Artificial Intelligence PhD Qualifying Examination

High-Fidelity Mock Examination 08

Choose three of four sections; maximum score: 100 points.

Name: _____ **Student ID:** _____ **[1 point]**

Instructions. This examination consists of four sections, each worth 33 points. Choose **three** sections and indicate your choices clearly at the beginning of your answer sheet. If you answer more than three sections, only the first three sections that you select will be graded. The additional point above is awarded for writing your name and student ID. Give complete reasoning unless a question states otherwise.

A. Machine Learning Theory [33 points]

Part I: Warm-Up Questions [3 points]. Select one answer per item.

(i) [0.5 points] Switching from average to sum loss while preserving a ridge objective requires

(a) rescaling regularization.

(b) changing labels.

(c) deleting the Hessian.

(d) a new VC definition.

(ii) [0.5 points] Affine parity adds to linear parity

(a) a binary offset.

(b) a PSD kernel only.

(c) a real threshold.

(d) no functions.

(iii) [0.5 points] Importance-weighted losses bounded by W have Hoeffding scale

(a) W^2.

(b) $1/W^2$ in the numerator.

(c) no W.

(d) 2^W necessarily.

(iv) [0.5 points] Averaging parameters cannot increase pointwise loss when the loss is

(a) convex in parameters.

(b) arbitrary.

(c) 0–1 necessarily.

(d) a hypothesis count.

(v) [0.5 points] The growth function of a union is at most

(a) the sum of component growth functions.

(b) their negative product.

(c) zero.

(d) the sample mean.

(vi) [0.5 points] Ridge with $\lambda > 0$ remains uniquely solvable when X is

(a) rank deficient.

(b) unlabeled only.

(c) nonnumeric.

(d) a hypothesis class.

Part II: Theoretical Exercises [30 points].

8 points For

$$J_\lambda(w) = \frac{1}{2m} \|Xw - y\|_2^2 + \frac{\lambda}{2} \|w\|_2^2, \qquad \lambda > 0,$$

derive the gradient, Hessian, normal equation, and minimizer. Prove uniqueness even if X is rank deficient. If the data-fit term is written as a sum rather than an average, state the regularization coefficient that gives the same minimizer.

6 points On $\{0,1\}^d$, define affine parities

$$h_{w,b}(x) = (w^\top x + b) \bmod 2, \qquad w \in \{0,1\}^d,\ b \in \{0,1\}.$$

Determine their VC dimension.

7 points Data are sampled from Q, target risk is under P, and $r = dP/dQ \le W$. For finite $\mathcal{H}$ of size N, empirical importance-weighted risk averages $r(Z_i)\ell(h, Z_i)$ with $\ell \in [0,1]$. Prove an explicit sufficient sample size for the importance-weighted ERM to have P-risk within ε of the best member of $\mathcal{H}$, with confidence $1 - \delta$.

5 points Let $\ell(w, z)$ be convex in w and let W be a random parameter with finite mean $\bar{w}$. Prove

$$L(\bar{w}) \le \mathbb{E}_W L(W).$$

State why this does not imply that majority vote under 0–1 loss always improves.

4 points Prove for arbitrary binary classes $\mathcal{H}, \mathcal{G}$ that

$$\tau_{\mathcal{H}\cup\mathcal{G}}(m) \le \tau_{\mathcal{H}}(m) + \tau_{\mathcal{G}}(m).$$

Give one reason the inequality can be strict.

B. Deep Learning and Reinforcement Learning [33 points]

(a) **A linear denoising autoencoder.** [10 points]

Let $X \in \mathbb{R}^d$ have mean zero and covariance $\Sigma \succeq 0$. Observe $Y = X + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ is independent of X. A linear denoiser predicts AY.

(a) [5 points] Derive the population minimizer of $\mathbb{E}\,\|X - AY\|_2^2$. State a valid answer when $\Sigma + \sigma^2 I$ is invertible.

