QIUZHEN QUALIFY EXAM FOR ARTIFICIAL INTELLIGENCE SPRING 2025

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

1. [7 pts.] Let $\mathcal{H}$ be a class of binary classifiers over a domain $\mathcal{X}$. Let $\mathcal{D}$ be an unknown distribution over $\mathcal{X}$, and let f be the target hypothesis in $\mathcal{H}$. Fix some $h \in \mathcal{H}$. Show that the variance of the empirical loss $L_S(h)$, over all possible samples S of size m drawn from $\mathcal{D}$, is inversely proportional to m. That is, show the following relationship:
$$\operatorname*{Var}_{S|x\sim D^m}\big[L_S(h)\big] \propto \frac{1}{m}.$$
More specifically, prove the exact expression:
$$\operatorname*{Var}_{S|x\sim D^m}\big[L_S(h)\big] = \frac{L_{\mathcal{D},f}(h)\cdot(1-L_{\mathcal{D},f}(h))}{m}.$$

2. [7 pts.] Let $\mathcal{X}$ be a domain, and let $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_m$ be a sequence of distributions over $\mathcal{X}$. Let $\mathcal{H}$ be a finite class of binary classifiers over $\mathcal{X}$, and let $f \in \mathcal{H}$. Suppose we obtain a sample $S = \{(x_1, y_1), ..., (x_m, y_m)\}$, where the $i-$th instance x_i is sampled from $\mathcal{D}_i$, and $y_i = f(x_i)$ Let $\overline{\mathcal{D}}_m$ denote the average as:
$$\overline{\mathcal{D}}_m = \frac{\mathcal{D}_1 + \cdots + \mathcal{D}_m}{m}.$$
Fix an accuracy parameter $\epsilon = \frac{1}{4}$. Show that
$$\mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } L_{(\overline{\mathcal{D}}_m,f)}(h) > \epsilon \text{ and } L_{(S,f)}(h) = 0\right] \leq |\mathcal{H}|e^{\frac{-m}{4}}.$$
Hint: $\ln\!\left(\frac{3}{4}\right) < -\frac{1}{4}$.

3. [9 pts.] Let H and H' be two families of functions mapping from X to $\{0,1\}$ with finite VC dimensions.
   (a) [5 pts.] Show that
$$\mathrm{VCdim}(H \cup H') \leqslant \mathrm{VCdim}(H) + \mathrm{VCdim}(H') + 1.$$
   (b) [4 pts.] Use this to determine the VC dimension of the hypothesis set formed by the union of axis-aligned rectangles and triangles in dimension 2. You may use the fact that the VC dimension of the hypothesis set formed by the union of triangles in dimension 2 is 7, without any proof.

4. [10 pts.] Prove the following two statements:
   (a) [5 pts.] Let $\mathcal{D}$ be a distribution. Let $S = (z_1, \ldots, z_m)$ be an *i.i.d.* sequence of examples. Let A be a learning algorithm that is on-average replace-one stable with rate $\epsilon(m)$. Then:
$$\mathop{\mathbb{E}}_{S \sim D^m}\big[L_D\big(A(S)\big) - L_S\big(A(S)\big)\big] \leqslant \epsilon(m).$$
   (b) [5 pts.] Let ℓ be a convex ρ-Lipschitz loss function. The regularised Empirial Risk Minimisation (ERM) satisfies:
$$\mathrm{E}_S\big[L_{\mathcal{D}}(A(S))\big] \leqslant L_{\mathcal{D}}(w^*) + \lambda \|w^*\|^2 + \frac{2\rho^2}{\lambda m},$$
   where
$$w^* = \arg\min_{w \in \mathcal{H}} L_{\mathcal{D}}(w).$$

B. Deep Learning and Reinforcement Learning

1. [18 pts.] Consider a classification problem: the training dataset is given as $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^d$ represents the input features, and $y_i \in \{1, 2, \ldots, C\}$ represents the class labels. A supervised deep learning pipeline typically includes preparing training data, defining a hypothesis space, designing a training scheme, and optimizing the network. Answer the following questions :
   (a) [4 pts.] Define the hypothesis space consisting of all neural networks structured with
   - a **feature extractor**: a multi-layer perceptron (MLP) that maps the inputs to a learned feature representation,
   - a **classifier**: a multi-layer perceptron (MLP) that maps the feature representation to class probabilities.

