Artificial Intelligence PhD Qualifying Examination

模拟考试 07：参考解答与评分要点

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]

(a) 给定 query x，按 metric distance 排序训练输入，取最近 k 个 labels 的 majority；平票可固定预测 0（任何预先声明的 deterministic rule 均可）。Inductive bias 是邻近输入有相似 conditional label distribution。增大 k 通常通过 averaging 降 variance，但纳入更远点会增 locality bias；这不是任意分布下的单调定理。Storage 是 $O(md)$；lazy training 可为 $O(1)$，naive query 计算全部距离为 $O(md)$，再作 selection。Normalization 与 metric 改变“邻近”本身，必须只用 training information 拟合。

Rubric: definition/tie 2 分；inductive bias 1 分；bias–variance 2 分；complexity 2 分；metric/leakage 1 分。

(b) 令 $\Delta = \sup_h |L(h) - \widehat{L}(h)|$。由 $e^{\lambda \max_j a_j} \le \sum_j e^{\lambda a_j}$，对 N 个 hypotheses 和两个 signs，

$$\mathbb{E}e^{\lambda\Delta} \le 2Ne^{\lambda^2/(8m)}.$$

Jensen 给

$$\mathbb{E}\Delta \le \frac{\log(2N)}{\lambda} + \frac{\lambda}{8m}.$$

取 $\lambda = \sqrt{8m\log(2N)}$ 得题给 bound。若 $\widehat{h}$ 是 ERM、$h^\star$ 最小化 population risk，则逐 sample $L(\widehat{h}) - L(h^\star) \le 2\Delta$，所以

$$\boxed{\mathbb{E}[L(\widehat{h}) - L(h^\star)] \le \sqrt{\frac{2\log(2N)}{m}}.}$$

Rubric: max-to-sum 2 分；MGF sum 1 分；Jensen/optimization 2 分；ERM comparison 2 分。

(c) Standard basis points $e_1, \dots, e_d$ 被 shattered：给 labels $y \in \{0,1\}^d$，取 $w = y$，则 $h_w(e_i) = w_i = y_i$。另一方面 class size 是 2^d，所以 VC dimension 至多 $\log_2(2^d) = d$。故答案为 d。

Rubric: 点集 2 分；任意 labeling 3 分；finite-size upper bound 2 分；结论 1 分。

(d) Integral representation 给

$$\nabla F(x) - \nabla F(y) = A_{x,y}(x-y), \qquad A_{x,y} = \int_0^1 \nabla^2 F(y + t(x-y))\, dt.$$

仍有 $\mu I \preceq A_{x,y} \preceq LI$。因此

$$T(x) - T(y) = (I - \eta A_{x,y})(x-y).$$

当 $\eta = 2/(L+\mu)$，对任意 eigenvalue $a \in [\mu, L]$，$|1 - \eta a| \le (L-\mu)/(L+\mu)$。取 operator norm 即得结论。

Rubric: integral Hessian 2 分；spectral interval 1 分；mapping difference 1 分；endpoint spectral maximum 2 分。

(e) 每个 $Z_i \in [0,1]$ 满足 $\mathrm{Var}(Z_i) \le 1/4$。Independence 给

$$\mathrm{Var}(\widehat{\mu}) = \frac{1}{m^2}\sum_{i} \mathrm{Var}(Z_i) \le \frac{1}{4m}.$$

Chebyshev 因而给 $\Pr\{|\widehat{\mu} - \bar{\mu}| > \varepsilon\} \le 1/(4m\varepsilon^2)$。充分条件为

$$\boxed{m \ge \frac{1}{4\delta\varepsilon^2}.}$$

Rubric: bounded variance 1 分；independent sum 1 分；Chebyshev 1 分；样本量 1 分。

B. 深度学习与强化学习：参考解答 [33 points]

(a) **Multi-head attention。** [10 points]

(a) $W_j^Q, W_j^K, W_j^V \in \mathbb{R}^{d\times d_k}$，$Q_j, K_j, V_j \in \mathbb{R}^{L\times d_k}$，$A_j \in \mathbb{R}^{L\times L}$，$A_jV_j \in \mathbb{R}^{L\times d_k}$。拼接后为 $L \times (hd_k) = L \times d$，输出投影 $W^O \in \mathbb{R}^{d\times d}$，最终仍为 $L \times d$。

(b) 对至少有一个 allowed key 的 row，masked logits 在 allowed 位置有限、forbidden 位置指数为零；row-softmax 因此满足 $A_{i\ell} \geq 0$ 且 $\sum_\ell A_{i\ell} = 1$。输出 $o_i = \sum_{\ell\in\mathcal{N}(i)} A_{i\ell}v_\ell$ 是 allowed values 的 convex combination。若整行全 masked，归一化分母为零（实现常出现 NaN），必须保证 self-key/valid key 或显式处理空行。

(c) 若 q_r, k_r 独立、均值零、方差一，则 $\mathrm{Var}(q^\top k) = \sum_{r=1}^{d_k} \mathrm{Var}(q_r k_r) = d_k$。除以 $\sqrt{d_k}$ 后方差约为 1，避免大维度 logits 使 softmax 饱和及 Jacobian 变小。

(d) 主时间为 $O(Ld^2 + L^2d)$，attention matrix memory 为 $O(hL^2)$（常数/实现可融合）。固定 d 增加 heads 会减小每头 d_k、提供不同子空间，但每头容量更小且可能头冗余。

Rubric: 全部维度 3 分；normalization/convexity/empty row 3 分；variance 与 scaling 2 分；complexity/tradeoff 2 分。

(b) **Affine coupling flow。** [13 points]

(a)

$$x_A = y_A, \qquad x_B = (y_B - t_\theta(y_A)) \odot \exp(-s_\theta(y_A)).$$

$\exp(s)$ 每坐标严格为正；给定未变的 $x_A = y_A$ 后即可反解。s, t 本身无需可逆。

(b)

$$\frac{\partial y}{\partial x} = \begin{bmatrix} I & 0 \\ * & \mathrm{diag}(\exp(s_\theta(x_A))) \end{bmatrix}, \quad \log\left|\det \frac{\partial y}{\partial x}\right| = \sum_j s_{\theta,j}(x_A).$$

左下块虽复杂，但不影响三角矩阵 determinant。

(c) 若 $x = f_L \circ \cdots \circ f_1(z_0)$，$z_0 \sim p_0$，则

$$\log p_\theta(x) = \log p_0(z_0) - \sum_{\ell=1}^{L} \log|\det J_{f_\ell}(z_{\ell-1})|.$$

平均 NLL 为

$$\frac{1}{N}\sum_i \left[-\log p_0(z_0^{(i)}) + \sum_\ell \log|\det J_{f_\ell}(z_{\ell-1}^{(i)})|\right],$$

其中训练数据先经各层 inverse 得到 $z_0^{(i)}$。

(d) 单层不改变 x_A；若每层固定同一 split，这些坐标不能被充分变换。层间可交换 split、permutation/channel shuffle，或加入可逆线性 mixing。

(e) 可限制 s 的范围（如 $s = c\tanh\widetilde{s}$）或正则化，以避免 $\exp(s)$ overflow、underflow 和病态 Jacobian。

Rubric: inverse 3 分；block Jacobian/logdet 4 分；density/NLL sign 3 分；mixing 2 分；稳定性 1 分。

(c) **DQN。** [10 points]

(a)

$$y = r + \gamma(1-d)\max_b Q_{\bar{\omega}}(s', b),$$
$$\mathcal{L}(\omega) = \frac{1}{2}\big(Q_\omega(s,a) - \mathrm{stopgrad}(y)\big)^2.$$

terminal 时 $y = r$。Target 不反传；否则成为不同的 residual-gradient 问题。

(b) 对每个固定 a，$\max_b \widehat{Q}_b \geq \widehat{Q}_a$。取期望得 $\mathbb{E}\max_b \widehat{Q}_b \geq \mathbb{E}\widehat{Q}_a = Q_a$，再对 a 取最大即得 $\mathbb{E}\max_b \widehat{Q}_b \geq \max_a Q_a$。若不同动作噪声使非最优动作偶尔取得最大，通常严格。

(c)

$$y_{\mathrm{DDQN}} = r + \gamma(1-d)Q_{\bar{\omega}}\left(s', \arg\max_b Q_\omega(s', b)\right).$$

online network 选择动作，target network 评价该动作，从而减少同一噪声同时 selection/evaluation 造成的偏差。

(d) Replay 随机抽取旧 transition，打破强时间相关、提高数据复用并混合访问分布；target network 在一段时间内固定 bootstrap 目标，减小 moving-target feedback。二者均不自动保证一般 nonlinear off-policy 学习收敛。

Rubric: loss/terminal/stopgrad 3 分；不等式证明与严格性 3 分；Double target 2 分；两个机制各 1 分。

C. 人工智能中的优化方法：参考解答 [33 points]

(1) [3 points] 有效学习率为 η/h_t。$\delta > 0$ 保证初期即使所有梯度为零也不会除零。又因

$$h_t^2 - h_{t-1}^2 = \|g^t\|^2 \geq 0,$$

所以 h_t 非递减，而有效学习率非递增。

(2) [5 points] 因 $u \in X$ 且投影非扩张，

$$\begin{aligned}\|x^{t+1} - u\|^2 &\leq \left\| x^t - \frac{\eta}{h_t} g^t - u \right\|^2 \\ &= \|x^t - u\|^2 - \frac{2\eta}{h_t}\langle g^t, x^t - u\rangle + \frac{\eta^2}{h_t^2}\|g^t\|^2.\end{aligned}$$

移项并乘 $h_t/(2\eta)$ 即得题设式。

(3) [5 points] 记 $a_t = \|x^t - u\|^2$。分部求和给出

$$\begin{aligned}\sum_{t=1}^{T} h_t(a_t - a_{t+1}) &= h_1 a_1 + \sum_{t=2}^{T}(h_t - h_{t-1})a_t - h_T a_{T+1} \\ &\leq D^2\left[h_1 + \sum_{t=2}^{T}(h_t - h_{t-1})\right] \\ &= D^2 h_T.\end{aligned}$$

这里使用 $a_t \leq D^2$、$h_t - h_{t-1} \geq 0$，并丢弃末尾非正项。

(4) [4 points]

$$h_t - h_{t-1} = \frac{h_t^2 - h_{t-1}^2}{h_t + h_{t-1}} = \frac{\|g^t\|^2}{h_t + h_{t-1}} \geq \frac{\|g^t\|^2}{2h_t}.$$

所以

$$\frac{\|g^t\|^2}{h_t} \leq 2(h_t - h_{t-1}).$$

对 t 求和得到 $2(h_T - h_0) = 2(h_T - \delta) \leq 2h_T$。

(5) [6 points] 凸性给出 $f_t(x^t) - f_t(u) \leq \langle g^t, x^t - u\rangle$。对第 (2) 问求和，再使用第 (3)–(4) 问：

$$\begin{aligned}\mathrm{Regret}_T(u) &\leq \frac{1}{2\eta}\sum_{t=1}^{T} h_t(a_t - a_{t+1}) + \frac{\eta}{2}\sum_{t=1}^{T}\frac{\|g^t\|^2}{h_t} \\ &\leq \frac{D^2 h_T}{2\eta} + \eta h_T.\end{aligned}$$

代入 h_T 定义即得目标式。

(6) [5 points] $D^2/(2\eta) + \eta$ 在

$$\eta = \frac{D}{\sqrt{2}}$$

处最小，最小值为 $\sqrt{2}D$。若 $\|g^t\| \leq G$，

$$\frac{\mathrm{Regret}_T(u)}{T} \leq \frac{\sqrt{2}D}{T}\sqrt{\delta^2 + TG^2}.$$

当 $f_t \equiv F$ 时，由 Jensen，

$$F(\bar{x}^T) - F(u) \le \frac{1}{T}\sum_{t=1}^{T}[F(x^t) - F(u)] \le \frac{\mathrm{Regret}_T(u)}{T}.$$

(7) [5 points] 时变损失下没有一个固定目标要求点列收敛；regret 只与固定 comparator 比较累积损失。其次，h_t 只增不减，早期异常大的梯度会长期压低之后的有效步长，这是无遗忘 accumulator 的典型缺点。最后，AdaGrad 仍需选择 base rate η 和稳定参数 δ；adaptive 指历史梯度控制缩放，而非不存在超参数。

Rubric: (1) 三项各 1 分；(2) projection expansion 3 分、移项常数 2 分；(3) summation by parts 3 分、diameter/符号 2 分；(4) rationalization 2 分、telescope 2 分；(5) convexity 1 分、两个已证界 3 分、最终常数 2 分；(6) 最优 η 2 分、bounded-gradient 1 分、online-to-batch 2 分；(7) 三项限制分别 2、2、1 分。

D. 自然语言处理：参考解答 [33 points]

1. Learning a reward model from preferences [8 分]

令 $\Delta = r_+ - r_-$。单对数据的 NLL 为

$$\ell = -\log \sigma(\Delta).$$

因为 $d[-\log\sigma(a)]/da = -\sigma(-a)$，

$$\boxed{\frac{\partial \ell}{\partial r_+} = -\sigma(-\Delta), \qquad \frac{\partial \ell}{\partial r_-} = \sigma(-\Delta).}$$

Gradient descent 提高 preferred reward、降低 dispreferred reward，且已大幅正确排序时梯度变小。

若 $r'(x,y) = r(x,y) + c(x)$，则 $r'(x,y^+) - r'(x,y^-) = \Delta$，likelihood 不变。因此仅有同一 prompt 内的 reward differences/order 可识别；绝对 offset 不可识别，除非加入额外 anchor/regularizer。

