Artificial Intelligence PhD Qualifying Examination

模拟考试 03：参考解答与评分要点

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) (a) 若 $K_j(x,x') = \langle \phi_j(x), \phi_j(x') \rangle$，则 $\phi(x) = (\sqrt{a}\,\phi_1(x), \sqrt{b}\,\phi_2(x))$ 给出 $aK_1 + bK_2$。在 tensor-product space 中，$\psi(x) = \phi_1(x) \otimes \phi_2(x)$ 满足

$$\langle \psi(x), \psi(x') \rangle = K_1(x,x') K_2(x,x'),$$

故 product 也 PSD。

(b) 展开 polynomial kernel 后可取

$$\boxed{\psi(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, \sqrt{2}x_1x_2, x_2^2)^\top.}$$

逐坐标内积给 $1 + 2x^\top x' + (x^\top x')^2$。

Rubric: direct-sum map 2 分；tensor map 3 分；展开 1 分；feature map 2 分。

(b) (a) 每个 $\widehat{p}_g$ 是 p_g 的 unbiased Bernoulli mean，且两组独立，所以

$$\mathbb{E}\widehat{L} = \pi_1 p_1 + \pi_2 p_2, \qquad \mathrm{Var}(\widehat{L}) = \frac{\pi_1^2 p_1(1-p_1)}{m_1} + \frac{\pi_2^2 p_2(1-p_2)}{m_2}.$$

(b) 记 $\sigma_g = \sqrt{p_g(1-p_g)}$。Lagrange multiplier 条件 $-\pi_g^2\sigma_g^2/m_g^2 + \nu = 0$ 给

$$\boxed{m_g = m\,\frac{\pi_g \sigma_g}{\pi_1\sigma_1 + \pi_2\sigma_2}}.$$

由题设 $p_g \in (0,1)$ 可知 $\sigma_g > 0$，所以这是 positive-real constraint 下的 interior minimizer。代回的 minimum 是 $(\pi_1\sigma_1 + \pi_2\sigma_2)^2/m$。

Rubric: unbiasedness 1 分；variance 1 分；Lagrange condition 1 分；allocation 1 分；interior/minimum 1 分。

(c) 该 class 至多有 $4d$ 个 parameter triples（即使不同 triples 可能代表同一函数也不影响 upper bound）。固定 true error 大于 ε 的 h，它在 m 个 i.i.d. examples 上仍 consistent 的概率至多

$$(1-\varepsilon)^m \le e^{-m\varepsilon}.$$

若 consistent learner 输出 bad hypothesis，至少有一个 bad class member 在样本上 consistent。Union bound 给失败概率至多 $4d\,e^{-m\varepsilon}$，故

$$\boxed{m \ge \frac{\log(4d) + \log(1/\delta)}{\varepsilon}}$$

充分。

Rubric: class-size upper bound 1 分；fixed bad survival 2 分；union event 1 分；样本量 2 分。

(d) 记 $u = w_S, v = w_{S'}$，并设被替换的 examples 为 z, z'。题给 lemma 分别用于 $F_S, F_{S'}$ 后相加：

$$\lambda \|u - v\|^2 \le \frac{1}{m}[\ell(v, z) - \ell(u, z) + \ell(u, z') - \ell(v, z')].$$

Lipschitzness 使右端至多 $(2\rho/m)\|u - v\|$。若 $u \neq v$，约去一次 norm 得

$$\|u - v\| \le \frac{2\rho}{\lambda m};$$

$u = v$ 时也成立。再用 Lipschitzness，任意 z_0 上

$$|\ell(u, z_0) - \ell(v, z_0)| \le \rho \|u - v\| \le \frac{2\rho^2}{\lambda m}.$$

Rubric: 两个 strong-convex inequalities 2 分；相加消去共同样本 2 分；移动界 2 分；loss stability 1 分。

