ADVANCED AI

Quiz 3

Friday 2 October 2026

ANSWER KEY

10 questions, one mark each. Circle one answer per question.

1. Why is the gradient $\nabla \ell$ computed backward from the loss rather than forward from each parameter?

(A) Forward-mode differentiation gives only approximate derivatives, while the backward sweep is exact.

(B) One sweep back from the scalar loss yields the derivative for every parameter.

(C) The chain rule holds only when multiplied out from the loss end.

(D) A backward sweep needs no stored forward values, so it saves memory.

(E) A backward sweep skips most of the graph, so depth adds no cost.

2. With unlimited data and a model family that can represent the conditional mean $\mu[x]$ exactly, which error term remains?

(A) The cross terms: they vanish in expectation but not for one dataset.

(B) Noise: the target still varies around $\mu[x]$ at that input.

(C) Squared bias: it comes from the data rather than from the model.

(D) Variance: a family that flexible also fits the noise in any sample.

(E) All three shrink toward zero once $\mu[x]$ is representable, noise included.

3. A ReLU saw pre-activations $f = [0.7, -0.3, 1.5]^T$. The sensitivity at its outputs is $[-2, 4, 1]^T$. What sensitivity does it pass back to f?

(A) $[-2, 4, 1]^T$

(B) $[0, 4, 1]^T$

(C) $[-1.4, 0, 1.5]^T$

(D) $[-2, 0, 1]^T$

(E) $[0, 4, 0]^T$

4. With the model family fixed, doubling the training set mainly reduces which error term?

(A) Squared bias: the model now overfits less to its training sample.

(B) Noise: more observations average out the variation in y.

(C) Squared bias and variance equally, since both depend on the sample.

(D) Variance: models fitted to different samples agree more closely.

(E) Squared bias: more examples let the family get closer to $\mu[x]$.

5. Layer k maps source activations h_k (64) to pre-activations with sensitivities δ_k (32). What is the loss gradient for its weights Ω_k?

(A) $\Omega_k \odot \delta_k h_k^T$, a 32 × 64 matrix

(B) $h_k \delta_k^T$, a 64 × 32 matrix

(C) $\delta_k h_k^T$, a 32 × 64 matrix

(D) δ_k, a 32 × 1 vector

(E) $\Omega_k^T \delta_k$, a 64 × 1 vector

6. After training error reaches zero, held-out cross-entropy rises while held-out error barely moves. Why?

(A) **Wrong predictions grow more confident while their predicted class stays the same.**

(B) Error reads only the argmax, so the rising loss is a numerical artefact.

(C) The held-out labels are noisy, so the loss is tracking label noise.

(D) The learning rate is too high, so the held-out loss is oscillating upward.

(E) Correct predictions grow overconfident, and the loss penalises that overconfidence.

7. Why not initialise every weight to zero?

(A) The optimiser needs random starting weights to explore the loss surface.

(B) The gradient for a weight is proportional to that weight's current value.

(C) Weights should start near their trained values, and zero is too far from them.

(D) Zero weights make the first loss too large for a stable first step.

(E) **Units that start identical receive identical updates, so they stay identical.**

8. The best of 20 models is chosen by test score, and that score is reported. What's wrong?

(A) **The test set helped choose the model, so its score is no longer untouched.**

(B) Nothing, provided no parameters were fitted on the test set.

(C) The test set is smaller than the training set, so its score is biased.

(D) One model's score is noisy; report the mean over all 20 instead.

(E) Twenty candidates is too few; the search should cover more settings.

9. In a ReLU network, a hidden layer has fan-in 200 and fan-out 50. What weight variance keeps the forward signal stable?

(A) 0.04

(B) 0.005

(C) **0.01**

(D) 0.016

(E) 0.1

10. What did the double-descent experiments show?

(A) Interpolating the training data is what causes held-out error to fall.

(B) Training error rose again once capacity passed the interpolation threshold.

(C) The error peak sits exactly where the parameter count equals the training-set size.

(D) Over-parameterised training reliably selects the smoothest of the interpolants that fit.

(E) **Held-out error fell again once capacity passed the interpolation threshold.**