QIUZHEN QUALIFY EXAM FOR ARTIFICIAL INTELLIGENCE
FALL 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

Part I: Warm-Up Questions [5 pts. + 1 bonus pts.]

This section contains multiple-choice questions. Please select **only one** answer for each question. No justification is required.

1. [0.5 pts.] Which of the following best describes the bias–variance tradeoff?
   (a) Increasing model bias reduces variance but increases approximation error
   (b) Increasing model bias reduces both bias and variance
   (c) Variance is always independent of bias
   (d) The tradeoff only applies to neural networks
2. [0.5 pts.] Why does the 'No Free Lunch' theorem matter in machine learning?
   (a) Every algorithm performs optimally on all data sets
   (b) All learning tasks require exponential time
   (c) There is no universally superior learning algorithm across all tasks
   (d) Only smooth loss functions can be optimised
3. [0.5 pts.] Which of the following is true about Rademacher complexity?
   (a) It measures how well a hypothesis class fits random labels
   (b) It always equals the VC dimension
   (c) It decreases with model complexity
   (d) It is independent of sample size
4. [0.5 pts.] In Empirical Risk Minimisation (ERM), what is being minimised?
   (a) The true risk on unseen data
   (b) The average loss on the training sample

(c) The VC dimension of the hypothesis class
(d) The variance of the hypothesis

5. [1 pts.] Which statement is true about the VC dimension?
   (a) A class with infinite VC dimension can never be PAC learnable
   (b) The VC dimension of halfspaces in $\mathbb{R}^d$ is exactly $d+1$
   (c) VC dimension always equals the number of parameters of the hypothesis class
   (d) Finite VC dimension implies zero generalisation error
6. [1 pts.] In the agnostic PAC learning model, what changes compared to the realisable PAC model?
   (a) The learner must always find a hypothesis with zero training error
   (b) The hypothesis class must contain the true labelling function
   (c) The goal is to compete with the best hypothesis in the class, even if labels are noisy
   (d) The sample complexity becomes independent of ϵ
7. [1 pts.] In stochastic gradient descent (SGD), why does using a decreasing learning rate help?
   (a) It avoids overfitting completely
   (b) It ensures convergence under certain convexity assumptions
   (c) It increases variance in gradient estimates
   (d) It eliminates the need for backpropagation
8. **[Optional: 1-Points Bonus!]** What does Sauer's Lemma imply about a hypothesis class H with VC-dimension d?
   (a) H can shatter any set of size larger than d
   (b) The growth function $\tau_H(m)$ is bounded polynomially in m once $m > d$
   (c) The empirical risk minimiser achieves zero risk for $m \leq d$
   (d) The sample complexity is independent of d

Part II: Theoretical Exercises [28 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.

9. [5 pts.] Let $\mathcal{H}$ be the class of signed intervals, that is,

$$\mathcal{H} = \{h_{a,b,s} : a \leq b,\ s \in \{-1, 1\}\},$$

where

$$h_{a,b,s}(x) = \begin{cases} s & \text{if } x \in [a, b] \\ -s & \text{if } x \notin [a, b] \end{cases}.$$

Calculate VCdim($\mathcal{H}$).

10. [6 pts.] **Lemma – show it holds. Strong Convexity Properties.** Show the following holds.

Let $f : \mathbb{R}^d \to \mathbb{R}$. Then:

(a) The function $f(\mathbf{w}) = \lambda\|\mathbf{w}\|^2$ is 2λ-strongly convex.

(b) If f is λ-strongly convex and g is convex, then $f+g$ is λ-strongly convex.
(c) If f is λ-strongly convex and $\mathbf{u}$ is a minimiser of f, then for any $\mathbf{w}$,

$$f(\mathbf{w}) - f(\mathbf{u}) \geqslant \frac{\lambda}{2}\|\mathbf{w}-\mathbf{u}\|^2.$$

11. [5 pts.] Let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{0,1\}$, and let $\mathcal{H}$ be the class of concentric circles in the plane, that is,

$$\mathcal{H} = \{h_r : r \in \mathbb{R}_+\}, \quad \text{where } h_r(x) = \mathbf{1}_{\{\|x\| \leq r\}}.$$

