

**QIUZHEN QUALIFY EXAM FOR ARTIFICIAL INTELLIGENCE
FALL 2026**

INSTRUCTIONS

- *This exam consists of four sections, each with a total of 33 points. You are required to choose three out of the four sections and answer the questions in the selected sections. You will earn an additional point for writing your name and student ID number on your answer sheet. The maximum possible score you can achieve is 100.*
- *Please clearly indicate your section choices at the very beginning of your answer sheet. If you answer questions in more than three sections, only the first three sections will be graded.*

A. MACHINE LEARNING THEORY

Part I: Warm-Up Questions 3 pts. — This section contains multiple-choice questions. Please select **only one** answer for each question. No justification is required.

- MQ1.** [0.5 pts.] Suppose a hypothesis class $\mathcal{H}$ satisfies uniform convergence. Which statement necessarily follows?
- (a) Every hypothesis in $\mathcal{H}$ has zero true risk.
 - (b) The empirical risk minimiser always achieves the Bayes risk.
 - (c) The difference between empirical and true risk is small simultaneously for all hypotheses in $\mathcal{H}$ with high probability.
 - (d) The hypothesis class contains the optimal classifier.
- MQ2.** [0.5 pts.] Suppose a hypothesis class $\mathcal{H}_1$ is enlarged to $\mathcal{H}_2$, where $\mathcal{H}_1 \subseteq \mathcal{H}_2$. Which statement correctly describes the effect of this enlargement?
- (a) The minimum achievable empirical risk cannot increase, but the complexity of the class may increase.
 - (b) The approximation error always increases.
 - (c) The generalisation error always decreases.
 - (d) The VC dimension must remain unchanged.
- MQ3.** [0.5 pts.] Let $\mathcal{H}_1 \subseteq \mathcal{H}_2$. Suppose an algorithm performs ERM over both classes using the same training sample. Which statement is always true?
- (a) The ERM solution over $\mathcal{H}_2$ has smaller true risk.

- (b) The empirical risk achieved over $\mathcal{H}_2$ is no larger than the empirical risk achieved over $\mathcal{H}_1$.
 - (c) The ERM solution over $\mathcal{H}_1$ has smaller empirical risk.
 - (d) The two ERM solutions must be identical.
- MQ4.** [0.5 pts.] Consider two hypothesis classes $\mathcal{H}_1$ and $\mathcal{H}_2$ with the same VC dimension. Which conclusion is valid?
- (a) They have the same number of hypotheses.
 - (b) They have identical generalisation behaviour for every distribution.
 - (c) ERM produces the same classifier for both classes.
 - (d) They have the same worst-case sample complexity order, but their performance on a particular distribution may differ.
- MQ5.** [0.5 pts.] Let K_1 and K_2 be valid kernels. Which of the following is necessarily a valid kernel?
- (a) $K(x, z) = K_1(x, z) - K_2(x, z)$.
 - (b) $K(x, z) = aK_1(x, z) + bK_2(x, z)$ for $a, b \geq 0$.
 - (c) $K(x, z) = \sqrt{K_1(x, z)}$.
 - (d) $K(x, z) = \frac{K_1(x, z)}{K_2(x, z)}$.
- MQ6.** [0.5 pts.] Consider a neural network using ReLU activations. A neuron has pre-activation $z(x) = w^\top x + b$ and satisfies $z(x) < 0$ for every input in the training set. Which statement is correct?
- (a) The neuron will necessarily become active after one gradient step.
 - (b) The neuron contributes a constant non-zero feature to the network.
 - (c) The gradient through this neuron can vanish, preventing its parameters from being updated.
 - (d) The neuron increases the effective VC dimension of the network.

Part II: Theoretical Exercises 30 pts. — This is the main part of the exam. Provide **detailed solutions and reasoning** for each question. Full marks are awarded only for complete and well-explained answers.

- (1) [7 pts.] Let $\mathcal{X}$ be a discrete domain. For a fixed integer $k \geq 1$, define the hypothesis class $\mathcal{H}_k = \{h_A : A \subseteq \mathcal{X}, |A| \leq k\}$, where

$$h_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$$

Assume that the realisability assumption holds. That is, there exists some unknown set $A^* \subseteq \mathcal{X}$ with $|A^*| \leq k$ such that the target hypothesis is h_{A^*} .

