Artificial Intelligence PhD Qualifying Examination

High-Fidelity Mock Examination 10

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] Symmetric label noise below rate 1/2

(a) preserves risk ordering.

(b) reverses every ordering.

(c) removes labels.

(d) makes every risk zero.

(ii) [0.5 points] An Occam bound with prior $\pi(h)$ penalizes a hypothesis by

(a) $\log(1/\pi(h))$.

(b) $\pi(h)^2$ only.

(c) no complexity.

(d) its Hessian rank.

(iii) [0.5 points] Axis-aligned boxes in $\mathbb{R}^d$ have VC dimension

(a) $2d$.

(b) 1 for all d.

(c) infinite.

(d) 2^d necessarily.

(iv) [0.5 points] Jensen's inequality applies directly to

(a) convex functions.

(b) arbitrary discontinuous losses.

(c) label counts only.

(d) VC dimensions.

(v) [0.5 points] At symmetric noise rate 1/2, observed labels carry

(a) no clean-label signal.

(b) perfect clean signal.

(c) a larger margin.

(d) a unique feature map.

(vi) [0.5 points] A countable-class union bound can allocate failure probability

(a) proportional to a prior mass.

(b) only equally.

(c) using test labels.

(d) without summability.

Part II: Theoretical Exercises [30 points].

8 points Let $\widetilde{Y} = Y \oplus B$, where $B \sim \text{Bernoulli}(\eta)$ is independent of (X, Y) and $\eta < 1/2$. For classifier h, define clean and noisy 0–1 risks $L(h)$ and $\widetilde{L}(h)$.

(a) [4 points] Prove $\widetilde{L}(h) = \eta + (1 - 2\eta)L(h)$.

(b) [4 points] Suppose

$$\widetilde{L}(h) - \inf_{g \in \mathcal{H}} \widetilde{L}(g) \leq \gamma.$$

Derive the corresponding clean-risk excess bound and explain its behavior as $\eta \uparrow 1/2$.

7 points Let $\mathcal{H}$ be countable and let π be a prior with $\pi(h) > 0$ and $\sum_h \pi(h) = 1$. Data are realizable. Prove that, with probability at least $1 - \delta$, every hypothesis consistent with all m examples satisfies

$$L(h) \leq \frac{\log(1/\pi(h)) + \log(1/\delta)}{m}.$$

8 points Let $\mathcal{B}_d$ be the indicators of axis-aligned boxes $\prod_{j=1}^{d}[a_j, b_j] \subset \mathbb{R}^d$. Determine $\text{VCdim}(\mathcal{B}_d)$.

7 points Let $F : \mathbb{R}^d \to \mathbb{R}$ be λ-strongly convex and let W be a random vector with finite second moment. Prove the strong Jensen inequality

$$\mathbb{E}F(W) \geq F(\mathbb{E}W) + \frac{\lambda}{2}\mathbb{E}\,\|W - \mathbb{E}W\|_2^2.$$

Interpret what it says about randomizing parameters under a strongly convex objective.

B. Deep Learning and Reinforcement Learning [33 points]

(a) **A decoder-only Transformer.** [9 points]

Let a length-T token sequence be represented by $X \in \mathbb{R}^{T\times d}$.

(a) [3 points] Define the causal attention mask and the dimensions of Q, K, V and the attention matrix for one d_k-dimensional head. State the softmax axis.

(b) [3 points] Explain why teacher-forced training can compute losses for all token positions in parallel, whereas exact autoregressive sampling remains sequential. Do not claim that causal masking makes sampling parallel.

(c) [3 points] Compare the per-layer time complexity of training, naive generation that recomputes every prefix, and generation with a KV cache. Give the cache memory order in T and d.

(b) **Autoregressive likelihood and normalization.** [14 points]

For a finite vocabulary, define

$$p_\theta(x_{1:T}) = \prod_{t=1}^{T} p_\theta(x_t \mid x_{<t}).$$

(a) [4 points] Prove by iterated summation that this is a normalized joint distribution when every conditional is normalized. Is the chain factorization itself an approximation?

(b) [3 points] Write the average maximum-likelihood training loss for a dataset of variable-length sequences, including a padding mask and a clear sum-versus-average convention.

(c) [2 points] Define exposure bias. Explain why its existence does not mean that teacher forcing fails to maximize the stated joint likelihood.

(d) [3 points] Give the exact sampling procedure and two distinct reasons why high test likelihood need not imply the best perceived sample quality.

(e) [2 points] Explain what changes if sequence scores are divided by sequence length during evaluation or decoding. Does this define the same joint probability?

(c) **Sequence policy gradient with a reference policy.** [10 points]

For a fixed prompt x, a policy generates $y = (y_1, \ldots, y_T)$ with $\pi_\eta(y \mid x) = \prod_t \pi_\eta(y_t \mid x, y_{<t})$. A frozen reference policy is π_{ref}. Assume the target policy has support only where the reference is positive.

(a) [3 points] Formulate token generation as an episodic MDP and prove the sequence log-ratio decomposition

$$\log \frac{\pi_\eta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} = \sum_t \log \frac{\pi_\eta(y_t \mid x, y_{<t})}{\pi_{\text{ref}}(y_t \mid x, y_{<t})}.$$

(b) [3 points] For an external terminal reward $R(x, y)$, derive the trajectory likelihood-ratio policy-gradient estimator and state how reward-to-go applies if rewards are available per token.