Prove that $\mathcal{H}$ is PAC learnable (assume realisability), and its sample complexity is bounded by

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \frac{\log(1/\delta)}{\epsilon}.$$

12. [7 pts.] We initialise $\mathbf{w}_1 \in \mathcal{W}$. At round $t = 1, 2, \ldots$, we obtain a random estimate $\hat{\mathbf{g}}_t$ of a subgradient $\mathbf{g}_t \in \partial F(\mathbf{w}_t)$ so that $\mathbb{E}[\hat{\mathbf{g}}_t] = \mathbf{g}_t$, and update the iterate $\mathbf{w}_t$ as follows:

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}(\mathbf{w}_t - \eta_t \hat{\mathbf{g}}_t),$$

where η_t is a suitably chosen step-size parameter, and $\Pi_{\mathcal{W}}$ denotes projection on $\mathcal{W}$. Assume F is λ -strongly convex, and that

$$\mathbb{E}\left[\|g_t\|^2\right] \leq G^2$$

for all t. Consider Stochastic Gradient Descent with step sizes $\eta_t = \frac{1}{\lambda t}$.
Show that for any $w \in W$, the following inequality holds:
$\mathbb{E}\left[\|w_{t+1} - w\|^2\right] \leq \mathbb{E}\left[\|w_t - w\|^2\right] - 2\eta_t \mathbb{E}\left[\langle g_t, w_t - w\rangle\right] + \eta_t^2 G^2$.

13. [5 pts.] **Neural Networks are universal approximators:** Let $f : [-1,1]^n \to [-1,1]$ be a ρ-Lipschitz function. Fix some $\epsilon > 0$. Construct a neural network $N : [-1,1]^n \to [-1,1]$, with the sigmoid activation function, such that for every $\mathbf{x} \in [-1,1]^n$ it holds that

$$|f(\mathbf{x}) - N(\mathbf{x})| \leq \epsilon.$$

Hint: Partition $[-1,1]^n$ into small boxes. Use the Lipschitzness of f to show that it is approximately constant at each box. Finally, show that a neural network can first decide which box the input vector belongs to, and then predict the averaged value of f at that box.

B. Deep Learning and Reinforcement Learning

1. [9 pts.] **Neural Network Architectures (CNNs & Transformers)**
   (a) [3 pts.] Derive the number of trainable parameters in a single convolutional layer with input size $H \times W \times C_{\text{in}}$, kernel size $k \times k$, and C_{out} output channels (assume bias), and compare it with a fully connected (dense) layer of the same input and output size.
   (b) [3 pts.] Write the scaled dot-product self-attention formula (define Q, K, V). Explain why positional information is necessary in Transformers and describe one method to inject positional information.
   (c) [3 pts.] State one key benefit of (i) convolution for vision and (ii) self-attention. Then design a minimal vision transformer for images that uses both structures: specify how to tokenize the image into patch embeddings, where self-attention is applied, and where convolution is introduced.

2. [14 pts.] **Generative Models and Likelihood-based Training**
   (a) [2 pts.] Show that maximizing the likelihood of a generative model $p_\theta(x)$ given data distribution $p_{\text{data}}(x)$ is equivalent to minimizing the KL divergence $\mathrm{KL}(p_{\text{data}} \,\|\, p_\theta)$.
   (b) [4 pts.]
   Consider a normalizing flow model composed of L invertible transformations

   $$z_0 \sim p(z_0), \qquad z_\ell = f_\ell(z_{\ell-1}), \;\; \ell = 1, \ldots, L, \qquad x = z_L,$$

   where each f_ℓ is bijective and differentiable, and $p(z_0)$ is a simple base density (e.g., standard Gaussian). Write down the training objective (loss) for normalizing flows on a dataset $\{x^{(i)}\}_{i=1}^N$ sampled from the data distribution.
   (c) [4 pts.] Explain why the variational autoencoder (VAE) uses the evidence lower bound (ELBO) to approximate maximum likelihood training. Write down the ELBO expression and explain the roles of the reconstruction term and the regularization term.
   (d) [4 pts.] We can interpret diffusion probabilistic models as a form of hierarchical variational autoencoders (VAEs). Let $x_0 \sim p_{\text{data}}$ denote a data sample. The forward process (the encoder) is defined by adding noise

   $$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}), \qquad q(x_t \mid x_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}\, x_{t-1},\; \beta_t I)\,,$$

   where $\alpha_t = 1 - \beta_t \in (0, 1)$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