- (a) Describe a learning algorithm that implements the ERM rule for $\mathcal{H}_k$.
- (b) Prove that $\mathcal{H}_k$ is PAC learnable.
- (c) Derive an upper bound on the sample complexity $m_{\mathcal{H}_k}(\epsilon, \delta)$.

- (2) [9 pts.] Let $G = (V, E)$ be a fixed feedforward directed acyclic graph representing a neural network architecture. Assume that every non-input neuron uses the sign activation function

$$\text{sign}(z) = \begin{cases} +1, & z > 0, \\ -1, & z \leq 0. \end{cases}$$

Let $\mathcal{H}_G$ denote the hypothesis class realised by all assignments of real-valued weights to the edges of G . Assume that the network has a single output neuron, and let $W = |E|$ be the total number of edges in the network.

For each neuron $v_{t,i}$, let $d_{t,i}$ denote its number of incoming edges. A bias is represented by an additional edge from a constant input whose value is 1.

- (a) Prove that the class of functions implementable by the neuron $v_{t,i}$ has VC dimension at most $d_{t,i}$.
 (b) Using Sauer's lemma, prove that the growth function of this neuron satisfies

$$\tau_{t,i}(m) \leq (em)^{d_{t,i}}.$$

- (c) Using product and composition arguments for layered networks, prove that the growth function of the entire network satisfies

$$\tau_{\mathcal{H}_G}(m) \leq (em)^W.$$

- (3) [8 pts.] Let $F(w) = \frac{1}{n} \sum_{i=1}^n f_i(w)$, where each $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex and differentiable. Assume that $\|\nabla f_i(w)\| \leq \rho$ for every $i \in \{1, \dots, n\}$ and every $w \in \mathbb{R}^d$. Consider stochastic gradient descent initialised at $w^{(1)} = 0$. At iteration t , choose i_t uniformly at random from $\{1, \dots, n\}$ and update

$$w^{(t+1)} = w^{(t)} - \eta \nabla f_{i_t}(w^{(t)}), \quad t = 1, \dots, T.$$

The algorithm outputs the averaged iterate $\bar{w} = \frac{1}{T} \sum_{t=1}^T w^{(t)}$. Let $u \in \mathbb{R}^d$ satisfy $\|u\| \leq B$.

- (a) Show that the stochastic gradient is unbiased:

$$\mathbb{E}_{i_t} [\nabla f_{i_t}(w^{(t)}) \mid w^{(t)}] = \nabla F(w^{(t)}).$$

- (b) You may use the following one-step inequality:

$$2\eta \mathbb{E} [F(w^{(t)}) - F(u)] \leq \mathbb{E} [\|w^{(t)} - u\|^2 - \|w^{(t+1)} - u\|^2] + \eta^2 \rho^2.$$

Prove that

$$\mathbb{E} [F(\bar{w})] - F(u) \leq \frac{B^2}{2\eta T} + \frac{\eta \rho^2}{2}.$$

- (c) Choose η to minimise the upper bound in part (b). Determine a sufficient number of iterations T such that

$$\mathbb{E} [F(\bar{w})] - F(u) \leq \epsilon.$$

- (4) [6 pts.] Let $S = \{(x_i, y_i)\}_{i=1}^m$, $y_i \in \mathbb{R}$. Let $\psi : \mathcal{X} \rightarrow \mathcal{H}$ be a feature mapping into a Hilbert space $\mathcal{H}$, with associated kernel $K(x, z) = \langle \psi(x), \psi(z) \rangle_{\mathcal{H}}$. For $\lambda > 0$, consider the regularised objective

$$J(w) = \lambda \|w\|_{\mathcal{H}}^2 + \frac{1}{2m} \sum_{i=1}^m (\langle w, \psi(x_i) \rangle_{\mathcal{H}} - y_i)^2.$$

- (a) Let $V = \text{span}\{\psi(x_1), \dots, \psi(x_m)\}$. Prove that every minimiser w^* of J belongs to V .
- (b) Let $G \in \mathbb{R}^{m \times m}$ be the Gram matrix defined by $G_{ij} = K(x_i, x_j)$, and let $\mathbf{y} = (y_1, \dots, y_m)^\top$. (b.1) Show that one minimising representation is $w^* = \sum_{j=1}^m \alpha_j^* \psi(x_j)$, where $\boldsymbol{\alpha}^* = (G + 2m\lambda I)^{-1} \mathbf{y}$. (b.2) Justify why $G + 2m\lambda I$ is invertible, and show that the prediction at a new point x can be computed without explicitly evaluating $\psi(x)$.

