Artificial Intelligence PhD Qualifying Examination

模拟考试 08：参考解答与评分要点

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)

$$\nabla J_\lambda = \frac{1}{m}X^\top(Xw-y)+\lambda w, \qquad \nabla^2 J_\lambda = \frac{1}{m}X^\top X + \lambda I \succ 0.$$

故 objective strongly convex，normal equation 与 solution 为

$$(X^\top X + m\lambda I)w = X^\top y, \qquad \boxed{\widehat{w} = (X^\top X + m\lambda I)^{-1}X^\top y}.$$

矩阵对任意 $\operatorname{rank}(X)$ 都 positive definite，所以唯一。若 data fit 改写为 $\frac{1}{2}\|Xw-y\|^2$，相同 minimizer 的 penalty 是 $(m\lambda/2)\|w\|^2$。

Rubric: gradient/Hessian 2 分；PD/uniqueness 2 分；equation/solution 2 分；sum scaling 2 分。

(b) 取 $d+1$ 点 $0, e_1, \dots, e_d$。给定这些点的 labels $y_0, y_1, \dots, y_d$，令 $b = y_0$、$w_i = y_i \oplus y_0$，则 $h(0) = y_0$ 且 $h(e_i) = y_i$，故下界为 $d+1$。Class size 为 2^{d+1}，所以 finite-size upper bound 也是 $d+1$。

Rubric: 点集 1 分；parameter construction 3 分；size upper bound 1 分；结论 1 分。

(c) 对 fixed h，变量 $U_i = r(Z_i)\ell(h, Z_i) \in [0, W]$，且 $\mathbb{E}_Q U_i = L_P(h)$。Hoeffding 给

$$\Pr\{|\widehat{L}_{\mathrm{IW}}(h) - L_P(h)| > \alpha\} \le 2e^{-2m\alpha^2/W^2}.$$

对 N 个 hypotheses union bound。在 uniform event 上，ERM excess 至多 2α。取 $\alpha = \varepsilon/2$，充分条件为

$$\boxed{m \ge \frac{2W^2}{\varepsilon^2}\log\frac{2N}{\delta}.}$$

Rubric: range/change of measure 2 分；Hoeffding 1 分；union 1 分；ERM comparison 1 分；常数 2 分。

(d) 对每个 fixed z，Jensen 给 $\ell(\mathbb{E}W, z) \le \mathbb{E}_W \ell(W, z)$。再对 z 取 population expectation，并用 Tonelli/Fubini，即得 $L(\bar{w}) \le \mathbb{E}_W L(W)$。0–1 loss 对参数或 predictions 通常不 convex，而 majority vote 也不等于 parameter mean，所以该 Jensen 论证不适用。

Rubric: pointwise Jensen 2 分；交换期望 1 分；0–1 caveat 2 分。

(e) 固定任意 m 点 sample。$\mathcal{H} \cup \mathcal{G}$ 在该 sample 上产生的 label-vector 集合是两个 component label-vector sets 的 union，其 cardinality 至多两者 cardinalities 之和。再对 samples 取 maximum 即得题式。若两个 classes 在 sample 上产生重复 labeling，inequality 严格。

Rubric: fixed-sample label sets 2 分；maximum argument 1 分；strictness 1 分。

B. 深度学习与强化学习：参考解答 [33 points]

(a) **Linear denoiser**。 [10 points]

(a) 展开 risk：

$$R(A) = \operatorname{tr}(\Sigma) - 2\operatorname{tr}(A\Sigma) + \operatorname{tr}\left(A(\Sigma + \sigma^2 I)A^\top\right).$$

矩阵一阶条件为 $2A(\Sigma + \sigma^2 I) - 2\Sigma = 0$，故

$$A^* = \Sigma(\Sigma + \sigma^2 I)^{-1}.$$

$\sigma > 0$ 时括号内正定。也可由 linear least-squares normal equation $A\operatorname{Cov}(Y) = \operatorname{Cov}(X, Y)$ 得出。

(b) 若 $\Sigma = U\operatorname{diag}(\lambda_i)U^\top$，

$$A^* = U\operatorname{diag}\left(\frac{\lambda_i}{\lambda_i + \sigma^2}\right)U^\top.$$