(e) 在任意 m 点上，所有 binary hypotheses 总共至多产生 2^m 种 label vectors；另一方面 $|\mathcal{C}_d| \le 2d$，所以

$$\tau_{\mathcal{C}_d}(m) \le \min\{2^m, 2d\}.$$

若该 class shatter m 点，则 $2^m \le 2d$，即 $m \le \log_2(2d)$。取最大整数得到所述 VC upper bound。

Rubric: 2^m 上界 1 分；class-size 上界 1 分；shattering implication 与取整 2 分。

B. 深度学习与强化学习：参考解答 [33 points]

(a) **Image restoration**。 [11 points]

(a) 记 $m(Y) = \mathbb{E}[X \mid Y]$。条件于 Y 展开

$$\|X - f(Y)\|^2 = \|X - m(Y)\|^2 + \|m(Y) - f(Y)\|^2 + 2\langle X - m(Y), m(Y) - f(Y)\rangle .$$

最后一项取期望为零，因为 $\mathbb{E}[X - m(Y) \mid Y] = 0$，而另一因子是 Y-可测的。因此

$$\mathbb{E}\,\|X - f(Y)\|^2 = \mathbb{E}\,\|X - m(Y)\|^2 + \mathbb{E}\,\|m(Y) - f(Y)\|^2 ,$$

第二项在且仅在 $f(Y) = m(Y)$ 几乎处处时最小。

(b) 标量 absolute loss 的 Bayes predictor 是任一 conditional median：$\mathbb{P}(X \le f(Y) \mid Y) \ge 1/2$ 且 $\mathbb{P}(X \ge f(Y) \mid Y) \ge 1/2$。多峰 posterior 的 conditional mean 可能位于几个清晰模式之间，因而产生并非任一真实模式的平均化、模糊图像。

(c) 代表性设计：输入 $Y \in \mathbb{R}^{C\times H\times W}$；编码器以卷积和下采样获得多尺度特征；解码器上采样；每个尺度把 encoder 特征 concat/add 到对应 decoder；最后 1×1 卷积输出 $\widehat{X} \in \mathbb{R}^{C\times H\times W}$。可最小化 $\left\|X - \widehat{X}\right\|_1 /(CHW)$，或像素 NLL；若使用混合损失必须给权重。skip 保留高分辨率局部信息并提供较短梯度通路。

(d) 例：真实噪声与训练假设不匹配；$\mathcal{A}$ 丢失信息导致多解；像素 loss 与感知质量不一致；训练/测试退化分布变化。评价可使用 PSNR/SSIM，并应在多解任务中补充感知或人工评价；仅报告训练 loss 不足。

Rubric: orthogonal decomposition 与唯一性 4 分；median 与模糊解释各 1 分；网络四要素和 skip 3 分；两个 failure 与 metric 2 分。

(b) **Dequantization 与 flow**。 [13 points]

(a) 模型诱导的离散概率质量为

$$P_\theta(Y = y) = \int_{[0,1)^d} p_\theta(y + u)\, du = \mathbb{E}_U[p_\theta(y + U)].$$

由于 log 为凹函数，Jensen 给出

$$\log P_\theta(Y = y) = \log \mathbb{E}_U p_\theta(y + U) \ge \mathbb{E}_U \log p_\theta(y + U).$$

所以随机 dequantization 上的 log-density 是离散 log-likelihood 的下界，而非上界。

(b) 令 $z = f_\theta^{-1}(x)$。变量替换给出

$$\log p_\theta(x) = \log p_0(z) - \log |\det J_{f_\theta}(z)| .$$

数据 $x^{(i)}$ 上的平均 NLL 为

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_i \left[-\log p_0(z^{(i)}) + \log \left| \det J_{f_\theta}(z^{(i)}) \right| \right] .$$

