Artificial Intelligence PhD Qualifying Examination

High-Fidelity Mock Examination 04

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] A support vector has a dual multiplier that is typically

(a) positive.

(b) imaginary.

(c) a label.

(d) a VC dimension.

(ii) [0.5 points] Importance weighting corrects a change in

(a) sampling distribution.

(b) matrix symmetry.

(c) activation differentiability.

(d) label alphabet only.

(iii) [0.5 points] A lower-left orthant classifier is monotone under

(a) coordinatewise order.

(b) arbitrary rotations.

(c) label flips.

(d) kernel subtraction.

(iv) [0.5 points] The gradient of a convex differentiable function is

(a) monotone.

(b) always constant.

(c) a probability.

(d) necessarily bounded.

(v) [0.5 points] Validation labels may be used to

(a) choose among a pre-fixed finite list.

(b) train unlimited new candidates without correction.

(c) fit preprocessing before splitting.

(d) report training loss only.

(vi) [0.5 points] The hard-margin SVM objective fixes functional margin scale by

(a) constraints $y_i f_i \geq 1$.

(b) setting all labels to zero.

(c) removing the norm.

(d) maximizing training error.

Part II: Theoretical Exercises [30 points].

9 points For separable data, consider

$$\min_{w,b} \frac{1}{2} \|w\|_2^2 \quad \text{s.t.} \quad y_i(w^\top x_i + b) \geq 1.$$

(a) [5 points] Derive the Lagrange dual and all stationarity constraints.

(b) [4 points] For $x_1 = (-1, 0), y_1 = -1$ and $x_2 = (1, 0), y_2 = 1$, find w, b, α_1, α_2, the geometric margin, and the common primal/dual value.

6 points Target risk is under distribution P, but data are sampled i.i.d. from Q. Assume $P \ll Q$, let $r(x, y) = dP/dQ \leq W$, and let $\ell(h(x), y) \in [0, 1]$. For fixed h, define

$$\widehat{L}_{\mathrm{IW}}(h) = \frac{1}{m} \sum_{i=1}^{m} r(X_i, Y_i) \ell(h(X_i), Y_i).$$

Prove it is unbiased for $L_P(h)$ and show $\mathrm{Var}(\widehat{L}_{\mathrm{IW}}(h)) \leq W L_P(h)/m$.

8 points On $\mathbb{R}^2$, let

$$h_{a,b}(x) = \mathbf{1}\{x_1 \leq a,\ x_2 \leq b\}.$$

Determine the VC dimension of this lower-left-orthant class.

4 points Let g be differentiable and convex and set $F(w) = g(w) + (\lambda/2) \|w\|_2^2$. Prove

$$\langle \nabla F(u) - \nabla F(v), u - v \rangle \geq \lambda \|u - v\|_2^2,$$

and deduce that F has at most one stationary point.

3 points For r pre-fixed candidates evaluated with an independent m_v-sample and loss in $[0, 1]$, state a valid high-probability excess-risk penalty and the independence condition needed for it.

B. Deep Learning and Reinforcement Learning [33 points]

(a) **Permutation structure of self-attention.** [11 points]

Let token rows form $X \in \mathbb{R}^{L\times d}$ and define one-head self-attention without positional information by

$$F(X) = A(X)V(X), \quad A(X) = \mathrm{softmax}_{\mathrm{row}}\left(\frac{Q(X)K(X)^\top}{\sqrt{d_k}}\right),$$

where $Q = XW_Q$, $K = XW_K$, and $V = XW_V$.

(a) [5 points] For any $L \times L$ permutation matrix P, prove $F(PX) = PF(X)$. Explicitly justify how row-wise softmax acts on $PSP^\top$.

(b) [2 points] Is F permutation invariant or equivariant? What happens if its token outputs are followed by a symmetric sum or mean pooling?

(c) [2 points] Explain how absolute positional embeddings and a causal mask alter the symmetry. Distinguish changing the sequence order from merely relabeling all positions consistently.

(d) [2 points] Give the dominant time and attention-memory complexity in L and d for a standard d-wide attention layer.

(b) **A masked autoregressive flow.** [13 points]

Let $x \in \mathbb{R}^d$. An inverse flow from data to base variables is

$$z_i = \big(x_i - m_i(x_{<i})\big)\exp\{-s_i(x_{<i})\}, \qquad i = 1, \ldots, d,$$

with base density $p_0(z)$. A masked network computes all m_i, s_i while respecting the indicated dependencies.

(a) [5 points] Derive the forward sampling equation for x_i, show that the Jacobian $\partial z/\partial x$ is triangular, and compute its log-determinant.

(b) [4 points] Write the exact log-density and the average NLL on a dataset. Check the sign of every s_i term.

(c) [2 points] Explain why density evaluation can be parallelized across coordinates with a masked network while exact ancestral sampling is sequential.

(d) [2 points] State conditions ensuring invertibility and normalized density.