高 variance signal directions 保留更多。$\sigma \downarrow 0$ 时在 Σ 的 range 上趋近 identity（零特征值方向仍为零）；$\sigma \uparrow \infty$ 时趋近零预测。

(c) 若 clean input=target 且网络容量足够，identity 使 reconstruction loss 为零，不必学到稳健结构。Denoising 使用 corrupted input、clean target，identity 不再最优；最优函数是 conditional mean $\mathbb{E}[X \mid Y]$，必须利用数据统计去噪。

Rubric: risk/normal equation、最优矩阵 5 分；eigen shrinkage 和两极限 3 分；identity failure 与 denoising 机制 2 分。

(b) **Conditional score**。 [14 points]

(a) Bayes 给出 $\log p_t(x \mid c) = \log p_t(x) + \log p_t(c \mid x) - \log p_t(c)$。最后一项对 x 为常数，故

$$\nabla_x \log p_t(x \mid c) = s_u(x, t) + g_c(x, t).$$

(b) Gaussian score 为

$$\nabla_{x_t} \log p_t(x_t \mid x_0) = -\frac{x_t - a_t x_0}{b_t^2} = -\frac{\epsilon}{b_t}, \quad x_t = a_t x_0 + b_t \epsilon.$$

可最小化

$$\mathbb{E}_{t, x_0, \epsilon}\left[\lambda(t)\left\|s_\theta(x_t, t) + \epsilon/b_t\right\|_2^2\right].$$

平方损失的 conditional-expectation identity 保证其 minimizer 学到 marginal score（在函数类充分时）。

(c) $w = 1$ 时 s_w 是精确 conditional score。一般

$$s_w = \nabla_x \log\left[p_t(x)p_t(c \mid x)^w\right].$$

归一化后的 density 为上述量除以只依赖 (c, t, w) 的 Z；$\nabla_x \log Z = 0$，故 score 不变。也可写 $s_w = (1 - w)s_u + w s_c$。

(d) 例如：classifier/conditional-score 在低密度区误差被放大；大 w 会降低 diversity、过分集中；离散化误差或 train/test condition shift 增大；条件模型可能被 adversarially exploited。任两项并说明机制。

Rubric: Bayes 与常数 4 分；Gaussian score/DSM/target 4 分；exact w、unnormalized density、normalizer 4 分；两个 tradeoff 2 分。

(c) **Entropy bandit。** [9 points]

(a) 对约束 $\sum_a \pi_a = 1$ 的 Lagrangian 求导：$r_a - \tau(\log \pi_a + 1) + \lambda = 0$。因此

$$\pi_a^* = \frac{\exp(r_a/\tau)}{\sum_b \exp(r_b/\tau)}.$$

严格凹 entropy 加线性 reward 在 simplex interior 给出唯一最优分布（并列 reward 也由公式唯一分配）。

(b) $\tau \downarrow 0$ 时质量集中在最大 reward 动作，并列最大时均匀分配；$\tau \uparrow \infty$ 时趋向所有动作均匀分布。

(c) $r_a \mapsto r_a + c$ 使分子分母同乘 $e^{c/\tau}$，policy 不变。τ 是 reward exploitation 与 entropy/exploration 的温度，单位应与 reward 相同。

Rubric: Lagrangian、softmax、normalizer/唯一性 5 分；两极限 2 分；平移不变与温度解释 2 分。

C. 人工智能中的优化方法：参考解答 [33 points]

(1) [3 points] 对任意 x, y，两项定义分别为

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

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

强凸性保证最小点唯一；无约束可微最小点满足 $\nabla f(x^\star) = 0$。

(2) [5 points] 令 $\zeta_{k,j} = g(x^k, \xi_{k,j}) - \nabla f(x^k)$。给定 $\mathcal{F}_k$ 后，x^k 可测且 $\mathbb{E}_k \zeta_{k,j} = 0$，故

$$\mathbb{E}_k G_k = \nabla f(x^k).$$

条件独立性与中心化给出，$i \neq j$ 时

$$\mathbb{E}_k \langle \zeta_{k,i}, \zeta_{k,j} \rangle = \langle \mathbb{E}_k \zeta_{k,i}, \mathbb{E}_k \zeta_{k,j} \rangle = 0.$$