   Write down the probabilistic model of the backward process (the decoder), and show that the ELBO for $\log p_\theta(x_0)$ can be written as

   $$\log p_\theta(x_0) \;\geqslant\; -\,\mathrm{KL}\big(q(x_T \mid x_0) \,\|\, p(x_T)\big)$$

$$
\begin{aligned}
&- \sum_{t=2}^{T} \mathbb{E}_q\big[\mathrm{KL}\big(q(x_{t-1} \mid x_t, x_0) \,\|\, p_\theta(x_{t-1} \mid x_t)\big)\big] \\
&+ \ \mathbb{E}_q\big[\log p_\theta(x_0 \mid x_1)\big].
\end{aligned}
$$

3. [10 pts.] **Policy Gradient Methods**

Consider a discounted Markov Decision Process (MDP) $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with states $s \in \mathcal{S}$, actions $a \in \mathcal{A}$, transition kernel $P(s' \mid s, a)$, reward $r(s, a)$ bounded, and discount $\gamma \in (0, 1)$. Let $\pi_\theta(a \mid s)$ be a differentiable, stochastic policy with parameters θ, and let s_0 be a fixed start state. Define the (discounted) return

$$
G_0 \ = \ \sum_{t=0}^{\infty} \gamma^t \, r(s_t, a_t),
$$

and the performance objective

$$
J(\theta) \ = \ V^{\pi_\theta}(s_0) \ = \ \mathbb{E}_{\tau \sim \pi_\theta}\big[\, G_0 \,\big],
$$

where a trajectory $\tau = (s_0, a_0, s_1, a_1, \ldots)$ is generated by $s_{t+1} \sim P(\cdot \mid s_t, a_t)$ and $a_t \sim \pi_\theta(\cdot \mid s_t)$.

Denote the value and action-value functions by

$$
\begin{aligned}
V^\pi(s) &= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s\right], \\
Q^\pi(s, a) &= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s, a_0 = a\right],
\end{aligned}
$$

and the advantage by $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$. Let $d^\pi(s)$ be the (unnormalized) γ-discounted state visitation distribution:

$$
d^\pi(s) \ = \ \sum_{t=0}^{\infty} \gamma^t \ \Pr(s_t = s \mid \pi).
$$

(a) [6 pts.] Prove the Policy Gradient Theorem.

$$
\begin{aligned}
\nabla_\theta J(\theta) \ &= \ \mathbb{E}_{\pi_\theta}\left[\, \sum_{t=0}^{\infty} \gamma^t \, \nabla_\theta \log \pi_\theta(a_t \mid s_t) \, Q^{\pi_\theta}(s_t, a_t) \,\right] \\
&= \ \frac{1}{1-\gamma} \, \mathbb{E}_{s \sim d^{\pi_\theta}, \, a \sim \pi_\theta}\big[\, \nabla_\theta \log \pi_\theta(a \mid s) \, Q^{\pi_\theta}(s, a) \,\big].
\end{aligned}
$$

(b) [4 pts.] Show that for any function $b : \mathcal{S} \to \mathbb{R}$,

$$
\mathbb{E}_{\pi_\theta}\left[\, \sum_{t=0}^{\infty} \gamma^t \, \nabla_\theta \log \pi_\theta(a_t \mid s_t) \, b(s_t) \,\right] \ = \ 0,
$$

and hence the policy gradient can be equivalently written as

$$\nabla_\theta J(\theta) \;=\; \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t \, \nabla_\theta \log \pi_\theta(a_t \mid s_t) \left(Q^{\pi_\theta}(s_t, a_t) - b(s_t)\right)\right].$$

In the advantage method, how should $b(s)$ be chosen, and what is the benefit of this choice?