若改用 inverse Jacobian，其 log-determinant 符号相反，不能混用。

(c) Dequantization 解决离散 probability mass 与连续 density 的测度不匹配；它不只是希望模型对标签保持不变的 augmentation。跨维度比较通常报告 bits per dimension：以 base-2 的 negative log-likelihood 除以数据维数，并说明像素缩放与 dequantization 约定。

(d) $z = (x-b)/a$，且 $dx/dz = a$，所以

$$\log p_X(x) = \log p_0\left(\frac{x-b}{a}\right) - \log|a|.$$

Rubric: cell mass、Jensen 方向、下界含义 5 分；变量替换、inverse 和 NLL 4 分；测度解释与可比指标 2 分；标量式 2 分。

(c) **Bellman residual。** [9 points]

(a) 对任意 V, W，

$$\|T^\pi V - T^\pi W\|_\infty \le \gamma \|V - W\|_\infty ,$$

因为 $P^\pi(\cdot \mid s)$ 是概率分布。

(b) 利用 $V^\pi = T^\pi V^\pi$，

$$\begin{aligned}\|V - V^\pi\|_\infty &\le \|V - T^\pi V\|_\infty + \|T^\pi V - T^\pi V^\pi\|_\infty \\ &\le \delta + \gamma \|V - V^\pi\|_\infty .\end{aligned}$$

移项即得 $\|V - V^\pi\|_\infty \le \delta/(1-\gamma)$。

(c) 数值为 $0.02/(1-0.95) = 0.4$。长时域中局部 Bellman 误差可在约 $1/(1-\gamma)$ 的有效 horizon 上累积，故该 worst-case bound 可能较松。

Rubric: contraction 3 分；三角不等式、fixed point、移项 4 分；数值和解释 2 分。

C. 人工智能中的优化方法：参考解答 [33 points]

(1) [3 points]

$$\nabla f(x) = A^\top (Ax - b).$$

因此

$$\|\nabla f(x) - \nabla f(y)\| = \|A^\top A(x-y)\| \le \|A^\top A\|_{\mathrm{op}} \|x-y\| = L\|x-y\|.$$

f 是光滑部分，$R = \lambda\|\cdot\|_1$ 允许不可微；L 只由光滑部分决定。

(2) [5 points] Prox 子问题按坐标分离，每个坐标最小化

$$\lambda|z| + \frac{1}{2\gamma}(z-u)^2.$$

最优性条件是 $0 \in \lambda\partial|z| + \gamma^{-1}(z-u)$。若 $z>0$，则 $z = u - \gamma\lambda$，并需 $u > \gamma\lambda$；若 $z<0$，则 $z = u + \gamma\lambda$，并需 $u < -\gamma\lambda$。若 $z = 0$，条件为 $u/\gamma \in [-\lambda, \lambda]$，即 $|u| \le \gamma\lambda$。合并三种情形得 soft thresholding 公式。

(3) [4 points] 记 $S_\tau(u)_i = \mathrm{sign}(u_i)(|u_i| - \tau)_+$，则

$$x^{k+1} = S_{\lambda/L}\left[x^k - \frac{1}{L}A^\top(Ax^k - b)\right].$$

$G_\gamma(x) = 0$ 当且仅当 $x = \mathrm{prox}_{\gamma R}(x - \gamma\nabla f(x))$。由 prox 最优性，这等价于

$$0 \in \nabla f(x) + \partial R(x) = \partial F(x),$$

而对 proper convex F，这是全局最小的充要条件。

(4) [5 points] 删去与 z 无关的常数，Q_γ 的最小化等价于

$$\min_z R(z) + \frac{1}{2\gamma}\|z - [x - \gamma\nabla f(x)]\|^2,$$

因而其唯一解正是 x^+。由 descent lemma，

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

当 $\gamma \le 1/L$ 时，$1/(2\gamma) \ge L/2$，故 $Q_\gamma(z;x) \ge F(z)$。又因 x^+ 最小化 Q_γ，

$$F(x^+) \le Q_\gamma(x^+;x) \le Q_\gamma(x;x) = F(x).$$

(5) [6 points] Prox 最优性给出

$$s^+ := \frac{x - x^+}{\gamma} - \nabla f(x) \in \partial R(x^+).$$

因此 $R(x^+) - R(z) \le \langle s^+, x^+ - z\rangle$。另一方面，smoothness 与 convexity 给出