(b) [3 points] Express the action of the optimal matrix in an eigenbasis of Σ and interpret the limits $\sigma \downarrow 0$ and $\sigma \uparrow \infty$.

(c) [2 points] Explain why an unconstrained deterministic autoencoder trained on clean input and target may learn the identity, and how denoising changes the learning problem.

(b) **Conditional scores and guidance.** [14 points]

At noise level t, suppose $p_t(x, c)$ is differentiable in x. Define $s_u(x, t) = \nabla_x \log p_t(x)$ and $g_c(x, t) = \nabla_x \log p_t(c \mid x)$.

(a) [4 points] Use Bayes' rule to derive the exact conditional score $\nabla_x \log p_t(x \mid c)$.

(b) [4 points] If $p_t(x_t \mid x_0) = \mathcal{N}(a_t x_0, b_t^2 I)$, derive its conditional score and give a denoising score-matching objective that does not require the marginal score as a target.

(c) [4 points] Define $s_w = s_u + w g_c$. For which w is this the exact conditional score? For general $w > 0$, identify an unnormalized density whose score is s_w and explain why an unknown normalization constant does not affect the score.

(d) [2 points] Give two reasons why increasing w need not monotonically improve sample quality.

(c) **Entropy-regularized decision making.** [9 points]

A one-state bandit has finite actions with rewards r_a. For $\tau > 0$, optimize

$$\max_{\pi \in \Delta} \left\{ \sum_a \pi_a r_a + \tau H(\pi) \right\}, \qquad H(\pi) = -\sum_a \pi_a \log \pi_a.$$

(a) [5 points] Derive the unique optimal policy, including its normalizing constant.

(b) [2 points] Analyze the limits $\tau \downarrow 0$ and $\tau \uparrow \infty$.

(c) [2 points] Show that adding the same constant to all rewards leaves the policy unchanged, and interpret τ.

C. Optimization Methods in Artificial Intelligence [33 points]

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, L-smooth, and μ-strongly convex, where $0 < \mu \le L$. Let $x^\star$ be its minimizer and

$$\Delta_k = f(x^k) - f(x^\star), \qquad \Delta_0 < \infty.$$

Assume the stochastic-gradient oracle satisfies

$$\mathbb{E}[g(x,\xi) \mid x] = \nabla f(x), \qquad \mathbb{E}[\|g(x,\xi) - \nabla f(x)\|^2 \mid x] \le \sigma^2.$$

Before iteration k, let $\mathcal{F}_k$ contain all previously revealed randomness. Conditionally on $\mathcal{F}_k$, draw $\xi_{k,1}, \dots, \xi_{k,\tau}$ independently and define

$$G_k = \frac{1}{\tau}\sum_{j=1}^{\tau} g(x^k, \xi_{k,j}), \qquad x^{k+1} = x^k - \eta G_k,$$

where $\tau \ge 1$ and $0 < \eta \le 1/L$ are fixed.

(1) [3 points] State the first-order definitions of L-smoothness and μ-strong convexity. Explain why $x^\star$ is unique and $\nabla f(x^\star) = 0$.

(2) [5 points] Prove the conditional mini-batch identities

$$\mathbb{E}_k[G_k] = \nabla f(x^k), \qquad \mathbb{E}_k\|G_k - \nabla f(x^k)\|^2 \le \frac{\sigma^2}{\tau}.$$

Identify precisely why all cross terms vanish.

(3) [5 points] Starting only from μ-strong convexity, prove

$$\|\nabla f(x)\|^2 \ge 2\mu[f(x) - f(x^\star)].$$

Hint: Minimize the strong-convexity lower model over its comparison point.

