Artificial Intelligence PhD Qualifying Examination

模拟考试 05：参考解答与评分要点

每个 Section 为 33 分；考生选四个 Section 中的三个作答，姓名与学号另计 1 分。以下给出完整参考推导。系统设计题中的方案均为 representative full-credit design；其他满足题面核心约束、论证完整且技术可行的方案同等给分。

A. 机器学习理论：参考解答 [33 points]

(a) 对 $i \in [d]$ 取 $x^{(i)} = \mathbf{1} - e_i$。给定 labels y_i，令 $S_y = \{i : y_i = 0\}$，则 $h_{S_y}(x^{(i)}) = y_i$，故下界为 d。另一方面 $|\mathcal{H}_d| = 2^d$，所以 VC dimension 至多 $\log_2 |\mathcal{H}_d| = d$。因此答案为 d。

Rubric: 点集 2 分；任意 labeling construction 3 分；逐点验证 1 分；upper bound 与结论 2 分。

(b) 记输出集合为 $\widehat{S}$。每个 target variable 在所有正样本上为 1，所以 $S^\star \subseteq \widehat{S}$；没有正样本时 $\widehat{S} = [d]$，仍成立。Observed positive 在所有 retained coordinates 上为 1，故输出 1。Observed negative 至少在某个 $j \in S^\star \subseteq \widehat{S}$ 上为 0，故输出 0；该论证也覆盖 no-positive case。扫描数据为 $O(md)$ time、$O(d)$ space，预测最坏 $O(d)$。

Rubric: subset relation 2 分；positive 1 分；negative 2 分；边界情形 1 分；复杂度 2 分。

(c) 对 fixed h，$\Pr\{|\widehat{L}(h) - L(h)| > \alpha\} \le 2e^{-2m\alpha^2}$。对 2^d 个 hypotheses union bound，并在 uniform event 上作 ERM comparison，得到

$$L(\widehat{h}) \le \min_h L(h) + 2\alpha.$$

取 $\alpha = \varepsilon/2$，充分条件为

$$\boxed{m \ge \frac{2}{\varepsilon^2}\left(d \log 2 + \log \frac{2}{\delta}\right).}$$

枚举全部 conjunctions 的实现需约 $O(md2^d)$ time；统计 sample bound 不自动提供有效算法。

Rubric: Hoeffding/union 2 分；ERM comparison 2 分；常数 2 分；计算解释 2 分。

(d) (a) 记 $\mu(x) = \mathbb{E}_S \widehat{f}_S(x)$。加减 $\mu(x)$ 与 $f(x)$：

$$\widehat{f}_S - Y = (\widehat{f}_S - \mu) + (\mu - f) - \epsilon.$$

平方取期望。第一项均值为零，ϵ 均值为零且与 S independent，因此所有 cross terms 消失；三个平方项依次为 variance、squared bias、noise variance，即得题式。

(b) 若 X 也随机，对 pointwise identity 两侧再取 $\mathbb{E}_X$ 即可，前提是 test noise 条件均值为零且与 training randomness 独立。Training error 是在 observed sample 上的 fit；更复杂模型可降低它，却可能增大 estimator 对换样本的敏感性（variance），也不保证改变 population approximation bias 的方向。

Rubric: 加减 mean/target 2 分；cross terms 2 分；三项识别 2 分；积分 1 分；training-error 解释 2 分。

B. 深度学习与强化学习：参考解答 [33 points]

(a) **Multi-view depth。** [11 points]

(a) 目标相机坐标中的 3D 点为

$$X_t = D_t(p)K^{-1}\widetilde{p}, \qquad X_s = RX_t + t.$$

令 $\widetilde{p}' \sim KX_s$，再除以第三坐标得到 $p' = (u', v')$。若 $M(p)$ 表示投影在图内、深度为正且未被遮挡，可取

$$\mathcal{L}_{\text{photo}} = \frac{\sum_p M(p)\rho\big(I_t(p) - I_s(p'(p))\big)}{\sum_p M(p) + \epsilon},$$

其中 ρ 可为 ℓ_1 或稳健 penalty。分母保证按有效像素平均。

(b) 若 $p' = (u', v')$ 位于四个整数像素之间，bilinear sampler 用四个邻点按横纵距离的乘积加权。权重和为 1，且在一个 cell 内对 (u', v') 分段线性；故梯度经 p'、projection、X_s、D_t 传回网络。在像素 cell 边界、遮挡 mask 的硬切换或越界处不经典可微，实践中使用几乎处处梯度或软 mask。

