Artificial Intelligence PhD Qualifying Examination

High-Fidelity Mock Examination 01

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 exactly one answer for each question. No justification is required.

(i) [0.5 points] A weighted average of independent Bernoulli losses is unbiased for their common mean when the weights

(a) sum to one.

(b) are all zero.

(c) depend on test labels.

(d) sum to m^2.

(ii) [0.5 points] Selecting among r fixed candidates on holdout data typically costs

(a) $\sqrt{\log r / m_v}$.

(b) $r^2 m_v$ in risk.

(c) no multiplicity term.

(d) $\sqrt{m_v}$.

(iii) [0.5 points] Basis pursuit replaces

(a) ℓ_0 minimization by ℓ_1 minimization.

(b) labels by kernels.

(c) ERM by SGD.

(d) a norm by VC dimension.

(iv) [0.5 points] The VC dimension of intervals on the real line is

(a) 1.

(b) 2.

(c) 3.

(d) infinite.

(v) [0.5 points] Stronger curvature converts a fixed objective gap into

(a) a smaller parameter-distance bound.

(b) larger VC dimension.

(c) an invalid loss.

(d) more holdout leakage.

(vi) [0.5 points] If $Z_1, \dots, Z_m$ are independent, then

(a) variances of their weighted sum add with squared weights.

(b) their means must be zero.

(c) every variance is one.

(d) weights are irrelevant.

Part II: Theoretical Exercises [30 points].

6 points Let $Z_1, \dots, Z_m$ be independent Bernoulli(p) losses for a fixed classifier. For deterministic weights $a_i \geq 0$ with $\sum_i a_i = 1$, define $\widehat{L}_a = \sum_i a_i Z_i$.

(a) [3 points] Compute its expectation and variance.

(b) [3 points] Among all such weights, find those minimizing the variance and prove optimality. State the minimum variance.

7 points A training sample is used to construct r possibly data-dependent classifiers $h_1, \dots, h_r$. Conditional on the training sample, an independent holdout sample of size m_v is drawn, and $\widehat{h}$ minimizes holdout 0–1 loss. Prove that, with probability at least $1 - \delta$ over the holdout sample,

$$L(\widehat{h}) \leq \min_{j \in [r]} L(h_j) + 2\sqrt{\frac{\log(2r/\delta)}{2m_v}}.$$

Explain why conditioning on training data is legitimate and why adaptively generating new candidates after reading holdout labels invalidates this proof.

5 points Let $A \in \mathbb{R}^{q \times d}$ and suppose $y = Ax$ for an s-sparse vector x.

(a) [2 points] State the ideal ℓ_0 recovery problem and its basis-pursuit relaxation.

(b) [3 points] Prove that if every nonzero $h \in \ker(A)$ has more than $2s$ nonzero coordinates, then no two distinct s-sparse vectors have the same measurement. Explain why this injectivity statement alone does not prove that basis pursuit succeeds.

7 points Let $\mathcal{I}_2$ be the class of indicators of unions of at most two disjoint closed intervals on $\mathbb{R}$ (empty intervals are allowed). Determine $\mathrm{VCdim}(\mathcal{I}_2)$. Your proof must give a shattered set and a matching upper bound.

5 points Let $F : \mathbb{R}^d \to \mathbb{R}$ be λ-strongly convex and let $w^\star$ minimize F. Suppose an optimization algorithm returns $\widehat{w}$ satisfying $F(\widehat{w}) - F(w^\star) \leq \xi$. Prove

$$\|\widehat{w} - w^\star\|_2 \leq \sqrt{\frac{2\xi}{\lambda}}.$$

If a prediction loss $\ell(w, z)$ is ρ-Lipschitz in w, deduce a uniform bound on $|\ell(\widehat{w}, z) - \ell(w^\star, z)|$ and interpret the roles of ξ and λ.

B. Deep Learning and Reinforcement Learning [33 points]

(a) **Convolutional networks for dense prediction.** [10 points]

An input image has shape $C_{\text{in}} \times H \times W$. Consider three convolutional layers with output channels C_1, C_2, C_3. Their one-dimensional kernel, stride, padding-per-side, and dilation tuples are respectively

$$(3, 1, 1, 1), \qquad (3, 2, 1, 1), \qquad (3, 1, 2, 2).$$

The same parameters are used in the two spatial dimensions, and every output channel has one bias.