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

两式相加，梯度项消去：

$$F(x^+) - F(z) \le \frac{1}{\gamma}\langle x - x^+, x^+ - z\rangle + \frac{L}{2}\|x^+ - x\|^2.$$

再用三点平方恒等式即得题设结论。注意 s^+ 位于新点 x^+ 的次微分，不是一般地位于 $\partial R(x)$。

(6) [6 points] 取 $x = x^k$、$x^+ = x^{k+1}$、$z = x^\star$、$\gamma = 1/L$，第 (5) 问给出

$$F(x^{k+1}) - F(x^\star) \le \frac{L}{2}\left(\|x^k - x^\star\|^2 - \|x^{k+1} - x^\star\|^2\right).$$

对 $k = 0, \dots, K-1$ 求和，

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

第 (4) 问已证明 $F(x^k)$ 单调不增，故和式左侧至少是 $K[F(x^K) - F(x^\star)]$，除以 K 得结论。

(7) [4 points] 稠密 A 下，计算 Ax、$A^\top(Ax-b)$ 的主导代价为 $O(md)$，thresholding 为 $O(d)$。若 A 稀疏，主导代价可写成 $O(\mathrm{nnz}(A))$。当梯度步后的第 i 坐标 u_i 满足 $|u_i| \le \gamma\lambda$ 时，soft threshold 把它精确置零。但 ℓ_1 只是 convex 而非 strongly convex；若 A 有非零 nullspace，F 可能不严格凸，因而稀疏性不自动推出最小点唯一。

Rubric: (1) gradient、L、非光滑部分各 1 分；(2) 三种坐标情况与边界齐全得 5 分；(3) update 2 分、最优性等价 2 分；(4) quadratic completion 2 分、majorization 2 分、单调性 1 分；(5) 新点次梯度 2 分、两个函数不等式 2 分、三点代数 2 分；(6) one-step 2 分、telescope 2 分、monotone last iterate 2 分；(7) 两种代价 2 分、置零机制 1 分、唯一性警告 1 分。

D. 自然语言处理：参考解答 [33 points]

1. Multiple-positive contrastive retrieval [8 分]

令

$$\pi_j = \frac{e^{s_j/\tau}}{\sum_{r=1}^{n} e^{s_r/\tau}}, \qquad \rho_p = \frac{e^{s_p/\tau}}{\sum_{r\in P} e^{s_r/\tau}} \quad (p \in P).$$

把 loss 写成两个 log-sum-exp 之差并求导，得

$$\boxed{\nabla_q \ell = \frac{1}{\tau}\left(\sum_{j=1}^{n} \pi_j d_j - \sum_{p\in P} \rho_p d_p\right).}$$

第一项是全 batch softmax 下的 document average，第二项是 positive-only softmax average；gradient descent 使 query 靠近当前有代表性的 positives，并远离被模型赋高分的非 positives。

若只保留一个 positive，其他确实相关的 passages 会进入分母却不进入分子，成为 false negatives，造成与任务目标相反的梯度。相同文档的重复 chunk、同一事实的 paraphrases 或共享 entity 的 passages 应按 relevance definition 合并进 P、从 denominator mask 掉，或使用 soft weights；不能仅因它们位于其他样本的 positive slot 就自动当负例。

评分要点 两组 normalized weights 2 分；完整梯度 3 分；几何解释 1 分；single-positive 比较及 duplicate 处理 2 分。

