🧠 Deep Learning Exam Worksheet

Coverage: 01-nn-fundamentals + 02-model-training-and-optimization
Format: Derivations, math, conceptual explanations
Total Points: 100

📐 Section 1: Matrix Ops, Vectorization, and Broadcasting (15 pts)

1. **[3 pts] Matrix Multiplication**
   Given:
   $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$,
   $B = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}$
   Compute $C = AB$ using:
   $$C_{ij} = \sum_k A_{ik} B_{kj}$$
2. **[3 pts] Broadcasting Conditions**
   Explain the rules of broadcasting and illustrate using tensors of shapes (2, 3) and (1, 3). Why does it work?
3. **[3 pts] Bias Addition via Broadcasting**
   Given $x \in \mathbb{R}^{n \times d}$, $W \in \mathbb{R}^{m \times d}$, and $b \in \mathbb{R}^m$, explain how broadcasting computes:
   $$y = xW^T + b$$
   What is the shape of y?
4. **[3 pts] Vectorization vs Looping**
   Define vectorization. Rewrite the following using vector notation:
   $$\text{sum} = \sum_{i=1}^{n} a_i b_i$$
5. **[3 pts] Tensor Shape Flow**
   Track shapes through:
   `Linear(100 → 64) → ReLU → Linear(64 → 32) → ReLU → Linear(32 → 1) → Sigmoid`
   Input batch shape: (16, 100). What are the output shapes after each layer?

🔥 Section 2: Activations & Gradients (20 pts)

1. **[4 pts] Activation Derivatives**
   Derive:
   - $\frac{d}{dx}\sigma(x)$ where $\sigma(x) = \frac{1}{1+e^{-x}}$
   - $\frac{d}{dx}\tanh(x)$
2. **[3 pts] GELU Approximation**
   Write the GELU approximation formula. Why is it preferred in Transformer models over ReLU?
3. **[3 pts] Vanishing vs Exploding Gradients**
   Define both problems and explain why they happen during backpropagation.
4. **[3 pts] ReLU vs GELU**
   Compare in terms of behavior on negatives, gradient flow, and impact on deep learning.
5. **[4 pts] Softmax + Cross-Entropy**
   Given logits $z = [2.0, 1.0, 0.1]$, compute:
   - Softmax outputs: $\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$

- Cross-entropy loss if true class is index 0: $-\log(\text{softmax}_0)$

Section 3: Loss Functions & Optimization (25 pts)

1. **[3 pts] BCE Derivation**
   Derive:
   $$\mathcal{L}_{\text{BCE}} = -\left[y \log(\hat{y}) + (1-y)\log(1-\hat{y})\right]$$
   and explain its connection to Maximum Likelihood Estimation.
2. **[3 pts] MSE Gradient**
   Given:
   $$\mathcal{L}_{\text{MSE}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
   Derive gradients w.r.t. weights w and bias b for a linear model.
3. **[3 pts] Cross-Entropy vs MSE**
   Why is cross-entropy preferred for classification problems? Use softmax output as an example.
4. **[4 pts] Adam Optimizer**
   Define and explain:
   - First moment estimate m_t
   - Second moment estimate v_t
   - Bias correction
   - Final update rule:
   $$\theta_t = \theta_{t-1} - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$
5. **[3 pts] Adam vs SGD**
   Compare in terms of speed, robustness, and generalization.
6. **[3 pts] AdamW Fix**
   How does AdamW decouple weight decay from gradients? Why is this better than L2 + Adam?
7. **[3 pts] LR Schedules**
   Describe cosine decay and linear warmup. Why are they useful in large model training?

Section 4: Regularization & Normalization (20 pts)

1. **[3 pts] L1 vs L2 Regularization**
   Define each penalty term:
   - L1: $\lambda \sum_i |w_i|$
   - L2: $\lambda \sum_i w_i^2$

   Which one induces sparsity and why?
2. **[3 pts] Dropout**
   Explain how dropout is applied to activations and why outputs are scaled by the keep probability.
3. **[4 pts] BatchNorm**
   Given:
   $$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}, \quad y_i = \gamma \hat{x}_i + \beta$$
   Explain each term and how BatchNorm helps training.
4. **[3 pts] LayerNorm vs BatchNorm**
   Complete:

Property	BatchNorm	LayerNorm
Normalizes over		
Needs batch stats?		
Best for	CNNs	Transformers
Works with batch size=1		

5. **[4 pts] Gradient Clipping**
   Provide the clipping formula:

$$\text{If } \|\mathbf{g}\|_2 > \text{max\_norm}, \quad \mathbf{g} \leftarrow \mathbf{g} \cdot \frac{\text{max\_norm}}{\|\mathbf{g}\|_2}$$

   Why is this useful in training deep or unstable networks?
6. **[3 pts] Compatibility Warnings**
   Explain why the following combinations can cause issues:
   - BatchNorm + Dropout
   - Adam + L2 Regularization (not AdamW)

📈 Section 5: Training & Debugging (20 pts)

1. **[3 pts] Overfitting**
   What signs appear in loss curves? List 3 techniques to reduce it.
2. **[3 pts] Underfitting**
   If both training and validation losses are high, what changes might help?
3. **[4 pts] Loss ↑, Accuracy ↑**
   Why can accuracy improve while loss increases in classification?
4. **[4 pts] Generalization Gap**
   Define:

$$\text{Gap} = \mathcal{L}_{\text{val}} - \mathcal{L}_{\text{train}}$$

   What does a growing gap suggest? How can it be fixed?
5. **[3 pts] Instability Diagnosis**
   What might cause:
   - Flat training loss early
   - NaNs or spikes in loss
     Suggest fixes.

In [ ]: