Artificial Intelligence PhD Qualifying Examination

模拟考试 01：参考解答与评分要点

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. 答案为 $\boxed{\text{(i)a, (ii)a, (iii)a, (iv)b, (v)a, (vi)a}}$，每题 0.5 分。

Part II.

(a) (a) 由 $\sum_i a_i = 1$，

$$\mathbb{E}\widehat{L}_a = p, \qquad \mathrm{Var}(\widehat{L}_a) = \sum_i a_i^2 \,\mathrm{Var}(Z_i) = p(1-p)\sum_i a_i^2,$$

其中第二式使用 independence。

(b) Cauchy–Schwarz 给 $1 = (\sum_i a_i)^2 \le m \sum_i a_i^2$，故 $\sum_i a_i^2 \ge 1/m$；等号当且仅当 $a_1 = \cdots = a_m = 1/m$。最小 variance 是 $p(1-p)/m$。

Rubric: unbiasedness 1 分；variance 2 分；Cauchy–Schwarz 1 分；等号条件 1 分；minimum 1 分。

(b) 条件于 training sample 后，$h_1, \ldots, h_r$ 是 fixed，而 holdout examples 仍 i.i.d. 且与它们 independent。对 fixed j，Hoeffding 给

$$\Pr\{|\widehat{L}_V(h_j) - L(h_j)| > \alpha\} \le 2e^{-2m_v\alpha^2}.$$

取 $\alpha = \sqrt{\log(2r/\delta)/(2m_v)}$ 并作 union bound，所有候选同时偏差不超过 α 的概率至少 $1-\delta$。令 $j^\star \in \arg\min_j L(h_j)$，则

$$L(\widehat{h}) \le \widehat{L}_V(\widehat{h}) + \alpha \le \widehat{L}_V(h_{j^\star}) + \alpha \le L(h_{j^\star}) + 2\alpha.$$

先 conditioning 再对 training randomness 积分仍保持同一概率保证。若候选随 holdout labels 反复改变，则 fixed finite list 的 union bound 不再覆盖该 adaptive search。

Rubric: conditioning/independence 1 分；Hoeffding 1 分；union bound 与 α 2 分；oracle chain 2 分；adaptive caveat 1 分。

(c) (a) 两个问题分别为

$$\min_z \|z\|_0 \ \text{s.t.}\ Az = y, \qquad \min_z \|z\|_1 \ \text{s.t.}\ Az = y.$$

(b) 若两个 s-sparse vectors x, z 满足 $Ax = Az$，则 $h = x - z \in \ker(A)$，且 $\|h\|_0 \le \|x\|_0 + \|z\|_0 \le 2s$。题设迫使 $h = 0$，所以 $x = z$。这只证明 A 在 s-sparse 集合上 injective；basis pursuit 还要排除一个非稀疏但 ℓ_1 norm 更小的可行向量，通常需要 null-space property 等更强条件。

Rubric: 两个 formulation 2 分；kernel-difference 论证 2 分；injectivity 与 ℓ_1 成功的区别 1 分。

(d) 任取四个递增点 $x_1 < x_2 < x_3 < x_4$。任意 binary labeling 的正点至多形成两个连续 blocks，每个 block 可由一个小区间恰好覆盖，故四点可 shattered。

反之，在任意五个递增点上，labeling $1, 0, 1, 0, 1$ 有三个彼此分离的正 blocks；两个区间不可能恰好实现它。因此五点不能 shattered，

$$\boxed{\mathrm{VCdim}(\mathcal{I}_2) = 4.}$$

Rubric: 四点构造 2 分；任意 labeling 的至多两个 blocks 2 分；五点交替反例 2 分；结论 1 分。

(e) 由 strong-convexity midpoint inequality 对 $(1-t)w^\star + tw$ 使用 minimizer property，再令 $t \downarrow 0$，得到 quadratic growth：

$$F(w) \geq F(w^\star) + \frac{\lambda}{2} \|w - w^\star\|^2 .$$

代入 $w = \widehat{w}$ 并使用 objective gap 至多 ξ，得到

$$\frac{\lambda}{2} \|\widehat{w} - w^\star\|^2 \leq \xi, \qquad \|\widehat{w} - w^\star\| \leq \sqrt{\frac{2\xi}{\lambda}}.$$

因此对所有 z，

$$|\ell(\widehat{w}, z) - \ell(w^\star, z)| \leq \rho \sqrt{\frac{2\xi}{\lambda}}.$$

较小 optimization error ξ 或较大 curvature λ 都使参数与 loss 更稳定。

Rubric: quadratic growth 2 分；距离界 1 分；Lipschitz 转换 1 分；解释 1 分。

B. 深度学习与强化学习：参考解答 [33 points]

(a) **卷积网络。** [10 points]

(a) 一维输出公式为

$$H' = \left\lfloor \frac{H + 2p - d(k-1) - 1}{s} \right\rfloor + 1,$$

宽度同理。因此三层空间尺寸依次为 $H \times W$、$\lceil H/2 \rceil \times \lceil W/2 \rceil$、$\lceil H/2 \rceil \times \lceil W/2 \rceil$。参数总数为

$$C_1(9C_{\text{in}} + 1) + C_2(9C_1 + 1) + C_3(9C_2 + 1).$$

这里 dilation 不增加参数量。

(b) 有效核 $\kappa_\ell = d_\ell(k_\ell - 1) + 1$，递推 $j_\ell = s_\ell j_{\ell-1}$、$r_\ell = r_{\ell-1} + (\kappa_\ell - 1) j_{\ell-1}$。故

$$(r_1, j_1) = (3, 1), \qquad (r_2, j_2) = (5, 2), \qquad (r_3, j_3) = (13, 2).$$

(c) 采用互相关记号

$$(K * X)(u) = \sum_v K(v) X(u + v), \qquad (T_\Delta X)(u) = X(u - \Delta).$$

于是

$$[K * (T_\Delta X)](u) = \sum_v K(v) X(u + v - \Delta) = [T_\Delta (K * X)](u).$$

逐通道求和不改变结论。stride 2 只在子采样格点输出，因此仅对输入的偶数像素平移保持相应输出平移；奇数平移会改变采样相位。零填充还会在边界破坏严格等变性。

(d) 可用 1×1 卷积产生 C 个类别 logits，再双线性上采样到 $C \times H \times W$，以逐像素平均 cross-entropy 训练。stride 2 丢失细粒度边界；可加入同尺度 skip connection 或学习型上采样，但不能保证完全恢复已丢失的信息。

Rubric: (a) 尺寸、三层参数各有分；(b) 每层递推各有分；(c) 等变证明与 stride/boundary 各占分；(d) 输出、损失、局限齐全得满分。

(b) **GAN。** [14 points]

(a) 对每个 x 最大化 $p_{\text{data}}(x) \log D + p_g(x) \log(1 - D)$。一阶条件给出

$$D^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_g(x)}.$$

若只有前者为正则 $D^* = 1$，若只有后者为正则 $D^* = 0$；两者均为零处目标不依赖 D。

(b) 令 $m = (p_{\text{data}} + p_g)/2$。代入后

$$\begin{aligned} V(D^*) &= \int p_{\text{data}} \log \frac{p_{\text{data}}}{2m} + p_g \log \frac{p_g}{2m} \\ &= -2 \log 2 + \mathrm{KL}(p_{\text{data}} \| m) + \mathrm{KL}(p_g \| m) \\ &= -\log 4 + 2\, \mathrm{JS}(p_{\text{data}}, p_g), \end{aligned}$$

其中 $\mathrm{JS}(p, q) = \frac{1}{2} \mathrm{KL}(p \| m) + \frac{1}{2} \mathrm{KL}(q \| m)$。

(c) 因 $D = \sigma(a)$，

$$\frac{\partial}{\partial a}\log(1-D) = -D, \qquad \frac{\partial}{\partial a}[-\log D] = -(1-D).$$

当判别器很强、$D \simeq 0$ 时，前者接近零而后者接近 -1；non-saturating loss 因而向生成器提供更强的 logit 梯度。还需经判别器与生成器 Jacobian 链式传播。

(d) Mode collapse 指生成分布只覆盖真实分布少数模式，即样本多样性系统性不足。少量漂亮样本不能检查 distribution coverage。合理缓解包括 minibatch discrimination、unrolled updates、Wasserstein 型 critic，或显式多样性正则；需说明其针对覆盖或梯度稳定性的作用。

