# Artificial Intelligence PhD Qualifying Examination

High-Fidelity Mock Examination 06

**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 threshold stump changes its sample labeling when the threshold

(a) crosses a sample value.

(b) changes a kernel.

(c) removes labels.

(d) changes VC definition.

(ii) [0.5 points] Averaging highly correlated predictors gives

(a) less variance reduction than averaging independent ones.

(b) zero bias always.

(c) infinite margin.

(d) no prediction.

(iii) [0.5 points] A one-sided threshold class on $\mathbb{R}$ has VC dimension

(a) 1.

(b) 2.

(c) $d + 1$.

(d) infinite.

(iv) [0.5 points] Sauer's lemma turns finite VC dimension into

(a) polynomial growth in sample size.

(b) zero risk.

(c) a unique kernel.

(d) exponential-time necessity.

(v) [0.5 points] No Free Lunch motivates

(a) inductive bias.

(b) a universally best learner.

(c) ignoring the data distribution.

(d) infinite validation reuse.

(vi) [0.5 points] Exact ERM training error over nested tree classes is

(a) nonincreasing with allowed depth.

(b) always increasing.

(c) equal to test error.

(d) independent of labels.

**Part II: Theoretical Exercises [30 points].**

8 points Let $\mathcal{H}_{\text{stump}}$ choose a coordinate $j \in [d]$, threshold $t$, and either polarity:

$$h_{j,t,+}(x) = \mathbf{1}\{x_j \le t\}, \qquad h_{j,t,-} = 1 - h_{j,t,+}.$$

Prove $\tau_{\mathcal{H}_{\text{stump}}}(m) \le 2d(m+1)$ and explain why equality need not hold on a given sample.

6 points For $M \ge 2$, random estimators $T_1, \dots, T_M$ have common variance $\sigma^2$ and common pairwise correlation $\rho \in [-1/(M-1), 1]$. Prove

$$\operatorname{Var}\left(\frac{1}{M}\sum_{j=1}^{M} T_j\right) = \sigma^2\left(\rho + \frac{1-\rho}{M}\right).$$

Interpret the independent and perfectly correlated cases for generic averaging.

7 points Let $\mathcal{H} = \{h_1, \dots, h_N\}$, where $N \ge 2$, and let every loss lie in $[0,1]$. Write $\widehat{L}(h)$ and $L(h)$ for empirical and population loss. For $\beta > 0$, define the Gibbs distribution

$$q_\beta(h) = \frac{\exp(-\beta \widehat{L}(h))}{\sum_{g \in \mathcal{H}} \exp(-\beta \widehat{L}(g))}.$$

For a distribution $q$ on $\mathcal{H}$, set

$$\widehat{L}(q) = \sum_h q(h)\widehat{L}(h), \qquad L(q) = \sum_h q(h) L(h).$$

For entropy $H(q) = -\sum_h q(h) \log q(h)$, you may use the variational identity

$$-\frac{1}{\beta}\log \sum_h e^{-\beta a_h} = \min_{q \in \Delta_N} \left\{ \sum_h q(h) a_h - \frac{H(q)}{\beta} \right\},$$

whose minimizer is proportional to $e^{-\beta a_h}$.

(a) Derive $\widehat{L}(q_\beta) \le \min_h \widehat{L}(h) + \log N/\beta$.

(b) On the event $\sup_h |L(h) - \widehat{L}(h)| \le \alpha$, prove

$$L(q_\beta) - \min_h L(h) \le 2\alpha + \frac{\log N}{\beta}.$$

(c) Using $\Pr\{\sup_h |L(h) - \widehat{L}(h)| > \alpha\} \le 2Ne^{-2m\alpha^2}$, give sufficient choices of $m$ and $\beta$ for the excess in (b) to be at most $\varepsilon$ with probability at least $1 - \delta$.

5 points You may use

$$\Pr\{\sup_{h \in \mathcal{H}} |L(h) - \widehat{L}(h)| > \alpha\} \le 4\tau_{\mathcal{H}}(2m) e^{-m\alpha^2/8}.$$

If $1 \le \operatorname{VCdim}(\mathcal{H}) = v \le 2m$, combine Sauer's lemma with ERM comparison to give an explicit implicit condition for agnostic excess risk at most $\varepsilon$ with confidence $1 - \delta$.