(a) [3 points] Give the output spatial size after every layer and the total number of trainable parameters.

(b) [3 points] Starting with receptive-field size $r_0 = 1$ and input jump $j_0 = 1$, compute (r_ℓ, j_ℓ) after every layer.

(c) [2 points] Prove that a stride-one convolution is translation equivariant away from boundary effects. Explain precisely what changes after the stride-two layer.

(d) [2 points] Add a minimal semantic-segmentation head. State its output shape, training loss, and one limitation caused by the downsampling layer.

(b) **The original GAN objective.** [14 points]

Let p_{data} and p_g be densities with respect to a common measure. For a fixed generator, the discriminator maximizes

$$V(D) = \mathbb{E}_{X \sim p_{\text{data}}} \log D(X) + \mathbb{E}_{X \sim p_g} \log(1 - D(X)), \qquad 0 \le D(x) \le 1,$$

with the usual limiting convention at 0 and 1.

(a) [4 points] Derive the pointwise optimal discriminator $D^*(x)$, including the cases in which one of the two densities vanishes.

(b) [4 points] Substitute D^* into V and show that the optimized value is $-\log 4 + 2\,\mathrm{JS}(p_{\text{data}}, p_g)$. Define the mixture used in the Jensen–Shannon divergence.

(c) [3 points] Write $D(G(z)) = \sigma(a)$, where a is the discriminator logit. Compare the derivatives with respect to a of the saturating generator loss $\log(1 - D(G(z)))$ and the non-saturating loss $-\log D(G(z))$. Explain the behavior when $D(G(z)) \simeq 0$.

(d) [3 points] Define mode collapse, explain why the value of a few generated samples cannot diagnose it, and give one principled mitigation.

(c) **Bellman contraction and value iteration.** [9 points]

For a finite discounted MDP with $\gamma \in (0, 1)$, define

$$(T^*V)(s) = \max_a \left\{ r(s, a) + \gamma \sum_{s'} P(s' \mid s, a) V(s') \right\}.$$

(a) [4 points] Prove that T^* is a γ-contraction in the sup norm.

(b) [3 points] Deduce that T^* has a unique fixed point V^* and that value iteration $V_{k+1} = T^* V_k$ satisfies an explicit error bound.

(c) [2 points] Give a sufficient number of iterations for $\|V_k - V^*\|_\infty \le \varepsilon$, in terms of γ, $\|V_0 - V^*\|_\infty$, and ε. Interpret the dependence on γ.

C. Optimization Methods in Artificial Intelligence [33 points]

Let $C \subseteq \mathbb{R}^d$ be nonempty, closed, and convex. Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with an L-Lipschitz gradient, and assume that a minimizer $x^\star \in \arg\min_{x \in C} f(x)$ exists. For $\gamma > 0$, consider projected gradient descent

$$x^{k+1} = P_C\big(x^k - \gamma \nabla f(x^k)\big), \qquad P_C(u) := \arg\min_{z \in C} \frac{1}{2}\|z - u\|^2,$$

and define the projected-gradient mapping

$$G_\gamma^C(x) := \frac{1}{\gamma}\left[x - P_C\big(x - \gamma \nabla f(x)\big)\right].$$

(1) [3 points] State the first-order definition of convexity and the gradient-Lipschitz definition of L-smoothness. Explain why $P_C(u)$ is single-valued under the assumptions on C.

(2) [4 points] Prove the projection variational inequality

$$p = P_C(u) \quad \Longrightarrow \quad \langle u - p, z - p\rangle \le 0 \quad \text{for every } z \in C.$$

Use it to show that P_C is nonexpansive.

(3) [4 points] Show that, for every $\gamma > 0$,

$$x^\star = P_C\big(x^\star - \gamma \nabla f(x^\star)\big).$$

Conversely, prove that any fixed point of this map minimizes f over C.

(4) [6 points] Show that $G_\gamma^C(x) = 0$ is the correct first-order stationarity condition for the constrained problem. Verify that $G_\gamma^{\mathbb{R}^d}(x) = \nabla f(x)$, and give a one-dimensional example in which a constrained minimizer has $\nabla f(x^\star) \neq 0$.