数据设计可随机交换两个 responses 的屏幕左右/上下位置并记录 order，使用多位 annotators 后按 annotator reliability 建模，或混入重复与 gold comparison 检测偏差。只要明确减少某一 bias 且可执行即可。

评分要点 NLL 与两个导数 4 分；不变性与 identifiable quantity 2 分；合理去偏设计 2 分。

2. KL-regularized policy optimization [9 分]

对 normalization 加 Lagrange multiplier λ：

$$\mathcal{J} = \sum_y \pi_y r_y - \beta \sum_y \pi_y \log \frac{\pi_y}{\pi_{\mathrm{ref},y}} + \lambda \left(\sum_y \pi_y - 1 \right).$$

内点一阶条件为

$$r_y - \beta \left(\log \frac{\pi_y}{\pi_{\mathrm{ref},y}} + 1 \right) + \lambda = 0.$$

故

$$\boxed{\pi^*(y \mid x) = \frac{\pi_{\mathrm{ref}}(y \mid x) \exp(r(x,y)/\beta)}{Z(x)}}, \qquad Z(x) = \sum_{y'} \pi_{\mathrm{ref}}(y' \mid x) e^{r(x,y')/\beta}.$$

目标对 π 严格凹（在支持内部），所以该 stationary point 为唯一最优解。

$\beta \to 0$ 时 policy 集中到 reference support 内的最高 reward responses（并列时按 reference mass 分配）；$\beta \to \infty$ 时 $e^{r/\beta} \to 1$，故 $\pi^* \to \pi_{\mathrm{ref}}$。若 reference 对某 response 概率为零，forward KL 使任何在该点分配正质量的 policy 代价无穷，reward 再高也无法产生该 response；这就是 support assumption 的作用。

评分要点 Lagrangian 与一阶条件 3 分；归一化 Gibbs policy 3 分；两个极限 2 分；support 解释 1 分。

3. Noisy preference labels [6 分]

观察为“y_1 胜”的两条互斥路径是 true label 未翻转或 true label 为负但翻转：

$$\boxed{\widetilde{p} = (1-\eta)p + \eta(1-p) = \eta + (1-2\eta)p.}$$

它把 p 向 1/2 收缩；当 $\eta \uparrow 1/2$ 时，$\widetilde{p} \to 1/2$，信号消失、有效样本复杂度上升。若忽略 noise，reward margin 通常被压小或 confidence calibration 失真。

可用重复 annotation/已知 gold pairs 估计 η，在 likelihood 中显式使用上述 noise channel；也可用 annotator-specific confusion matrices、majority/EM aggregation 或过滤低 agreement pairs。

评分要点 全概率推导与公式 3 分；收缩/极限 2 分；可行估计或修正 1 分。

4. Auditing an aligned language model [10 分]

Representative protocol：构建互不泄漏的 in-domain、temporal holdout、adversarial/red-team 和 long-tail multilingual prompts；比较 base、SFT、aligned policy 与 reference。指标至少分为：(i) blind pairwise human preference/Bradley–Terry win rate；(ii) task correctness、factuality、citation support；(iii) safety/toxicity/privacy；另报 KL-to-reference、length、refusal、latency 与 reward-model score。关键是检查 human/correctness 提升是否与 reward score 同步；仅 reward 升而 blind human 或 factuality 降是 reward hacking 信号。

Distribution shift test 使用训练截止日之后的 prompts、未见 domain/language 和 counterfactual templates，并按 shift bucket 报 calibration。Diversity test 对同一 prompt 多次采样，报告 distinct-n、self-BLEU/embedding dispersion、semantic mode coverage，并控制 answer quality。

Human A/B 顺序随机、model identity 与 reward 隐藏、同 prompt 同 decoding budget；多 annotators、重复题和 inter-rater agreement 用于估计噪声，不把 annotator 身份差异静默平均。

Rollback/abstain triggers 可包括：shift set 的 safety upper confidence bound 超阈值；reward–human disagreement 显著；calibrated uncertainty/ensemble disagreement 高；KL、mode collapse 或 privacy incidents 超预算。触发后回到稳定 checkpoint、路由人工或输出安全 abstention，并保留审计日志。

评分要点 datasets/baselines/三类指标 4 分；shift 与 diversity 测试 2 分；randomized blind human eval 2 分；含 uncertainty 的两个 trigger 2 分。其他能区分 reward hacking 的完整方案同等给分。