4 points Two hypotheses disagree on a set of probability mass $q$. The target is chosen uniformly from the two hypotheses. Show that any learner which sees no sample point in the disagreement set has conditional expected error at least $q/2$, and derive a lower bound on its unconditional expected error after $m$ examples.

# B. Deep Learning and Reinforcement Learning [33 points]

(a) **Backpropagation through a branched residual block.** [9 points]

Let $x \in \mathbb{R}^d$ and

$$y = x + F_\theta(x) + G_\phi(x), \qquad J = \mathcal{L}(y).$$

Write column gradients and let $J_F(x) = \partial F_\theta(x)/\partial x$ and $J_G(x) = \partial G_\phi(x)/\partial x$.

(a) [4 points] Derive $\nabla_x J$, $\nabla_\theta J$, and $\nabla_\phi J$ in terms of local Jacobians. Explain why gradients must be added at the branch junction.

(b) [3 points] Explain the role of the identity term in gradient propagation through a stack of such blocks. State one reason it does not guarantee perfectly stable gradients.

(c) [2 points] Compare the work of reverse-mode backpropagation for all parameters with separately forming a full input–output Jacobian for every parameter.

(b) $\beta$**-VAE and posterior collapse.** [15 points]

Let $p(z) = \mathcal{N}(0, I)$, $q_\phi(z \mid x) = \mathcal{N}(\mu_\phi(x), \mathrm{diag}(\sigma^2_\phi(x)))$, and $p_\theta(x \mid z)$ be a decoder.

(a) [4 points] Derive the standard VAE ELBO and its gap to $\log p_\theta(x)$.

(b) [4 points] Give the reparameterization and the exact diagonal-Gaussian KL to $\mathcal{N}(0, I)$. Be explicit about whether the network outputs variance, standard deviation, or log-variance.

(c) [3 points] Consider $\mathcal{L}_\beta = \mathbb{E}_q \log p_\theta(x \mid z) - \beta\,\mathrm{KL}(q\|p)$. For which values of $\beta$ is this expression automatically a lower bound by comparison with the standard ELBO? Explain the reconstruction–regularization tradeoff.

(d) [2 points] Define posterior collapse and give a quantitative diagnostic.

(e) [2 points] State the unconditional generation procedure and one incorrect procedure that must be avoided.

(c) $n$**-step actor–critic.** [9 points]

For a discounted episodic MDP, an actor is $\pi_\eta(a \mid s)$ and a critic is $V_w(s)$. Define

$$G_t^{(n)} = \sum_{\ell=0}^{n-1} \gamma^\ell R_{t+\ell} + \gamma^n V_w(S_{t+n}), \qquad \widehat{A}_t^{(n)} = G_t^{(n)} - V_w(S_t),$$

with bootstrapping omitted after termination.

(a) [4 points] Give critic and actor updates, including the required stop-gradient convention and the discount convention for the actor objective.

(b) [3 points] If $V_w = V^\pi$, prove that the conditional expectation of the one-step TD error given $(S_t = s, A_t = a)$ is $A^\pi(s, a)$.

(c) [2 points] Explain the bias–variance effect of changing $n$.

## C. Optimization Methods in Artificial Intelligence [33 points]

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. Assume coordinate smoothness: for $L_i > 0$,

$$f(x + ue_i) \le f(x) + u\nabla_i f(x) + \frac{L_i}{2}u^2.$$

Define $S = \sum_{i=1}^d L_i$ and $p_i = L_i/S$. Independently sample $i_k$ with probability $p_i$ and update

$$x^{k+1} = x^k - \frac{\nabla_{i_k} f(x^k)}{L_{i_k}} e_{i_k}.$$

Assume $x^\star \in \arg\min f$ exists and

$$R = \sup_{x: f(x) \le f(x^0)} \|x - x^\star\| < \infty.$$

(1) [3 points] Minimize the coordinate upper model and prove that updating coordinate $i$ decreases $f$ by at least $[\nabla_i f(x)]^2/(2L_i)$.

(2) [4 points] Take conditional expectation over $i_k$ and show

$$\mathbb{E}_k[f(x^{k+1})] \le f(x^k) - \frac{1}{2S}\|\nabla f(x^k)\|^2.$$

(3) [4 points] Is $\nabla_I f(x) e_I$ an unbiased estimator of $\nabla f(x)$? Give the correctly normalized unbiased coordinate estimator under a general distribution $p_i > 0$. Determine whether the present $p_i = L_i/S$ update can then be written as SGD with one common stepsize, and state that stepsize.