B. DEEP LEARNING AND REINFORCEMENT LEARNING

1. Basics of Deep Learning [14 pts.]

- (a) [3 pts.] Consider three observations from a supervised-learning project. (i) A large model reaches almost zero training error but has high validation error; adding representative data and regularization reduces this gap. (ii) A linear hypothesis class cannot represent the nonlinear decision boundary, even when its empirical loss is minimized. (iii) An expressive neural network initially has high training error, but changing only the optimizer and learning-rate schedule substantially lowers that error. For each observation, identify the dominant issue as *approximation*, *optimization*, or *generalization*, and briefly justify your choice.
- (b) [3 pts.] The same neural network and dataset can be trained using full-batch gradient descent or mini-batch stochastic gradient descent. Explain why mini-batches are commonly used for large datasets. State one potentially helpful and one potentially harmful effect of gradient noise. Finally, explain how momentum can reduce oscillations and make training more stable. No update equations are required.
- (c) [3 pts.] Consider a deep network

$$h_\ell = \sigma(W_\ell h_{\ell-1}), \quad \sigma(z) = \frac{1}{1 + e^{-z}}.$$

Use the derivative of the sigmoid and the chain rule to explain why gradients can vanish as the depth increases, even when the operator norms of the weight matrices are close to one. Give one modification that can reduce this problem and explain why it helps.

- (d) [3 pts.] Why is the softmax activation commonly used in multi-class classification? Why is the cross-entropy loss generally preferred over the mean squared error (MSE) loss for classification? Explain why minimizing the cross-entropy loss is equivalent to maximizing the likelihood under a categorical distribution.
- (e) [2 pts.] A residual block has the form $h_{\ell+1} = h_{\ell} + F_{\ell}(h_{\ell})$. Explain the main idea of the identity skip connection and why it can make optimization and backpropagation easier in very deep networks than in corresponding plain networks.

2. Generative Models [12 pts.]

Let $\beta > 0$ be a constant and consider the variance-preserving forward SDE

$$dX_t = -\frac{1}{2}\beta X_t dt + \sqrt{\beta} dW_t, \quad X_0 \sim p_{\text{data}}.$$

Its transition can be written as

$$X_t = a_t X_0 + \sigma_t Z, \quad a_t = e^{-\beta t/2}, \quad \sigma_t^2 = 1 - e^{-\beta t}, \quad Z \sim \mathcal{N}(0, I).$$

Let p_t be the density of X_t , and define

$$s_t(x) = \nabla_x \log p_t(x).$$

For discrete sampling, let $T = K\Delta$, use a grid $t_k = k\Delta$, and define

$$\alpha = e^{-\beta\Delta}, \quad \bar{\alpha}_k = \alpha^k = e^{-\beta t_k}.$$

Thus

$$X_k = \sqrt{\bar{\alpha}_k} X_0 + \sqrt{1 - \bar{\alpha}_k} Z.$$

(a) Reverse SDE and Score Learning [4 pts.]

- (i) [1 pt.] Starting from

$$X_T \sim p_T \approx \mathcal{N}(0, I),$$

write the reverse-time SDE, using the original time variable t and integrating from T down to 0. Use negative time increments $dt < 0$, and denote reverse-time Brownian motion by $\bar{W}_t$.

- (ii) [3 pts.] In the reverse-time SDE, identify the quantity that is unknown and must be learned from data. Suppose a neural network $\epsilon_{\theta}(x, t)$ is used to predict the Gaussian noise Z in

$$X_t = a_t X_0 + \sigma_t Z.$$

Derive the corresponding approximation of the unknown quantity in the reverse SDE, and write the squared-error training objective for ϵ_{θ} . For the objective, sample $t \sim \rho$, where ρ is a specified distribution supported on $(0, T]$.

Briefly explain why training samples at time t can be generated directly from the transition formula above without simulating the entire forward SDE trajectory from 0 to t .

(b) **Equivalent ODE and DDIM Discretization [4 pts.]**

Consider deterministic sampling starting from

$$X_T \sim p_T \approx \mathcal{N}(0, I).$$

- (i) [2 pts.] Write the deterministic ODE whose marginal distributions are identical to those of the reverse-time SDE. Express this ODE in terms of the score $s_t(X_t)$, and state the direction in which it is solved for data generation.
- (ii) [2 pts.] Substitute the noise-prediction model $\epsilon_\theta(X_t, t)$ into this ODE. Using an integrating-factor discretization on the grid $\{t_k\}$, and holding the predicted noise fixed at $\epsilon_\theta(X_k, t_k)$ over $[t_{k-1}, t_k]$, derive the deterministic DDIM update from X_k to X_{k-1} .

