Artificial Intelligence PhD Qualifying Examination

模拟考试 09：参考解答与评分要点

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. (i)a, (ii)a, (iii)a, (iv)a, (v)a, (vi)a, (vii)a, (viii)a。

Part II.

(a) 令 $D = \mathrm{diag}(p_i(1-p_i))$。直接微分得

$$\nabla J(\theta) = \frac{1}{m}\widetilde{X}^\top(p-y), \qquad \nabla^2 J(\theta) = \frac{1}{m}\widetilde{X}^\top D\widetilde{X} \succeq 0.$$

故 convex。Finite θ 下 $0 < p_i < 1$，所以 $D \succ 0$；Hessian PD 当且仅当 $\widetilde{X}$ full column rank。若 design rank deficient，存在非零 v 使 $\widetilde{X}v = 0$，quadratic form 沿该方向为零。

Rubric: gradient 2 分；Hessian 2 分；PSD convexity 1 分；PD iff rank condition 2 分。

(b) (a) 非空 interval 在 ordered sample 上恰好选择一个 contiguous positive block。选择其 left/right endpoints 的 index 有 $n(n+1)/2$ 种；再加 empty pattern 得所述数量。不同 blocks 给不同 label vectors。

(b) 两点可实现四种 patterns。三个 ordered points 上 pattern $1, 0, 1$ 不是 contiguous block，不能实现。因此 VC dimension 是 2。

Rubric: block characterization 2 分；count 2 分；两点下界 2 分；三点 upper bound 2 分。

(c) ℓ_1–ℓ_∞ duality 给

$$\widehat{\mathfrak{R}}_{S_X}(\mathcal{G}_1(B)) = \frac{B}{m}\mathbb{E}_\sigma\left\|\sum_i \sigma_i x_i\right\|_\infty.$$

令 $v_j = (x_{1j}, \ldots, x_{mj})$，并取 $\mathcal{A} = \{\pm v_j : j \in [d]\}$。则 maximum norm 等于 $\max_{a\in\mathcal{A}}\langle\sigma, a\rangle$，而 $|\mathcal{A}| = 2d$、$\|v_j\| \le R\sqrt{m}$。Massart lemma 给

$$\widehat{\mathfrak{R}} \le \frac{B}{m}R\sqrt{m}\sqrt{2\log(2d)} = BR\sqrt{\frac{2\log(2d)}{m}}.$$

Rubric: duality 2 分；$2d$ vectors 2 分；radius 1 分；Massart/constants 2 分。

(d) 定义右侧为 $q(x)$。在最左侧 region，所有 ReLU terms 为零，所以取 a 使 $a + s_0 x$ 与 g 在该 region 相同。每经过 knot t_j，term $\Delta_j[x - t_j]_+$ 的 derivative 从 0 变为 Δ_j，故 q 的 slope jumps 与 g 完全一致。两函数在左侧相同、每段 slopes 相同，并都 continuous，归纳得处处相同。Constant 可作为 output bias，而 $x = [x]_+ - [-x]_+$，所以 linear term 也可由两个额外 ReLU units 表示；加上 K 个 shifted units 即构成 standard finite one-hidden-layer representation。若 g 在 knot 有 jump discontinuity，有限 ReLU composition 仍 continuous，不能由该公式表示。

Rubric: 左侧 matching 1 分；slope-jump calculation 3 分；global induction 1 分；network interpretation 1 分；continuity caveat 1 分。

B. 深度学习与强化学习：参考解答 [33 points]

(a) **Siamese graph。** [9 points]

(a) 令 $r = \|h\|_2$、$z = h/r$，微分为

$$dz = \frac{1}{r}\left(I - zz^\top\right) dh, \qquad J_{\text{norm}}(h) = \frac{I - zz^\top}{\|h\|_2}.$$

矩阵为对称的 tangent-space projection 除以输入 norm。

(b) 由 $\nabla_{z_1}\ell = -z_2$、$\nabla_{z_2}\ell = -z_1$，

$$\nabla_{h_1}\ell = -\frac{(I - z_1 z_1^\top)z_2}{\|h_1\|_2}, \qquad \nabla_{h_2}\ell = -\frac{(I - z_2 z_2^\top)z_1}{\|h_2\|_2}.$$