所以

$$\mathbb{E}_k \left\| \frac{1}{\tau} \sum_{j=1}^{\tau} \zeta_{k,j} \right\|^2 = \frac{1}{\tau^2} \sum_{j=1}^{\tau} \mathbb{E}_k \|\zeta_{k,j}\|^2 \le \frac{\sigma^2}{\tau}.$$

(3) [5 points] 强凸性给出，对任意 y，

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

右侧 lower model 关于 y 的最小点为 $y = x - \nabla f(x)/\mu$，最小值为 $f(x) - \|\nabla f(x)\|^2/(2\mu)$。特别地，在 $y = x^\star$ 代入强凸不等式，再用该 model 值不小于其最小值，得到

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

移项即为题设 PL inequality。

(4) [7 points] 记 $\nabla f_k = \nabla f(x^k)$。descent lemma 给出

$$f(x^{k+1}) \le f(x^k) - \eta \langle \nabla f_k, G_k \rangle + \frac{L\eta^2}{2}\|G_k\|^2.$$

第 (2) 问及条件 bias–variance decomposition 给出

$$\mathbb{E}_k \|G_k\|^2 = \|\nabla f_k\|^2 + \mathbb{E}_k \|G_k - \nabla f_k\|^2 \le \|\nabla f_k\|^2 + \frac{\sigma^2}{\tau}.$$

因此

$$\begin{aligned} \mathbb{E}_k[\Delta_{k+1}] &\le \Delta_k - \eta \left(1 - \frac{L\eta}{2}\right) \|\nabla f_k\|^2 + \frac{L\eta^2\sigma^2}{2\tau} \\ &\le \Delta_k - \frac{\eta}{2}\|\nabla f_k\|^2 + \frac{L\eta^2\sigma^2}{2\tau} \\ &\le (1 - \mu\eta)\Delta_k + \frac{L\eta^2\sigma^2}{2\tau}. \end{aligned}$$

第二行使用 $\eta \le 1/L$，第三行使用第 (3) 问。

(5) [5 points] 令 $a_k = \mathbb{E}\Delta_k$、$r = 1-\mu\eta \in [0,1)$、$b = L\eta^2\sigma^2/(2\tau)$。tower property 给出 $a_{k+1} \le ra_k + b$，故

$$a_K \le r^K\Delta_0 + b\sum_{j=0}^{K-1} r^j = r^K\Delta_0 + \frac{b}{1-r}(1-r^K).$$

由于 $1-r = \mu\eta$，

$$\mathbb{E}\Delta_K \le (1-\mu\eta)^K\Delta_0 + \frac{L\eta\sigma^2}{2\mu\tau}[1-(1-\mu\eta)^K].$$

相应上界的 limiting noise floor 是 $L\eta\sigma^2/(2\mu\tau)$。

(6) [5 points] 若 $\sigma = 0$，可取 $\tau = 1, \eta = 1/L$。若 $\sigma > 0$，任取整数 $\tau \ge 1$，再取

$$\eta = \min\left\{\frac{1}{L}, \frac{\mu\tau\varepsilon}{L\sigma^2}\right\}.$$

两种情形下都保证 noise-floor 上界不超过 $\varepsilon/2$。若 $\Delta_0 \le \varepsilon/2$，瞬态要求在 $K = 0$ 已满足。否则利用 $(1-\mu\eta)^K \le e^{-\mu\eta K}$，充分取

$$K \ge \left\lceil \frac{1}{\mu\eta}\log\frac{2\Delta_0}{\varepsilon}\right\rceil.$$

于是 transient 与 noise-floor 上界各不超过 $\varepsilon/2$。

(7) [3 points] K 次迭代使用 τK 次 stochastic-gradient evaluations。减小 η 线性降低 noise floor，却减慢 contraction；增大 τ 以 $1/\tau$ 降低平台而不改变每步 contraction，却增加串行 oracle cost。增大 K 只削减 transient，不能突破固定 η, τ 的平台。batch 内梯度可并行，故较大的 τ 可能降低 wall-clock time。