(5) [6 points] Let $x^+ = P_C(x - \gamma \nabla f(x))$. For any $z \in C$ and $0 < \gamma \le 1/L$, prove the three-point inequality

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

Hint: Use $2\langle x - x^+, x^+ - z\rangle = \|x - z\|^2 - \|x - x^+\|^2 - \|x^+ - z\|^2$.

(6) [5 points] Set $\gamma = 1/L$. Prove that the function values are nonincreasing and that, for every integer $K \ge 1$,

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

(7) [5 points] Suppose that one gradient evaluation costs T_g and one projection costs T_P. Give the iteration and arithmetic-cost bounds needed to obtain function error at most ε. Briefly explain why $\|\nabla f(x)\|$ alone can be a misleading stopping criterion when the solution lies on the boundary of C.

D. Natural Language Processing [33 points]

1. Interpolated language modeling. [6 points] Let $C(u,v,w)$, $C(u,v)$, $C(v,w)$, and $C(w)$ be corpus counts, and let N be the total number of tokens. Consider

$$p_\lambda(w \mid u,v) = \lambda_3(u,v)\frac{C(u,v,w)}{C(u,v)} + \lambda_2(u,v)\frac{C(v,w)}{C(v)} + \lambda_1(u,v)\frac{C(w)}{N},$$

where the three weights are nonnegative and sum to one. Assume the displayed denominators are nonzero.

2 points Show that $p_\lambda(\cdot \mid u,v)$ is a probability distribution.

2 points For a word w, suppose

$$C(u,v) = 10,\ C(u,v,w) = 0,\ C(v) = 40,\ C(v,w) = 4,\ C(w) = 50,\ N = 1000$$

and $(\lambda_3, \lambda_2, \lambda_1) = (0.6, 0.3, 0.1)$. Compute $p_\lambda(w \mid u,v)$ and explain the role of interpolation here.

2 points Describe how to choose the weights using held-out text. What should the model do when one of the conditioning counts is zero?

2. A globally normalized sequence model. [8 points] For an input $x_{1:L}$ and labels $y_t \in \{1,\dots,K\}$, a linear-chain CRF has score

$$S(y,x) = \sum_{t=1}^{L} s_t(y_{t-1}, y_t, x), \qquad p(y \mid x) = \frac{e^{S(y,x)}}{Z(x)},$$

where y_0 is a distinguished start label.

3 points Give a forward recursion that computes $Z(x)$.

3 points Give the Viterbi recursion for the MAP label sequence and state its time and memory complexity.

2 points Explain one way in which global normalization addresses the label-bias problem of locally normalized transition models.

3. Failure modes of translational knowledge-graph embeddings. [7 points] TransE represents an entity e by a vector $\mathbf{e}$ and a relation r by $\mathbf{r}$, and favors triples (h,r,t) with $\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|_2$ small.

3 points Suppose a symmetric relation contains both (h,r,t) and (t,r,h), each with score at most ϵ. Prove that $\|\mathbf{r}\|_2 \le \epsilon$ and $\|\mathbf{h}-\mathbf{t}\|_2 \le 2\epsilon$.

2 points Show the analogous collapse caused by fitting two one-to-many triples (h,r,t_1) and (h,r,t_2) to error at most ϵ.

2 points Propose a relation-scoring family that can model a symmetric relation without forcing linked entity vectors to coincide, and explain why.

4. Citation-grounded domain QA. [12 points] An organization has an unlabeled, daily changing collection of M technical documents. It wants a QA system that returns an answer together with sentence-level citations, may abstain when evidence is insufficient, and has an 800 ms online latency budget.

4 points Specify a complete indexing and inference pipeline, including chunking, retrieval, reranking, answer generation, citation attachment, and abstention.

3 points Give training or adaptation objectives that do not require manually labeled question–answer pairs. State how negative examples are obtained.

3 points Let d be the embedding dimension, k the number of retrieved chunks, and L_c their total token length. Identify the dominant online costs and give two concrete ways to meet the latency budget while supporting daily updates.

2 points Give a layered evaluation plan and two failure modes that the plan can distinguish.