Artificial Intelligence PhD Qualifying Examination

模拟考试 06：参考解答与评分要点

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]

Part I. $\boxed{\text{(i)a, (ii)a, (iii)a, (iv)a, (v)a, (vi)a}}$。

Part II.

(a) 固定 sample 和 coordinate j。排序该坐标后，threshold 至多产生 $m+1$ 个 cuts；每个 cut 有两种 polarity，故每个 coordinate 至多 $2(m+1)$ 个 label vectors。对 d 个 coordinates 取 union 得 $2d(m+1)$。Duplicate values、不同 coordinates 产生同一 labeling，或 extreme cuts 与相反 polarity 重复，都可能使实际数量更小。

Rubric: 固定坐标/排序 2 分；$m+1$ 计数 2 分；polarity/coordinate 2 分；非等号原因 2 分。

(b) 展开 covariance：

$$\begin{aligned}\operatorname{Var}\left(\frac{1}{M}\sum_j T_j\right) &= \frac{1}{M^2}\left(M\sigma^2 + M(M-1)\rho\sigma^2\right)\\ &= \sigma^2\left(\rho + \frac{1-\rho}{M}\right).\end{aligned}$$

$\rho = 0$ 时 variance 为 σ^2/M；$\rho = 1$ 时仍为 σ^2，增加完全相同的 estimators 没有收益。这说明 averaging 的收益依赖 component diversity，而非任何特定模型家族。

Rubric: diagonal terms 1 分；off-diagonal terms 2 分；化简 1 分；两个 cases 2 分。

(c) 令 $\widehat{h} \in \arg\min_h \widehat{L}(h)$。把题给 variational identity 用于 $a_h = \widehat{L}(h)$，并将最优的 q_β 与 point mass $\delta_{\widehat{h}}$ 比较：

$$\widehat{L}(q_\beta) - \frac{H(q_\beta)}{\beta} \le \widehat{L}(\widehat{h}) - \frac{H(\delta_{\widehat{h}})}{\beta} = \widehat{L}(\widehat{h}).$$

又因 $0 \le H(q_\beta) \le \log N$，

$$\boxed{\widehat{L}(q_\beta) \le \min_h \widehat{L}(h) + \frac{\log N}{\beta}.}$$

在 uniform-deviation event 上，对任意 distribution q，

$$|L(q) - \widehat{L}(q)| \le \sum_h q(h)|L(h) - \widehat{L}(h)| \le \alpha.$$

若 $h^\star \in \arg\min_h L(h)$，则

$$\begin{aligned}L(q_\beta) &\le \widehat{L}(q_\beta) + \alpha\\ &\le \widehat{L}(h^\star) + \frac{\log N}{\beta} + \alpha\end{aligned}$$

$$\leq L(h^\star) + 2\alpha + \frac{\log N}{\beta}.$$

这给出所需 population excess bound。

取

$$\alpha = \frac{\varepsilon}{4}, \qquad \beta \geq \frac{2\log N}{\varepsilon}.$$

此时 $2\alpha + \log N/\beta \leq \varepsilon$。而

$$2Ne^{-2m\alpha^2} = 2Ne^{-m\varepsilon^2/8} \leq \delta$$

只要

$$\boxed{m \geq \frac{8}{\varepsilon^2}\log\frac{2N}{\delta}.}$$

Rubric: variational comparison 1 分；entropy bound 与 empirical soft-min 2 分；uniform event 上两次 comparison 2 分；α, β, m 的充分选择及常数 2 分。

(d) Sauer 给 $\tau_{\mathcal{H}}(2m) \leq (e(2m)/v)^v$。在 uniform deviation α 事件上，ERM excess 至多 2α。取 $\alpha = \varepsilon/2$，失败概率至多

$$4\left(\frac{2em}{v}\right)^v e^{-m\varepsilon^2/32}.$$

因此充分的 implicit condition 是

$$\boxed{\frac{m\varepsilon^2}{32} \geq v\log\frac{2em}{v} + \log\frac{4}{\delta}.}$$

