

Artificial Intelligence PhD Qualifying Examination

模拟考试 04: 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. (i)a, (ii)a, (iii)a, (iv)a, (v)a, (vi)a。

Part II.

(a) (a) Lagrangian 为

$$\mathcal{L} = \frac{1}{2} \|w\|^2 + \sum_i \alpha_i [1 - y_i(w^\top x_i + b)], \quad \alpha_i \geq 0.$$

Stationarity 给 $w = \sum_i \alpha_i y_i x_i$ 与 $\sum_i \alpha_i y_i = 0$ 。代回得

$$\max_{\alpha \geq 0} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^\top x_j, \quad \sum_i \alpha_i y_i = 0.$$

(b) 两个 primal constraints 均要求 $w_1 \geq 1$ ；最小 norm 给 $w = (1, 0), b = 0$ 。几何 margin 是 $1/\|w\| = 1$ 。Dual equality 令 $\alpha_1 = \alpha_2 = a$ ，而 $w = 2a(1, 0)$ ，故 $a = 1/2$ 。Primal 与 dual values 均为 $1/2$ 。

Rubric: Lagrangian 1 分；stationarity 2 分；dual 2 分；toy primal 1 分；dual variables 1 分；margin/value 2 分。

(b) 记 $U = r(X, Y)\ell(h(X), Y)$ 。Change of measure 给

$$\mathbb{E}_Q U = \mathbb{E}_P \ell = L_P(h),$$

故 sample mean unbiased。又因 $0 \leq r \leq W$ 、 $0 \leq \ell \leq 1$ ，

$$U^2 = r^2 \ell^2 \leq W r \ell.$$

所以 $\mathbb{E}_Q U^2 \leq W L_P(h)$ ，从而

$$\text{Var}(\widehat{L}_{\text{IW}}) = \frac{\text{Var}(U)}{m} \leq \frac{\mathbb{E} U^2}{m} \leq \frac{W L_P(h)}{m}.$$

Rubric: change of measure 2 分；unbiasedness 1 分；pointwise square bound 2 分；variance scaling 1 分。

(c) 两点 $u = (1, 0), v = (0, 1)$ 可 shattered：阈值 (a, b) 分别选择低于/高于对应坐标即可实现四种 labeling，故 VC dimension 至少 2。

任取三个点。若其中 $x \preceq y$ coordinatewise，则所有 lower orthants 满足 $h(y) = 1 \Rightarrow h(x) = 1$ ，该 pair 不能任意 labeling。若三点 pairwise incomparable，按第一坐标递增排列后第二坐标严

格递减；令中间点为 $x^{(2)}$ 。任何 orthant 若包含两个端点，就必须有 $a \geq x_1^{(3)}$ 且 $b \geq x_2^{(1)}$ ，因而也包含中间点。所以端点标 1、中间标 0 无法实现。故任何三点都不能 shattered，

$$\boxed{\text{VCdim} = 2.}$$

Rubric: 两点下界 2 分；dominated case 2 分；antichain ordering 2 分；不可实现 labeling 与结论 2 分。

(d) Convex differentiable g 的 gradient monotonicity 给

$$\langle \nabla g(u) - \nabla g(v), u - v \rangle \geq 0.$$

Quadratic 项贡献 $\lambda \|u - v\|^2$ ，相加即得所求。若 $\nabla F(u) = \nabla F(v) = 0$ ，左端为 0，题式迫使 $u = v$ 。

Rubric: convex-gradient monotonicity 2 分；quadratic contribution 1 分；唯一性 1 分。

(e) 可取 penalty

$$2\sqrt{\frac{\log(2r/\delta)}{2m_v}},$$

即 validation ERM 的 true risk 至多候选最佳 true risk 加该项，概率至少 $1 - \delta$ 。候选列表必须在查看 validation labels 前固定（可条件于独立 training data）。

Rubric: penalty 2 分；independence 1 分。

B. 深度学习与强化学习：参考解答

[33 points]

(a) Self-attention 的置换结构。 [11 points]

- (a) 有 $Q(PX) = PQ(X)$, $K(PX) = PK(X)$, $V(PX) = PV(X)$, 故 score matrix 变为 $PSP^\top$ 。对任一行, 右乘 $P^\top$ 只重排该行的 key logits, 左乘 P 再重排行; softmax 对坐标重排满足 $\text{softmax}(uP^\top) = \text{softmax}(u)P^\top$ 。因此

$$\text{softmax}_{\text{row}}(PSP^\top) = P \text{softmax}_{\text{row}}(S)P^\top.$$