Hint: First introduce the predicted clean sample

$$\hat{X}_0 = \frac{X_k - \sqrt{1 - \bar{\alpha}_k} \epsilon_\theta(X_k, t_k)}{\sqrt{\bar{\alpha}_k}},$$

and then obtain the update for X_{k-1} . State whether any additional Gaussian noise is sampled in this update.

(c) **X_0 -Prediction Parameterization [4 pts.]**

Suppose instead that a neural network $m_\theta(x, t)$ is trained to predict the clean sample X_0 . At the population optimum,

$$m_\theta(x, t) \approx m_t(x) := \mathbb{E}[X_0 \mid X_t = x].$$

- (i) [2 pts.] Derive the relationship between $m_t(x)$ and the score $s_t(x)$. Then, starting from the X_0 -prediction viewpoint, derive a squared-error training objective for m_θ using samples
- $$X_0 \sim p_{\text{data}}, \quad t \sim \rho, \quad Z \sim \mathcal{N}(0, I), \quad X_t = a_t X_0 + \sigma_t Z.$$
- (ii) [1 pt.] Using the relationship above, write the reverse-time SDE directly in terms of $m_\theta(X_t, t)$. Then give one Euler–Maruyama step from t_k to t_{k-1} , identifying the new Gaussian random variable.
 - (iii) [1 pt.] State one potential practical advantage of X_0 -prediction compared with ϵ -prediction, and briefly explain why it can be useful.

3. Reinforcement Learning [7 pts.]

- (a) [3 pts.] Consider the following deterministic discounted MDP with $\gamma = 1/2$:

	a	b
s_1	$(0, s_1)$	$(1, s_2)$
s_2	$(2, s_2)$	$(0, s_1)$

Each entry is (immediate reward, next state). Write the Bellman optimality equations for $V^*(s_1)$ and $V^*(s_2)$. Starting from $V_0(s_1) = V_0(s_2) = 0$, perform two synchronous value-iteration updates to obtain V_1 and V_2 , and give the greedy action in each state after the second update.

- (b) [2 pts.] When the state space is too large to store a table, suppose $Q_\theta(s, a)$ is a neural-network approximation to the optimal action-value function for a discounted MDP with discount factor γ . A transition (s, a, r, s') is sampled from experience, s' is nonterminal, and a second network $Q_{\bar{\theta}}$ is held fixed for several updates. Write a one-step Bellman target and a squared training loss for Q_θ .
- (c) [2 pts.] Consider a differentiable stochastic policy $\pi_\theta(a \mid s)$ in an episodic discounted MDP, with objective

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{H-1} \gamma^t r_t \right].$$

For a sampled trajectory, define the reward-to-go from time t by

$$G_t = \sum_{j=t}^{H-1} \gamma^{j-t} r_j.$$

Write an unbiased Monte Carlo policy-gradient estimator used by REINFORCE.

C. OPTIMIZATION METHODS IN ARTIFICIAL INTELLIGENCE

- [6 pts.] Provide three necessary and sufficient conditions for $f : \mathbb{R}^n \rightarrow \mathbb{R}$ to be μ -strongly convex and L -smooth, respectively.
- [5 pts.] Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be differentiable and strictly convex. For two fixed distinct points $u, v \in \mathbb{R}^n$, define

$$d = u - v, \quad g = \nabla f(u) - \nabla f(v), \quad P = I + \frac{gg^T}{g^T d} - \frac{dd^T}{d^T d}.$$

Show that there exist two positive constants $m \leq M$ such that $F(x) = x^T P x$ is m -strongly convex and M -smooth on $\mathbb{R}^n$.

- [4 pts.] Let $z \in \mathbb{R}^n$ be a random vector, $\mathbb{E}$ and Var denote the expectation and variance, respectively. Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is an L -smooth function. Show that

$$\mathbb{E}[f(z)] - f(\mathbb{E}[z]) \leq \frac{L}{2} \text{Var}[z].$$

- [10 pts.] Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is L -smooth and convex. Let x^* be an optimal solution of

$$\min_{x \in \mathbb{R}^n} f(x).$$

Consider the gradient descent iteration $x^{k+1} = x^k - \alpha \nabla f(x^k)$, and x^0 is an initial iteration point.

