博士资格考试复习资料

人工智能博士资格考试复习资料

Machine Learning Theory ·Deep Learning and Reinforcement Learning
Optimization Methods in AI ·Natural Language Processing

本资料面向博士资格考试复习。在读者具备基础微积分、线性代数和概率知识的前提下，可独立用于系统复习、证明训练和历年真题准备。本书按四个正式考试 Part 组织，覆盖三期正式试卷、75 道大题与 132 个最细作答单元。理论章负责独立学习，真题章保留唯一完整解答位置；六份未核实答案候选不作为权威来源。

2026 年 7 月

使用说明

本书面向已经学过相关课程、但需要系统恢复知识体系的读者。四个 Part 分别建立学习理论、深度学习与强化学习、AI 优化和自然语言处理主线。跨 Part 内容只在一个主归属章节完整展开：Transformer 统一在第 15 章，凸优化语言统一在第 10–12 章，其他章节使用交叉引用。

考试范围提醒

正式题面以原始试卷和冻结正式版为准。除特别注明外，完整解答是独立核验的复习稿，不冒充官方评分标准。原题存在歧义时保留原文并显式说明。

目录

第一部分

Machine Learning Theory

机器学习理论

第一章　学习问题、风险与经验原则

学习目标

学完本章，读者应能从一个自然语言任务写出样本空间、数据分布、假设类、损失、总体风险和经验风险；区分 Bayes 最优、类内最优与算法输出；解释近似误差、估计误差和优化误差；并能判断一个经验均值论证何时必须要求“假设固定”。

1.1　学习问题是一个带量词的统计实验

监督学习的基本对象不是某个网络或某条更新式，而是五元组

$$(\mathcal{X}, \mathcal{Y}, \mathcal{D}, \mathcal{H}, \ell).$$

输入 $X \in \mathcal{X}$、标签 $Y \in \mathcal{Y}$，联合分布 $\mathcal{D}$ 未知；假设类 $\mathcal{H}$ 规定学习器允许输出的预测函数；损失 $\ell(h(x), y)$ 规定错误的代价。训练样本

$$S = (Z_1, \dots, Z_m), \qquad Z_i = (X_i, Y_i)$$

通常假设独立同分布于 $\mathcal{D}$。算法是从样本到假设的映射 $A : S \mapsto A(S) \in \mathcal{H}$。

定义 1.1 (总体风险与经验风险)**.** 若损失可积，定义

$$R(h) = \mathbb{E}_{Z\sim\mathcal{D}}\ell(h; Z), \qquad \widehat{R}_S(h) = \frac{1}{m}\sum_{i=1}^{m}\ell(h; Z_i).$$

前者是新样本上的期望代价，后者是对同一 h 的样本平均。

这里有三个容易被省略的量词：

1. $\mathcal{D}$ 在训练前固定但未知；
2. 固定 h 时，$\widehat{R}_S(h)$ 是独立随机变量的均值；
3. $A(S)$ 依赖于样本，因此不能把 $\widehat{R}_S(A(S))$ 直接当作“固定 h”的经验均值。

例题 1.1 (贯穿例：一维阈值)**.** 令 $\mathcal{X} = \mathbb{R}$、$\mathcal{Y} = \{0, 1\}$，

$$h_t(x) = \mathbf{1}_{\{x\ge t\}}, \qquad t \in \mathbb{R}.$$

在 0-1 损失下，给定样本后经验风险只会在相邻排序样本之间改变。ERM 因而可在 $m + 1$ 个候

选区间中搜索；但选择出的阈值依赖整个样本，其总体风险不能只靠固定阈值的点态大数律控制。后两章将把这个差距分别用 VC 维和统一收敛填平。

1.2 Bayes 规则、类内最优与三种误差

二分类时记

$$\eta(x) = \mathbb{P}(Y = 1 \mid X = x).$$

对 0-1 损失，在给定 $X = x$ 后预测 1 和 0 的条件风险分别为 $1 - \eta(x)$ 与 $\eta(x)$。

定理 1.1 (Bayes 分类器与超额风险公式)**.** 令

$$h_B(x) = \mathbf{1}_{\{\eta(x) \geq 1/2\}}.$$

则 h_B 在所有可测二分类器中最小化 0-1 风险，并且对任意可测 h,

$$R(h) - R(h_B) = \mathbb{E}\left[|2\eta(X) - 1|\mathbf{1}_{\{h(X) \neq h_B(X)\}}\right].$$