于是

$$F(PX) = PA(X)P^\top PV(X) = PF(X).$$

- (b) 这是 permutation equivariance: 输入 token 重排, 输出按相同方式重排, 而非保持不变。若随后作 sum/mean pooling, 则 $\mathbf{1}^\top PF(X) = \mathbf{1}^\top F(X)$, 整体 pooled representation 才是 invariant。
- (c) 加绝对位置 E 后通常 $PX + E \neq P(X + E)$, 故任意内容重排不再保持上述对称性。Causal mask 由位置顺序决定, 也只保留与该有序图相容的对称。若 token、位置编码和 mask 同时被一致地重新标号, 那只是计算图 relabeling, 不代表模型忽略顺序。
- (d) 线性投影主项为 $O(Ld^2)$, score 与 value aggregation 为 $O(L^2d)$; attention matrix memory 为 $O(L^2)$, 另有 activation/projection memory。

Rubric: 三个张量变换、softmax 重排、最终等式共 5 分; equivariant/invariant 2 分; position 与 mask 2 分; time/memory 2 分。

(b) Masked autoregressive flow。 [13 points]

(a) 逆式为

$$x_i = m_i(x_{<i}) + \exp\{s_i(x_{<i})\}z_i.$$

生成时已有 $x_{<i}$ 后可顺序求 x_i 。因为 z_i 只依赖 $x_{\leq i}$, $\partial z/\partial x$ 为下三角, 且对角项 $\partial z_i/\partial x_i = \exp\{-s_i(x_{<i})\}$ 。故

$$\log \left| \det \frac{\partial z}{\partial x} \right| = - \sum_i s_i(x_{<i}).$$

(b) 变量替换给出

$$\log p_\theta(x) = \log p_0(z) - \sum_{i=1}^d s_i(x_{<i}).$$

因此

$$\mathcal{L}_{\text{NLL}} = \frac{1}{N} \sum_{n=1}^N \left[-\log p_0(z^{(n)}) + \sum_i s_i(x_{<i}^{(n)}) \right].$$

这里 flow 已直接给 inverse Jacobian, 故 density 中是 $-\sum_i s_i$ 。

- (c) 给定完整数据 x , 所有 prefix 信息均已知, masked feed-forward network 可一次并行输出全部 m_i, s_i 。采样时 m_i, s_i 依赖尚未生成的 $x_{<i}$, 故必须 ancestral 顺序生成。
- (d) m_i, s_i 应为有限可测/可微函数, $\exp s_i > 0$ 保证每个坐标映射可逆; masked dependency 保证整体双射。若 p_0 已归一化, 变量替换保证 p_θ 归一化。数值上还应控制过大的 s_i 。

Rubric: inverse、三角性、logdet 5 分；density 与平均 NLL 4 分；并行/顺序 2 分；可逆和归一化 2 分。

(c) **Policy improvement.** [9 points]

(a) 由 greedy 定义，逐状态有

$$(T^{\pi'} V^{\pi})(s) = \sum_a \pi'(a | s) Q^{\pi}(s, a) = \max_a Q^{\pi}(s, a) \geq V^{\pi}(s).$$

$T^{\pi'}$ 单调，故反复应用得到 $V^{\pi} \leq T^{\pi'} V^{\pi} \leq (T^{\pi'})^2 V^{\pi} \leq \dots$ 。由 $T^{\pi'}$ 的 contraction，右侧收敛到其唯一不动点 $V^{\pi'}$ ，所以 $V^{\pi'} \geq V^{\pi}$ 。

(b) 多个动作可并列最大，任意混合或选择均可成为 π' ，故不唯一。一次 greedy step 只相对当前 Q^{π} 改进；新策略改变未来访问与价值，通常还需再次 evaluation/improvement。

(c) Policy iteration: 精确解 $V^{\pi_k} = T^{\pi_k} V^{\pi_k}$ ；令 π_{k+1} 对 Q^{π_k} greedy；若各状态原策略已为 greedy（或策略不再变化），停止，此时满足 Bellman optimality，得到最优策略。

Rubric: 一步不等式、单调迭代、极限各有分共 5 分；非唯一/非一步最优 2 分；完整算法与停止条件 2 分。

C. 人工智能中的优化方法：参考解答

[33 points]