   Please specify the input, output and the parameters for training.
   (b) [4 pts.] Specify an appropriate loss function for this classification problem and explain how stochastic gradient descent (SGD) is used to optimize the neural network.
   (c) [4 pts.] To stabilize training, batch normalization is often applied. Describe how batch normalization works during the training and testing phase.

(d) [3 pts.] If the training error is unsatisfactory, describe what adjustments you can make to improve the expressivity of the neural network. Discuss at least two approaches.

(e) [3 pts.] If the training error is low but the testing error is high, propose strategies to reduce overfitting. Discuss at least two approaches.

2. [15 pts.] Generative models aim to train a neural network generator to produce samples similar to the training data. The forward process of a diffusion model progressively adds noise to the data, transforming the data distribution into a normal distribution. The reverse process gradually denoises the data, reverting the normal distribution back to the data distribution. In score-based diffusion models, the forward process can be represented as a Stochastic Differential Equation (SDE). Consider the Variance Preserving (VP) SDE:

$$d\mathbf{x} = -\frac{1}{2}\beta(t)\mathbf{x}\,dt + \sqrt{\beta(t)}\,d\mathbf{W}_t,$$

where $\mathbf{W}_t$ is a standard Wiener process, and $\beta(t)$ is a time-dependent noise schedule.

(a) [4 pts.] Write down the reverse-time SDE corresponding to the forward process and explain why training the diffusion model requires score matching as follows

$$\min_\theta \mathbb{E}_t \lambda(t) \mathbb{E}_{\mathbf{x}_t} \|s_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)\|_2^2,$$

where $\lambda(t)$ is a weighting function, $s_\theta(\mathbf{x}_t, t)$ is the neural network to train and p_t is the marginal distribution of $\mathbf{x}_t$.

(b) [6 pts.] Prove that

$$\mathbb{E}_{\mathbf{x}_t} \|s(\theta, t) - \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)\|_2^2 = \mathbb{E}_{(\mathbf{x}_0, \mathbf{x}_t)} \|s_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p_{t|0}(\mathbf{x}_t|\mathbf{x}_0)\|_2^2 + C,$$

where C is a constant, $p_t(\mathbf{x}_t)$ is the marginal distribution of $\mathbf{x}_t$, and $p_{t|0}(\mathbf{x}_t|\mathbf{x}_0)$ is the conditional distribution of $\mathbf{x}_t$ given the original data $\mathbf{x}_0$.

(c) [5 pts.] For the given VP SDE, what is the conditional distribution $p_{t|0}(\mathbf{x}_t|\mathbf{x}_0)$? Based on this, derive the final denoising score-matching loss function for training score-based diffusion models.

C. Optimization Methods in Artificial Intelligence

Let $f : \mathbb{R}^d \to \mathbb{R}$ be an L-smooth and μ-strongly convex function. The stochastic gradient is defined as:

$$g(x, \xi) = \nabla f(x) + \xi,$$

where ξ is a zero-mean random variable with $\mathbb{E}[\|\xi\|^2] \leq \sigma^2$. Consider the SGD method:

$$x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g(x^k, \xi^k)),$$

used to solve the optimization problem:

$$\min_x f(x) + R(x).$$

1. [5 pts.] *L-Smoothness and μ-Strong Convexity:* Provide the definitions of L-smoothness and μ-strong convexity, respectively.

