Artificial Intelligence PhD Qualifying Examination

High-Fidelity Mock Examination 05

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]

Let $X = \{0,1\}^d$ and $\mathcal{H}_d = \{h_S(x) = \bigwedge_{j \in S} x_j : S \subseteq [d]\}$, with the empty conjunction equal to one.

8 points Determine $\mathrm{VCdim}(\mathcal{H}_d)$ by giving a shattered set and a matching finite-class upper bound.

8 points A learner retains exactly the coordinates that are one on every positive training example; with no positive example it retains all coordinates.

(a) [6 points] Under realizability by some $h_{S^\star}$, prove consistency on all positive and negative sample points, including the no-positive case.

(b) [2 points] Give training and prediction complexity.

8 points For arbitrary labeled data, let $\widehat{h}$ be an empirical 0–1 risk minimizer over $\mathcal{H}_d$. Using Hoeffding and a union bound, give an explicit sufficient m for

$$L(\widehat{h}) \le \min_{h \in \mathcal{H}_d} L(h) + \varepsilon$$

with probability at least $1 - \delta$. Explain why the brute-force ERM implementation does not give a polynomial-time guarantee in d.

9 points Fix an input x. A test response is $Y = f(x) + \epsilon$, where $\mathbb{E}\epsilon = 0$, $\mathrm{Var}(\epsilon) = \sigma^2$, and ϵ is independent of the training sample S. The learned prediction $\widehat{f}_S(x)$ is random through S.

(a) [6 points] Prove the bias–variance–noise decomposition

$$\mathbb{E}_{S,\epsilon}(\widehat{f}_S(x) - Y)^2 = (\mathbb{E}_S \widehat{f}_S(x) - f(x))^2 + \mathrm{Var}_S(\widehat{f}_S(x)) + \sigma^2.$$

(b) [3 points] Explain how to integrate the identity over random X, and why lower training error need not lower either bias or variance on new inputs.

B. Deep Learning and Reinforcement Learning [33 points]

(a) **Multi-view depth and reprojection.** [11 points]

Let I_t be a target image and I_s a source image of the same static scene. The camera intrinsic matrix is K. A network predicts target-view depth $D_t(p) > 0$. The rigid transform from the target camera to the source camera is (R, t). For a homogeneous target pixel $\widetilde{p} = (u, v, 1)^\top$, answer the following.

(a) [4 points] Write the back-projection to 3D, the transform to the source camera, and the projection to a source pixel p'. Define a photometric reprojection loss using a valid-pixel mask.

(b) [3 points] Explain how a differentiable bilinear sampler evaluates $I_s(p')$ and why gradients can reach the depth network. Identify the points at which the sampler is not classically differentiable.

(c) [2 points] Give two conditions under which brightness-based reprojection is invalid or ambiguous.

(d) [2 points] State the network input/output and one depth-evaluation metric. Mention one scale ambiguity that must be fixed by data or supervision.

(b) **A conditional VAE for depth uncertainty.** [13 points]

Let I denote the available images and calibration, and D the ground-truth depth during training. Consider

$$p_\psi(z \mid I)p_\theta(D \mid z, I), \qquad q_\phi(z \mid D, I).$$

(a) [5 points] Derive a conditional ELBO for $\log p_{\theta,\psi}(D \mid I)$ and state the direction of the KL divergence and of the ELBO gap.

(b) [3 points] For diagonal-Gaussian prior and posterior, give a valid reparameterization and the closed-form KL formula.

(c) [3 points] Describe training and test-time sampling. Explain how multiple samples may represent ambiguity that a squared-error regressor would average.

(d) [2 points] Give one calibration test and one failure mode of the learned uncertainty.

(c) **Active next-view selection.** [9 points]

A robot has a finite set of candidate camera poses and a budget of B additional views. It should select views that improve a reconstruction while paying acquisition cost.

(a) [4 points] Formulate this problem as a finite-horizon MDP. Specify a sufficient state, the action set, transition, and a reward that uses reconstruction gain and cost.

(b) [3 points] Write a Bellman optimality equation and a model-free Q-learning update. Handle the terminal budget correctly.

(c) [2 points] Give an evaluation protocol that prevents test-scene leakage and compares policies under the same view budget.

C. Optimization Methods in Artificial Intelligence [33 points]

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and L-smooth, and assume $x^\star \in \arg\min f$ exists. Initialize $x^0 = v^0$ and define

$$\theta_k = \frac{2}{k+2}, \qquad y^k = (1-\theta_k)x^k + \theta_k v^k, \qquad g^k = \nabla f(y^k),$$

$$x^{k+1} = y^k - \frac{1}{L} g^k, \qquad v^{k+1} = v^k - \frac{1}{L\theta_k} g^k.$$

Write $\Delta_k = f(x^k) - f(x^\star)$.