(c) 任何两项即可：遮挡/新显露区域；光照或曝光变化；非 Lambertian 反射；动态物体；无纹理区域；重复纹理；相机位姿错误。

(d) 输入可为 (I_t, I_s, K, R, t)，输出 $D_t \in \mathbb{R}_+^{H\times W}$。可报告 AbsRel、RMSE 或阈值准确率。纯单目联合 depth–pose 学习存在整体尺度 ambiguity；需 metric depth、已知 baseline、stereo 或其他尺度监督固定。

Rubric: 几何链与规范化 loss 4 分；sampler/gradient/非光滑点 3 分；两个 failure 2 分；I/O、metric、scale 2 分。

(b) **Conditional VAE。** [13 points]

(a) 插入 posterior 得

$$\log p_{\theta,\psi}(D \mid I) \geq \mathbb{E}_{q_\phi(z|D,I)} \log p_\theta(D \mid z, I) - \mathrm{KL}\big(q_\phi(z \mid D, I) \| p_\psi(z \mid I)\big) =: \mathcal{L}_{\text{ELBO}}.$$

gap 为 $\mathrm{KL}(q_\phi(z \mid D, I) \| p_{\theta,\psi}(z \mid D, I)) \geq 0$。

(b) 若 $q = \mathcal{N}(\mu_q, \mathrm{diag}(\sigma_q^2))$，则 $z = \mu_q + \sigma_q \odot \epsilon$，$\epsilon \sim \mathcal{N}(0, I)$。若 $p = \mathcal{N}(\mu_p, \mathrm{diag}(\sigma_p^2))$，

$$\mathrm{KL}(q\|p) = \frac{1}{2}\sum_j \left[\frac{\sigma_{q,j}^2 + (\mu_{q,j} - \mu_{p,j})^2}{\sigma_{p,j}^2} - 1 + \log\frac{\sigma_{p,j}^2}{\sigma_{q,j}^2}\right].$$

(c) 训练时用 (I, D) 得到 q，重参数化采样并优化 reconstruction log-likelihood 减 KL。测试时没有 D，必须从 $p_\psi(z \mid I)$ 采样并由 decoder 生成 D。不同 z 可产生不同但合理的深度模式；平方误差单点回归趋向 conditional mean，可能平均这些模式。

(d) 可检查 prediction interval coverage、PIT/rank histogram，或比较预测方差与实际误差的相关性。failure 例：prior/posterior mismatch、posterior collapse、out-of-distribution 场景仍过度自信。

Rubric: ELBO 与 gap 5 分；reparameterization 与一般 Gaussian KL 3 分；训练/测试与多峰 3 分；calibration/failure 2 分。

(c) **Active view MDP。** [9 points]

(a) 代表性满分模型：S_k 包含已观测 view 集、当前重构状态与 uncertainty、未选 pose 及剩余预算 $B-k$。A_k 是未选 candidate pose（也可含 stop）。Transition 获取新图像并更新重构；

$$r_k = Q(R_{k+1}) - Q(R_k) - \lambda c(a_k),$$

其中 Q 是训练时可计算的 reconstruction quality，c 是 acquisition/运动 cost。状态必须包含足以使转移近似 Markov 的信息。

(b)

$$Q^*(s,a) = \mathbb{E}\left[r + \gamma \max_{a'} Q^*(s',a') \mid s,a\right],$$

样本更新为 $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q_{\text{target}}(s',a') - Q(s,a)]$。预算用尽或 stop 后为 terminal，bootstrap 项为零。

(c) 按 scene 划分 train/validation/test，不能让同一场景相邻 view 横跨划分；所有策略使用相同初始 view、candidate set 与最大 view/cost budget，报告 quality–cost curve 及固定预算质量，并与 random/coverage heuristic 比较。

Rubric: state/action/transition/reward 4 分；Bellman、update、terminal 3 分；无泄漏与公平预算 2 分。其他满足约束的 MDP 同样可得满分。

C. 人工智能中的优化方法：参考解答 [33 points]

(1) [4 points] 凸性和 smoothness 分别给出

$$f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle,$$

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2.$$

$\theta_0 = 1$。对 $k \ge 1$，

$$\frac{1-\theta_k}{\theta_k^2} = \frac{k(k+2)}{4} \le \frac{(k+1)^2}{4} = \frac{1}{\theta_{k-1}^2}.$$