(c) **Policy improvement.** [9 points]

In a finite discounted MDP, let Q^π be the action-value function of a policy π. Define a policy π' supported only on $\arg\max_a Q^\pi(s, a)$ at every state.

(a) [5 points] Prove the policy-improvement result $V^{\pi'}(s) \geq V^\pi(s)$ for all states.

(b) [2 points] Explain why the improved policy need not be unique and why one greedy improvement step need not produce an optimal policy.

(c) [2 points] Describe exact policy iteration, including its evaluation, improvement, and stopping conditions.

C. Optimization Methods in Artificial Intelligence [33 points]

Consider

$$f(x) = \frac{1}{2}x^\top Q x + c^\top x, \qquad Q = Q^\top \succ 0,$$

$$\mu = \lambda_{\min}(Q), \qquad L = \lambda_{\max}(Q), \qquad 0 < \mu < L.$$

Heavy-Ball iteration with $x^{-1} = x^0$ is

$$x^{k+1} = x^k - \gamma \nabla f(x^k) + \beta(x^k - x^{k-1}), \qquad \gamma > 0, \quad 0 \le \beta < 1.$$

(1) [4 points] Find $x^\star$ and derive the recurrence for $e^k = x^k - x^\star$.

(2) [5 points] Diagonalize Q and show that an eigencoordinate z^k associated with $\lambda \in [\mu, L]$ satisfies

$$z^{k+1} = (1 + \beta - \gamma\lambda)z^k - \beta z^{k-1}.$$

Derive its characteristic polynomial.

(3) [6 points] You may use the Jury criterion: the roots of $r^2 - ar + \beta$ are strictly inside the unit disk exactly when

$$|\beta| < 1, \qquad 1 - a + \beta > 0, \qquad 1 + a + \beta > 0.$$

Prove that all eigendirections are stable if and only if

$$0 < \gamma < \frac{2(1+\beta)}{L}, \qquad 0 \le \beta < 1.$$

(4) [6 points] Define

$$q = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}, \qquad \gamma_\star = \frac{4}{(\sqrt{L} + \sqrt{\mu})^2}, \qquad \beta_\star = q^2.$$

For $a_\lambda = 1 + \beta_\star - \gamma_\star \lambda$, verify $a_\mu = 2q$, $a_L = -2q$, and $|a_\lambda| \le 2q$ for every $\lambda \in [\mu, L]$.

(5) [5 points] Show that the roots of $r^2 - a_\lambda r + q^2$ have modulus q for interior eigenvalues. Describe the repeated roots at $\lambda = \mu, L$ and state the worst-case asymptotic spectral radius.

(6) [4 points] With $\kappa = L/\mu$, compare $q = (\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)$ with the optimal fixed-step gradient descent factor $(\kappa - 1)/(\kappa + 1)$. State the resulting asymptotic condition-number dependence.

(7) [3 points] Explain why stability need not make $f(x^k)$ decrease at every step, and why this quadratic spectral proof is not a convergence theorem for every smooth nonconvex objective.

D. Natural Language Processing [33 points]

1. Tokenization and perplexity. [7 points] Two language models assign the same probability 0.01 to a character string. Model A tokenizes it into two tokens; Model B tokenizes it into five tokens.

3 points Compute the per-token perplexity assigned by each model to this one string.

2 points Explain why this does not imply that Model B is the better language model.

2 points Propose a normalization that permits a more meaningful comparison across tokenizers, and state one remaining limitation.

2. HMM sequence labeling. [10 points] An HMM for named-entity recognition has labels $y_t \in \{1, \ldots, K\}$, transition probabilities $A_{ij} = p(y_t = j \mid y_{t-1} = i)$, and emissions $B_j(x_t) = p(x_t \mid y_t = j)$.

3 points Write the joint probability of $x_{1:L}, y_{1:L}$, including an initial-state distribution.

4 points Give the log-domain Viterbi recursion and its time complexity.

3 points A test token was never observed with any label in training. Give two principled ways to define its emission probabilities and explain the inductive bias of each.

3. Evaluation under class imbalance. [6 points] A two-class entity tagger has the following entity-level counts:

	TP	FP	FN
PERSON	90	10	10
CHEMICAL	2	1	8

Compute the per-class F1 scores, macro-F1, and micro-F1. Which summary better exposes the rare-class failure, and why?

4. Low-resource multilingual NER. [10 points] You have abundant labeled NER data in English, unlabeled text in a morphologically rich target language, and only 200 labeled target-language sentences. Design a system that uses all three resources.

4 points Specify the representation, prediction architecture, and supervised objectives.

2 points Give one self-training or consistency objective on unlabeled target text, with a safeguard against confirmation bias.

2 points Explain how morphology and script differences affect tokenization and alignment, and give a targeted mitigation.

2 points Specify evaluation splits and diagnostics that would reveal negative transfer from English.