Artificial Intelligence PhD Qualifying Examination

High-Fidelity Mock Examination 03

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 valid kernel requires every finite Gram matrix to be

(a) PSD.

(b) diagonal.

(c) invertible.

(d) entrywise positive.

(ii) [0.5 points] The pointwise product of two PSD kernels is

(a) PSD.

(b) always indefinite.

(c) nonsymmetric.

(d) a loss.

(iii) [0.5 points] For a fixed classifier, empirical-loss variance scales as

(a) $1/m$.

(b) m.

(c) 2^m.

(d) a constant necessarily.

(iv) [0.5 points] A realizable consistent learner needs a probability bound on

(a) bad hypotheses surviving the sample.

(b) Hessian rank only.

(c) kernel uniqueness.

(d) test-set training.

(v) [0.5 points] Strong convexity in regularized ERM is useful for

(a) controlling how much the minimizer moves.

(b) increasing VC dimension.

(c) making losses unbounded.

(d) deleting the sample.

(vi) [0.5 points] Validation candidates must be constructed

(a) without the validation labels.

(b) from the test labels.

(c) after unlimited validation adaptation.

(d) only after deployment.

Part II: Theoretical Exercises [30 points].

8 points (a) [5 points] Let K_1, K_2 be PSD kernels. Prove that $aK_1 + bK_2$ is PSD for $a, b \geq 0$, and that the pointwise product K_1K_2 is PSD.

(b) [3 points] Construct $\psi : \mathbb{R}^2 \to \mathbb{R}^6$ such that $\langle \psi(x), \psi(x') \rangle = (1 + x^\top x')^2$.

5 points A population has two strata with known masses $\pi_1, \pi_2 > 0$, $\pi_1 + \pi_2 = 1$. A fixed classifier has conditional errors $p_1, p_2 \in (0, 1)$. From stratum g we draw m_g independent examples and observe empirical error $\widehat{p}_g$. For $\widehat{L} = \pi_1\widehat{p}_1 + \pi_2\widehat{p}_2$:

(a) [2 points] Prove unbiasedness and compute the variance.

(b) [3 points] Treating m_1, m_2 as positive reals with $m_1 + m_2 = m$, derive the variance-minimizing allocation.

6 points On $X = \{0, 1\}^d$, consider one-variable decision lists

$$h_{j,b,c}(x) = \begin{cases} b, & x_j = 1, \\ c, & x_j = 0, \end{cases} \qquad j \in [d], \quad b, c \in \{0, 1\}.$$

Assume realizability. Prove a sufficient sample size for any consistent learner over this class to have error at most ε with probability at least $1 - \delta$.

7 points For a sample $S = (z_1, \ldots, z_m)$ define

$$F_S(w) = \frac{1}{m} \sum_{i=1}^{m} \ell(w, z_i) + \frac{\lambda}{2} \|w\|_2^2 ,$$

where $w \mapsto \ell(w, z)$ is convex and ρ-Lipschitz. Let w_S minimize F_S. If S, S' differ in one example, prove

$$\|w_S - w_{S'}\|_2 \leq \frac{2\rho}{\lambda m}, \qquad |\ell(w_S, z) - \ell(w_{S'}, z)| \leq \frac{2\rho^2}{\lambda m}.$$

You may use: if F is λ-strongly convex and minimized at u, then $F(v) \geq F(u) + (\lambda/2)\|v - u\|^2$.

4 points Let $\mathcal{C}_d = \{x \mapsto x_j,\ x \mapsto 1 - x_j : j \in [d]\}$ on $\{0, 1\}^d$. Prove $\tau_{\mathcal{C}_d}(m) \leq \min\{2^m, 2d\}$ and deduce $\mathrm{VCdim}(\mathcal{C}_d) \leq \lfloor \log_2(2d) \rfloor$.