证明. 条件于 $X = x$。若 $h_B(x) = 1$，则 $\eta(x) \geq 1/2$，把预测从 1 改成 0 会使条件风险增加 $\eta(x) - (1 - \eta(x)) = 2\eta(x) - 1$。若 $h_B(x) = 0$，增加量为 $1 - 2\eta(x)$。两种情形合并为 $|2\eta(x) - 1|\mathbf{1}_{\{h(x) \neq h_B(x)\}}$。对 X 积分即得结论。 □

Bayes 规则可能不在 $\mathcal{H}$ 中。定义类内最优

$$h^*_{\mathcal{H}} \in \arg\min_{h \in \mathcal{H}} R(h),$$

若最小元不存在，则用任意 $\gamma > 0$ 的 γ-近似最小元代替。算法输出的误差可拆成

$$R(A(S)) - R(h_B) = \underbrace{R(h^*_{\mathcal{H}}) - R(h_B)}_{\text{近似误差}} + \underbrace{R(\widehat{h}) - R(h^*_{\mathcal{H}})}_{\text{统计/估计误差}} + \underbrace{R(A(S)) - R(\widehat{h})}_{\text{优化误差}}, \tag{1.1}$$

其中 $\widehat{h}$ 是精确 ERM。增大模型通常降低近似误差，却会扩大搜索难度和统计复杂度；正则化正是在这几项之间做受控交换。

1.3 经验风险最小化、结构风险与正则化

定义 1.2 (ERM 与 SRM)**.** 经验风险最小化器满足

$$\widehat{h} \in \arg\min_{h \in \mathcal{H}} \widehat{R}_S(h).$$

若有嵌套类 $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \cdots$，结构风险最小化选择

$$(\widehat{k}, \widehat{h}) \in \arg\min_{k, h \in \mathcal{H}_k} \left\{\widehat{R}_S(h) + \mathrm{pen}(k, m, \delta)\right\}.$$

惩罚项不是为了让训练误差更小，而是偿还“从更大集合中挑最小经验风险”产生的选择偏差。正则化形式

$$\min_{w} \ \widehat{R}_S(w) + \lambda \Omega(w)$$

常可理解为 SRM 的连续版本，也可由约束问题 $\min \widehat{R}_S(w)$ s.t. $\Omega(w) \le r$ 的 Lagrange 形式得到。这种等价需要适当的凸性与约束资格；不能对任意非凸问题无条件声称一一对应。

命题 1.2 (双侧统一偏差如何控制 ERM)**.** *若某事件上*

$$\sup_{h \in \mathcal{H}} |R(h) - \widehat{R}_S(h)| \le \varepsilon,$$

则任意精确 ERM 满足

$$R(\widehat{h}) - R(h^*_{\mathcal{H}}) \le 2\varepsilon.$$

证明.

$$R(\widehat{h}) \le \widehat{R}_S(\widehat{h}) + \varepsilon \le \widehat{R}_S(h^*_{\mathcal{H}}) + \varepsilon \le R(h^*_{\mathcal{H}}) + 2\varepsilon.$$

第一、三步使用偏差的相反方向，所以单侧上界不足以证明该结论。 □

1.4 损失函数：统计目标与计算代理

1.4.1 0-1、平方、对数与 hinge

二分类常把标签写成 $Y \in \{-1, 1\}$，模型输出分数 $f(x)$，分类器为 $\operatorname{sign} f(x)$。常用损失为

$$\ell_{01}(yf) = \mathbf{1}_{\{yf \le 0\}}, \quad \ell_{\text{hinge}}(yf) = (1 - yf)_+, \quad \ell_{\log}(yf) = \log(1 + e^{-yf}).$$

hinge 与 logistic 都是 0-1 损失的凸代理，但作用不同：hinge 在间隔超过 1 后变为零；logistic 持续惩罚低置信度，并具有概率建模解释。若 $p_f(x) = \sigma(f(x))$，交叉熵就是 Bernoulli 负对数似然。

回归的平方损失有一个特别干净的总体最优解。

定理 1.3 (平方损失的条件均值投影)**.** *若 $\mathbb{E}Y^2 < \infty$，则在所有平方可积预测函数中，*

$$f^*(x) = \mathbb{E}[Y \mid X = x]$$

最小化 $\mathbb{E}(Y - f(X))^2$，且

$$\mathbb{E}(Y - f(X))^2 = \mathbb{E}(Y - f^*(X))^2 + \mathbb{E}(f(X) - f^*(X))^2.$$

证明. 写

$$Y - f(X) = Y - f^*(X) + f^*(X) - f(X).$$

平方后交叉项的期望为

$$2\mathbb{E}[(f^*(X) - f(X))\mathbb{E}(Y - f^*(X) \mid X)] = 0.$$

故得到正交分解和最优性。 □

1.4.2 分类校准的考试级理解

一个代理损失训练得很好，并不自动保证分类错误小。需要代理条件风险的最优分数与 $\eta(x)-1/2$ 同号。对 logistic 损失，固定 x 的条件风险为

$$\eta \log(1+e^{-f}) + (1-\eta)\log(1+e^{f}).$$

求导并令零得到 $f^* = \log(\eta/(1-\eta))$，其符号与 $\eta - 1/2$ 相同，因此它是分类校准的。考试中若只写“交叉熵可导，所以可替代 0-1”是不充分的；可导性解释优化便利，校准性才连接统计目标。

1.5 偏差–方差分解与模型选择

设训练集随机，回归估计器为 $\widehat{f}_S(x)$，数据模型 $Y = f_0(X) + \varepsilon$，且 $\mathbb{E}(\varepsilon \mid X) = 0$、$\mathrm{Var}(\varepsilon \mid X = x) = \sigma^2(x)$。对固定 x，

$$\mathbb{E}_S\mathbb{E}\Big[(Y - \widehat{f}_S(x))^2 \mid X = x, S\Big] = \sigma^2(x) + \big(\mathbb{E}_S\widehat{f}_S(x) - f_0(x)\big)^2 + \mathrm{Var}_S(\widehat{f}_S(x)).$$

三项依次为不可约噪声、平方偏差和方差。这个分解依赖平方损失；对 0-1 或交叉熵不能机械照搬同一代数。

交叉验证估计的是某个训练流程的外样本表现。若超参数选择和最终性能报告使用同一验证集，验证结果也会产生选择偏差。稳妥流程是训练集拟合、验证集选模型、测试集只做一次最终报告；数据少时使用嵌套交叉验证。

1.6 正式真题：固定假设的经验风险方差

正式真题：2025 春 MLT Q1（原卷第 1 页，7 分）

设 $\mathcal{H}$ 是 $\mathcal{X}$ 上的二分类器类，D 为未知输入分布，目标假设 $f \in \mathcal{H}$。固定 $h \in \mathcal{H}$。证明当 $S \sim D^m$ 时，

$$\mathrm{Var}_S[L_S(h)] = \frac{L_{D,f}(h)(1 - L_{D,f}(h))}{m}.$$

完整解答（独立核验稿） 令

$$Z_i = \mathbf{1}_{\{h(X_i) \neq f(X_i)\}}, \qquad p = L_{D,f}(h).$$

因为 h 在抽样前固定且 X_i 独立同分布，所以 Z_i 独立同分布为 Bernoulli(p)，并且

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} Z_i.$$

独立性使交叉协方差为零，故

$$\mathrm{Var}\, L_S(h) = \frac{1}{m^2}\sum_{i=1}^{m} \mathrm{Var}\, Z_i = \frac{mp(1-p)}{m^2} = \frac{p(1-p)}{m}.$$

