Artificial Intelligence PhD Qualifying Examination

模拟考试 02：参考解答与评分要点

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) 由 Euclidean duality 与 Jensen，

$$\begin{aligned}\widehat{\mathfrak{R}}_{S_X}(\mathcal{G}_2(B)) &= \mathbb{E}_\sigma \sup_{\|w\|_2\le B} \frac{1}{m}\left\langle w, \sum_i \sigma_i x_i\right\rangle \\ &= \frac{B}{m}\mathbb{E}_\sigma \left\|\sum_i \sigma_i x_i\right\|_2 \\ &\le \frac{B}{m}\sqrt{\mathbb{E}_\sigma \left\|\sum_i \sigma_i x_i\right\|_2^2} = \frac{B}{m}\sqrt{\sum_i \|x_i\|_2^2}.\end{aligned}$$

最后一步用 $\mathbb{E}\sigma_i\sigma_j = \mathbf{1}\{i=j\}$。若每个 norm 至多 R，即得 $BR/\sqrt{m}$。

Rubric: 定义 1 分；dual norm 2 分；Jensen 1 分；交叉项 2 分；结论 1 分。

(b) 对 fixed h，Hoeffding 给

$$\Pr\{|\widehat{L}(h) - L(h)| > \alpha\} \le 2e^{-2m\alpha^2}.$$

Union bound 后，所有 h 同时偏差不超过 α 的失败概率至多

$$2|\mathcal{H}|e^{-2m\alpha^2}.$$

在该事件上，若 $h^\star \in \arg\min_h L(h)$，则

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

取 $\alpha = \varepsilon/2$，充分条件是

$$\boxed{m \ge \frac{2}{\varepsilon^2}\log\frac{2|\mathcal{H}|}{\delta}.}$$

Rubric: pointwise Hoeffding 2 分；union bound 1 分；ERM chain 2 分；常数 2 分。

(c) Strong convexity 的 first-order form 对 $w, w^\star$ 相加，且 $\nabla F(w^\star) = 0$，得到

$$\langle \nabla F(w), w - w^\star\rangle \ge \lambda \|w - w^\star\|^2.$$

Cauchy–Schwarz 给第一式。另一方面，把 strong-convexity lower bound 从 w 写到 $w^\star$：

$$F(w^\star) \ge F(w) + \langle \nabla F(w), w^\star - w\rangle + \frac{\lambda}{2}\|w - w^\star\|^2.$$

令 $r = \|w - w^\star\|$，右侧 comparison 给

$$F(w) - F(w^\star) \le \|\nabla F(w)\|\, r - \frac{\lambda}{2} r^2 \le \frac{\|\nabla F(w)\|^2}{2\lambda}.$$

因此 gradient norm 至多 γ 同时保证距离至多 γ/λ、objective gap 至多 $\gamma^2/(2\lambda)$。

Rubric: strong monotonicity 2 分；距离界 1 分；quadratic maximization 1 分；解释 1 分。

(d) 取 d 个 standard basis vectors $e_1, \dots, e_d$。对任意 labels $y \in \{0,1\}^d$，令 $S = \{i : y_i = 1\}$，则 $h_S(e_i) = 1$ 当且仅当 $i \in S$，故这 d 点被 shattered。另一方面 $|\mathcal{D}_d| = 2^d$，finite-class counting bound 给 $\mathrm{VCdim}(\mathcal{D}_d) \le \log_2 |\mathcal{D}_d| = d$。所以

$$\boxed{\mathrm{VCdim}(\mathcal{D}_d) = d.}$$

Rubric: 点集 1 分；任意 labeling construction 3 分；counting upper bound 1 分；结论 1 分。

(e) ReLU coordinatewise 是 1-Lipschitz：$\|\sigma(u) - \sigma(v)\|_2 \le \|u - v\|_2$。线性层满足 $\|W_\ell u - W_\ell v\|_2 \le \|W_\ell\|_2 \|u - v\|_2$。从第一层到输出层逐层复合即得

$$|f(x) - f(x')| \le \prod_{\ell=1}^{L} \|W_\ell\|_2 \, \|x - x'\|_2 .$$

控制 product 可限制 input sensitivity/robustness，并进入某些 norm-based capacity 或 generalization bounds；若只增大相互抵消的层 norm，function value 未必改变，但该上界会变松。

