Quiz 1: Gradient Descent — Answer Key

E190BO *Instructor copy — 15 points total*

1. (3 pts) *Reason to prefer gradient descent over the analytical solution.*

Any one of:

- A closed-form solution only exists for simple models; gradient descent works for complex models (e.g. neural networks) where no closed form is available.
- The analytical solution can be computationally expensive or intractable at scale, even when it exists.

Rubric. 3 pts for either idea stated clearly. 2 pts if the idea is right but vague ("it's more general", "closed form doesn't always work") without saying *what* it doesn't work for. 0 pts for describing what gradient descent *is* without giving a reason to prefer it.

2. (3 pts) *The three steps.*

1. Start with a random guess for the weights.
2. Take a step in the direction of the negative gradient.
3. Repeat until convergence.

Rubric. 1 pt each. On step 2, accept "step downhill" / "step in the direction that decreases the loss." Do **not** require the word "negative" if the direction is otherwise clearly downhill. On step 3, accept "repeat until the loss stops changing" or "repeat for a fixed number of iterations."

3. (2 pts) *What is γ?*

γ is the **learning rate**. It controls the **size of the step** taken on each iteration.

Rubric. 1 pt for the name, 1 pt for step size. "How fast it learns" earns the second point; "γ is the thing we multiply the gradient by" alone does not — that restates the formula without saying what it controls.

4. (2 pts) *Loss increases wildly — is γ too large or too small?*

Too large. The steps overshoot the minimum, landing on the far side at a point with higher loss; each iteration overshoots further and the loss diverges.

Rubric. 1 pt for "too large," 1 pt for a reason invoking overshooting / stepping past the minimum / diverging. A student who says "too large" with no reason or a wrong reason gets 1 pt. A student who says "too small" gets 0 even if the explanation is fluent — too small produces slow progress, not increasing loss.

5. (5 pts) $e(w) = (w-4)^2$.

(a) (2 pts) $\dfrac{de}{dw} = 2(w-4) = 2w - 8.$

(b) (2 pts) At $w^{(0)} = 0$: gradient $= 2(0 - 4) = -8$.

$$w^{(1)} = w^{(0)} - \gamma \frac{de}{dw} = 0 - (0.1)(-8) = \mathbf{0.8}$$

(c) (1 pt) Yes. The minimum is at $w = 4$, and 0.8 is closer to 4 than 0 is — the step moved in the right direction.

Rubric for 5.

- (a) 2 pts. 1 pt if the chain rule is applied but a constant is dropped.
- (b) 2 pts. 1 pt for correctly evaluating the gradient at $w^{(0)}$ even if the update is then botched. **Watch for** $w^{(1)} = -0.8$ — this is the sign error from dropping the minus in the update rule, and it is the single most common failure mode on this problem. Award 1 pt and mark it as a sign error rather than an arithmetic slip; track how many students land here, since it tells you whether the update rule itself needs re-teaching.
- (c) 1 pt. Requires identifying $w = 4$ as the minimum, explicitly or implicitly. A student who answers "yes, because gradient descent always decreases the loss" has asserted the conclusion rather than checked it — award the point only if they reference where the minimum actually is.
- Carry-through: a student with a wrong (a) who then uses it consistently in (b) and (c) loses points only in (a).