若把 h 换成数据依赖的 $A(S)$，则各错误指示量通过共同的 $A(S)$ 耦合，以上独立 Bernoulli 论证不再成立。

易错点

“经验风险是无偏的”只对固定 h 直接成立。ERM 的训练误差通常向下偏；泛化分析必须控制选择后的假设，而不是把固定假设的方差公式直接套给 $\widehat{h}$。

答题方法

把自然语言学习题翻译成数学时依次写：

$$\text{数据分布} \to \text{允许的假设} \to \text{损失} \to \text{总体目标} \to \text{经验目标} \to \text{选择规则}.$$

然后检查：题目讨论的是固定假设还是数据依赖输出？比较的是 Bayes 最优、类内最优还是某个算法？所需偏差是一侧还是双侧？

自测

1. 证明绝对损失下任意条件中位数最小化条件风险。
2. 构造一个 Bayes 分类器不属于给定线性假设类的例子，并指出近似误差。
3. 说明为何 L_2 正则化可降低算法敏感性，却不保证训练误差降低。
4. 若损失取值于 $[0, B]$，固定 h 的经验风险方差可得到什么上界？
5. 在 命题 1.2 中，若只有 $\sup_h(R(h) - \widehat{R}(h)) \le \varepsilon$，哪一步无法完成？

本节结论

本章的核心不是背四种损失，而是分清三个层次：分布决定 Bayes 目标，假设类决定可逼近范围，数据与算法决定估计和优化误差。后续 VC、Rademacher 和稳定性分别给出控制“数据依赖选择”的不同机制。

1.7 风险分解：近似、估计与优化不能混为一谈

设所有可测预测器中的 Bayes 最优风险为 R^*，假设类内最优为

$$h^*_{\mathcal{H}} \in \arg\min_{h \in \mathcal{H}} R(h),$$

经验风险最小化器为 $\widehat{h}$，实际算法输出为 $\widetilde{h}$。恒等分解

$$R(\widetilde{h}) - R^* = \underbrace{R(h^*_{\mathcal{H}}) - R^*}_{\text{近似误差}} + \underbrace{R(\widehat{h}) - R(h^*_{\mathcal{H}})}_{\text{估计误差}} + \underbrace{R(\widetilde{h}) - R(\widehat{h})}_{\text{优化误差}}$$

把三种改进手段分开：扩大模型类主要降低近似误差，却可能增加估计误差；增加样本主要降低估计误差；更强优化器只直接控制第三项。若 $\arg\min$ 不存在，可用 ε-近似极小元替代，所有比较多出可控的 ε。

例题 1.2. 用常数 c 预测平方损失。风险

$$R(c) = \mathbb{E}(Y-c)^2 = \mathrm{Var}(Y) + (\mathbb{E}Y - c)^2$$

在 $c = \mathbb{E}Y$ 处最小。若类限制为 $c \in \{0,1\}$，类内最优是距 $\mathbb{E}Y$ 最近的端点；即使有无限样本，也不能消除这个近似误差。

1.8 Bayes 分类器与条件风险

二分类 $Y \in \{0,1\}$，记 $\eta(x) = \mathbb{P}(Y=1 \mid X=x)$。在 0-1 损失下，给定 $X=x$:

$$R_x(a) = \begin{cases} 1-\eta(x), & a=1, \\ \eta(x), & a=0. \end{cases}$$

故 Bayes 分类器为

$$h^*(x) = \mathbf{1}\{\eta(x) \geq 1/2\}, \qquad R^* = \mathbb{E}\min\{\eta(X), 1-\eta(X)\}.$$

进一步,

$$R(h) - R^* = \mathbb{E}\left[|2\eta(X)-1|\,\mathbf{1}\{h(X) \neq h^*(X)\}\right].$$