Rubric: (1) 两定义及唯一/驻点 3 分；(2) 条件无偏 2 分、交叉项与 $1/\tau$ 3 分；(3) lower model 2 分、最小化与 PL 常数 3 分；(4) descent 2 分、二阶矩 2 分、$\eta \le 1/L$ 与 PL 3 分；(5) tower/geometric sum 3 分、精确 floor 2 分；(6) η, τ 选择 2 分、两个 K 情形 2 分、误差拆分 1 分；(7) oracle count 与 tradeoff 3 分。

D. 自然语言处理：参考解答 [33 points]

1. Autoregressive and masked pre-training objectives [10 分]

Autoregressive model 写成

$$p_\theta(x_{1:T}) = \prod_{t=1}^{T} p_\theta(x_t \mid x_{<t}), \qquad \mathcal{L}_{\mathrm{AR}} = -\sum_{t=1}^{T} \log p_\theta(x_t \mid x_{<t}).$$

对 $x_T, x_{T-1}, \ldots, x_1$ 依次求和，每一步都使用 $\sum_{x_t} p_\theta(x_t \mid x_{<t}) = 1$，最终得到 $\sum_{x_{1:T}} p_\theta(x_{1:T}) = 1$；chain rule 本身不是 approximation。

若 $M \sim q(M)$，一种 MLM objective 为

$$\mathcal{L}_{\mathrm{MLM}} = \mathbb{E}_{M\sim q}\left[-\sum_{t\in M} \log p_\theta(x_t \mid x_{\overline{M}}, M)\right].$$

它训练许多 bidirectional conditionals，但没有给出生成完整 sequence 的顺序和 joint normalizer；不同 mask patterns 的 conditionals 也未必彼此 compatible。因此不能直接把这些 token probabilities 相乘声称得到一个 normalized joint。Pseudo-likelihood 或 Gibbs-style sampling 可使用这些 conditionals，但那是额外定义的评分/生成 procedure。

AR 的 token t 只看 prefix，天然适合 left-to-right generation；teacher forcing 时所有位置仍可用 causal mask 一次并行训练。MLM 同时使用左右上下文，通常更适合 bidirectional understanding/representation。一个 AR forward pass 可同时给出 held-out sequence 的所有 conditional log-probs；标准 MLM 若要每个位置都在“该位置被 mask”条件下评分，通常需 T 次 masked passes，或使用一次 multi-mask approximation。

AR 的 pitfall 是 training 总看到 gold prefix，而 sampling 看到自身 prefix，形成 exposure mismatch；不能用未来 token 构造未遮蔽 feature。MLM 若 masking 太容易（例如 target 仍通过复制、subword overlap 或 data leakage 可见）会学习 shortcut；pre-training 的 artificial masks 与真实无 mask inference 也有 mismatch。任何准确的一对比较均可。

评分要点 AR factorization/NLL 1 分、归一化证明 2 分；MLM objective 2 分、非 joint 解释 1 分；information/use 2 分、held-out cost 与两类 pitfall 2 分。

2. Subword-composed word representations [10 分]

Softmax 非负且

$$\sum_{c=1}^{V} p(c \mid w) = \frac{\sum_c e^{\mathbf{h}_w^\top \mathbf{v}_c}}{\sum_{c'} e^{\mathbf{h}_w^\top \mathbf{v}_{c'}}} = 1.$$

记 $p_c = p(c \mid w)$。Cross-entropy 对 composed representation 的 gradient 为

$$\nabla_{\mathbf{h}_w} \ell = \sum_{c=1}^{V} p_c \mathbf{v}_c - \mathbf{v}_{c^*}.$$

由于 $\partial \mathbf{h}_w / \partial \mathbf{z}_g = I/|G(w)|$，

$$\boxed{\nabla_{\mathbf{z}_g} \ell = \frac{1}{|G(w)|}\left(\sum_c p_c \mathbf{v}_c - \mathbf{v}_{c^*}\right), \quad g \in G(w),}$$

且

$$\boxed{\nabla_{\mathbf{v}_c}\ell = \big(p_c - \mathbf{1}\{c = c^*\}\big)\mathbf{h}_w.}$$