2. Calibrated hybrid retrieval [7 分]

由 Bayes rule 和条件独立性，

$$\begin{aligned}\log \frac{p(R=1 \mid S_1, S_2)}{p(R=0 \mid S_1, S_2)} &= \log \frac{p(R=1)}{p(R=0)} + \log \frac{p(S_1, S_2 \mid R=1)}{p(S_1, S_2 \mid R=0)} \\ &= \boxed{\log \frac{\pi}{1-\pi} + \ell_1(S_1) + \ell_2(S_2)},\end{aligned}$$

其中 $\pi = p(R=1)$。因此在 calibrated log-evidence 空间中相加有概率解释；实际可用 held-out relevance data 做 Platt/isotonic calibration，或直接训练 logistic fusion。

Raw BM25 与 cosine 的尺度、偏移和 query-dependent range 不同，直接相加会让某一分量任意支配。两 signals 往往还因共享 lexical cues 而相关；若错误假设独立，会 double count evidence 并产生过度自信。可使用带 interaction features 的 learned fusion，并用 reliability diagram/ expected calibration error 检查。

评分要点 Bayes 推导与 prior 3 分；fusion 含义 1 分；raw-scale 问题 1 分；依赖导致重复计数及合理修正 2 分。

3. Global entity linking [8 分]

定义

$$V_1(e) = a_1(e), \qquad V_t(e) = a_t(e) + \max_{e' \in C_{t-1}} \left\{V_{t-1}(e') + \lambda b(e', e)\right\}.$$

对每个 (t, e) 保存取得最大值的 e'。从 $\arg\max_e V_T(e)$ 反向沿 backpointer 即得全局最优 assignment。时间 $O(TK^2)$，score rolling arrays 为 $O(K)$，恢复路径所需 backpointers 为 $O(TK)$。

若所有 mention pairs 都有 factor，则选择 e_t 会同时影响多个尚未处理和已处理变量，状态不能只由 e_{t-1} 概括；对应 graph 一般有大量环，chain DP 不再 exact。可用 loopy belief propagation、mean-field、beam search、integer linear programming 的松弛，或先构造 maximum spanning tree 再做 tree inference。应说明近似的代价或可能非全局最优。

评分要点 正确 recursion 3 分；终止/回溯和复杂度 2 分；全连接失效原因 2 分；合理 approximate inference 1 分。

4. Cross-lingual entity-linking system [10 分]

下面是一种 representative full-credit design。Candidate generation 联合使用 alias table、字符/音译 encoder、跨语言 dense retrieval 和 mention context translation，输出带先验的 top-K KB IDs，并始终加入 NIL。Local cross-encoder 输入原语言 mention context、candidate description 和类型，输出 $a_t(e)$。Document inference 再加入 entity co-occurrence、KB relation 和类型一致性分数，可用上一题的 chain/beam 或稀疏 factor graph MAP，输出 entity ID 或 NIL 及 calibrated confidence。

Weak supervision 可来自跨语言 Wikipedia hyperlinks、同一新闻事件的 comparable documents、high-precision exact/unique alias，以及 round-trip transliteration agreement。用这些 positives 配合 in-batch/hard negatives 训练 bi-encoder contrastive loss，再训练 local pairwise or multiclass cross-entropy；低置信伪标签应 down-weight，并保留 language-balanced sampling。

若 $\max_e p(e \mid m, x) < \tau$、top-1/top-2 margin 小，或 KB candidate coverage detector 判为低，则输出 NIL。除 micro accuracy 外，报告 macro-F1、按 entity frequency/language 分桶 recall、candidate recall@K、NIL precision/recall 和 calibration。

两类示例：音译碰撞导致候选错误，可加原文上下文、地名/职业 type features 与 document-level coherence；KB 缺失导致强行链接，可显式训练 NIL、open-set calibration 并允许创建 provisional entity。其他成对出现的 failure–targeted mitigation 同等给分。

评分要点 完整数据流和 I/O 4 分；可行 weak supervision/objectives 2 分；NIL 与长尾指标 2 分；两项针对性 failure mitigation 2 分。