这说明在 $\eta(x) \approx 1/2$ 的区域分类标签虽可能频繁改变，超额风险却很小；margin/noise 条件正是控制边界附近质量。

1.9 代理损失与分类校准

令 $Y \in \{-1,1\}$，实值分数为 $f(x)$。条件代理风险

$$C_\eta(t) = \eta\,\phi(t) + (1-\eta)\phi(-t).$$

若任一条件极小元 t_η^* 的符号与 $2\eta-1$ 相同，则该代理损失在分类意义下校准。以 logistic 损失 $\phi(u) = \log(1+e^{-u})$ 为例,

$$C_\eta'(t) = -\frac{\eta}{1+e^t} + \frac{1-\eta}{1+e^{-t}},$$

解得

$$t_\eta^* = \log\frac{\eta}{1-\eta}.$$

其符号与 $\eta - 1/2$ 相同。平方损失若把标签编码为 $\{-1,1\}$，条件最优分数为 $\mathbb{E}[Y \mid X=x] = 2\eta(x)-1$，同样校准。

易错点

“凸”只说明优化结构，不自动保证分类校准；“校准”只连接代理风险与分类风险，也不说明有限样本优化一定成功。考试答案应分别交代统计目标、函数类与优化算法。

1.10 经验过程中的量词顺序

点态大数律陈述：对每个固定 h，$\widehat{R}_m(h) \to R(h)$。统一收敛要求

$$\sup_{h\in\mathcal{H}} |\widehat{R}_m(h) - R(h)| \to 0.$$

二者量词不同：

$$\forall h\ \mathbb{P}(\text{该}h\ \text{偏差小}) \geq 1-\delta$$

不能直接推出

$$\mathbb{P}(\forall h\ \text{偏差都小}) \geq 1-\delta.$$

当 $\mathcal{H}$ 有限，可用 union bound；当其无限，必须使用覆盖数、增长函数、Rademacher 复杂度或稳定性等结构。后续所有 PAC 证明都应先指出随机样本外层概率，再写假设类上的上确界。

答题方法

风险题先写随机对象与量词，再写最优对象属于哪个集合。比较两个算法时，把近似、估计和优化误差逐项定位；不要把训练误差低直接解释为泛化更好。

自测

1. 推导 Bayes 分类器超额风险恒等式。
2. 为什么扩大假设类可能同时改善近似误差并恶化估计误差？
3. 证明 logistic 条件风险的最优分数为 log-odds。
4. 给出“对每个固定假设”与“同时对所有假设”的量词区别。

第二章　打散、增长函数与 VC 维

学习目标

本章建立从“有限样本上能产生多少种标记”到 VC 维的组合语言。读者应能给出打散集作为下界证书、给出线性依赖或计数作为上界证书，完整证明 Sauer–Shelah 引理，并处理补类、并类、交类和几何复合类。

2.1　Restriction、trace 与打散

令 $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$。对有限点集 $C = \{x_1, \dots, x_m\}$，真正参与经验风险的不是函数在整个 $\mathcal{X}$ 上的行为，而是它们在 C 上留下的标记向量。

定义 2.1 (限制类与 trace)**.**

$$\mathcal{H}|_C = \{(h(x_1), \dots, h(x_m)) : h \in \mathcal{H}\} \subseteq \{0,1\}^m.$$

这称为 $\mathcal{H}$ 在 C 上的 restriction 或 trace。若 $\mathcal{H}|_C = \{0,1\}^m$，称 $\mathcal{H}$ 打散 C。