Rubric: ReLU 1-Lipschitz 1 分；单层 spectral bound 1 分；复合证明 2 分；解释 1 分。

B. 深度学习与强化学习：参考解答 [33 points]

(a) **BN 与 dropout。** [9 points]

(a) 训练时

$$\mu_B = \frac{1}{B}\sum_i x_i, \qquad v_B = \frac{1}{B}\sum_i (x_i - \mu_B)^2,$$
$$\widehat{x}_i = \frac{x_i - \mu_B}{\sqrt{v_B + \epsilon}}, \qquad y_i = \gamma_{\mathrm{BN}}\widehat{x}_i + \beta_{\mathrm{BN}}.$$

例如可用 $\mu_{\mathrm{run}} \leftarrow \rho\mu_{\mathrm{run}} + (1-\rho)\mu_B$、$v_{\mathrm{run}} \leftarrow \rho v_{\mathrm{run}} + (1-\rho)v_B$ 更新 running statistics。方差是否作无偏修正可依软件约定，但须前后一致。

(b) 推理时 $y = \gamma_{\mathrm{BN}}(x-\mu_{\mathrm{run}})/\sqrt{v_{\mathrm{run}}+\epsilon} + \beta_{\mathrm{BN}}$。若使用当前测试 batch，同一输入会依赖 batch 中其他样本，且 batch 很小时噪声很大。

(c) 条件于 h，

$$\mathbb{E}[\widetilde{h} \mid h] = h, \qquad \mathrm{Var}(\widetilde{h} \mid h) = \frac{1-q}{q}h^2.$$

标准 inverted dropout 在推理时关闭，直接使用 h，不再缩放。

(d) 小 batch 使均值、方差噪声大，训练和运行统计可能失配。可冻结可靠的 running statistics，或改用 LayerNorm/GroupNorm；仅说“增大网络”不能解决统计量问题。

Rubric: 训练公式与 running update 3 分；推理统计及原因 2 分；均值、方差各 1 分；failure 与针对性处理各 1 分。

(b) **Conditional VAE。** [15 points]

(a) 插入任意 $q_\phi(z \mid x, c)$ 并用 Jensen：

$$\log p_\theta(x \mid c) = \log \mathbb{E}_{q_\phi} \frac{p(z \mid c)p_\theta(x \mid z, c)}{q_\phi(z \mid x, c)}$$
$$\geq \mathbb{E}_{q_\phi} \log p_\theta(x \mid z, c) - \mathrm{KL}\big(q_\phi(z \mid x, c) \| p(z \mid c)\big).$$

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

(b) 令 $\epsilon \sim \mathcal{N}(0, I)$，则 $z = \mu + \sigma \odot \epsilon$。闭式 KL 为

$$\frac{1}{2}\sum_j \left[\frac{\sigma_j^2 + (\mu_j - m_{c,j})^2}{\tau_{c,j}^2} - 1 + \log\frac{\tau_{c,j}^2}{\sigma_j^2}\right].$$

所有方差必须为正；实现中常由 log-variance 参数化。

(c) 训练时采样 (x, c)，encoder 输出 (μ, σ)，重参数化采样 z，计算 reconstruction log-likelihood 与 KL，反向传播更新 (θ, ϕ)。生成时给定 c，先采 $z \sim p(z \mid c)$，再采 $x \sim p_\theta(x \mid z, c)$；不能从训练 posterior 无条件生成。