Rubric: Sauer substitution 2 分；ERM factor 2 1 分；常数正确条件 2 分。

(e) 事件 E 为 sample 未进入 disagreement set，故 $\Pr(E) = (1-q)^m$。条件于 E，两种 targets 产生完全相同的 observed sample distribution，learner 无法分辨；在 disagreement set 上，它对两种 targets 的 errors 逐点相加为 1。均匀 target prior 下 conditional expected error 至少 $q/2$。所以 unconditional expected error 至少

$$\boxed{\frac{q}{2}(1-q)^m.}$$

Rubric: no-hit probability 1 分；indistinguishability 1 分；conditional error 1 分；unconditional lower bound 1 分。

B. 深度学习与强化学习：参考解答 [33 points]

(a) **Branched residual backprop。** [9 points]

(a) 令 $g = \nabla_y J$。则

$$\nabla_x J = [I + J_F(x) + J_G(x)]^\top g.$$

若 $J_{F,\theta} = \partial F_\theta(x)/\partial\theta$、$J_{G,\phi} = \partial G_\phi(x)/\partial\phi$，则

$$\nabla_\theta J = J_{F,\theta}^\top g, \qquad \nabla_\phi J = J_{G,\phi}^\top g.$$

同一节点 x 影响三条后继边；总微分的线性性要求把三条 vector–Jacobian product 相加，而不是任选一条路径。

(b) 单块含 $I^\top g$ 的直接通路；多块反传是 $\prod_\ell (I + J_{F_\ell} + J_{G_\ell})^\top$，相较纯 Jacobian 连乘提供接近恒等的路径。但若 branch Jacobian 很大、特征尺度失控或各因子产生不利谱方向，仍可爆炸/消失；identity 不是无条件保证。

(c) Reverse mode 保存前向中间量，并对标量 loss 只传播一个 adjoint；计算全部参数梯度通常与一次前向同阶（常数倍）。逐参数显式形成完整 Jacobian 会重复公共链并产生大矩阵，成本随参数数额外放大。

Rubric: 三类梯度和汇合解释 4 分；identity 与 caveat 3 分；reverse-mode 复用/复杂度 2 分。

(b) **β-VAE。** [15 points]

(a)

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)} \log p_\theta(x \mid z) - \mathrm{KL}(q_\phi(z \mid x) \| p(z)) = \mathcal{L}_{\mathrm{ELBO}}.$$

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

(b) 若网络输出 $a = \log\sigma^2$，则 $\sigma = \exp(a/2)$，并采样 $z = \mu + \exp(a/2) \odot \epsilon$，$\epsilon \sim \mathcal{N}(0, I)$。KL 为

$$\frac{1}{2}\sum_j \left(\mu_j^2 + \exp(a_j) - a_j - 1\right) = \frac{1}{2}\sum_j (\mu_j^2 + \sigma_j^2 - \log\sigma_j^2 - 1).$$

(c) 当 $\beta \geq 1$ 时，因 KL 非负，$\mathcal{L}_\beta \leq \mathcal{L}_{\mathrm{ELBO}} \leq \log p(x)$，故自动仍为 lower bound；$0 \leq \beta < 1$ 时一般不保证。增大 β 强迫 aggregate/posterior 更接近 prior、限制信息率，但可能牺牲 reconstruction；减小则相反。

(d) Collapse 指 $q_\phi(z \mid x) \approx p(z)$ 且 decoder 几乎不依赖 z。可报告每维 KL、active units，或 mutual-information estimate，并与 reconstruction 同时观察。

(e) 生成时采 $z \sim p(z)$，再采 $x \sim p_\theta(x \mid z)$。不能在没有观测 x 时从 $q_\phi(z \mid x)$ 采样；那是 inference posterior。

Rubric: ELBO/gap 4 分；参数化、reparameterization、KL 4 分；bound 范围与 tradeoff 3 分；collapse/diagnostic 2 分；正确/错误生成流程 2 分。

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

(a) Critic 可最小化 $\frac{1}{2}\big(V_w(S_t) - \mathrm{stopgrad}(G_t^{(n)})\big)^2$，等价 semi-gradient 更新 $w \leftarrow w + \alpha_w \widehat{A}_t^{(n)} \nabla_w V_w(S_t)$。若目标 $J = \mathbb{E}[\sum_t \gamma^t R_t]$，actor ascent 为

$$\eta \leftarrow \eta + \alpha_\eta \gamma^t \,\mathrm{stopgrad}(\widehat{A}_t^{(n)}) \nabla_\eta \log \pi_\eta(A_t \mid S_t).$$