- (a) Let $\alpha = \frac{1}{L}$ and show that

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2L} \|\nabla f(x^k)\|^2, \quad f(x^k) - f(x^*) \leq \frac{L \|x^0 - x^*\|^2}{2k}.$$

- (b) Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be defined by $f(x, y) = 10x^2 + y^2$. Show that $f(x, y)$ is L -smooth with $L = 20$ and if $\alpha = \frac{1}{11}$, then the iteration complexity of gradient descent is $\mathcal{O}(10 \log(1/\epsilon))$.
- (c) Provide an adaptive gradient descent method such that it can find the optimal solution $(0, 0)$ with $f(x, y)$ given in (b) after one iteration from any starting point.
5. [8 pts.] Consider the composite optimization problem

$$(\text{CO}) : \min_{x \in \mathbb{R}^n} \phi(x) := f(x) + h(x),$$

where $f, h : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ is proper, closed and convex. For $z \in \mathbb{R}^n$, define

$$G(z) = \text{prox}_{th}(z) - \text{prox}_{tf}(2\text{prox}_{th}(z) - z),$$

where $\text{prox}_{th}(x)$ is the proximal operator of h defined by

$$\text{prox}_{th}(x) = \arg \min_{u \in \mathbb{R}^n} \left\{ h(u) + \frac{1}{2t} \|u - x\|_2^2 \right\} \quad (t > 0).$$

- (a) Show that $\text{prox}_{th}(z^*)$ is an optimal solution to (CO) if $G(z^*) = 0$.
- (b) Given the constant $0 < \rho < 2$ and consider the iteration: $z^k = z^{k-1} - \rho G(z^{k-1})$. Show that there exists $C > 0$ such that

$$\|z^k - z^*\|_2^2 - \|z^{k-1} - z^*\|_2^2 \leq -C \|G(z^{k-1})\|_2^2, \quad \lim_{k \rightarrow \infty} \|G(z^k)\|_2 = 0,$$

where z^* satisfies $G(z^*) = 0$.

(Hint: first show $\langle G(z_1) - G(z_2), z_1 - z_2 \rangle \geq \|G(z_1) - G(z_2)\|_2^2, \forall z_1, z_2 \in \mathbb{R}^n$)

D. NATURAL LANGUAGE PROCESSING

D.1. Symmetric Dirichlet Priors in LDA [4 pts.] In LDA, θ_k is the proportion of topic k in a document. A symmetric Dirichlet prior uses a to favor sparse or diffuse topic mixtures; its normalizing constant is irrelevant here.

Let $\theta = (\theta_1, \dots, \theta_K)$ denote the topic proportions. Under the symmetric prior $\theta \sim \text{Dir}(a, \dots, a)$,

$$p(\theta \mid a) \propto \prod_{k=1}^K \theta_k^{a-1}, \quad \theta_k \geq 0, \quad \sum_{k=1}^K \theta_k = 1.$$

Using the density, compare qualitatively how the prior mass is distributed when $a < 1$, $a = 1$, and $a > 1$. Identify the regime appropriate when a typical document is expected to use only a few of the K topics, and justify your choice. [4 pts.]

D.2. Ordered Relation Composition in Knowledge Graph Embeddings [7 pts.] For a triple (h, r, t) (head, relation, tail), TransE models r as a translation in $\mathbb{R}^d$, so exact fit means $v_t = v_h + v_r$. RotatE models r as a coordinate-wise complex rotation in $\mathbb{C}^d$, so exact fit means $v_t = v_h \circ v_r$, where $\circ$ is elementwise multiplication and $|(v_r)_j| = 1$.

Let a, b, c, d, e be entities and require $v_c \neq v_e$. Assume the following triples are fitted exactly:

$$(a, r_1, b), \quad (b, r_2, c), \quad (a, r_2, d), \quad (d, r_1, e).$$

The path to c applies r_1 then r_2 ; the path to e reverses the order. Such compositions may have different meanings.

- (a) For both TransE and RotatE, prove that the four exact-fit equations force $v_c = v_e$, and identify the common algebraic obstruction. [5 pts.]
- (b) For the matrix model $v_t = M_r v_h$, construct invertible 2×2 real matrices M_{r_1}, M_{r_2} and a nonzero v_a such that

$$M_{r_2} M_{r_1} v_a \neq M_{r_1} M_{r_2} v_a,$$

and explain why this permits the paths to have distinct endpoints. [2 pts.]

