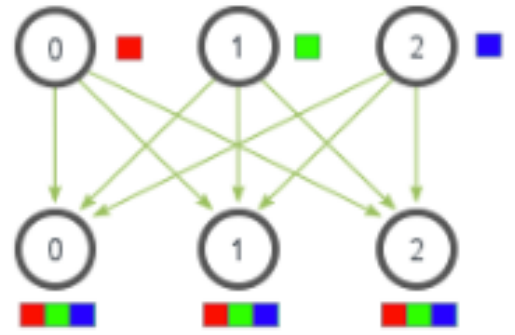
Written Oct 31, 2025 2:08 PM - Oct 31, 2025 2:30 PM

Attempt Score **14 / 25 - F**

Overall Grade (Highest Attempt) **14 / 25 - F**

Question 1 1 / 1 point

What collective can you use to accomplish the following task?

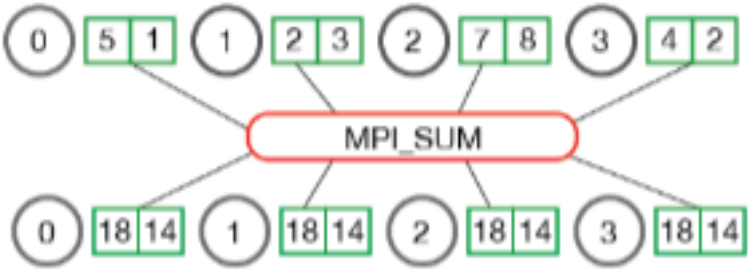


- ○ Gather
- ○ Scatter
- ○ Broadcast
- ✓ ◉ All-gather

- Reduce

Question 2 1 / 1 point

What collective can you use to accomplish the following task?



- ✔ All-Reduce
- Reduce
- Scatter
- Gather
- Broadcast

Question 3 0 / 3 points

Arrange the following distributed training synchronous SGD algorithms with P learners in terms of their increasing expected training time for the **same number of epochs**. Here 1<K<P.

Enter 1, 2, 3 in the blanks.

1. _____ Fully-sync

2. _____ K-batch sync

3. _____ K-sync

Answer for blank # 1: Fully-sync ✖ **(3)**

Answer for blank # 2: K-sync ✖ **(1)**

Answer for blank # 3: K-batch sync ✗ **(2)**

Question 4 0 / 5 points

Enter True or False for each of the following statements. Let P be the number of learners and K is a number between and including 1 and P. Consider six distributed training algorithms: Sync, K-sync, K-batch sync, Async, K-async, K-batch async. When all the other factors are same,

1. When K=P, K-batch sync is same as Sync _____

2. When K=P, K-sync is same as K-batch sync _____

3. When K=P, K-batch async is same as Async _____

4. When K=1, K-batch async and K-async both behave same as Async _____

5. When K>1, stale gradients problem can happen both in K-batch sync and K-batch async _____

Answer for blank # 1: True ✗ **(False, false)**
Answer for blank # 2: True ✗ **(False, false)**
Answer for blank # 3: True ✗ **(False, false)**
Answer for blank # 4: False ✗ **(True, true)**
Answer for blank # 5: True ✗ **(False, false)**

Question 5 1 / 2 points

Write the correct torch.distributed function against each of the following statements. Please make sure the answer is in the format **dist. ‹functionname›**. There is only one answer out of four options for each blank.

1. Sending bytes: [a]

2. Non-blocking send: [b]

3. Receiving bytes: [c]

4. Non-blocking receive[d]

a.

dist.send()

dist.send

send

send()

b.

dist.isend()

dist.isend

isend

isend()

c.

dist.recv()

dist.recv

recv

recv()

d.

dist.irecv()

dist.irecv

irecv()

irecv

Answer for blank # 1: dist.send() ✔(25 %)

Answer for blank # 2: dist.isend() ✔(25 %)

Answer for blank # 3: dist.recv() ✖ **(dist.recv)**

Answer for blank # 4: dist.irecv() ✖ **(dist.irecv)**

Question 6 2 / 2 points

What role does the scaling factor play in neural network quantization? Please choose all valid options.

- ✔ ☐ It determines the learning rate during training
- ✔ ☑ It controls the trade-off between precision and dynamic range
- ✔ ☐ It specifies the number of bits used for each weight in the network
- ✔ ☐ It defines the size of the neural network input layer

Question 7 1 / 1 point

What is the primary advantage of quantization-aware training over post-training quantization?