(d) Posterior collapse 指 $q_\phi(z \mid x, c) \approx p(z \mid c)$，decoder 基本忽略 z。KL 权重过大时，减少 KL 的收益压过 latent 所带来的 reconstruction 改善。诊断应同时报告每维 KL/active units 与 reconstruction term，不能只看总 loss。KL annealing、free bits 或限制过强 decoder 均是合理处理。

Rubric: ELBO 方向与 gap 5 分；reparameterization 1 分、KL 3 分；训练/生成各项 3 分；collapse 定义、机制、诊断 3 分。

(c) **TD targets**。 [9 points]

(a) 非终止时

$$\begin{aligned} y_{\text{SARSA}} &= r + \gamma Q(s', a'), \\ y_{\text{Exp}} &= r + \gamma \sum_b \mu(b \mid s') Q(s', b), \\ y_Q &= r + \gamma \max_b Q(s', b). \end{aligned}$$

终止时三者均为 r，不得 bootstrap。

(b) 已实现 $a' = a_2$，故

$$\begin{aligned} y_{\text{SARSA}} &= 1 + 0.9(2) = 2.8, \\ y_{\text{Exp}} &= 1 + 0.9(0.25 \cdot 4 + 0.75 \cdot 2) = 3.25, \\ y_Q &= 1 + 0.9(4) = 4.6. \end{aligned}$$

(c) SARSA 和以行为策略取期望的 Expected SARSA 评估/改进当前行为策略，属于 on-policy。Q-learning 的 target 对下一动作取 greedy max，可由另一行为策略收集数据，属于 off-policy。SARSA 有下一动作采样噪声，Expected SARSA 对此取期望但仍有转移/奖励噪声，Q-learning 还含 maximization bias。Off-policy 不代表可以不探索；若关键 (s, a) 从不出现，其值无法被数据更新。

Rubric: 三个 target 及 terminal 3 分；三个数值各 1 分；policy 分类、噪声和 coverage 各 1 分。

C. 人工智能中的优化方法：参考解答 [33 points]

(1) [3 points] $g \in \partial F(x)$ 指对所有 $y \in C$，

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

取 $y = x^\star$ 并移项得 $F(x) - F(x^\star) \leq \langle g, x - x^\star \rangle$。

(2) [5 points] 因 $x^\star \in C$，$P_C(x^\star) = x^\star$。投影非扩张性与平方展开给出

$$\begin{aligned}\|x^{k+1} - x^\star\|^2 &\leq \|x^k - \eta g^k - x^\star\|^2 \\ &= \|x^k - x^\star\|^2 - 2\eta\langle g^k, x^k - x^\star\rangle + \eta^2\|g^k\|^2 \\ &\leq \|x^k - x^\star\|^2 - 2\eta[F(x^k) - F(x^\star)] + \eta^2 G^2.\end{aligned}$$

(3) [7 points] 将上式移项并对 $k = 0, \ldots, K-1$ 求和：

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

故

$$\frac{1}{K}\sum_{k=0}^{K-1}[F(x^k) - F(x^\star)] \leq \frac{R^2}{2\eta K} + \frac{\eta G^2}{2}.$$

平方距离差是 telescope 项，末项 $\|x^K - x^\star\|^2$ 非负，因而可丢弃。

(4) [5 points] 由 Jensen 不等式，

$$F(\bar{x}^K) \leq \frac{1}{K}\sum_{k=0}^{K-1} F(x^k),$$

所以

$$F(\bar{x}^K) - F(x^\star) \leq \frac{R^2}{2\eta K} + \frac{\eta G^2}{2}.$$

一般次梯度法不保证函数值单调，上述求和只控制平均 gap，因而不能无条件地把 $\bar{x}^K$ 换成 x^K。

(5) [4 points] 对 $B(\eta) = R^2/(2\eta K) + \eta G^2/2$ 求导，$B'(\eta) = -R^2/(2K\eta^2) + G^2/2$。最优值为 $\eta = R/(G\sqrt{K})$，两项各等于 $RG/(2\sqrt{K})$。因此取