定义 2.2 (增长函数与 VC 维)**.**

$$\Pi_{\mathcal{H}}(m) = \sup_{x_1,\dots,x_m \in \mathcal{X}} \left|\mathcal{H}|_{\{x_1,\dots,x_m\}}\right|,$$

$$\mathrm{VCdim}(\mathcal{H}) = \sup\{m : \Pi_{\mathcal{H}}(m) = 2^m\}.$$

重复点不可能增加 trace，故求上确界时可限制为互异点。若任意有限 m 都可被打散，则 VC 维为无穷。

证明 $\mathrm{VCdim}(\mathcal{H}) \geq d$，必须构造同一个 d 点集并说明对其每一种标记都有一个可选假设。证明 $\mathrm{VCdim}(\mathcal{H}) \leq d$，必须排除所有 $d+1$ 点集；只展示一个不能打散的点集没有上界意义。

2.2　基础类：阈值、区间与矩形

例题 2.1 (一维阈值)**.** 对 $\mathcal{H} = \{x \mapsto \mathbf{1}_{\{x \geq t\}} : t \in \mathbb{R}\}$，任意一点可被打散。两个点 $x_1 < x_2$ 的标记 $(1,0)$ 无法由右阈值产生，故 $\mathrm{VCdim}(\mathcal{H}) = 1$。在 m 个有序点上，阈值只产生 $m+1$ 个后缀标记，所以 $\Pi_{\mathcal{H}}(m) = m + 1$。

例题 2.2 (实线上的区间)**.** 闭区间指示类能打散任意两个不同点。对三个有序点 $x_1 < x_2 < x_3$，标记 $(1,0,1)$ 不可能由单个区间产生。因此 VC 维为 2。在 m 个有序点上，正例必须形成连续块，

再加全零标记，故

$$\Pi_{\mathcal{H}}(m) = 1 + \frac{m(m+1)}{2}.$$

例题 2.3 (二维轴对齐矩形)**.** 取上、下、左、右四个极值点，它们可被轴对齐矩形打散：需要删除某个点时，从相应方向收紧边界。任意五点中至少有一点不是四个坐标极值之一；若要保留所有极值点而排除该内部点，任何包含四个极值点的轴对齐矩形也会包含它。故二维轴对齐矩形类 VC 维为 4。

上例中的“内部”指位于由四个方向极值确定的包围盒内，不要求它是欧氏凸包内部。考试作图时必须标清是哪一种几何结构。

2.3 仿射半空间的上下界证书

定理 2.1 (仿射半空间的 VC 维)**.** $\mathbb{R}^p$ *中仿射半空间指示类*

$$\mathcal{H}_{\text{hs}}^{(p)} = \{x \mapsto \mathbf{1}_{\{w^\top x + b \geq 0\}} : (w, b) \in \mathbb{R}^{p+1}\}$$

的 VC 维为 $p+1$。

证明. **下界。**取 $p+1$ 个仿射无关点 $x_1, \ldots, x_{p+1}$，其增广向量 $\widetilde{x}_i = (x_i^\top, 1)^\top$ 线性无关。对任意标记 $y_i \in \{-1, 1\}$，线性方程

$$\widetilde{X}\widetilde{w} = y$$

有解。于是 $\operatorname{sign}(\widetilde{w}^\top \widetilde{x}_i) = y_i$，这 $p+1$ 点被打散。

上界。任取 $p+2$ 个点。其增广向量线性相关，存在不全为零的 α_i 使

$$\sum_i \alpha_i \widetilde{x}_i = 0.$$

最后一个坐标给 $\sum_i \alpha_i = 0$，故正、负系数都出现。把 $\alpha_i > 0$ 的点标为 1，把 $\alpha_i < 0$ 的点标为 0；对 $\alpha_i = 0$ 的点任意指定标记，例如标为 0。