也可把绝对折扣吸收到 return/occupancy 定义中，但不能重复乘 γ^t。

(b) 真实 critic 下

$$\begin{aligned}
&\mathbb{E}[R_t + \gamma V^\pi(S_{t+1}) - V^\pi(S_t) \mid S_t = s, A_t = a] \\
&\quad = r(s,a) + \gamma \sum_{s'} P(s' \mid s,a) V^\pi(s') - V^\pi(s) \\
&= Q^\pi(s,a) - V^\pi(s) = A^\pi(s,a).
\end{aligned}$$

(c) 小 n 更多 bootstrap，通常方差低但受 critic 误差产生较大 bias；大 n 更接近 Monte Carlo，bootstrap bias 降低但累积更多 reward/transition 噪声。终止前的完整 return 无 bootstrap。

Rubric: critic、actor、stopgrad、discount 4 分；条件期望链 3 分；bias–variance 2 分。

C. 人工智能中的优化方法：参考解答 [33 points]

(1) [3 points] 坐标上界中与 u 有关的部分为 $u\nabla_i f(x) + (L_i/2)u^2$，其最小点是 $u = -\nabla_i f(x)/L_i$。代回即得

$$f\left(x - \frac{\nabla_i f(x)}{L_i} e_i\right) \le f(x) - \frac{[\nabla_i f(x)]^2}{2L_i}.$$

(2) [4 points] 给定 x^k，第 (1) 问和 $p_i = L_i/S$ 给出

$$\begin{aligned}\mathbb{E}_k[f(x^{k+1})] &\le f(x^k) - \frac{1}{2}\sum_i \frac{p_i}{L_i}[\nabla_i f(x^k)]^2 \\ &= f(x^k) - \frac{1}{2S}\sum_i [\nabla_i f(x^k)]^2.\end{aligned}$$

(3) [4 points]

$$\mathbb{E}[\nabla_I f(x) e_I] = \sum_i p_i \nabla_i f(x) e_i,$$

一般不等于 $\nabla f(x)$；均匀抽样时它等于 $\nabla f(x)/d$。正确的无偏估计为

$$\widetilde{g}(x, I) = \frac{\nabla_I f(x)}{p_I} e_I, \qquad \mathbb{E}[\widetilde{g}(x, I)] = \nabla f(x).$$

由于本题 $p_I = L_I/S$，

$$\frac{\nabla_I f(x)}{L_I} e_I = \frac{1}{S}\,\widetilde{g}(x, I).$$

因此在正确归一化后，本更新确可写成使用共同步长 $1/S$ 的 SGD；不能无偏地这样解释的是未归一化向量 $\nabla_I f(x) e_I$。

(4) [5 points] 凸性给出

$$f(x^k) - f(x^\star) \le \langle \nabla f(x^k), x^k - x^\star \rangle \le \|\nabla f(x^k)\| \|x^k - x^\star\|.$$

第 (1) 问说明每条样本路径上 $f(x^{k+1}) \le f(x^k)$，故所有迭代位于初始 sublevel set，距离项不超过 R。于是

$$\|\nabla f(x^k)\|^2 \ge \frac{[f(x^k) - f^\star]^2}{R^2}.$$

代入第 (2) 问：

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

(5) [6 points] 对第 (4) 问取全期望，并用 $\mathbb{E}[Z^2] \ge (\mathbb{E}Z)^2$，得到题设标量递推。若某个 $\delta_{k+1} = 0$，此后结论显然。否则

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

其中最后一步用 $\delta_{k+1} \le \delta_k$。从 1 到 $K-1$ 求和并丢弃非负初始倒数项，得

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

(6) [6 points] 强凸性给出任意 y 下

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

对右侧关于 y 取下确界，得到

$$f(x^\star) \geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2,$$

即所需 PL inequality。代入第 (2) 问：

$$\mathbb{E}_k[f(x^{k+1}) - f^\star] \leq \left(1 - \frac{\mu}{S}\right)[f(x^k) - f^\star].$$