(1) [4 points] State the first-order convexity inequality and the descent lemma for an L-smooth function. Verify $\theta_0 = 1$ and

$$\frac{1-\theta_k}{\theta_k^2} \le \frac{1}{\theta_{k-1}^2} \qquad (k \ge 1).$$

(2) [5 points] Apply the descent lemma to the x-update and prove

$$f(x^{k+1}) \le f(y^k) - \frac{1}{2L}\|g^k\|^2.$$

(3) [6 points] Use convexity at y^k to prove

$$f(y^k) - f(x^\star) \le (1-\theta_k)\Delta_k + \theta_k \langle g^k, v^k - x^\star \rangle.$$

Hint: Split the gap with weights $1-\theta_k, \theta_k$ and use $y^k = (1-\theta_k)x^k + \theta_k v^k$.

(4) [7 points] You may use the following identity, obtained by expanding the v-update:

$$\frac{L\theta_k^2}{2}\big(\|v^{k+1} - x^\star\|^2 - \|v^k - x^\star\|^2\big) = -\theta_k \langle g^k, v^k - x^\star \rangle + \frac{1}{2L}\|g^k\|^2.$$

Combine parts (2)–(3) with this identity, identify both cancellations, and show

$$\frac{\Delta_{k+1}}{\theta_k^2} + \frac{L}{2}\|v^{k+1} - x^\star\|^2 \le \frac{1-\theta_k}{\theta_k^2}\Delta_k + \frac{L}{2}\|v^k - x^\star\|^2.$$

(5) [7 points] Telescope the potential, treating the base case $k = 0$ explicitly, and prove that for every $K \ge 1$,

$$f(x^K) - f(x^\star) \le \frac{2L\|x^0 - x^\star\|^2}{(K+1)^2}.$$

(6) [4 points] Give a sufficient K for error at most ε, compare the dependence on ε with ordinary gradient descent, and explain why replacing g^k by an unbiased gradient with persistent variance does not preserve this proof unchanged.

D. Natural Language Processing [33 points]

1. Latent Dirichlet allocation. [7 points] For documents $d = 1, \dots, D$ and topics $k = 1, \dots, K$, let $\theta_d \sim \mathrm{Dir}(\alpha)$ and $\phi_k \sim \mathrm{Dir}(\eta)$. Each token first draws $z_{dn} \sim \mathrm{Cat}(\theta_d)$ and then $w_{dn} \sim \mathrm{Cat}(\phi_{z_{dn}})$.

4 points Write the full joint probability $p(W, Z, \Theta, \Phi \mid \alpha, \eta)$ with all products explicit.

3 points Explain the exchangeability assumptions within a document and name one linguistic phenomenon that this bag-of-words model cannot represent.

2. Collapsed topic update. [8 points] Assume symmetric priors $\alpha > 0$ and $\eta > 0$. After integrating out Θ and Φ, let n_{dk}^{-i} be the number of tokens in document d assigned to topic k, excluding token i; let n_{kv}^{-i} be the corpus count of vocabulary item v in topic k, and $n_{k\cdot}^{-i} = \sum_v n_{kv}^{-i}$. Derive, up to a normalizing constant over k,

$$p(z_i = k \mid Z_{-i}, W, \alpha, \eta) \propto (n_{dk}^{-i} + \alpha) \frac{n_{k,w_i}^{-i} + \eta}{n_{k\cdot}^{-i} + V\eta}.$$

Interpret the two factors and explain why the $-i$ convention is necessary.

3. Adding document labels. [7 points] Let $\bar{z}_d \in \mathbb{R}^K$ be the empirical topic-frequency vector of document d and let

$$p(y_d = 1 \mid \bar{z}_d, \beta) = \sigma(\beta^\top \bar{z}_d), \qquad y_d \in \{0, 1\}.$$

3 points Write the label log-likelihood and its gradient with respect to β.

2 points Explain how this label term changes posterior topic assignments.

2 points Give one failure mode when labels are highly imbalanced and one corresponding remedy.

4. Topic discovery under temporal drift. [11 points] A platform uses topics to route support tickets. Vocabulary and issue prevalence change monthly, but only a small subset of tickets receives a routing label.

4 points Design a probabilistic or optimization-based topic system that uses both unlabeled text and the available labels while allowing topics to evolve over time. Define its main variables and objective or factorization.

3 points Describe an inference/training procedure and how new vocabulary is handled without rebuilding everything from scratch.

2 points Give quantitative diagnostics that distinguish topic drift, label shift, and a general degradation of the text encoder.

2 points Explain how human-readable topic names are produced and audited.