Artificial Intelligence PhD Qualifying Examination

High-Fidelity Mock Examination 09

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 [4 points]. Select one answer per item.

(i) [0.5 points] Logistic negative log-likelihood is

(a) convex in linear-model parameters.

(b) 0–1 loss.

(c) always strongly concave.

(d) a kernel matrix.

(ii) [0.5 points] An interval labels a contiguous block of ordered sample points

(a) positive.

(b) with arbitrary alternation.

(c) only real-valued.

(d) independently of endpoints.

(iii) [0.5 points] ℓ_1–ℓ_∞ duality produces a maximum over

(a) coordinates.

(b) sample sizes.

(c) labels only.

(d) Hessians.

(iv) [0.5 points] A one-dimensional ReLU network represents a

(a) continuous piecewise-linear function.

(b) nonmeasurable function.

(c) polynomial necessarily.

(d) constant necessarily.

(v) [0.5 points] If a design matrix is rank deficient, logistic Hessian may be

(a) PSD but not PD.

(b) negative definite.

(c) nonsymmetric.

(d) integer-valued.

(vi) [0.5 points] Massart's finite-class lemma introduces

(a) $\sqrt{\log |\mathcal{A}|}$.

(b) $|\mathcal{A}|^2$ necessarily.

(c) no radius.

(d) a margin constraint.

(vii) [0.5 points] Two points on the real line can be shattered by

(a) intervals.

(b) one fixed constant.

(c) no binary class.

(d) a single label.

(viii) [0.5 points] A slope jump in a piecewise-linear function can be represented by

(a) a shifted ReLU.

(b) a constant kernel only.

(c) a label flip.

(d) a VC upper bound.

Part II: Theoretical Exercises [29 points].

7 points For binary logistic regression with augmented design $\widetilde{X}$, parameter θ, and $p_i = \sigma(\theta^\top \widetilde{x}_i)$, derive the gradient and Hessian of average negative log-likelihood. Prove convexity and state exactly when this Hessian is positive definite at finite θ.

8 points Let $\mathcal{I}$ be indicators of closed intervals on $\mathbb{R}$, with the empty interval allowed.

(a) [4 points] On n distinct ordered points, prove $\tau_{\mathcal{I}}(n) = 1 + n(n+1)/2$.

(b) [4 points] Determine $\mathrm{VCdim}(\mathcal{I})$.

7 points Let $\mathcal{G}_1(B) = \{x \mapsto \langle w, x\rangle : \|w\|_1 \le B\}$ and assume $\|x_i\|_\infty \le R$. You may use Massart's lemma:

$$\mathbb{E}_\sigma \max_{a \in \mathcal{A}} \langle \sigma, a\rangle \le r\sqrt{2\log|\mathcal{A}|} \quad \text{if } \|a\|_2 \le r.$$

Prove

$$\widehat{\mathfrak{R}}_{S_X}(\mathcal{G}_1(B)) \le BR\sqrt{\frac{2\log(2d)}{m}}.$$

7 points Let $g : \mathbb{R} \to \mathbb{R}$ be continuous and piecewise affine with knots $t_1 < \cdots < t_K$. Its slope is s_0 to the left of t_1, and the slope jumps by Δ_j at t_j. Prove that for a suitable constant a,

$$g(x) = a + s_0 x + \sum_{j=1}^{K} \Delta_j [x - t_j]_+.$$

Explain how this yields a finite one-hidden-layer ReLU representation and why continuity is essential.

B. Deep Learning and Reinforcement Learning [33 points]

(a) **A shared Siamese computation graph.** [9 points]

Two views x_1, x_2 are passed through the same encoder:

$$h_i = f_\theta(x_i) \neq 0, \qquad z_i = \frac{h_i}{\|h_i\|_2}, \qquad \ell = -\langle z_1, z_2 \rangle .$$

(a) [3 points] Derive the Jacobian of $h \mapsto h/\|h\|_2$.

(b) [4 points] Compute $\nabla_{h_1}\ell$, $\nabla_{h_2}\ell$, and $\nabla_\theta\ell$. Make explicit why the shared parameter receives two branch contributions.