沿每个坐标方向比较 strong-convexity 下界与 coordinate-smoothness 上界可得 $\mu \leq L_i$，故 $0 \leq 1 - \mu/S < 1$。用 tower property 递归即得结论。

(7) [5 points] 均匀抽样时

$$\mathbb{E}_k[f(x^{k+1})] \leq f(x^k) - \frac{1}{2d}\sum_i \frac{[\nabla_i f(x^k)]^2}{L_i} \leq f(x^k) - \frac{1}{2dL_{\max}}\|\nabla f(x^k)\|^2.$$

由于 $S = \sum_i L_i \leq dL_{\max}$，$p_i \propto L_i$ 的常数 S 不劣于该 uniform worst-case constant。一次 coordinate update 只需一个偏导和一个坐标写入，通常远低于 full gradient；但比较时需区分 coordinate iterations 与 full passes。构造分布需 $O(d)$ 预处理，alias table 可实现 $O(1)$ 抽样，前缀和/树结构通常为 $O(\log d)$。

Rubric: (1) 最优坐标步与常数 3 分；(2) 条件期望与 S 4 分；(3) bias 1 分、归一化 2 分、算法解释 1 分；(4) convex/Cauchy 2 分、sublevel 1 分、gap-square decrease 2 分；(5) Jensen 2 分、倒数递推 3 分、零情形 1 分；(6) PL 3 分、tower recurrence 3 分；(7) uniform bound 2 分、常数比较 1 分、计算/抽样 2 分。

D. 自然语言处理：参考解答 [33 points]

1. Weighted co-occurrence embeddings [9 分]

X_{ij} 统计 word i 与 context j 在窗口内共同出现的次数，因此模型使用的是全局 co-occurrence signal。取 logarithm 可压缩极端长尾 count，并把乘性比例变成可由 inner product 拟合的加性量；b_i, c_j 吸收 word/context 的 marginal frequency。f 应降低极低 count 的高方差影响，同时避免极高频 pairs 完全支配 objective，例如随 x 增长后截断。

令

$$r_{ij} = \mathbf{u}_i^\top \mathbf{v}_j + b_i + c_j - \log X_{ij}.$$

只保留含 $\mathbf{u}_i, b_i$ 的项，得到

$$\boxed{\nabla_{\mathbf{u}_i} J = 2\sum_j f(X_{ij}) r_{ij} \mathbf{v}_j}, \qquad \boxed{\frac{\partial J}{\partial b_i} = 2\sum_j f(X_{ij}) r_{ij}}.$$

对任意 invertible A，令

$$\mathbf{u}_i' = A\mathbf{u}_i, \qquad \mathbf{v}_j' = A^{-\top}\mathbf{v}_j.$$

则 $(\mathbf{u}_i')^\top \mathbf{v}_j' = \mathbf{u}_i^\top \mathbf{v}_j$，故所有 residual 及 J 不变。这说明单个 coordinate 或方向没有唯一语义；需要固定 gauge、加入额外约束，或只解释在相同变换下不变的预测量。另有 $b_i' = b_i + a, c_j' = c_j - a$ 的 bias shift ambiguity。

评分要点 signal、log/bias/weight 3 分；两个 gradient 各 1.5 分；正确的 $A, A^{-\top}$ 变换 2 分，identifiability consequence 1 分。

2. Nucleus sampling [10 分]

将 token 按 $p_{(1)} \geq \cdots \geq p_{(V)}$ 排序，并以固定 token ID 等 deterministic rule 处理相同概率。令

$$m = \min\left\{ r : \sum_{i=1}^{r} p_{(i)} \geq \rho \right\}, \qquad Z_m = \sum_{i=1}^{m} p_{(i)}.$$

Nucleus distribution 为

$$\boxed{q_{(i)} = \begin{cases} p_{(i)}/Z_m, & i \leq m, \\ 0, & i > m. \end{cases}}$$

其非负且总和为 1。若希望 boundary tie 全部纳入，也可纳入所有与 $p_{(m)}$ 相等的 tokens 并重新计算 Z；关键是规则预先确定且最终重新归一化。