(2) [5 points] 由 $x^{k+1} - y^k = -g^k/L$，

$$\begin{aligned} f(x^{k+1}) &\le f(y^k) + \langle g^k, x^{k+1} - y^k \rangle + \frac{L}{2}\|x^{k+1} - y^k\|^2 \\ &= f(y^k) - \frac{1}{L}\|g^k\|^2 + \frac{1}{2L}\|g^k\|^2, \end{aligned}$$

即得结论。

(3) [6 points] 将 gap 拆为

$$f(y^k) - f^\star = (1-\theta_k)[f(y^k) - f^\star] + \theta_k[f(y^k) - f^\star].$$

在 y^k 使用凸性，

$$f(y^k) \le f(x^k) + \langle g^k, y^k - x^k \rangle, \qquad f(y^k) - f^\star \le \langle g^k, y^k - x^\star \rangle.$$

因此

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

由 $y^k = (1-\theta_k)x^k + \theta_k v^k$ 直接整理，

$$(1-\theta_k)(y^k - x^k) + \theta_k(y^k - x^\star) = \theta_k(v^k - x^\star),$$

即得所需式。

(4) [7 points] 第 (2)–(3) 问先给出

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

与题给距离恒等式相加时，$+\theta_k\langle g^k, v^k - x^\star\rangle$ 与其负项消去，$-\|g^k\|^2/(2L)$ 与其正项消去，得到

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

除以 θ_k^2 即为题设 potential recurrence。

(5) [7 points] $k = 0$ 时 $\theta_0 = 1$，右侧的 gap 项为零，故

$$\frac{\Delta_1}{\theta_0^2} + \frac{L}{2}\|v^1 - x^\star\|^2 \le \frac{L}{2}\|v^0 - x^\star\|^2.$$

对 $k \ge 1$，第 (1) 问保证

$$\frac{1-\theta_k}{\theta_k^2}\Delta_k \le \frac{\Delta_k}{\theta_{k-1}^2}.$$

逐步 telescope 得

$$\frac{\Delta_K}{\theta_{K-1}^2} + \frac{L}{2}\|v^K - x^\star\|^2 \le \frac{L}{2}\|v^0 - x^\star\|^2.$$

丢弃非负距离项，使用 $v^0 = x^0$ 与 $\theta_{K-1} = 2/(K+1)$：

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

(6) [4 points] 若 $R_0 = \|x^0 - x^\star\|$，充分取

$$K \ge \max\left\{1, \left\lceil \sqrt{\frac{2LR_0^2}{\varepsilon}} - 1 \right\rceil\right\}.$$

其确定性复杂度为 $O(\varepsilon^{-1/2})$，而普通凸光滑 GD 为 $O(\varepsilon^{-1})$。若使用有持续方差的随机梯度，descent 与距离展开会留下额外二阶矩项，题中精确的 squared-gradient cancellation 不再闭合；无偏性本身不能删除 variance term。

Rubric: (1) 两定义 2 分、θ 代数 2 分；(2) descent lemma 与常数 5 分；(3) 两次凸性 3 分、向量合并 3 分；(4) 初始 gap 式 2 分、两项消去 3 分、缩放 recurrence 2 分；(5) base case 2 分、telescope 3 分、最终常数 2 分；(6) K、比较、随机噪声警告共 4 分。

D. 自然语言处理：参考解答 [33 points]

1. Latent Dirichlet allocation [7 分]

完整 joint 为

$$p(W, Z, \Theta, \Phi \mid \alpha, \eta) = \left[\prod_{k=1}^{K} \mathrm{Dir}(\phi_k; \eta)\right] \prod_{d=1}^{D} \left[\mathrm{Dir}(\theta_d; \alpha) \prod_{n=1}^{N_d} \theta_{d,z_{dn}} \phi_{z_{dn},w_{dn}}\right].$$

这同时写清每个 latent/observed variable 的生成来源。

给定 θ_d 后，document 内 topic assignments 独立同分布；再给定 z_{dn} 与 Φ 后，words 条件独立。对 token positions 的置换不改变 likelihood，因此 LDA 是 bag-of-words：它不能表示 word order、negation scope、syntax 或相邻实体关系，任一具体例子均可。