(c) [2 points] Explain how applying stop-gradient to z_2 changes the update, and why the stated positive-pair loss alone admits a collapsed global solution.

(b) **Implicit score matching.** [15 points]

Let p be a positive differentiable density on $\mathbb{R}^d$ with score $s_p(x) = \nabla_x \log p(x)$. Let $s_\theta : \mathbb{R}^d \to \mathbb{R}^d$ be differentiable. Assume the integrability and boundary conditions required for integration by parts; in particular, $p(x)s_{\theta,j}(x)$ vanishes at the boundary for every coordinate.

(a) [4 points] Expand $\frac{1}{2}\mathbb{E}_p \|s_\theta(X) - s_p(X)\|_2^2$ and identify which term is independent of θ.

(b) [5 points] Use integration by parts, coordinate by coordinate, to show

$$\mathbb{E}_p \langle s_\theta(X), s_p(X) \rangle = -\mathbb{E}_p[\nabla \cdot s_\theta(X)].$$

(c) [3 points] Deduce an objective that can be estimated from samples of p without evaluating p or s_p.

(d) [2 points] For $p = \mathcal{N}(0, I)$ and $s_\theta(x) = -x$, compute the value of the implicit objective from part (c).

(e) [1 point] State one computational difficulty in high dimension.

(c) **Adaptive early exit as an MDP.** [9 points]

A network has exits after layers $1, \ldots, L$. At exit ℓ, a controller observes the current representation and confidence and chooses either *stop* or *continue*.

(a) [3 points] Specify a Markov state, actions, transition, and reward that trade prediction utility against computation.

(b) [2 points] Write a Bellman optimality equation that displays the stop and continue alternatives.

(c) [3 points] For a differentiable stochastic controller π_η, give a REINFORCE estimator with a state baseline. State your discount convention.

(d) [1 point] Give one evaluation rule that prevents a policy from appearing better only because it uses more computation.

C. Optimization Methods in Artificial Intelligence [33 points]

Let $C \subseteq \mathbb{R}^d$ be nonempty, closed, and convex, and let $F : C \to \mathbb{R}$ be convex. Assume that $x^\star \in \arg\min_{x\in C} F(x)$ exists. Before iteration k, let $\mathcal{F}_k$ contain x^0 and all oracle randomness revealed before the current call. The stochastic oracle returns H_k satisfying

$$\mathbb{E}[H_k \mid \mathcal{F}_k] = g^k \in \partial F(x^k), \qquad \mathbb{E}[\|H_k\|^2 \mid \mathcal{F}_k] \le G^2,$$

where $G > 0$. For deterministic stepsizes $\eta_k > 0$, consider

$$x^{k+1} = P_C(x^k - \eta_k H_k), \qquad P_C(u) = \arg\min_{z\in C} \frac{1}{2}\|z - u\|^2.$$

For $K \ge 1$, define

$$S_K = \sum_{k=0}^{K-1} \eta_k, \qquad \widehat{x}^K = \frac{1}{S_K}\sum_{k=0}^{K-1} \eta_k x^k, \qquad R = \|x^0 - x^\star\| > 0.$$

(1) [4 points] State the subgradient inequality. Prove the projection variational inequality

$$p = P_C(u) \quad \Longrightarrow \quad \langle u - p, z - p\rangle \le 0 \quad (\forall z \in C),$$

and use it to state why P_C is nonexpansive.

(2) [6 points] Prove the conditional one-step inequality

$$\mathbb{E}[\|x^{k+1} - x^\star\|^2 \mid \mathcal{F}_k] \le \|x^k - x^\star\|^2 - 2\eta_k[F(x^k) - F(x^\star)] + \eta_k^2 G^2.$$

Every use of conditional measurability or unbiasedness should be explicit.

(3) [6 points] Sum the recurrence, use the tower property and convexity, and prove the general weighted-averaging bound

$$\mathbb{E}[F(\widehat{x}^K) - F(x^\star)] \le \frac{R^2 + G^2\sum_{k=0}^{K-1}\eta_k^2}{2S_K}.$$

Identify the nonnegative terminal term that is discarded.