B. Deep Learning and Reinforcement Learning [33 points]

(a) **Image restoration as conditional prediction.** [11 points]

Let (X, Y) be square-integrable random variables, where a degraded image satisfies $Y = \mathcal{A}(X) + N$. A network f predicts X from Y.

(a) [4 points] Prove that the population minimizer of $\mathbb{E}\,\|X - f(Y)\|_2^2$ is $f^*(Y) = \mathbb{E}[X \mid Y]$. Your proof should include an orthogonal decomposition of the risk.

(b) [2 points] In the scalar case, state the corresponding Bayes predictor under absolute loss $\mathbb{E}|X - f(Y)|$. Explain why squared loss may produce a visually blurred prediction when $p(x \mid y)$ is multimodal.

(c) [3 points] Specify a minimal U-Net-like restoration network, including input, output, skip connections, and a training objective. Explain the role of the skips.

(d) [2 points] Give two distinct failure modes of the degradation model or loss, and one appropriate evaluation metric beyond training loss.

(b) **Dequantization and normalizing flows.** [13 points]

An image Y takes values in $\{0, \dots, K-1\}^d$, whereas a flow defines a continuous density $p_\theta(x)$. Let $U \sim \mathrm{Unif}([0,1)^d)$ and $X = Y + U$.

(a) [5 points] Define the discrete mass induced by p_θ on each quantization cell. Prove that training on uniformly dequantized samples maximizes a lower bound on the discrete log-likelihood. State the direction of Jensen's inequality.

(b) [4 points] For an invertible differentiable map $x = f_\theta(z)$ with base density p_0, derive $\log p_\theta(x)$ and the dataset-average negative log-likelihood. Make the sign of the log-determinant explicit.

(c) [2 points] Explain why dequantization is not merely ordinary data augmentation. What quantity should be reported when comparing image likelihoods of different dimensions?

(d) [2 points] For the scalar flow $x = az + b$, $a \neq 0$, write the exact log-density of x in terms of $z = (x - b)/a$ and p_0.

(c) **A Bellman-residual certificate.** [9 points]

For a fixed policy π, let

$$(T^\pi V)(s) = r^\pi(s) + \gamma \sum_{s'} P^\pi(s' \mid s) V(s'), \qquad \gamma \in (0, 1).$$

(a) [3 points] Prove that T^π is a γ-contraction in the sup norm.

(b) [4 points] If an approximate value function satisfies $\|V - T^\pi V\|_\infty \leq \delta$, prove $\|V - V^\pi\|_\infty \leq \delta/(1-\gamma)$.

(c) [2 points] Evaluate the certificate for $\delta = 0.02$ and $\gamma = 0.95$. Explain why a small residual can still give a loose guarantee when γ is close to one.

C. Optimization Methods in Artificial Intelligence [33 points]

Let $A \in \mathbb{R}^{m\times d}$ be nonzero, $b \in \mathbb{R}^m$, and $\lambda > 0$. Consider the sparse least-squares problem

$$\min_{x\in\mathbb{R}^d} F(x) := f(x) + R(x), \qquad f(x) = \frac{1}{2}\|Ax - b\|^2, \qquad R(x) = \lambda\|x\|_1.$$

Let $L = \|A\|_{\mathrm{op}}^2$, $\gamma = 1/L$, and define

$$x^+ = \mathrm{prox}_{\gamma R}\big(x - \gamma\nabla f(x)\big), \qquad G_\gamma(x) = \frac{1}{\gamma}(x - x^+).$$

Assume that a minimizer $x^\star$ exists.

(1) [3 points] Compute $\nabla f(x)$ and prove that it is L-Lipschitz. State which part of F is allowed to be nonsmooth.

(2) [5 points] Starting from the definition of a proximal operator, derive the coordinatewise soft-thresholding formula

$$[\mathrm{prox}_{\gamma\lambda\|\cdot\|_1}(u)]_i = \mathrm{sign}(u_i)(|u_i| - \gamma\lambda)_+.$$

Include the case $|u_i| \le \gamma\lambda$.