(c) [3 points] Consider $J(\eta) = \mathbb{E}_{y\sim\pi_\eta}[R(x, y) - \beta \log(\pi_\eta(y \mid x)/\pi_{\text{ref}}(y \mid x))]$. Give an unbiased score-function form of its gradient and explain why the direct derivative of the log-policy contributes zero in expectation.

(d) [1 point] State the condition under which a state baseline preserves unbiasedness.

C. Optimization Methods in Artificial Intelligence [33 points]

Let $h : \mathbb{R}^d \to \mathbb{R}$ be convex and L_0-smooth. Assume that $x^\star \in \arg\min h$ exists and $\|x^\star\| \le R$, where $R > 0$. For $\lambda > 0$, define

$$F_\lambda(x) = h(x) + \frac{\lambda}{2}\|x\|^2, \qquad x_\lambda = \arg\min_x F_\lambda(x).$$

Run gradient descent

$$x^{k+1} = x^k - \frac{1}{L_0 + \lambda}[\nabla h(x^k) + \lambda x^k].$$

(1) [5 points] Prove that F_λ is $(L_0 + \lambda)$-smooth and λ-strongly convex. Show that its minimizer exists and is unique, state its condition number, and give its first-order optimality condition.

(2) [5 points] Compare $F_\lambda(x_\lambda)$ with $F_\lambda(x^\star)$ and prove

$$0 \le h(x_\lambda) - h(x^\star) \le \frac{\lambda R^2}{2}.$$

(3) [7 points] Using the descent lemma and strong convexity, prove

$$F_\lambda(x^K) - F_\lambda(x_\lambda) \le \left(1 - \frac{\lambda}{L_0 + \lambda}\right)^K [F_\lambda(x^0) - F_\lambda(x_\lambda)].$$

Identify exactly why the rate is linear.

(4) [6 points] Prove the original-objective bound

$$h(x^K) - h(x^\star) \le [F_\lambda(x^K) - F_\lambda(x_\lambda)] + \frac{\lambda R^2}{2},$$

and identify the optimization and regularization-bias terms.

(5) [6 points] Given $\varepsilon > 0$, set $\lambda = \varepsilon/R^2$. With $\Delta_\lambda = F_\lambda(x^0) - F_\lambda(x_\lambda)$, give a sufficient integer K for $h(x^K) - h(x^\star) \le \varepsilon$, including the case $\Delta_\lambda \le \varepsilon/2$. Explain the speed–bias tradeoff.

(6) [4 points] For $h(x) = \frac{1}{2}\|Ax - b\|^2$, give x_λ and the Hessian condition number in terms of the extreme singular values of A. Explain the rank-deficient case.

D. Natural Language Processing [33 points]

1. Orthogonal cross-lingual alignment. [8 points] Let $X, Y \in \mathbb{R}^{d\times n}$ contain paired source- and target-language anchor embeddings as columns. Consider

$$\min_{W^\top W=I} \|WX - Y\|_F^2.$$

5 points If $YX^\top = U\Sigma V^\top$ is an SVD, derive an optimal $W^* = UV^\top$. State what can be non-unique when the cross-covariance is rank deficient.

3 points Explain what geometry orthogonality preserves and give one reason why a single linear map may fail for distant languages.

2. Estimating a scaling exponent. [7 points] Suppose a family of language models follows

$$\mathcal{L}(N) = \mathcal{L}_\infty + AN^{-\alpha}, \qquad A, \alpha > 0,$$

over a measured range of parameter counts N.

2 points Derive a log-linear relation whose slope identifies α when $\mathcal{L}_\infty$ is known.

3 points The measured excess losses at sizes N and $4N$ are 0.20 and 0.10. Estimate α and predict the excess loss at $16N$ under the model.

2 points Explain why estimating or mis-specifying $\mathcal{L}_\infty$ matters, and give one reason the extrapolation may fail outside the measured range.

3. Length bias in sequence decoding. [7 points] An autoregressive model scores a completed sequence, including its end token, by $\log p(y) = \sum_{t=1}^{T} \log p(y_t \mid y_{<t})$.

3 points Candidate A has length 2 and conditional probabilities $(0.7, 0.7)$; candidate B has length 4 and probabilities $(0.8, 0.8, 0.8, 0.8)$. Which candidate wins under raw log probability and which wins under average log probability $T^{-1}\log p(y)$?

2 points Explain the source of length bias. Does ranking by average log probability preserve the original normalized joint distribution?

2 points Explain why finite-width beam search need not find the global MAP sequence and why increasing beam width need not improve human-perceived quality.

4. Multilingual, privacy-preserving RAG. [11 points] An international organization wants a QA system over multilingual internal documents. Answers must respect per-user access controls, cite evidence in its original language, and support deletion and permission-change requests.

4 points Specify the offline and online architecture, including cross-lingual retrieval, reranking, authorization, generation, and original-language citations.

2 points Define top-k retrieval with authorization as a hard constraint. Explain why filtering only after a global approximate-nearest-neighbor search can hurt both security and authorized recall.

3 points Describe propagation of deletion/permission changes through indices and caches. Give a translation/abstention policy when the only relevant evidence is in a language the generator handles poorly.

2 points Give quantitative end-to-end metrics for answer/evidence quality, privacy, and deletion compliance.