Rubric: D^* 及零密度边界 4 分；JS 展开与常数 4 分；两个导数和解释 3 分；collapse 定义、诊断、合理缓解 3 分。

(c) **Bellman contraction。** [9 points]

(a) 使用 $|\max_a u_a - \max_a v_a| \le \max_a |u_a - v_a|$，对任意 s，

$$\begin{aligned}|(T^*V)(s) - (T^*W)(s)| &\le \gamma \max_a \sum_{s'} P(s' \mid s,a)|V(s') - W(s')| \\ &\le \gamma \|V - W\|_\infty .\end{aligned}$$

对 s 取最大即得结论。

(b) 有界价值函数空间在 sup norm 下完备，Banach 不动点定理给出唯一不动点 V^*，且

$$\|V_k - V^*\|_\infty \le \gamma^k \|V_0 - V^*\|_\infty .$$

(c) 若初始误差为 $E_0 > 0$，充分条件是

$$k \ge \left\lceil \frac{\log(\varepsilon/E_0)}{\log \gamma} \right\rceil .$$

若 $E_0 \le \varepsilon$ 可取 $k = 0$。$\gamma \uparrow 1$ 时收缩变慢，所需迭代数增大。

Rubric: contraction 4 分；唯一性与误差界 3 分；迭代数、边界及解释 2 分。

C. 人工智能中的优化方法：参考解答 [33 points]

(1) [3 points] 可微凸函数满足

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \qquad x, y \in \mathbb{R}^d,$$

而 L-smoothness 指 $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$。对固定 u，目标 $z \mapsto \frac{1}{2}\|z - u\|^2$ 是强凸函数；C 非空、闭保证最小值取到，C 凸与严格凸性保证解唯一。

(2) [4 points] 令 $p = P_C(u)$。对任意 $z \in C$，$p + t(z - p) \in C$。函数 $\phi(t) = \frac{1}{2}\|p + t(z - p) - u\|^2$ 在 $t = 0$ 右侧取最小，故

$$0 \leq \phi'(0+) = \langle p - u, z - p \rangle,$$

即得变分不等式。

记 $p = P_C(u), q = P_C(v)$。分别在两个变分不等式中取 $z = q, p$，相加得

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

若 $p \neq q$ 可约去 $\|p - q\|$；$p = q$ 时结论显然。

(3) [4 points] 凸可微约束问题的一阶最优性条件为

$$\langle \nabla f(x^\star), z - x^\star \rangle \geq 0, \qquad z \in C.$$

将 $u = x^\star - \gamma \nabla f(x^\star)$、$p = x^\star$ 代入投影变分不等式，正好得到上式，故 $x^\star = P_C(x^\star - \gamma \nabla f(x^\star))$。反之，fixed point 通过变分不等式给出同一一阶条件；再用凸性，

$$f(z) \geq f(x^\star) + \langle \nabla f(x^\star), z - x^\star \rangle \geq f(x^\star),$$

所以该 fixed point 是全局约束最小点。

(4) [6 points] 由定义，$G_\gamma^C(x) = 0$ 当且仅当 $x = P_C(x - \gamma \nabla f(x))$，再由第 (3) 问等价于约束一阶最优性。当 $C = \mathbb{R}^d$ 时 P_C 是恒等映射，因而

$$G_\gamma^{\mathbb{R}^d}(x) = \gamma^{-1}[x - (x - \gamma \nabla f(x))] = \nabla f(x).$$

例如在 $C = [0, \infty)$ 上最小化 $f(x) = x$，唯一解为 $x^\star = 0$，但 $\nabla f(0) = 1 \neq 0$，而 $G_\gamma^C(0) = 0$。

(5) [6 points] 投影变分不等式给出

$$\langle x - \gamma \nabla f(x) - x^+, z - x^+ \rangle \leq 0,$$

从而

$$\langle \nabla f(x), x^+ - z \rangle \leq \frac{1}{\gamma} \langle x - x^+, x^+ - z \rangle.$$