2. [6 pts.] *Combining Smoothness and Strong Convexity:* Suppose that $f(x)$ is continuously differentiable L-smooth and μ-convex.
   - Show that $g(x) = f(x) - \frac{\mu}{2}\|x\|^2$ is continuously differentiable, convex and $(L-\mu)$-smooth.
   - Using the fact
   $$\frac{1}{L-\mu}\|\nabla g(x) - \nabla g(y)\|^2 \le \langle \nabla g(x) - \nabla g(y), x - y\rangle,$$
   show that
   $$\mu\|x-y\|^2 + \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \le \left(1 + \frac{\mu}{L}\right)\langle \nabla f(x) - \nabla f(y), x - y\rangle.$$

3. [6 pts.] *Recurrence Relation:* Using the non-expansiveness property of prox and the optimality condition of $\text{prox}_{\gamma R}$, derive the recurrence relation for $\|x^{k+1} - x^\star\|^2$:
$$\|x^{k+1} - x^\star\|^2 \le \|x^k - x^\star - \gamma(\nabla f(x^k) - \nabla f(x^\star) + \xi^k)\|^2.$$

4. [6 pts.] *Variance Decomposition:* What is variance decomposition? Explain how it is applied to $\|x^{k+1} - x^\star\|^2$, and derive:
$$\mathbb{E}_k[\|x^{k+1} - x^\star\|^2] \le \|x^k - x^\star - \gamma(\nabla f(x^k) - \nabla f(x^\star))\|^2 + \gamma^2\sigma^2,$$
where $\mathbb{E}_k[\cdot]$ denotes the conditional expectation given $\xi^k, \dots, \xi^0$.

5. [5 pts.] *Simplified Recurrence:* Using the tower property, prove that when $\gamma = \frac{2}{\mu+L}$, the recurrence simplifies to:
$$\mathbb{E}[\|x^{k+1} - x^\star\|^2] \le (1-\rho)\mathbb{E}[\|x^k - x^\star\|^2] + \gamma^2\sigma^2,$$
where $\rho = \frac{4\mu L}{(\mu+L)^2}$.

6. [5 pts.] *Complexity:* Prove that for any desired precision $\varepsilon > 0$, there exists a step size γ such that:
$$k \ge \frac{L/\mu + 3}{4}\log\frac{1}{\varepsilon}$$
implies:
$$\mathbb{E}[\|x^k - x^\star\|^2] \le \varepsilon\|x^0 - x^\star\|^2 + \frac{\gamma\sigma^2}{\mu}.$$

D. Natural Language Processing

Short Answer Questions on Concepts.

1. [3 pts.] In a trigram language model, how is $p(w_3 \mid w_1, w_2)$ learned from a training corpus? In real-world applications, we may encounter situations where (w_1, w_2) , or even w_1 or w_2 , do not appear in the training corpus. How can we estimate $p(w_3 \mid w_1, w_2)$ in such cases?

2. [3 pts.]What is the central idea behind representation learning? Which algorithms implement this concept?

3. [3 pts.] Why do current large language models (LLMs) use top-P sampling during inference instead of greedy or beam search? Additionally, why is top-P sampling considered superior to other methods, such as top-K sampling?

4. [3 pts.] List several key points that contribute to the success of current LLM based approach of Artificial General Intelligence and explain their importance.

5. [3 pts.] Identify at least two advantages of the Transformer architecture compared to recurrent-based models (e.g., LSTM, GRU) for sequence-to-sequence tasks.

Questions on Algorithm Analysis.

6. [5 pts.] Explain why the data scaling law, which describes the relationship between performance and sample size, holds true from the perspective of statistical learning. Please provide a detailed analysis process, avoiding overly simplistic assumptions (such as assuming the data follows a Gaussian distribution).

7. [5 pts.] We are implementing the Reinforcement Learning with Human Feedback (RLHF) procedure to train a Large Language Model (LLM). Assuming that a reward model has already been trained, explain how reinforcement learning is used to align the model with human preferences. The explanation should include details about data usage, the reinforcement learning loss function, and the gradient of the loss.

Questions on Applied System Design.

8. [8 pts.] We have a general-purpose Large Language Model (LLM), but it performs poorly on question-answering tasks in a specific problem domain due to a lack of domain-specific knowledge. Although we have collected a large corpus of documents in this domain, we do not have labeled question-answering pairs. Design a system that can accurately answer

questions within this domain. You may incorporate additional components, such as embedding models, into the system.