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

足以保证精度 ε。

(6) [5 points] Φ 为 μ-strongly convex，唯一最小点为 0。令

$$a := \frac{\eta}{2 - \eta\mu} > 0.$$

在 $a > 0$ 处取次梯度 $1 + \mu a$，则

$$a - \eta(1 + \mu a) = -a.$$

在 $-a<0$ 处次梯度为 $-1-\mu a$，下一步为

$$-a-\eta(-1-\mu a)=a.$$

故 $a,-a,a,-a,\dots$ 是非零二周期。强凸性本身不能消除常步长对非光滑尖点的跨越。

(7) [4 points] 一般凸非光滑一阶 oracle 的标准平均点保证为 $O(K^{-1/2})$，而凸 L-smooth GD 可利用 descent lemma 得到 $O(K^{-1})$。本题缺失的正是 gradient Lipschitz/smoothness；有界次梯度不能代替它。平均把可能振荡的迭代点转化为可由 convexity 控制的输出，是证明的必要组成部分，不只是工程 heuristic。

Rubric: (1) 定义与代入最优点 3 分；(2) projection、平方展开、subgradient 不等式分别计分；(3) 移项 2 分、telescope 3 分、丢弃非负末项 2 分；(4) Jensen 3 分、last iterate 局限 2 分；(5) 步长与复杂度各 2 分；(6) 构造 a、两步验证、解释分别计分；(7) 两种 rate、缺失假设、averaging 解释共 4 分。

D. 自然语言处理：参考解答 [33 points]

1. Skip-gram with negative sampling [7 分]

令 $a_c = \mathbf{u}_w^\top \mathbf{v}_c$、$a_i = \mathbf{u}_w^\top \mathbf{v}_{n_i}$。利用 $\frac{d}{da}[-\log\sigma(a)] = \sigma(a) - 1$ 和 $\frac{d}{da}[-\log\sigma(-a)] = \sigma(a)$，得到

$$\boxed{\nabla_{\mathbf{u}_w}\ell = (\sigma(a_c) - 1)\mathbf{v}_c + \sum_{i=1}^{m}\sigma(a_i)\mathbf{v}_{n_i}.}$$

第一项在 gradient descent 下把 center 拉向 positive context；每个负项按当前混淆程度 $\sigma(a_i)$ 把 center 推离 negative，容易混淆的负例权重更大。

直接按 unigram frequency 抽样会让极高频 function words 占据大量 negative slots，提供的区分信号有限。可用 $q(w) \propto C(w)^{3/4}$ 后归一化、按频率分层抽样，或结合模型检索 hard negatives。False negative 实际与 center 语义相关，却被 loss 推远，会破坏局部语义结构并产生有偏梯度；可通过同义词/实体 ID mask、multiple-positive loss 或降低可疑负例权重缓解。

评分要点 完整梯度及符号 4 分；频率问题、合理修改、false-negative 后果 3 分。

2. Causal attention and autoregressive decoding [7 分]

取 $Q, K \in \mathbb{R}^{L\times d_k}$、$V \in \mathbb{R}^{L\times d_v}$，定义

$$A = \mathrm{softmax}_{\text{key axis}}\left(\frac{QK^\top}{\sqrt{d_k}} + M\right), \qquad AV,$$

其中 $M_{ij} = 0$ 当 $j \le i$，$M_{ij} = -\infty$ 当 $j > i$。Softmax 对每一 query 所在行的 key index j 归一化。

若第 t 步把长度 t 的 prefix 全部重新前向计算，一层的 dense projections 与 self-attention 分别为 $O(td^2)$ 与 $O(t^2d)$，故总计 $O(T^2d^2 + T^3d)$。KV cache 保存旧 K, V，第 t 步只生成新 query/key/value，代价为 $O(d^2)$，新 query 与 t 个 keys 的 attention 为 $O(td)$；总计 $O(Td^2 + T^2d)$，cache 为 $O(Td)$（固定层数/heads 下）。若只统计 attention score/value products，上述两种时间分别简化为 $O(T^3d)$ 与 $O(T^2d)$。它去掉了对旧 prefix 的重复计算并把总时间降一阶，但 exact dense causal attention 仍需让每个新 token 查看所有历史 token，所以总 pairwise attention 并未变成线性时间。