由 smoothness 与 convexity，

$$\begin{aligned} f(x^+) - f(z) &\leq \langle \nabla f(x), x^+ - z \rangle + \frac{L}{2}\|x^+ - x\|^2 \\ &\leq \frac{1}{\gamma} \langle x - x^+, x^+ - z \rangle + \frac{L}{2}\|x^+ - x\|^2. \end{aligned}$$

代入题示三点恒等式即得所需结论。

(6) [5 points] 在第 (5) 问中取 $z = x^k$，$\gamma = 1/L$，得

$$f(x^{k+1}) - f(x^k) \le -\frac{L}{2}\|x^{k+1} - x^k\|^2 \le 0.$$

再取 $z = x^\star$ 并对 $k = 0, \dots, K-1$ 求和：

$$\sum_{k=0}^{K-1} [f(x^{k+1}) - f(x^\star)] \le \frac{L}{2}\|x^0 - x^\star\|^2.$$

函数值单调不增，左侧至少为 $K[f(x^K) - f(x^\star)]$，故得目标上界。

(7) [5 points] 令 $R_0 = \|x^0 - x^\star\|$。取

$$K \ge \left\lceil \frac{LR_0^2}{2\varepsilon} \right\rceil$$

足够，总时间上界为 $O(K(T_g + T_P))$。这里不能隐去 projection cost：某些复杂凸集上 T_P 可能比梯度还贵。边界最优点往往有 $\nabla f(x^\star) \neq 0$，因而应使用 $\|G_\gamma^C(x)\|$ 而非普通梯度范数作约束驻点残差。

Rubric: (1) 两定义及投影唯一性 3 分；(2) 变分不等式与非扩张各 2 分；(3) 两个方向各 2 分；(4) 驻点等价 3 分、无约束化简 1 分、边界反例 2 分；(5) 投影不等式、smooth/convex 合并、常数各 2 分；(6) 单调性 2 分、telescope 与 last-iterate 界 3 分；(7) 迭代数 2 分、总代价 1 分、正确约束残差解释 2 分。

D. 自然语言处理：参考解答 [33 points]

1. Interpolated language modeling [6 分]

对任一固定上下文 (u, v)，三种 maximum-likelihood 分布各自对 w 求和为 1。因此

$$\sum_{w} p_\lambda(w \mid u, v) = \lambda_3 + \lambda_2 + \lambda_1 = 1,$$

且每项非负，故它是合法分布。

题给数字下

$$p_\lambda(w \mid u, v) = 0.6 \cdot 0 + 0.3 \cdot \frac{4}{40} + 0.1 \cdot \frac{50}{1000} = \boxed{0.035}.$$

未见过的 trigram 仍从 lower-order statistics 获得非零概率。

可在 held-out corpus 上最小化平均 negative log-likelihood（或 perplexity）来选 λ；权重也可按 context count 分桶或由一个受 simplex 约束的 gating model 产生。若 $C(u, v) = 0$，应把对应高阶权重置零并将剩余权重重新归一化；若 $C(v) = 0$，继续 back off 到 unigram。Unigram 本身还应使用基本 smoothing，以免 $C(w) = 0$ 时仍给零概率。

评分要点 归一化与非负性 2 分；数值 0.035 及解释 2 分；held-out 目标与正确 back-off/renormalization 2 分。

2. A globally normalized sequence model [8 分]

定义

$$\alpha_1(j) = \exp s_1(y_0, j, x), \qquad \alpha_t(j) = \sum_{i=1}^{K} \alpha_{t-1}(i) \exp s_t(i, j, x).$$

则 $Z(x) = \sum_{j=1}^{K} \alpha_L(j)$。实际计算宜在 log domain 使用 log-sum-exp。

MAP decoding 把求和换成最大值：

$$\delta_1(j) = s_1(y_0, j, x), \qquad \delta_t(j) = \max_i \{\delta_{t-1}(i) + s_t(i, j, x)\}.$$

保存取得最大值的 backpointer，最后从 $\arg\max_j \delta_L(j)$ 回溯。时间为 $O(LK^2)$；仅求最优分数可用 $O(K)$ 工作内存，恢复路径需 $O(LK)$ backpointers。

局部归一化模型在每个前一标签处分别归一化，出边少的状态容易把概率质量“困住”，即使后续全局证据不支持也难纠正。CRF 的 $Z(x)$ 对完整 label sequence 统一归一化，不同路径的总分可直接竞争，因而消除了这一特定的局部归一化来源的 label bias；这不等于 CRF 不会有建模偏差。