C. Optimization Methods in Artificial Intelligence

Consider the regularized finite-sum problem

$$\min_{x \in \mathbb{R}^d} \; F(x) \;:=\; f(x) + R(x), \qquad f(x) \;:=\; \frac{1}{n} \sum_{i=1}^{n} f_i(x),$$

where each $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex and L_i-smooth, the aggregate f is L-smooth and μ-strongly convex ($\mu > 0$), and $R : \mathbb{R}^d \to (-\infty, +\infty]$ is proper, closed, and convex. For a probability vector $q = (q_1, \ldots, q_n)$ with $q_i > 0$ and $\sum_i q_i = 1$, define a categorical random variable $s \in \{1, \ldots, n\}$ with $\mathbb{P}(s = i) = q_i$. Fix a minibatch size $\tau \in \{1, 2, \ldots\}$, and draw i.i.d. copies $s_1, \ldots, s_\tau$ of s. Define the multisampling gradient estimator

$$g(x) \;:=\; \frac{1}{\tau} \sum_{t=1}^{\tau} \frac{1}{n \, q_{s_t}} \nabla f_{s_t}(x),$$

and the proximal SGD iteration

$$x^{k+1} \;=\; \operatorname{prox}_{\gamma R}\!\left(x^k - \gamma \, g^k\right), \qquad g^k \;:=\; g(x^k).$$

We denote the Bregman divergence of f by

$$D_f(x, y) \;:=\; f(x) - f(y) - \langle \nabla f(y),\, x - y \rangle.$$

1. [3 pts.] *Basic Definitions:* Give the definitions of L-smoothness and μ-strong convexity.

2. [2 pts.] *Uniqueness of the Minimizer:* Show that F admits a unique minimizer $x^\star$.

3. [5 pts.] *Unbiasedness of the Gradient Estimator:* Prove that $g(x)$ is unbiased, i.e., $\mathbb{E}[g(x)] = \nabla f(x)$.

4. [8 pts.] *Expected Smoothness Bound:* Assuming each f_i is L_i-smooth and convex, and f is L-smooth, show that for all $x, y \in \mathbb{R}^d$,

$$\mathbb{E}\left[\|g(x) - g(y)\|^2\right] \;\leqslant\; 2\, A''(\tau, q)\, D_f(x, y),$$

where the *expected-smoothness constant* A'' (depending on τ and q) is

$$A''(\tau, q) \;:=\; \frac{1}{\tau}\Big(\max_i \frac{L_i}{nq_i}\Big) \;+\; \Big(1 - \frac{1}{\tau}\Big)L.$$

Hint: Expand the square, separate diagonal and cross terms using independence, and use smoothness to bound gradient differences via Bregman divergences.

5. [4 pts.] *Extremes, Monotonicity, and Interpolation:*
   (a) Evaluate $A''(\tau, q)$ at $\tau = 1$ and as $\tau \to +\infty$. Identify the limiting algorithms (SGD-NS vs. GD) and the corresponding constants.
   (b) Show that $A''(\tau, q)$ is nonincreasing in τ, and interpret how the minibatch size τ interpolates between SGD-NS and GD.

   Hint: Use $\frac{1}{n}\sum_{i=1}^n L_i \geqslant L$.

6. [4 pts.] *Design of Importance Sampling q:* For a fixed τ, minimize the first term of $A''(\tau, q)$, i.e.,

$$\min_{q \in \Delta_n} \max_i \frac{L_i}{nq_i}, \qquad \text{where } \Delta_n = \{q \in \mathbb{R}^n_{++} : \sum_i q_i = 1\}.$$

   Derive the optimal $q^\star$ and the attained value of $\max_i \frac{L_i}{nq_i^\star}$. Compare with uniform sampling $q_i^{\text{uni}} = \frac{1}{n}$.

   Hint: Use $\frac{1}{n}\sum_{i=1}^n L_i \leqslant \max_i L_i$.