若 $J_i = \partial h_i/\partial\theta$，共享参数梯度为

$$\nabla_\theta \ell = J_1^\top \nabla_{h_1}\ell + J_2^\top \nabla_{h_2}\ell.$$

两个调用使用同一 θ，总微分必须累加两条路径。

(c) 对 z_2 stop-gradient 会删除第二项，只用 branch 1 更新 encoder；前向 loss 值不变但优化向量场改变。若所有输入都映到同一个单位向量，则任意正 pair 的 inner product 为 1，$\ell = -1$ 达到最小；所以仅此 symmetric positive loss 不能排除 collapse，需要 negatives、variance/ covariance constraint、predictor/asymmetry 等额外机制。

Rubric: normalization Jacobian 3 分；两输入梯度、共享参数求和 4 分；stopgrad 与 collapse 2 分。

(b) **Implicit score matching。** [15 points]

(a)

$$\frac{1}{2}\mathbb{E}_p \|s_\theta - s_p\|^2 = \frac{1}{2}\mathbb{E}_p \|s_\theta\|^2 - \mathbb{E}_p \langle s_\theta, s_p\rangle + \frac{1}{2}\mathbb{E}_p \|s_p\|^2 .$$

最后一项只依赖 p，与 θ 无关。

(b) 因 $s_{p,j} = \partial_j \log p = (\partial_j p)/p$，

$$\begin{aligned}\mathbb{E}_p[s_{\theta,j} s_{p,j}] &= \int s_{\theta,j}(x)\, \partial_j p(x)\, dx \\ &= [p s_{\theta,j}]_{\text{boundary}} - \int p(x)\partial_j s_{\theta,j}(x)\, dx \\ &= -\mathbb{E}_p[\partial_j s_{\theta,j}(X)].\end{aligned}$$

对 j 求和得到题设等式。边界项消失及微积分换序是假设，不可省略。

(c) 去掉与 θ 无关的常数，可最小化

$$\mathcal{J}_{\text{ISM}}(\theta) = \mathbb{E}_{X\sim p}\left[\frac{1}{2}\|s_\theta(X)\|_2^2 + \nabla\cdot s_\theta(X)\right].$$

它只需样本、network score 和其 divergence。

(d) 标准 Gaussian 的真实 score 为 $-x$，且 $\nabla\cdot(-x) = -d$。因此

$$\mathcal{J}_{\mathrm{ISM}} = \frac{1}{2}\mathbb{E}\,\|X\|^2 - d = \frac{d}{2} - d = -\frac{d}{2}.$$

目标可为负；它与非负 Fisher divergence 相差常数。

(e) 精确 divergence 是 Jacobian trace，朴素计算在高维昂贵；可使用 trace estimator，但会引入额外随机方差。

Rubric: 展开与常数 4 分；逐坐标分部积分及假设 5 分；最终 objective 3 分；Gaussian sanity 2 分；计算困难 1 分。

(c) **Early-exit MDP。** [9 points]

(a) 代表性设计：$S_\ell = (\ell, h_\ell, p_\ell, c_{\mathrm{used}})$，其中 p_ℓ 是当前 predictive distribution；动作是 stop/continue；continue 确定或随机地执行下一层，stop 进入 terminal；可把每层计算记为负 reward $-\lambda c_\ell$，终止时加 prediction utility（训练时如正确性或负 loss）。若 confidence 不足以 Markov，state 应包含 representation/history summary。

(b) 写 stop utility 为 $U_\ell(s)$，continue cost 为 c_ℓ，则

$$V^*(s_\ell) = \max\left\{U_\ell(s_\ell), -\lambda c_\ell + \gamma\mathbb{E}[V^*(S_{\ell+1}) \mid s_\ell]\right\},$$

且第 L 层必须 stop。

(c) 若 return 定义为 $G_t = \sum_{k=t}^{T-1}\gamma^{k-t}R_k$，对应 $J = \mathbb{E}\sum_t \gamma^t R_t$ 的 estimator 为

$$\widehat{g} = \sum_t \gamma^t \nabla_\eta \log \pi_\eta(A_t \mid S_t)\,\mathrm{stopgrad}(G_t - b(S_t)).$$

b 不依赖当前动作，故不改变期望；若使用 absolute-discount return，则不再额外乘 γ^t。

(d) 在相同数据与硬 compute/latency budget 下比较 accuracy，或报告完整 accuracy–compute Pareto curve；不能只比未匹配计算量的 accuracy。