评分要点 forward 与 Z 3 分；Viterbi、回溯和复杂度 3 分；全局/局部归一化的准确比较 2 分。

3. Failure modes of TransE [7 分]

令

$$a = \mathbf{h} + \mathbf{r} - \mathbf{t}, \qquad b = \mathbf{t} + \mathbf{r} - \mathbf{h}, qquad\|a\|, \|b\| \le \epsilon.$$

相加得 $a + b = 2\mathbf{r}$，故 $\|\mathbf{r}\| \le \epsilon$。又 $\mathbf{h} - \mathbf{t} = a - \mathbf{r}$，所以 $\|\mathbf{h} - \mathbf{t}\| \le \|a\| + \|\mathbf{r}\| \le 2\epsilon$。

对 one-to-many，令 $a_i = \mathbf{h} + \mathbf{r} - \mathbf{t}_i$。则

$$\|\mathbf{t}_1 - \mathbf{t}_2\| = \|a_2 - a_1\| \le \|a_1\| + \|a_2\| \le 2\epsilon,$$

因此不同 tails 被迫靠近。

一种可接受替代是 bilinear score $s_r(h,t) = \mathbf{h}^\top W_r \mathbf{t}$。对 symmetric relation 可约束 $W_r = W_r^\top$，于是 $s_r(h,t) = s_r(t,h)$，但高分不要求 $\mathbf{h} = \mathbf{t}$；同一 head 也可与多个不同 tail 取得高分。其他能明确打破单一平移限制的 full-credit model 也可。

评分要点 对称关系的两个界 3 分；one-to-many 界 2 分；合理替代模型及针对性解释 2 分。

4. Citation-grounded domain QA [12 分]

下面给出一个 representative full-credit design；满足同样约束的其他架构也可得满分。

离线端先按标题、段落和 overlap window 切分，并保留 document ID、版本、ACL 和句子 offset。为 chunk 建 BM25 倒排索引和 dense embedding ANN index。在线端把问题做轻量 query rewriting，同时执行 sparse/dense retrieval，融合后取较大的 candidate set，用 cross-encoder reranker 压到 top-k。Generator 只接收带稳定句子编号的 evidence，输出 atomic claims 与 citation IDs。一个 verifier 检查每个 claim 是否被引用句支持；若最大支持概率或 retrieval coverage 低于阈值，则 abstain。权限过滤必须在 retrieval 之前和之后各执行一次。

无人工 QA 时，可从标题、定义句、表格说明等自动产生 pseudo queries，以对应 chunk 为 positive；同文档相邻但不含答案的 chunk、BM25 high-rank nonpositive 以及旧模型误检作为 hard negatives。Retriever 可最小化 contrastive loss，reranker 用 pairwise logistic loss。Generator 可先做 continued pretraining，再以“从 chunk 生成问题/答案/引用”的 synthetic triples 做 SFT；verifier 用 entailment-style synthetic positive/negative claims 训练。伪标签需经过规则过滤，避免把生成错误循环放大。

ANN query 的典型代价近似为索引相关的 sublinear search 加 $O(kd)$ reranking input preparation；cross-encoder 和 generation 的主要代价随 L_c 增长，Transformer attention 若直接拼接 evidence 可达 $O(L_c^2 d)$。可采用小 candidate set 后的两级 reranking、限制 k/L_c、cache 高频 query/embedding、量化 ANN，以及按文档增量写入索引；后台周期性 compaction 保持检索质量而不阻塞 daily update。

评价分层进行：retriever 报 Recall@k/MRR 与 latency；reranker 报 pairwise accuracy；answer 报 EM/token-F1 或人工 correctness；citation 报 claim-level precision/recall；另报 abstention calibration、权限泄漏和更新新鲜度。这样能区分 retrieval miss（gold evidence 未进入 top-k）与 generation/grounding failure（证据已到但答案或引用错误），也能识别 stale-index 和 prompt-injection 等失败。

评分要点 数据流及 citation/abstention 4 分；无标注训练目标与 negatives 3 分；复杂度判断及两项可执行 latency/update 措施 3 分；分层指标和可区分失败 2 分。