评分要点 mask、维度和 softmax axis 3 分；两种时间及 cache memory 3 分；正确解释 asymptotic 限制 1 分。

3. Compute-constrained scaling [8 分]

由 $D = C/(\kappa N)$，忽略 $\mathcal{L}_\infty$ 后最小化

$$f(N) = AN^{-\alpha} + B\left(\frac{\kappa N}{C}\right)^{\beta}.$$

一阶条件为

$$\alpha A N^{-\alpha} = \beta B \kappa^\beta N^\beta C^{-\beta}.$$

因此

$$\boxed{N^*(C) = \left(\frac{\alpha A}{\beta B \kappa^\beta}\right)^{1/(\alpha+\beta)} C^{\beta/(\alpha+\beta)}}$$

且

$$\boxed{D^*(C) = \frac{C}{\kappa N^*(C)} \propto C^{\alpha/(\alpha+\beta)}}.$$

二阶行为或两端发散说明该 stationary point 是全局最小点。最优点的平衡关系是

$$\alpha\, A(N^*)^{-\alpha} = \beta\, B(D^*)^{-\beta};$$

两项不一定数值相等，除非 $\alpha = \beta$。远距离外推会因 data quality/重复数据改变有效 D、optimization error 未包含在式中、architecture/optimizer 变化、irreducible loss 改变或硬件利用率使 $C \not\propto ND$ 而失效，任答两点即可。

评分要点 约束代入与一阶条件 2 分；两个闭式/幂次 3 分；balance 1 分；两项有效外推限制 2 分。

4. Evidence-preserving long-document summarization [11 分]

一种 full-credit 方案是先按自然段切为 M 个 overlapping chunks，得到 embedding z_i、metadata 和句子 offsets。Issue query 得到 z_q，粗排用 $a_i = z_q^\top z_i$，再用轻量 cross-encoder 产生 b_i。在约束 $\sum_i m_i \ell_i \le B$、$m_i \in \{0,1\}$ 下选择兼顾 relevance、diversity 与 chronology 的 chunks，例如最大化

$$\sum_i m_i b_i - \rho \sum_{i<j} m_i m_j \,\mathrm{sim}(z_i, z_j).$$

对入选 chunk 先生成带 span IDs 的局部 proposition，再由 generator 合并；claim verifier 删除或改写没有 supporting span 的结论。

只有 document-level summary 时，可把 summary sentences 当 query，以 lexical/dense match 形成 noisy positive chunks，训练 selector 的 contrastive 或 listwise ranking loss；也可把 selector 当 latent variable，用 summary likelihood 的 variational/REINFORCE surrogate。为避免 selector 总选开头或重复 chunk，可加入 coverage/diversity regularization、minimum section coverage、hard negatives，并在一小批人工 span labels 上校准。

预先计算 chunk embeddings 后，ANN selection 近似 sublinear search，精排约 $O(kd)$ 到 cross-encoder 的 token cost；若扫描全部 embeddings 则为 $O(Md)$。Generator 的 dominant attention 约为 $O(B^2 d)$（dense case），输入 memory 为 $O(Bd)$；索引 memory 为 $O(Md)$。

Coverage 可用 gold key-span recall 或 question-conditioned content-unit recall；faithfulness 用 atomic-claim support rate、citation precision/recall 和人工 factuality。Adversarial test 可插入一段措辞高度相关但法律结论相反或已废止的文本，检查 selector、version filter 和 verifier 是否会被 lexical overlap 欺骗。

评分要点 分层选择、变量与预算约束 4 分；可训练目标及防退化 3 分；复杂度 2 分；coverage/faithfulness 指标和 adversarial test 2 分。其他满足约束的设计同等给分。