(4) [5 points] Suppose the horizon K is known. Optimize the bound over a constant stepsize and show that

$$\eta_k = \frac{R}{G\sqrt{K}} \quad \Longrightarrow \quad \mathbb{E}[F(\widehat{x}^K) - F(x^\star)] \le \frac{RG}{\sqrt{K}}.$$

Give a sufficient K for expected error at most ε.

(5) [4 points] If the horizon is unknown, take $\eta_k = R/(G\sqrt{k+1})$. Using

$$\sum_{j=1}^{K} j^{-1/2} \ge 2(\sqrt{K+1} - 1), \qquad \sum_{j=1}^{K} j^{-1} \le 1 + \log K,$$

derive an explicit bound for $\mathbb{E}[F(\widehat{x}^K) - F(x^\star)]$.

(6) [4 points] Explain why $\mathcal{F}_k$ must exclude the current oracle sample. Also prove the conditional bias–variance identity

$$\mathbb{E}[\|H_k\|^2 \mid \mathcal{F}_k] = \|g^k\|^2 + \mathbb{E}[\|H_k - g^k\|^2 \mid \mathcal{F}_k],$$

and explain why a second-moment bound is not the same statement as a variance bound.

(7) [4 points] Explain why the proof controls the weighted averaged iterate rather than automatically controlling the last iterate. Compare the $O(K^{-1/2})$ result with deterministic gradient descent on a smooth convex objective, and state why strong convexity alone does not make a constant-step nonsmooth stochastic subgradient method converge exactly at a linear rate.

D. Natural Language Processing [33 points]

1. Biased walks for node representations. [8 points] Node2Vec has just moved from node t to node v. For a neighbor x of v, let

$$\pi_{vx} = w_{vx}\alpha_{pq}(t,x), \qquad \alpha_{pq}(t,x) = \begin{cases} 1/p, & d(t,x)=0, \\ 1, & d(t,x)=1, \\ 1/q, & d(t,x)=2, \end{cases}$$

where $w_{vx} > 0$ and d is shortest-path distance.

3 points Normalize the next-step probabilities and explain the effect of p.

3 points Relate small versus large q to depth-first versus breadth-first exploration and to structural-equivalence versus community representations.

2 points State the learning objective used after walks are sampled and one bias inherited from the walk distribution.

2. Inside inference in a probabilistic grammar. [8 points] Consider the following normalized PCFG:

S rule	$S \to AB$ (0.6)	$S \to BA$ (0.4)
A rule	$A \to a$ (0.8)	$A \to b$ (0.2)
B rule	$B \to a$ (0.1)	$B \to b$ (0.9).

4 points For a general Chomsky-normal-form PCFG, define inside values and give the base case and recurrence for the probability of a sentence.

2 points Compute the probability of the sentence ab under the displayed grammar and the posterior probability that its root rule was $S \to AB$.

2 points State time and memory complexity, and explain the change needed for Viterbi parsing with recovery of the best parse.

3. A most-reliable relation path. [7 points] Each directed knowledge-base edge (u, r, v) has a calibrated marginal probability $p(u, r, v) \in (0, 1]$. For a fixed relation pattern $r_1, \ldots, r_H$, the score of a path is the product of its edge probabilities.

5 points Give a dynamic program that finds the highest-scoring path from a source s to every entity using exactly this relation pattern. State time and memory complexity in terms of H, $|V|$, and the matching edges E_h at layer h.

2 points Explain why the product need not be a calibrated path probability even when each edge is marginally calibrated, and give one remedy.

4. Evidence-aware knowledge-base question answering. [10 points] Design a system that answers multi-hop questions using both a knowledge graph and supporting text passages. The graph may contain missing or wrong edges.

4 points Specify semantic parsing, candidate path generation, text retrieval, joint scoring, and answer production with evidence.

2 points Give a training objective using positive and hard-negative paths.

2 points Define a calibrated abstention rule and metrics for reasoning and evidence quality.

2 points Give two failure modes and a targeted mitigation for each.