Rubric: MDP 四要素 3 分；stop/continue Bellman 2 分；score、baseline、discount 3 分；公平评价 1 分。其他自洽设计同样可得满分。

C. 人工智能中的优化方法：参考解答 [33 points]

(1) [4 points] $g \in \partial F(x)$ 指

$$F(y) \geq F(x) + \langle g, y - x \rangle, \qquad y \in C.$$

若 $p = P_C(u)$，则对任意 $z \in C$，线段 $p + t(z - p)$ 可行。投影目标在 $t = 0$ 处的右导数非负，故

$$0 \leq \langle p - u, z - p \rangle,$$

即得题设变分不等式。分别对 $p = P_C(u)$、$q = P_C(v)$ 取测试点 q, p 并相加，得到

$$\|p - q\|^2 \leq \langle p - q, u - v \rangle \leq \|p - q\| \|u - v\|,$$

所以 $\|P_C(u) - P_C(v)\| \leq \|u - v\|$。

(2) [6 points] x^k 是 $\mathcal{F}_k$-可测，且 $x^\star \in C$。由投影非扩张性，

$$\begin{aligned} \|x^{k+1} - x^\star\|^2 &\leq \|x^k - \eta_k H_k - x^\star\|^2 \\ &= \|x^k - x^\star\|^2 - 2\eta_k \langle H_k, x^k - x^\star \rangle + \eta_k^2 \|H_k\|^2. \end{aligned}$$

给定 $\mathcal{F}_k$ 取条件期望，并使用 $\mathbb{E}[H_k \mid \mathcal{F}_k] = g^k$：

$$\mathbb{E}_k \langle H_k, x^k - x^\star \rangle = \langle g^k, x^k - x^\star \rangle \geq F(x^k) - F(x^\star),$$

最后一步是 subgradient inequality。再用条件二阶矩上界即得结论。

(3) [6 points] 对第 (2) 问取全期望、移项并求和：

$$\begin{aligned} 2 \sum_{k=0}^{K-1} \eta_k \mathbb{E}[F(x^k) - F(x^\star)] &\leq R^2 - \mathbb{E}\|x^K - x^\star\|^2 + G^2 \sum_{k=0}^{K-1} \eta_k^2 \\ &\leq R^2 + G^2 \sum_{k=0}^{K-1} \eta_k^2. \end{aligned}$$

被丢弃的是非负项 $\mathbb{E}\|x^K - x^\star\|^2$。对每条样本路径使用 convexity，

$$F(\widehat{x}^K) \leq \frac{1}{S_K} \sum_{k=0}^{K-1} \eta_k F(x^k).$$

再取期望并除以 $2S_K$，即得题设 bound。

(4) [5 points] 常步长 η 时，上界为

$$B(\eta) = \frac{R^2}{2K\eta} + \frac{\eta G^2}{2}.$$

$B'(\eta) = 0$ 给出 $\eta = R/(G\sqrt{K})$，且此时两项均为 $RG/(2\sqrt{K})$。因此

$$\mathbb{E}[F(\widehat{x}^K) - F(x^\star)] \leq \frac{RG}{\sqrt{K}}.$$

充分取

$$K \geq \left\lceil \frac{R^2 G^2}{\varepsilon^2} \right\rceil.$$

(5) [4 points] 记 $H_K = \sum_{j=1}^{K} 1/j$。题给步长满足

$$S_K = \frac{R}{G}\sum_{j=1}^{K} j^{-1/2} \geq \frac{2R}{G}(\sqrt{K+1}-1),$$

$$G^2 \sum_{k=0}^{K-1} \eta_k^2 = R^2 H_K \leq R^2(1+\log K).$$

代入第 (3) 问得到显式的 horizon-free 界

$$\mathbb{E}[F(\widehat{x}^K) - F(x^\star)] \leq \frac{RG(2+\log K)}{4(\sqrt{K+1}-1)}.$$