Rare word 可从在常见屈折词中学到的 stems/affixes n-grams 共享统计；OOV word 只要能分解成已知 n-grams 就能得到 representation。失败包括不同语言/词义共享相同短 n-grams 导致 collision、极长词的平均稀释关键 morpheme、script normalization 错误。可使用 boundary markers、语言特定的 n 范围、learned attention over morphemes、hash-bucket 扩容或真正的 morphological segmenter；mitigation 应对应 failure。

组合 $\mathbf{h}_w$ 为 $O(|G(w)|d)$，full logits/normalizer 为 $O(Vd)$；若有 N_g 个 subword types，主要 parameter memory 为 $O((N_g + V)d)$。Sampled softmax 可用 sampled correction 近似原 softmax likelihood；negative sampling 将其改成一组 binary logistic objectives，并非精确计算同一个 full-softmax normalizer。两者单 pair 常降到约 $O((|G(w)| + k)d)$。

评分要点 normalization 1 分；h, z_g, v_c gradients 4 分；rare/OOV 解释 2 分、failure/mitigation 1 分；time/memory 与 approximation distinction 2 分。

3. A bounded external memory [13 分]

递归展开为

$$\boxed{m_J = \sum_{i=1}^{J}(1 - g_i)\left(\prod_{r=i+1}^{J} g_r\right)h_i.}$$

空乘积为 1。这些系数 telescoping 后的总和为

$$\sum_{i=1}^{J}(1 - g_i)\prod_{r=i+1}^{J} g_r = 1 - \prod_{r=1}^{J} g_r.$$

缺少的 $\prod_r g_r$ 是 m_0 的系数，而本题 $m_0 = 0$，故一般不是 h_i 的 convex combination。若 $g_j = g$，h_i 的 coefficient 为 $(1 - g)g^{J-i}$，随 age 指数衰减，近 $g = 1$ 时 effective horizon 约为 $1/(1 - g)$。

一个不引入额外 projection 的 read 为

$$a = \mathrm{softmax}_{\mathrm{slot}}\left(\frac{Mq}{\sqrt{d}}\right) \in \mathbb{R}^m, \qquad r = M^\top a \in \mathbb{R}^d.$$

Softmax 在 m 个 slots 上归一化。Time 为 $O(md)$；persistent slot memory 为 $O(md)$；除输入 M 外的 weights/output working memory 为 $O(m + d)$。若使用 dense query/key/value projections，应另计相应 d^2 或预计算成本。

每个 slot 保存 content、embedding、entities/topic、source ID、immutable version、valid/ingest time、confidence 与 supersedes pointer。新 update 先找同 claim 的 active slot：若是 correction，append 新 record 并把旧 version 标为 superseded；互补 evidence 可 merge summary 但保留所有 source pointers；否则写新 slot。容量满时按预先固定的 $U_i = \text{recency} + \text{expected-use} + \text{confidence} - \lambda\text{redundancy}$ 排序，先 merge 再 evict 最低 utility 的非 protected slot，并用 deterministic tie-break/audit log。

设 writer $M_\phi(H)$、reader $R_\phi(q, M)$，主训练目标可为

$$-\log p_\theta(y \mid q, R_\phi(q, M_\phi(H))).$$

离散写入可用 straight-through relaxation 或 policy-gradient surrogate。为避免所有输入写同一 generic summary，可加入 source-span reconstruction、question–evidence contrastive loss、

slot-diversity/ coverage，以及需要细节证据的 hard questions；不能只加入单边 storage penalty，否则会鼓励空 memory。

生成器输出 atomic claims 与 active source-version IDs；若关键 claim 无 active support、两个高可信 active versions 冲突或 verifier confidence 低则 abstain/报告冲突。Catastrophic forgetting 用“注入大量后续 updates 后的 old-but-still-valid recall”测量；staleness 用 correction 后 superseded-source citation rate 测量，并可另报 latest-version accuracy。

评分要点 expansion 2 分、coefficient sum/decay 2 分；normalized read 1.5 分、三类 complexity 1.5 分；四种操作和 deterministic budget 2.5 分、objective/anti-degeneracy 1.5 分；versioned abstention 1 分、两项针对性 metrics 1 分。