(3) [4 points] Write the resulting ISTA update explicitly. Prove that $G_\gamma(x) = 0$ if and only if x minimizes F.

(4) [5 points] Show that x^+ minimizes the local quadratic model

$$Q_\gamma(z; x) = f(x) + \langle\nabla f(x), z - x\rangle + \frac{1}{2\gamma}\|z - x\|^2 + R(z).$$

Explain why $Q_\gamma(z; x) \ge F(z)$ for every z when $\gamma \le 1/L$, and deduce that $F(x^+) \le F(x)$.

(5) [6 points] For arbitrary $z \in \mathbb{R}^d$, prove

$$F(x^+) - F(z) \le \frac{\|x - z\|^2 - \|x^+ - z\|^2}{2\gamma} - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\|x^+ - x\|^2.$$

Your proof should identify the subgradient of R at the new point x^+.

(6) [6 points] Apply the preceding inequality to the ISTA iterates and show that, for all $K \ge 1$,

$$F(x^K) - F(x^\star) \le \frac{L\|x^0 - x^\star\|^2}{2K}.$$

Explain where monotonicity of $F(x^k)$ is used.

(7) [4 points] Give the dominant per-iteration cost for dense and sparse A. Explain precisely how the threshold $\gamma\lambda$ can create zero coordinates, and state why this fact alone does not imply that the minimizer is unique.

D. Natural Language Processing [33 points]

1. Multiple-positive contrastive retrieval. [**8 points**] Let a query encoder produce $q \in \mathbb{R}^d$. In a batch of document embeddings $d_1, \ldots, d_n$, the set $P \subseteq \{1, \ldots, n\}$ contains all known positives. With $s_j = q^\top d_j$, consider

$$\ell(q) = -\log \frac{\sum_{p \in P} e^{s_p/\tau}}{\sum_{j=1}^{n} e^{s_j/\tau}}.$$

5 points Compute $\nabla_q \ell$ and interpret it as the difference of two weighted document averages.

3 points Compare this objective with treating only one member of P as positive. Explain how duplicate passages in a minibatch should be handled and why.

2. Calibrated hybrid retrieval. [**7 points**] For a query–document pair, a sparse retriever and a dense retriever output signals S_1, S_2. Let $R \in \{0, 1\}$ denote relevance and suppose each calibrated signal is a log-likelihood ratio

$$\ell_i(S_i) = \log \frac{p(S_i \mid R = 1)}{p(S_i \mid R = 0)}.$$

Assuming $S_1 \perp S_2 \mid R$, derive the posterior log-odds of relevance. Explain how the formula motivates score fusion, and discuss what can go wrong if raw BM25 and cosine scores are added without calibration or if the conditional-independence assumption fails.

3. Global entity linking. [**8 points**] A document contains mentions $m_1, \ldots, m_T$. Each mention has at most K candidate entities. A global assignment $e_{1:T}$ is scored by

$$F(e_{1:T}) = \sum_{t=1}^{T} a_t(e_t) + \lambda \sum_{t=2}^{T} b(e_{t-1}, e_t),$$

where a_t is a local mention–entity score and b is a coherence score.

5 points Give an exact dynamic program for the maximizing assignment and its time and memory complexity, including backtracking.

3 points If coherence is instead included between every pair of mentions, explain why the same recursion no longer applies and give one reasonable approximate inference strategy.

4. Cross-lingual entity-linking system. [**10 points**] You must link names in news written in two languages to an English knowledge base. There are no manually labeled links in the second language; transliterations are ambiguous and the KB is incomplete.

4 points Specify candidate generation, local scoring, and a document-level inference method. State the inputs and outputs of each component.

2 points Propose weak-supervision and training objectives without assuming a complete bilingual dictionary.

2 points Define an abstention/NIL decision and metrics that do not hide failure on rare entities.

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