它仍趋于零，但相对预先知道 K 的最优常步长多出对数因子。

(6) [4 points] $\mathcal{F}_k$ 表示调用当前 oracle 之前的信息。若把 H_k 也放入条件 sigma-field，则 H_k 已成为可测量，不能再使用 $\mathbb{E}[H_k \mid \mathcal{F}_k] = g^k$ 消去当前噪声。令 $\zeta_k = H_k - g^k$。条件均值为零，所以

$$\begin{aligned}\mathbb{E}_k\|H_k\|^2 &= \|g^k\|^2 + 2\langle g^k, \mathbb{E}_k\zeta_k\rangle + \mathbb{E}_k\|\zeta_k\|^2 \\ &= \|g^k\|^2 + \mathbb{E}_k\|H_k - g^k\|^2.\end{aligned}$$

左侧 second moment 同时包含信号平方和 variance；只给 variance 上界时，还必须另外控制 $\|g^k\|^2$ 才能替代本题假设。

(7) [4 points] 求和递推只控制 $\sum_k \eta_k[F(x^k) - F^\star]$，而非单独的末项；一般 nonsmooth stochastic subgradient 迭代也不保证函数值单调，因此需用 convexity 将加权平均转成可控输出。光滑凸确定性 GD 可利用 descent lemma 达到 $O(1/K)$，这里缺少 gradient Lipschitzness。即使目标 strongly convex，固定步长和持续 oracle noise 仍会造成稳态邻域；非光滑尖点还可导致跨越最优点的振荡。因此 strong convexity 本身既不能消除 noise floor，也不能给出精确的固定步长 linear convergence。

Rubric: (1) subgradient 1 分、投影 VI 2 分、非扩张 1 分；(2) 投影平方展开 2 分、条件无偏 2 分、subgradient/二阶矩 2 分；(3) tower/telescope 3 分、丢弃末项 1 分、weighted Jensen 2 分；(4) 优化步长 3 分、精度复杂度 2 分；(5) 两个和式 2 分、显式常数 2 分；(6) filtration 解释 2 分、bias–variance 恒等式与区别 2 分；(7) averaging、smooth comparison、strong-convexity/noise caveat 共 4 分。

D. 自然语言处理：参考解答 [33 points]

1. Biased walks for node representations [8 分]

条件于上一节点与当前节点 (t, v)，

$$\boxed{P(X_{j+1} = x \mid X_{j-1} = t, X_j = v) = \frac{w_{vx}\alpha_{pq}(t, x)}{\sum_{z \in N(v)} w_{vz}\alpha_{pq}(t, z)}}.$$

p 控制立即返回：p 大使 $1/p$ 小，抑制 backtracking；p 小则鼓励回到 t。

$q < 1$ 放大 distance-two moves，倾向向外的 DFS-like exploration，较容易捕捉不同社区中的 structural equivalence；$q > 1$ 抑制外移，更接近局部 BFS，偏向 homophily/community similarity。这是由 graph 与 weights 共同决定的定性倾向，不是严格保证。

采样后的 walks 被当作 token sequences，以 Skip-gram（常配 negative sampling）预测 window 内的 context nodes。Degree、edge weights 和 p, q 会过采样特定节点或 motif，因此 learned similarity 继承 sampling bias，不等同于对所有 nodes/relations 均匀优化。

评分要点 normalization 2 分、p 1 分；q/DFS/BFS 2 分、两类 representation bias 1 分；objective 与一个具体 sampling bias 2 分。

2. Inside inference in a probabilistic grammar [8 分]

对 nonterminal A 与 span $x_{i:j}$，令 $\beta_A(i, j)$ 为 A 生成该 span 的总概率。Base case 为

$$\beta_A(i, i) = P(A \to x_i).$$

对 $i < j$，CNF recurrence 为

$$\boxed{\beta_A(i, j) = \sum_{A \to BC} P(A \to BC) \sum_{k=i}^{j-1} \beta_B(i, k)\beta_C(k + 1, j).}$$

Sentence probability 为 $P(x_{1:n}) = \beta_S(1, n)$。两层求和恰好 marginalize root rule 与 split point，而不是选择一个 parse。