7. [3 pts.] *Variance at the Optimum and Minibatch Scaling:* Let $\xi(x) := g(x) - \nabla f(x)$. Show that

$$\mathbb{E}\big[\|\xi(x^\star)\|^2\big] = \frac{1}{\tau}\,\mathrm{Var}\left(\frac{1}{nq_s}\nabla f_s(x^\star)\right),$$

   i.e., the variance at the optimum scales as $1/\tau$. Express your answer in terms of q and $\{\nabla f_i(x^\star)\}_{i=1}^n$.

8. [2 pts.] *AC inequality and computing (A, C):* Consider the following classical results:
   AC inequality: There exist constants $A \geqslant 0$ and $C \geqslant 0$ such that for all $k \geqslant 0$,

$$\mathbb{E}\big[\, \|g^k - \nabla f(x^\star)\|^2 \;\big|\; x^k\big] \;\leqslant\; 2A\, D_f(x^k, x^\star) + C.$$

   Implication from expected smoothness: If $g(x)$ is an unbiased estimator of $\nabla f(x)$ and, for all x, y,

$$\mathbb{E}\big[\|g(x) - g(y)\|^2\big] \;\leqslant\; 2A''\, D_f(x, y) + C''(y),$$

   then for $G(x, y) := \mathbb{E}\|g(x) - \nabla f(y)\|^2$ one has the AC inequality

$$G(x, y) \;\leqslant\; 2A\, D_f(x, y) \;+\; C,$$

   where $A = 2A''$ and $C = 2\big(\mathrm{Var}[g(y)] + C''(y)\big)$.

Specialize the above implication to obtain the AC constants (A, C), and write an explicit formula for $\mathrm{Var}[g(x^\star)]$.

9. [2 pts.] *Stepsize and convergence:* Given the classical convergence theorem of SGD: Assume f is μ-convex, g^k is unbiased, and the AC inequality holds with constants (A, C). Then for any stepsize $0 < \gamma \leqslant \frac{1}{A}$, the iterates satisfy

$$\mathbb{E}\|x^k - x^\star\|^2 \;\leqslant\; (1-\gamma\mu)^k\|x^0 - x^\star\|^2 \;+\; \frac{\gamma C}{\mu}.$$

Using the (A, C) obtained in question 8 to compute:
- the admissible stepsize range;
- the explicit convergence bound with the noise floor written in closed form.

D. Natural Language Processing

Part I: Questions on Concepts and Algorithm Analysis [13 pts.]

1. [3 pts.] **Embedding for Texts and Graphs.** Embedding methods obtain vector representations of individual items by exploiting the relationships among them within a collection. **GloVe** and **Skip-Gram** are used to learn representations of words in text; **Node2Vec** learns representations of nodes in a network; and **TransE** learns representations of entities and relations in a knowledge graph. In *2–3 sentences each*, succinctly describe the *data signal*, the *learning objective type*, and the key *inductive bias* (the model's built-in assumptions) for the **four** paradigms: **GloVe**, **Skip-gram**, **Node2Vec**, **TransE**.

2. [5 pts.] **Scaling Laws and Architectural Bias (LSTM vs. Transformer).** Consider language models trained autoregressively on the same corpus with identical tokenization, context length, optimizer, and training pipeline. For an architecture $\mathrm{arch} \in \{\mathrm{LSTM}, \mathrm{Transformer}\}$, assume the test loss obeys

$$L_{\mathrm{arch}}(P, T) = L_\infty + A_{\mathrm{arch}} P^{-\alpha_{\mathrm{arch}}} + B_{\mathrm{arch}} T^{-\beta_{\mathrm{arch}}}, \qquad \alpha_{\mathrm{arch}}, \beta_{\mathrm{arch}} > 0,$$

where P is the parameter count and T is the number of training tokens. Assume L_∞ is the same across architectures (same task/data).

(a) **Fixed-data regime.** Fix a large but finite $T = T_0$. On a log–log plot of $L(P, T_0) - L_\infty$ versus P, state the expected qualitative relationships between $(\alpha_{\mathrm{arch}}, A_{\mathrm{arch}})$ for LSTM vs. Transformer and the resulting relative positions and slopes of their P–L curves.