D.3. Relative Position and RoPE [7 pts.] In self-attention, let q, k be content vectors and $R_i q, R_j k$ their position-encoded forms. Joint-shift invariance means that positional interaction depends only on relative displacement.

Let d be even. Let $R_t \in \mathbb{R}^{d \times d}$ be orthogonal for every $t \in \mathbb{Z}$, with $R_0 = I$. Suppose joint shifts leave every encoded inner product invariant:

$$\langle R_{i+s} q, R_{j+s} k \rangle = \langle R_i q, R_j k \rangle \quad \text{for all } q, k \in \mathbb{R}^d \text{ and } i, j, s \in \mathbb{Z}.$$

Prove that $R_i^\top R_j = R_{j-i}$ and deduce that $R_{t+s} = R_t R_s$, so the maps form an orthogonal representation of $(\mathbb{Z}, +)$. Now let

$$B(\alpha) = \begin{pmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{pmatrix}, \quad R_t = \text{diag}(B(t\omega_1), \dots, B(t\omega_{d/2})).$$

Verify that these RoPE matrices realize such a representation, write the encoded attention inner product explicitly as a function of $j - i$, and explain why the score depends on relative displacement although the query and key are encoded at absolute positions. [7 pts.]

D.4. Domain-Size Effects in a Markov Logic Network [7 pts.] In a Markov logic network, each satisfied ground formula contributes its weight to the unnormalized log-probability. Enlarging the domain therefore replicates features even when no new evidence is observed.

Let the domain be $\mathcal{D}_n = \{1, \dots, n\}$. Consider an MLN with Boolean predicate $A(x)$, Boolean atom T , and the formula schema

$$A(x) \Rightarrow T$$

with weight w , grounded once for each $x \in \mathcal{D}_n$. No atom is observed. A world is a complete truth assignment to $T, A(1), \dots, A(n)$; hence there are 2^{n+1} worlds.

If $T = 1$, all n groundings are satisfied. If $T = 0$, grounding i is satisfied exactly when $A(i) = 0$. Let $N(X)$ be the number of satisfied groundings in world X , so that $P(X) = Z_n(w)^{-1} \exp\{wN(X)\}$.

- (a) Without enumerating all worlds, derive $Z_n(w)$ and a closed form for $P(T = 1)$. [2 pts.]
- (b) Analyze the domain-size behavior. Prove that $P(T = 1) \rightarrow 1$ as $n \rightarrow \infty$ for every fixed $w > 0$. Then set $w_n = c/n$ for fixed $c \in \mathbb{R}$ and prove

$$\lim_{n \rightarrow \infty} P_{w_n}(T = 1) = \frac{1}{1 + e^{-c/2}}.$$

Explain why the contrast between these limits is a domain-size effect rather than accumulated observed evidence. [5 pts.]

D.5. KL-Regularized RLHF and Reward Hacking [8 pts.] For a finite response set $\mathcal{Y}$, let $\pi_0(y) > 0$ be the SFT reference policy. Given a learned reward $\hat{r}$, RLHF maximizes

$$J(\pi) = \sum_{y \in \mathcal{Y}} \pi(y) \hat{r}(y) - \beta D_{\text{KL}}(\pi \| \pi_0)$$

over the probability simplex; $\beta > 0$ controls adherence to π_0 .

The learned reward is only a proxy for the unknown true utility u . Assume

$$\hat{r}(y) = u(y) + \varepsilon(y), \quad \|\varepsilon\|_\infty \leq \delta.$$

Let s be a safe response and h a reward-hacking response. Suppose

$$g = u(s) - u(h) > 0,$$

but reward-model error may favor h . Define the reference-policy log-odds

$$a = \log \frac{\pi_0(s)}{\pi_0(h)}.$$

Thus $a > 0$ favors s , while $a < 0$ favors h .

- (a) Derive the unique maximizer, denoted $\hat{\pi}_\beta$. [3 pts.]
- (b) Prove the worst-case identity

$$\max_{\|\varepsilon\|_\infty \leq \delta} \log \frac{\hat{\pi}_\beta(h)}{\hat{\pi}_\beta(s)} = -a + \frac{2\delta - g}{\beta}.$$

Use this identity to give a necessary and sufficient condition for some admissible ε to yield $\hat{\pi}_\beta(h) > \hat{\pi}_\beta(s)$. Analyze the effect of increasing β when $a > 0$, $a = 0$, and $a < 0$, and state whether it repairs the reward model or only changes adherence to π_0 . [5 pts.]