题给 cumulative masses 为 $0.40, 0.65, 0.80, 0.90, 1$，故 $m = 3$、$Z_m = 0.80$，

$$\boxed{q = (0.50, 0.3125, 0.1875, 0, 0)}.$$

Greedy 只取 argmax，deterministic 且容易产生单调重复；top-k 始终保留固定个数，而 nucleus 随 uncertainty 自适应 support size：分布尖锐时集合小，平坦时集合大。Nucleus 仍可能在平坦或 miscalibrated distribution 下保留大量低质量 token，也可能在过小 ρ 时剪掉关键 token/ EOS。针对性措施包括 temperature calibration 后再截断、设置合理的 min/max support、保留 EOS、repetition penalty，或在 safety-critical generation 中加入 constrained decoding；答出一组对应 failure/safeguard 即可。

评分要点 sorting/minimal prefix 2 分，renormalization/ties 2 分；数值集合与三项概率 3 分；两种比较 2 分，针对性 failure/safeguard 1 分。

3. Temporal reasoning with an MLN [14 分]

一组可行的 grounded formulas 为

$$\mathrm{HighTime}(e,t) \Rightarrow \mathrm{Time}(e,t) \quad [w_{\mathrm{local}} > 0],$$

$$\mathrm{Time}(e,t) \Rightarrow \mathrm{CandTime}(e,t) \quad [\text{hard}],$$

$$\bigvee_{t:\mathrm{CandTime}(e,t)} \mathrm{Time}(e,t) \quad [\text{hard}],$$

$$\neg\,\mathrm{Time}(e,t_1) \vee \neg\,\mathrm{Time}(e,t_2), \quad t_1 \neq t_2 \quad [\text{hard}],$$

以及

$$\mathrm{HighBefore}(e_1,e_2) \wedge \mathrm{Time}(e_1,t_1) \wedge \mathrm{Time}(e_2,t_2) \Rightarrow \mathrm{Less}(t_1,t_2) \quad [w_{\mathrm{order}} > 0].$$

Candidate feasibility 与 exactly-one 通常为 hard；local 与 ordering classifiers 可能出错，故应为 soft。Soft weights 可由 held-out conditional likelihood/pseudolikelihood 学习，或由 calibrated log-odds 初始化。

四个 assignments 的 log-score 与 mass 为

(T_A,T_B)	$w_o n_o + w_A n_A + w_B n_B$	$\widetilde{P}$
$(1,1)$	0.8	$e^{0.8}$
$(1,2)$	$1.2+0.8+0.4=2.4$	$e^{2.4}$
$(2,1)$	0	1
$(2,2)$	0.4	$e^{0.4}$

因此

$$\boxed{Z = 1 + e^{0.4} + e^{0.8} + e^{2.4}}, \qquad \boxed{(T_A,T_B)_{\mathrm{MAP}} = (1,2)},$$

且

$$\boxed{P(T_A=1,T_B=2) = \frac{e^{2.4}}{1+e^{0.4}+e^{0.8}+e^{2.4}} \approx 0.7003}.$$

Time/Before logits 应先在 temporally separated validation set 上做 temperature scaling 或 isotonic calibration，并用 Brier score、ECE/reliability diagram 检查，再转为 soft evidence；统一 0.5 threshold 并无一般保证。Correction 应作为 immutable 新 version 写入，保留 source、valid-time 与 supersedes pointer，将旧 evidence 标为 inactive 而非物理抹除，再重跑受影响 component 的 inference。

若每个事件有 K 个 bins，无额外结构时 exact enumeration 为 $O(K^E)$。当 event graph 是 chain/tree 且 factors 为 pairwise 时，可用 max-product 或 sum-product dynamic programming，chain time $O(EK^2)$；一般 loopy graph 可用 ILP/weighted Max-SAT、loopy belief propagation、beam search 或 variational approximation，并报告 optimality gap 或 convergence diagnostics。

评分要点 local/candidate/exactly-one/order formulas 4 分，hard/soft 1 分；四个 masses 2 分，Z/MAP/probability 3 分；calibration 1 分，versioned correction 1 分，complexity 与合理 alternative 2 分。