(4) [7 points] Apply the function-value descent lemma at $x^{k+1} = x^k - \eta G_k$, take conditional expectation, and prove

$$\mathbb{E}_k[\Delta_{k+1}] \le (1 - \mu\eta)\Delta_k + \frac{L\eta^2\sigma^2}{2\tau}.$$

Display $\mathbb{E}_k\|G_k\|^2$, use $\eta \le 1/L$ with the correct constant, and then invoke part (3).

(5) [5 points] Use the tower property and solve the recurrence to show

$$\mathbb{E}[\Delta_K] \le (1 - \mu\eta)^K \Delta_0 + \frac{L\eta\sigma^2}{2\mu\tau}[1 - (1 - \mu\eta)^K].$$

Identify the limiting function-value noise floor.

(6) [5 points] Given $\varepsilon > 0$, give explicit sufficient choices of η, τ, and an integer K that ensure $\mathbb{E}[\Delta_K] \le \varepsilon$. Handle $\sigma = 0$ and the case in which the transient requirement is already satisfied at $K = 0$.

(7) [3 points] Count stochastic-gradient evaluations. Explain the tradeoff between decreasing η, increasing τ, and increasing K, including the possible benefit of parallel mini-batch evaluation.

D. Natural Language Processing [33 points]

1. Autoregressive and masked pre-training objectives. [10 points] Let $x_{1:T}$ be a sequence from a finite vocabulary.

3 points Write the autoregressive factorization of $p_\theta(x_{1:T})$ and its token-level negative log-likelihood. Prove briefly that normalized conditionals define a normalized joint distribution.

3 points Let a random set M of positions be masked. Write a masked-language- modeling objective that predicts x_t for $t \in M$ from $x_{\overline{M}}$ and M. Explain why this objective does not, in general, define one normalized joint probability over complete sequences.

4 points Compare the information available to the two objectives, their natural downstream uses, and the cost of assigning a score to every token of a held-out sequence. State one train–test mismatch or leakage pitfall for each objective.

2. Subword-composed word representations. [10 points] For a word w, let $G(w)$ be its nonempty set of character n-grams and define

$$\mathbf{h}_w = \frac{1}{|G(w)|} \sum_{g \in G(w)} \mathbf{z}_g, \qquad p(c \mid w) = \frac{\exp(\mathbf{h}_w^\top \mathbf{v}_c)}{\sum_{c'=1}^{V} \exp(\mathbf{h}_w^\top \mathbf{v}_{c'})}.$$

For one observed pair (w, c^*), let $\ell = -\log p(c^* \mid w)$.

5 points Show that $p(\cdot \mid w)$ is normalized and derive $\nabla_{\mathbf{z}_g} \ell$ for every $g \in G(w)$ and $\nabla_{\mathbf{v}_c} \ell$ for every context c.

3 points Explain why this parameterization helps rare or unseen words in a morphologically rich language. Give one failure mode and a targeted mitigation.

2 points State the time and parameter-memory costs of the full-softmax computation in terms of $|G(w)|$, V, and dimension d. Name one scalable training approximation and what objective change it introduces.

3. A bounded external memory. [13 points] Chunk representations $h_1, \ldots, h_J \in \mathbb{R}^d$ are compressed by scalar gates $g_j \in (0, 1)$:

$$m_0 = 0, \qquad m_j = g_j m_{j-1} + (1 - g_j) h_j.$$

A separate store contains at most m slot rows in $M \in \mathbb{R}^{m \times d}$.

4 points Expand m_J as a sum of the h_i, compute the sum of its coefficients, and characterize decay with age when every $g_j = g$.

3 points Define a normalized attention read for a query $q \in \mathbb{R}^d$, including the softmax axis. State read time, persistent memory, and additional working memory.

4 points For an unbounded stream in which later updates may correct earlier statements, specify write, merge, supersede, and eviction operations under a deterministic m-slot budget. Give an end-to-end answer-training objective and one mechanism that prevents a degenerate generic memory.

2 points Give an abstention rule tied to source versions and two metrics that separately expose catastrophic forgetting and stale-evidence use.