- (1) [4 points]
- $\nabla f(x) = Qx + c$
- 且
- $Q \succ 0$
- , 故

$$x^* = -Q^{-1}c.$$

利用 $Qx^* + c = 0$, 更新式减去 x^* 得

$$e^{k+1} = [(1 + \beta)I - \gamma Q]e^k - \beta e^{k-1}.$$

- (2) [5 points] 取
- $Q = U\Lambda U^\top$
- , 令
- $z^k = U^\top e^k$
- . 每个坐标独立满足

$$z_i^{k+1} = (1 + \beta - \gamma\lambda_i)z_i^k - \beta z_i^{k-1}.$$

代入试探解 $z_i^k = r^k$, 得到

$$r^2 - (1 + \beta - \gamma\lambda_i)r + \beta = 0.$$

- (3) [6 points] 置
- $a = 1 + \beta - \gamma\lambda$
- . Jury 的三项化为

$$|\beta| < 1, \quad 1 - a + \beta = \gamma\lambda > 0, \quad 1 + a + \beta = 2 + 2\beta - \gamma\lambda > 0.$$

在 $0 \leq \beta < 1, \lambda > 0$ 下, 单一方向稳定等价于 $0 < \gamma\lambda < 2(1 + \beta)$. 要同时覆盖 $\lambda \in [\mu, L]$, 上端最紧, 故充要条件为

$$0 < \gamma < \frac{2(1 + \beta)}{L}, \quad 0 \leq \beta < 1.$$

- (4) [6 points] 记
- $s = \sqrt{L}, t = \sqrt{\mu}$
- , 则
- $q = (s - t)/(s + t)$
- . 直接计算

$$a_\mu = 1 + \frac{(s - t)^2}{(s + t)^2} - \frac{4t^2}{(s + t)^2} = 2q,$$

$$a_L = 1 + \frac{(s - t)^2}{(s + t)^2} - \frac{4s^2}{(s + t)^2} = -2q.$$

 a_λ 关于 λ 线性, 所以其在区间上的值介于两个端点, 即 $|a_\lambda| \leq 2q$.

- (5) [5 points] 对内点若
- $|a_\lambda| < 2q$
- , 判别式
- $a_\lambda^2 - 4q^2 < 0$
- , 两根为共轭复数; 其乘积为
- q^2
- , 所以两根模长均为
- q
- . 在
- $\lambda = \mu$
- 处多项式为
- $(r - q)^2$
- ; 在
- $\lambda = L$
- 处为
- $(r + q)^2$
- . 端点解可带
- kq^k
- 或
- $k(-q)^k$
- 的多项式前因子, 但渐近谱半径仍为
- q
- .

- (6) [4 points]

$$q_{\text{HB}} = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} = 1 - \Theta(\kappa^{-1/2}), \quad q_{\text{GD}} = \frac{\kappa - 1}{\kappa + 1} = 1 - \Theta(\kappa^{-1}).$$

因此在固定 SPD 二次型类上, 达到相对精度的渐近迭代数分别为 $O(\sqrt{\kappa} \log(1/\varepsilon))$ 与 $O(\kappa \log(1/\varepsilon))$.

- (7) [3 points] 稳定根可以为负数或共轭复数, 因此迭代可越过最优点或振荡收缩, 函数值未必逐步下降. 本证明依赖固定 Hessian 的同一特征分解; 一般非二次目标的 Hessian 随位置变化, 非凸时还可能有负曲率, 故不能把这里的参数区间直接推广成普适非凸收敛定理。

Rubric: (1) 最优点 1 分、误差递推 3 分; (2) 特征分解 2 分、标量递推与多项式 3 分; (3) 三个 Jury 条件 3 分、统一区间 3 分; (4) 两端点各 2 分、线性区间结论 2 分; (5) 内点模长 2 分、端点重根 2 分、谱半径 1 分; (6) 两因子与复杂度各 2 分; (7) 两项解释共 3 分。

D. 自然语言处理：参考解答**[33 points]****1. Tokenization and perplexity [7 分]**

对长度 n 的 token sequence, 单例 perplexity 为 $p(x)^{-1/n}$ 。因此

$$\text{PPL}_A = 0.01^{-1/2} = \boxed{10}, \quad \text{PPL}_B = 0.01^{-1/5} = 10^{0.4} \approx \boxed{2.512}.$$

两模型对完整 character string 给出完全相同的概率, 差异只来自 denominator 中 token 数; 故不能据较低 per-token PPL 判 B 更好。