对 ab 有两条 derivations：

$$P(S \to AB, A \to a, B \to b) = 0.6(0.8)(0.9) = 0.432,$$

$$P(S \to BA, B \to a, A \to b) = 0.4(0.1)(0.2) = 0.008.$$

故

$$\boxed{P(ab) = 0.44}, \qquad \boxed{P(S \to AB \mid ab) = \frac{0.432}{0.44} = \frac{54}{55} \approx 0.9818}.$$

若 binary rules 数量为 $|R_2|$，直接 inside time 为 $O(n^3|R_2|)$，chart memory 为 $O(n^2|\mathcal{N}|)$。Viterbi parsing 在 log domain 将对 rules/splits 的 sum-product 改成 max-sum，并为每个 chart cell 保存 rule 与 split backpointer，最后从 root 回溯。

评分要点 定义/base 1.5 分、完整 recurrence/sentence probability 2.5 分；两项数值 2 分；time/memory 1 分、Viterbi/backpointer 1 分。

3. A most-reliable relation path [7 分]

在 log domain 中令 $D_0(s) = 0$ 且 $D_0(v) = -\infty$（$v \neq s$）。第 h 个 relation 为 r_h 时，

$$\boxed{D_h(v) = \max_{(u,r_h,v)\in E_h} \{D_{h-1}(u) + \log p(u, r_h, v)\}.}$$

保存 argmax predecessor 后，从每个 reachable final entity 可回溯 best path。Dense arrays 每层初始化 $|V|$ 个新 states 后扫描 E_h，总时间

$$O\left(H|V| + \sum_{h=1}^{H} |E_h|\right),$$

rolling score memory 为 $O(|V|)$，保存所有 backpointers 为 $O(H|V|)$。Sparse maps/ timestamped arrays 可使计算部分近 $O(\sum_h |E_h|)$，但最终输出所有 entity scores 仍需 $O(|V|)$。

Marginally calibrated edges 通常不条件独立：多条 edge 可来自同一错误 extractor/source，直接相乘会 double count evidence，且 calibration 未必转移到 conjunction/path level。可在 held-out paths 上训练含 length 与 shared-provenance features 的 path calibrator，或用 factor graph/learned reranker 建模 correlation，再做 path-level calibration。

评分要点 initialization/recurrence 3 分；backtrack、time、memory 2 分；dependence 问题与对应 remedy 2 分。

4. Evidence-aware knowledge-base question answering [10 分]

一种 full-credit design：semantic parser 产生 seed entities、answer type、relation constraints 与 relation-pattern distribution；在 KG 上用 bounded beam/上一题 DP 产生 candidate paths，并为每条 edge/entity 检索原始 provenance passages。Joint scorer 结合 parser log-prob、edge confidence、type compatibility、textual entailment 与 length penalty。Aggregator 合并到达同一 answer entity 的 paths，输出 answer、explicit path 与 sentence citations；仅有 graph edge 而无 source support 时应降低置信度。

有 gold paths 时可最大化其 conditional likelihood；只有 gold answer 时，将到达 gold answer 的 paths 作为 latent positives，优化 marginal log-likelihood 或 listwise ranking。Hard negatives 包括到达 type-compatible wrong answer、只含一条高分错误 edge、交换 relation order 或依赖 spurious shortcut 的 paths。

在 held-out set 上联合校准 answer posterior、best-path margin 与 citation entailment；低于阈值，或多条高分 paths 给出冲突 answers 时 abstain。指标包括 answer EM/F1、path precision/recall、edge provenance coverage、citation precision/recall、ECE/Brier score 与 abstention risk–coverage curve。

Missing edge 可由 text retrieval 补充 candidate，但标记为 unverified；wrong high-confidence edge 用 multi-source contradiction/provenance 降权。Entity ambiguity 用 alias+type-aware linking；spurious path 用 counterfactual edge removal 与 hard negatives。任答两组一一对应的 failure/mitigation。

评分要点 完整 graph+text data flow 4 分；positive/hard-negative objective 2 分；calibrated abstention 与 reasoning/evidence metrics 2 分；两组 targeted mitigation 2 分。其他满足证据约束且可审计的方案同等给分。