(4) [5 points] Use convexity, Cauchy–Schwarz, and monotonicity of the iterates' function values to prove

$$f(x^k) - f(x^\star) \le R\|\nabla f(x^k)\|.$$

Then derive a conditional decrease involving the square of the function gap.

(5) [6 points] Let $\delta_k = \mathbb{E}[f(x^k) - f(x^\star)]$. Prove

$$\delta_{k+1} \le \delta_k - \frac{\delta_k^2}{2SR^2},$$

and deduce, for $K \ge 2$,

$$\delta_K \le \frac{2SR^2}{K-1}.$$

(6) [6 points] Now assume additionally that $f$ is $\mu$-strongly convex in the Euclidean norm. Prove

$$\|\nabla f(x)\|^2 \ge 2\mu[f(x) - f(x^\star)]$$

and deduce the linear expected bound

$$\mathbb{E}[f(x^K) - f(x^\star)] \le \left(1 - \frac{\mu}{S}\right)^K [f(x^0) - f(x^\star)].$$

(7) [5 points] Under uniform sampling, derive a valid decrease bound using $L_{\max} = \max_i L_i$ and compare its constant with $S$. Then discuss coordinate-update cost, full-gradient cost, and the cost of sampling from $p$.

## D. Natural Language Processing [33 points]

**1. Weighted co-occurrence embeddings.** [**9 points**] For positive word–context counts $X_{ij}$, a GloVe-style model minimizes

$$J = \sum_{i,j} f(X_{ij})\big(\mathbf{u}_i^\top \mathbf{v}_j + b_i + c_j - \log X_{ij}\big)^2,$$

where $f(x) \geq 0$ is a prescribed weighting function.

3 points Explain the statistical signal represented by $X_{ij}$ and the roles of the log target, the two biases, and $f$.

3 points Derive $\nabla_{\mathbf{u}_i} J$ and $\partial J / \partial b_i$.

3 points Show that the fitted inner products are not identifiable: for any invertible matrix $A$, give a simultaneous transformation of all word and context vectors that leaves $J$ unchanged. Explain one consequence for interpreting individual embedding coordinates.

**2. Nucleus sampling.** [**10 points**] At one decoding step a language model produces a probability vector $p = (p_1, \ldots, p_V)$. Let the nucleus threshold be $\rho \in (0, 1]$.

4 points Define nucleus (top-$p$) sampling precisely, including sorting, the smallest retained prefix, and the normalized sampling distribution. State how ties are handled.

3 points For sorted probabilities

$$(0.40, 0.25, 0.15, 0.10, 0.10)$$

and $\rho = 0.75$, find the retained set and every nonzero probability after renormalization.

3 points Compare nucleus sampling with greedy decoding and fixed top-$k$ sampling. Give one failure mode of nucleus sampling and one targeted safeguard.

**3. Temporal reasoning with a Markov logic network.** [**14 points**] An event extractor proposes events $e$, candidate time bins $t$, and pairwise ordering signals. Evidence predicates are CandTime$(e, t)$, HighTime$(e, t)$, HighBefore$(e_1, e_2)$, and Less$(t_1, t_2)$. The unknown predicate is Time$(e, t)$. An MLN assigns

$$P(X \mid E) = \frac{1}{Z(E)} \exp\left(\sum_j w_j n_j(X, E)\right).$$

5 points Write weighted formulas encoding local time evidence, exactly one candidate time per event, and compatibility between a high-confidence before signal and the chosen times. State which constraints should normally be hard and which soft.

5 points There are two events $A, B$ and bins $1 < 2$. Exactly one bin is chosen for each event. Define

$$n_o = \mathbf{1}\{T_A < T_B\}, \quad n_A = \mathbf{1}\{T_A = 1\}, \quad n_B = \mathbf{1}\{T_B = 2\},$$

with weights $(w_o, w_A, w_B) = (1.2, 0.8, 0.4)$. Compute the unnormalized mass of all four assignments, the partition function, the MAP assignment, and $P(T_A = 1, T_B = 2)$.

4 points Explain how classifier scores should be calibrated before becoming soft evidence, how a later correction is incorporated without destroying provenance, and the complexity of exact inference for $E$ events with $K$ candidate bins. Give one tractable special case or approximate alternative.