- ○ Easier implementation in various frameworks
- ○ Faster quantization process
- ○ Smaller model size
- ✔ ◉ Better preservation of model accuracy

Question 8 1 / 1 point

How does a quantization scheme adapt to the distribution of weights in a neural network layer?

- ○ By keeping all weights as 32-bit floating point numbers

By ignoring the distribution and using a fixed precision

By using a fixed set of intervals for all layers

By binning weights based on their closest interval edge

Question 9 0 / 1 point

Which of the following is a characteristic of Cookbook Quantization?

It adapts to each layer's distribution

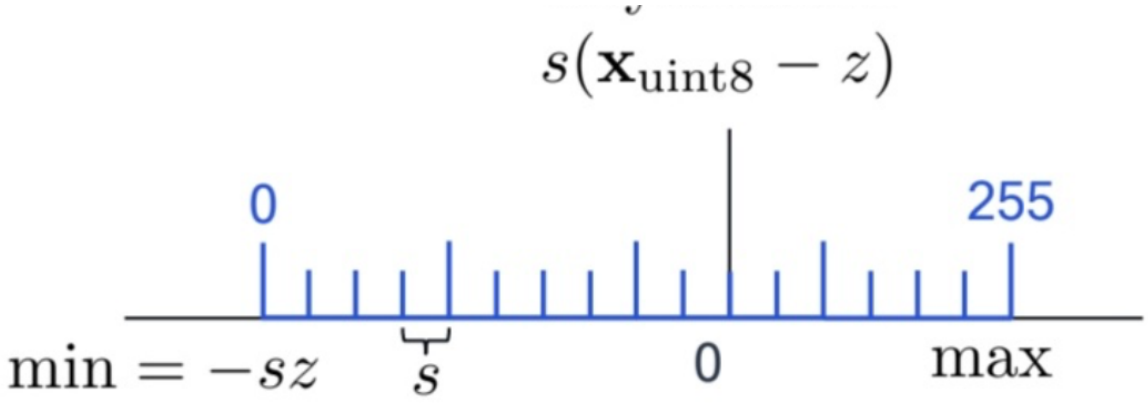
It involves random quantization

It is a one-size-fits-all approach

It only quantizes weights, not activations

Question 10 1 / 1 point

The figure shows a quantization grid for bit width of 8. *s* is the scaling factor. *z* is the zero point. The floating point grid is in black and the integer quantized grid in blue.



What type of quantization do you see in this figure?

Symmetric non-uniform quantization

Symmetric uniform quantization

✓ Asymetric uniform quantization

Asymetric non-uniform quantization

Question 11 1 / 1 point

What is the purpose of iterative fine-tuning after pruning?

To add more layers to the pruned network

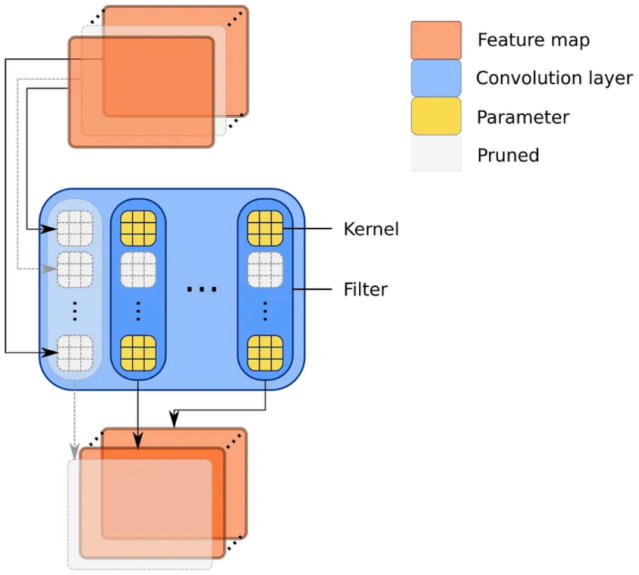
To speed up the pruning process

✓ To retrain the pruned model to recover lost accuracy

To increase the sparsity of the pruned model

Question 12 2 / 2 points

What type of pruning is shown in the figure below:

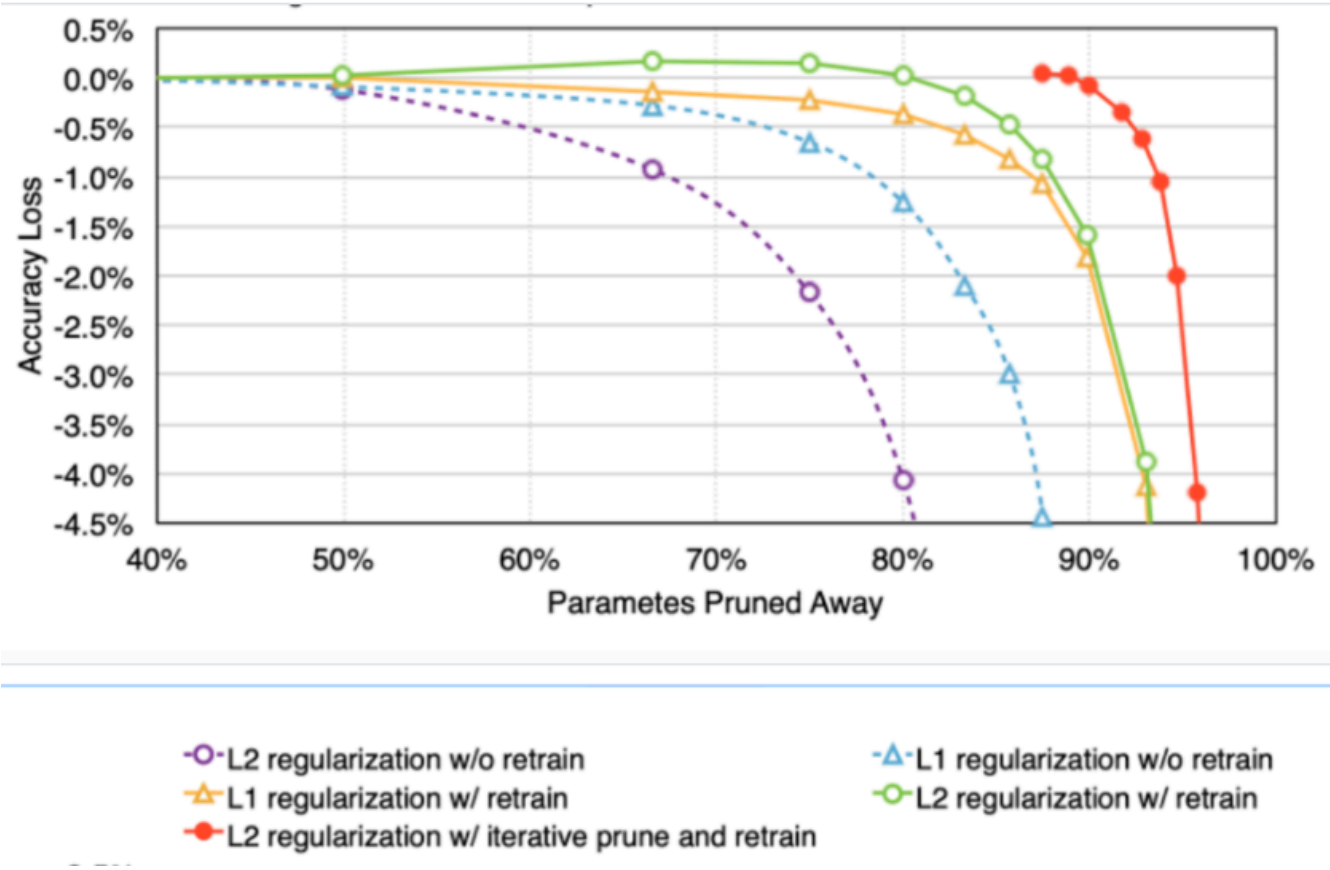


None of the options are correct

- ○ Unstructured pruning
- ✓ ◉ Structured filter and kernel pruning
- ○ Structured layer pruning

Question 13 3 / 4 points

The figure below shows the Trade-off curve for parameter reduction and loss in top-5 accuracy for a given model. We experimented with L1 and L2 regularization, with and without retraining, together with iterative pruning to give five trade off lines. Use the figure to help you answer the following True/False questions. Select the correct option.



1. L1 regularization penalizes non-zero parameters resulting in more parameters near zero. **True** or **False**
2. L1 regularization always results in better accuracy after retraining. **True** or **False**
3. Overall, L2 regularization gives the best pruning results. **True** or **False**
4. The best performance comes from L2 regularization without retraining. **True** or **False**

Answer for blank # 1: True ✓(25 %)

Answer for blank # 2: True ✗ **(False, false)**

Answer for blank # 3: True ✔(25 %)

Answer for blank # 4: False ✔(25 %)

Done