评分要点 priors、documents、tokens 的完整 product 4 分；exchangeability 2 分；一个确切遗漏现象 1 分。

2. Collapsed topic update [8 分]

Dirichlet–categorical conjugacy 给出 posterior predictive：在 document d 中下一 topic 为 k 的概率与 $n_{dk}^{-i} + \alpha$ 成正比；在 topic k 中生成 word w_i 的概率为

$$\frac{n_{k,w_i}^{-i} + \eta}{n_{k\cdot}^{-i} + V\eta}.$$

其乘积给出

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

更严格地，也可从 collapsed joint 中只保留将 z_i 放入 topic k 时变化的 Gamma-function ratios，并用 $\Gamma(a+1)/\Gamma(a) = a$ 得到同式。

第一因子偏好符合该 document 已有 topic mixture 的 assignments；第二因子偏好该 topic 已经解释得好的 word。$-i$ 表示先移除当前 token，否则该 token 会给自己的候选 topic 额外加一，造成 self-counting，所得不再是正确 full conditional。

评分要点 document predictive 2 分；topic-word predictive 2 分；乘积与 normalization 说明 2 分；两因素解释及 $-i$ 2 分。

3. Adding document labels [7 分]

对 documents 求和，label log-likelihood 为

$$\ell(\beta) = \sum_{d} \Big[y_d \log \sigma(\beta^\top \bar{z}_d) + (1 - y_d) \log(1 - \sigma(\beta^\top \bar{z}_d))\Big],$$

且

$$\boxed{\nabla_\beta \ell = \sum_{d} \big(y_d - \sigma(\beta^\top \bar{z}_d)\big)\bar{z}_d.}$$

在 posterior 中，改变某个 z_{dn} 也改变 $\bar{z}_d$ 及 label likelihood；于是除了解释 words，topics 还会被推动去区分 labels。若 label term 权重过大，也可能得到 predictive 但语义不连贯的 topics。

高度不平衡时，模型可能几乎总预测 majority，topics 也主要服务该类。可做 class-weighted log-likelihood、balanced minibatches、hierarchical prior 或在验证集上调阈值，并同时报告 macro-F1/ minority recall。

评分要点 likelihood 1 分；正确 gradient 2 分；posterior coupling 2 分；failure 与对应 remedy 2 分。

4. Topic discovery under temporal drift [11 分]

Representative full-credit design：每月 t 设 topic logits $\psi_{k,t} \in \mathbb{R}^V$，$\phi_{k,t} = \mathrm{softmax}(\psi_{k,t})$，并用随机游走 prior

$$\psi_{k,t} \sim \mathcal{N}(\psi_{k,t-1}, \sigma^2 I).$$

Document 仍有 θ_d, z_{dn}；少量 label 使用 $p(y_d \mid \bar{z}_d, \beta_t)$，且可对 β_t 加 temporal smoothness。Joint 由这些 priors、LDA token factors 和 label factors 相乘；等价的 optimization 版本可最大化

$$\mathrm{ELBO} - \frac{1}{2\sigma^2} \sum_{k,t} \|\psi_{k,t} - \psi_{k,t-1}\|^2,$$

即 smoothness 项在最大化目标中应作为负 penalty。

可用 amortized variational inference 或上一月 posterior warm-start 的 online variational EM；mini-batch 更新 local $q(\theta, z)$ 与 global ψ, β。Vocabulary 用 stable subwords 或保留 UNK bucket；新词追加 embedding/logit columns，以 prior 或近邻词初始化，再做增量索引和有限步更新，避免每月从零训练。

Topic drift 可测 matched topics 间 Jensen–Shannon divergence、top-word overlap 和 assignment transition；label shift 测 $p_t(y)$ 的变化并在固定 conditional performance 下比较；encoder degradation 则会同时降低 reconstruction/perplexity、held-out retrieval consistency 和多个 labels 的性能。应使用固定 anchor set 以免三者混淆。

命名可用 top words、代表性 tickets 和与受控 taxonomy label 的相似度生成候选，再由人工 reviewer 确认。审计记录跨月 name stability、代表样本、coherence 与敏感属性泄漏；不可仅让 LLM 任意命名而不提供 evidence。

评分要点 明确 temporal variables 与 joint/objective 4 分；可行 inference 和增量词表 3 分；能区分三类 drift 的诊断 2 分；有依据且可审核的命名 2 分。其他严谨设计同等给分。