若存在半空间实现该标记，记 $a_i = w^\top x_i + b$。对正系数点有 $a_i \geq 0$，对负系数点有 $a_i < 0$。但线性依赖给

$$0 = \sum_i \alpha_i a_i.$$

左边按正负系数分组后，一组非负而另一组严格为正，不能相等，矛盾。因此任意 $p+2$ 点都不能被打散。 □

易错点

线性依赖中可能出现 $\alpha_i = 0$。这些点不参与矛盾，可任意标记；若证明把它们误归入正组或负组并要求严格符号，就留下了缺口。

2.4 Sauer–Shelah 引理

VC 维只记录“何时第一次不能打散”，增长函数却需要控制任意 m 上的 trace 数量。Sauer–Shelah 引理把二者连接起来。

定理 2.2 (Sauer–Shelah)**.** *若* $\mathrm{VCdim}(\mathcal{H}) = d < \infty$*，则对任意* $m \geq 0$*,*

$$\Pi_{\mathcal{H}}(m) \leq \sum_{j=0}^{\min(d,m)} \binom{m}{j}.$$

特别地，当 $m \geq d \geq 1$ *时,*

$$\Pi_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d.$$

证明. 固定 m 点 $C = \{x_1, \ldots, x_m\}$，记

$$\mathcal{A} = \mathcal{H}|_C \subseteq \{0,1\}^m.$$

把最后一位删掉后得到前 $m-1$ 位的 trace 集 $\mathcal{A}_0$。再令 $\mathcal{A}_1 \subseteq \mathcal{A}_0$ 为那些既能接 0 又能接 1 的前缀。每个前缀若只有一种末位贡献 1 个 trace，若两种末位都出现则贡献 2 个，因此

$$|\mathcal{A}| = |\mathcal{A}_0| + |\mathcal{A}_1|.$$

$\mathcal{A}_0$ 的 VC 维至多 d。若 $\mathcal{A}_1$ 能打散前 $m-1$ 个坐标中的某个 d 元子集，则每种标记都能同时接 0 和 1，加入第 m 个坐标后便打散 $d+1$ 个点，矛盾。因此 $\mathcal{A}_1$ 的 VC 维至多 $d-1$。

令 $G(m,d)$ 表示 VC 维至多 d 的 trace 家族在 m 点上的最大大小，则

$$G(m,d) \leq G(m-1,d) + G(m-1,d-1).$$

边界为 $G(m,0) \leq 1$、$G(0,d) = 1$。对 $m+d$ 归纳并使用 Pascal 恒等式,

$$\begin{aligned} G(m,d) &\leq \sum_{j=0}^{d} \binom{m-1}{j} + \sum_{j=0}^{d-1} \binom{m-1}{j} \\ &= \sum_{j=0}^{d} \binom{m}{j}. \end{aligned}$$

这证明组合和上界。

当 $m \geq d \geq 1$ 时，对任意 $a > 0$,

$$\sum_{j=0}^{d} \binom{m}{j} \leq a^{-d} \sum_{j=0}^{d} \binom{m}{j} a^j \leq a^{-d}(1+a)^m.$$

取 $a = d/m$，再用 $(1+d/m)^m \leq e^d$，即得 $(em/d)^d$。 □

注 2.1. 当 $m \leq d$ 时，右侧组合和等于 2^m，没有非平凡压缩；这与类仍可能打散 m 点一致。使用 $(em/d)^d$ 前必须检查 $m \geq d$。

2.5 复合假设类

补类 $\neg\mathcal{H} = \{1 - h : h \in \mathcal{H}\}$ 在每个有限点集上的 trace 只是逐位翻转，故 VC 维不变。交类和并类可用 De Morgan 互换。

定理 2.3 (两个类之并的 VC 上界)**.** *若 $d = \mathrm{VCdim}(\mathcal{H})$、$d' = \mathrm{VCdim}(\mathcal{H}')$，则*

$$\mathrm{VCdim}(\mathcal{H} \cup \mathcal{H}') \le d + d' + 1.$$

证明. 令 $m = d + d' + 2$。若某个 m 点集被并类打散，则

$$2^m \le \Pi_{\mathcal{H}}(m) + \Pi_{\mathcal{H}'}(m) \le \sum_{j=0}^{d} \binom{m}{j} + \sum_{j=0}^{d'} \binom{m}{j}.$$

因 $d' = m - d - 2$，第二个和式利用 $\binom{m}{j} = \binom{m}{m-j}$ 可写成高阶二项系数之和；两部分之间至少遗漏一层，因此严格小于 2^m，矛盾。 □