可报告每 character/byte 的 negative log-likelihood, 例如 bits per byte $-\log_2 p(x)/(\# \text{bytes})$, 并保证相同 text normalization。它仍受 evaluation domain、Unicode encoding、模型是否对同一字符串空间正规化以及 context truncation 的影响, 故不是消除所有差异的万能指标。

评分要点 两个数值 3 分; token denominator 解释 2 分; 合理跨 tokenizer normalization 及 limitation 2 分。

2. HMM sequence labeling [10 分]

若初始分布为 π , 则

$$p(x_{1:L}, y_{1:L}) = \pi_{y_1} B_{y_1}(x_1) \prod_{t=2}^L A_{y_{t-1}, y_t} B_{y_t}(x_t).$$

Log-domain Viterbi 为

$$\delta_1(j) = \log \pi_j + \log B_j(x_1),$$

$$\delta_t(j) = \log B_j(x_t) + \max_i \{\delta_{t-1}(i) + \log A_{ij}\}.$$

保存 argmax backpointer 并回溯; dense transition 的时间是 $O(LK^2)$ 。

未知 token 的两种合理处理包括:

- 把低频词统一映射到 UNK, 对 $B_j(\text{UNK})$ 做 additive smoothing; 偏置是同一 label 下所有未知词近似可交换, 稳健但丢失形态;
- 用 character/subword features 建 $B_j(x) \propto \exp(w_j^\top \phi(x))$ 或按 suffix/case/script backoff; 偏置是共享词形规律, 使未见词可泛化, 但可能把表面相似误当语义相似。

还可答 class-conditional neural emission, 只要概率归一化和 inductive bias 说明清楚。

评分要点 joint 3 分; Viterbi、backpointer 与复杂度 4 分; 两种未知词方案及偏置 3 分。

3. Evaluation under class imbalance [6 分]

Person 的 precision=recall=90/100 = 0.9, 故 $F1_P = 0.9$ 。Chemical 的 precision=2/3、recall=2/10 = 0.2, 所以

$$F1_C = \frac{2TP}{2TP + FP + FN} = \frac{4}{13} \approx 0.3077.$$

于是

$$\text{macroF1} = \frac{0.9 + 4/13}{2} \approx \boxed{0.6038}.$$

合并 counts 得 $TP = 92, FP = 11, FN = 18$,

$$\text{microF1} = \frac{184}{184 + 11 + 18} = \frac{184}{213} \approx \boxed{0.8638}.$$

Macro-F1 对两类等权，因此更清楚揭示 rare Chemical class 的失败；micro-F1 被 Person 的大量样本主导。

评分要点 两个 class F1 共 2 分；macro 1 分；micro 2 分；解释 1 分。

4. Low-resource multilingual NER [10 分]

一种 full-credit design 使用共享 multilingual subword/character encoder，并为 target language 加入 character-CNN 或 morphology embeddings。Encoder 后接 token classifier 或 linear-chain CRF；在 English 和 200 个 target sentences 上最小化 masked token cross-entropy/CRF NLL，对 target language oversample，并可使用 language-specific adapter 避免完全共享参数。

对 unlabeled target text，teacher 产生 span-level pseudo labels；仅保留 calibrated high-confidence、跨两个 augmentations (dropout、轻微字符/上下文扰动) 一致的 spans，student 最小化 consistency loss。用 EMA teacher、class-balanced threshold 和每轮人工小集校准可抑制 confirmation bias，而非无条件接受所有预测。

Morphologically rich language 可能把一个实体表达为多种屈折形式；不同 script 使 English subword vocabulary 过度切分或产生 unknown pieces。可学习 target-specific subword vocabulary 并与 shared encoder 做 embedding alignment，或加强 character/byte encoder；label projection 必须按 character offsets 而不是假定 token 一一对应。

Evaluation 应有 target-language dev/test，按 entity type、seen/unseen surface form、morphological pattern 和 sentence length 分桶报告 span-level micro/macro-F1。与“仅 200 target labels”、“English transfer”及“English+unlabeled”做 ablation；若加入 English 后 unseen-target 或某些类别下降，即可诊断 negative transfer。

评分要点 表示/架构/双语 supervised objective 4 分；unlabeled objective 与 safeguard 2 分；形态/文字问题和针对性修正 2 分；可发现 negative transfer 的评测 2 分。其他完整设计同等给分。