(b) **Fixed-parameter regime.** Fix a parameter budget $P = P_0$ and vary T. On a log–log plot of $L(P_0, T) - L_\infty$ versus T, state the expected qualitative relationships between $(\beta_{\mathrm{arch}}, B_{\mathrm{arch}})$ for LSTM vs. Transformer and the relative positions and slopes of their T–L curves.

(c) **Justification.** Justify your answers in (a)–(b) from the viewpoints of: (i) long-range dependency handling, (ii) parallelizability/throughput and optimization dynamics, and (iii) inductive bias and sample efficiency.

3. [5 pts.] **Analysis of a Markov Logic Network (MLN).** A Markov Logic Network (MLN) defines a probability distribution over possible worlds. It is specified by a set of weighted first-order logic formulas, (ϕ_i, w_i). For a given world x (i.e., a truth assignment to all possible ground atoms), its probability is given by:

$$P(X = x) = \frac{1}{Z} \exp\left(\sum_i w_i n_i(x)\right),$$

where w_i is the weight of the i-th formula ϕ_i, $n_i(x)$ is the number of true groundings of ϕ_i in world x, and Z is the partition function (normalization constant).

Consider a simple MLN designed to analyze topics of academic papers, consisting of two weighted formulas:

(a) $\forall p_1, p_2\, \text{Cites}(p_1, p_2) \wedge \text{InTopic}(p_1, \text{“AI”}) \Rightarrow \text{InTopic}(p_2, \text{“AI”})$, with weight $w_1 = 1.5$;
(Interpretation: If a paper in the AI topic cites another paper, the cited paper is also likely in the AI topic.)

(b) $\forall p\, \text{InTopic}(p, \text{“AI”})$, with weight $w_2 = -1.0$.
(Interpretation: There is a general prior that a paper is less likely to be in the AI topic.)

Assume our domain contains only two papers, P1 and P2, and one topic, “AI”. Answer the following questions:

(a) **Model Structure Analysis** Please briefly describe the structure of the ground Markov network corresponding to this MLN. What are the nodes? What are the cliques and why?

(b) **Probability Calculation** Consider a specific possible world x_1 where: P1 cites P2, P1 is in the AI topic, but P2 is not. Furthermore, P2 does not cite P1. (i.e., Cites(P1, P2) is true, Cites(P2, P1) is false, InTopic(P1, “AI”) is true, and InTopic(P2, “AI”) is false.) 1) For this world x_1, calculate the number of true groundings for each of the two formulas (i.e., find the values of $n_1(x_1)$ and $n_2(x_1)$). 2) Write down the un-normalized probability of world x_1 (i.e., the $\exp(\dots)$ term).

(c) **Parameter Impact Analysis** 1) Suppose we change the weight of the first formula, w_1, from 1.5 to -1.5. In one sentence, what kind of academic citation phenomenon does the model now favor? 2) Without re-calculating specific probabilities, what qualitative effect (e.g., significant increase, significant decrease, or little change) does this modification have on the probability of a world where ‘P1’ and ‘P2’ are both AI papers and ‘P1’ cites ‘P2’? Briefly explain your reasoning.

(d) **Maximum a Posteriori (MAP) Inference.** Given that $\text{Cites}(P_1, P_2) = \text{true}$ and all other atoms are unobserved, and using the original weights $w_1 = 1.5$ and $w_2 = -1.0$, determine the MAP truth assignments for $\text{InTopic}(P_1, \text{“AI”})$ and $\text{InTopic}(P_2, \text{“AI”})$. Provide a 1–2 sentence justification.

Part II: Questions on Algorithm and System Design [20 pts.]

4. [10 pts.] **Topic–Ontology LDA for Fact Triples.** A document d is represented by a multiset of fact triples

$$\mathcal{F}_d = \{f = (s, r, o)\},$$

where $s \in \mathcal{V}_s$ and $o \in \mathcal{V}_o$ are subject/object *mentions* (surface phrases), and $r \in \mathcal{V}_r$ is a relation *mention*. Each triple also carries latent *ontology* variables: subject type $c_s \in \mathcal{C}$, object type $c_o \in \mathcal{C}$, and relation type $t \in \mathcal{R}$. Assume there are K topics. The goal is to uncover (i) document–level topic mixtures and (ii) ontology assignments/types for entities and relations using an LDA-style generative approach. Unless stated otherwise, use symmetric Dirichlet priors.

(a) **Generative process & model specification.** Design an LDA-style generative model that jointly produces fact triples. Your model should at least include: document-level topic mixtures $\theta_d \sim \text{Dir}(\boldsymbol{\alpha})$; topic-specific distributions over ontology variables $\pi_k^{(s)}$ on $\mathcal{C}$, $\pi_k^{(o)}$ on $\mathcal{C}$, $\pi_k^{(r)}$ on $\mathcal{R}$; and type-specific surface-form distributions $\phi_c^{(s)}$ on $\mathcal{V}_s$, $\phi_c^{(o)}$ on $\mathcal{V}_o$, $\phi_t^{(r)}$ on $\mathcal{V}_r$.

(b) **Joint probability.** Write the factorized form of the full joint

$$p(\Theta, \Pi, \Phi, Z, C_s, C_o, T, S, O, R \mid \text{hyperparameters}),$$

where $\Theta = \{\theta_d\}_d$, $\Pi = \{\pi_k^{(s)}, \pi_k^{(o)}, \pi_k^{(r)}\}_k$, $\Phi = \{\phi_c^{(s)}, \phi_c^{(o)}\}_{c\in\mathcal{C}} \cup \{\phi_t^{(r)}\}_{t\in\mathcal{R}}$ and $Z = \{z_f\}$, $C_s = \{c_{s,f}\}$, $C_o = \{c_{o,f}\}$, $T = \{t_f\}$, and S, O, R the observed mentions.

(c) **Automatic naming of discovered categories.** After inference, suppose you obtain several latent categories for ontologies (entity types and relation types). Design an automatic naming method that gives a reasonable name for each category.

5. [10 pts.] **Design and Analyze a Text-to-Image Diffusion System.** You will design a text-to-image generation system. Given a text prompt T , the system should generate an image I. We adopt a diffusion framework with a Transformer backbone for the image denoiser and a Transformer for the text encoder.

(a) **Forward process, ELBO, and closed-form posteriors.** Let the clean image be $\mathbf{x}_0 \in \mathbb{R}^{H\times W\times C}$ and define the forward noising process

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\big(\sqrt{\alpha_t}\,\mathbf{x}_{t-1},\, \beta_t\mathbf{I}\big), \quad \beta_t = 1 - \alpha_t, \quad \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s,$$

with prior $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{0}, \mathbf{I})$.

(i) Show that $q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t}\,\mathbf{x}_0,\, (1-\bar{\alpha}_t)\mathbf{I}\big)$.

(ii) Express the evidence lower bound (ELBO) on $\log p_\theta(\mathbf{x}_0)$ as the sum of a reconstruction term and KL divergence terms. You may state the final expression directly; if you provide a derivation, include only the two key steps.

(iii) Show that the exact posterior $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)$ is Gaussian with variance $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\,\beta_t$ and provide its mean.

(b) **Noise-prediction parameterization and training loss.** Assume $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\big(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t),\, \sigma_t^2\mathbf{I}\big)$ with fixed $\sigma_t^2 = \tilde{\beta}_t$.

(i) Show that the optimal mean can be parameterized by a noise-prediction network $\boldsymbol{\epsilon}_\theta$ as

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\,\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right).$$

(ii) Prove that, up to constant and per-timestep weights, training reduces to

$$\mathcal{L}_t = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\Big[\big\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},\, t)\big\|_2^2\Big],$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and t is sampled from a specified distribution over $\{1, \ldots, T\}$.

(c) **Text-conditional modeling with a Transformer.** Design a conditional reverse process $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \textit{Text})$ using Transformer architecture. Your answer should specify:

(i) How to implement the image denoiser using a Transformer architecture;

(ii) How to condition the image denoiser on the input text.