若类由 k 个基类的投票构成，不能只把各 VC 维相加，因为还要计数投票结构。常见路线是先在固定样本上计数每个基类 trace，再计数组合规则，最后用增长函数反推 VC 上界。结论通常带 $k \log k$ 而不是简单 k。

2.6 正式真题：并类与几何类

正式真题：2025 春 MLT Q3（原卷第 2 页，9 分）

设 $\mathcal{H}, \mathcal{H}'$ 是有限 VC 维的二分类函数族。

1. 证明
$$\mathrm{VCdim}(\mathcal{H} \cup \mathcal{H}') \le \mathrm{VCdim}(\mathcal{H}) + \mathrm{VCdim}(\mathcal{H}') + 1.$$
2. 将结论用于 $\mathbb{R}^2$ 中轴对齐矩形类与三角形类的并。可直接使用：三角形类的 VC 维为 7。

完整解答（独立核验稿） 第一问就是 定理 2.3。为使答卷自足，设两类 VC 维为 d, d'，取 $m = d + d' + 2$。若并类打散 m 点，则全部 2^m 种标记来自两类之一；Sauer 上界与二项系数对称性说明两类 trace 总数严格小于 2^m，矛盾。因此上界为 $m - 1 = d + d' + 1$。

第二问由 例题 2.3 知矩形类 VC 维为 4，题给三角形类 VC 维为 7，故

$$7 \le \mathrm{VCdim}(\mathcal{R} \cup \mathcal{T}) \le 4 + 7 + 1 = 12.$$

下界来自 $\mathcal{T} \subseteq \mathcal{R} \cup \mathcal{T}$。仅凭题面给出的事实与第一问只能推出区间，不能证明精确等于 12。原题的 “determine” 因此存在已记录的轻微歧义；考试中应报告上界 12，并明确现有信息不足以把上界写成等号。

答题方法

VC 维题的标准书写顺序：

1. 下界：固定一组点，给任意标记构造参数；
2. 上界：任取更多点，找一个不可能标记或作 trace 计数；

3. 复合类：先把无限函数降为固定样本上的有限 trace；

4. 使用 Sauer 时写清 $m > d$ 和严格不等式来自哪里。

自测

1. 求一维至多两个区间之并的 VC 维，并分别给上下界证书。

2. 证明齐次半空间在 $\mathbb{R}^p$ 中的 VC 维为 p，并比较仿射情形。

3. 证明 $\Pi_{\mathcal{H}\cap\mathcal{H}'}(m) \le \Pi_{\mathcal{H}}(m)\Pi_{\mathcal{H}'}(m)$。

4. Sauer 递推中，为什么“两种末位都可接”的前缀类 VC 维下降 1？

5. 构造一个例子说明并类 VC 上界通常不能直接写成等号。

本节结论

VC 维的本质是有限样本上的组合自由度，而不是参数个数或训练复杂度。Sauer–Shelah 把“不能再打散”转化成多项式增长；下一章再把这种增长转化成泛化概率。

第三章　从组合复杂度到泛化

学习目标

本章把增长函数转成概率界。读者应能完整复现有限类 union bound、ghost sample 对称化、双侧 VC 统一收敛和 realizable 双样本证明；区分 ε^{-1} 与 ε^{-2} 的来源；并掌握 Rademacher 复杂度和 uniform stability 两条替代路线。

3.1　集中不等式与有限类基准

本章损失先取值于 $[0,1]$。对固定 h，Hoeffding 给

$$\mathbb{P}\Big(|\widehat{R}_S(h) - R(h)| > \varepsilon\Big) \le 2e^{-2m\varepsilon^2}.$$

固定 h 是关键；若从样本中选出经验风险最小的 h，必须同时控制所有候选。

定理 3.1 (有限类的双侧统一收敛)**.** *若 $|\mathcal{H}| = M < \infty$，则*

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} |R(h) - \widehat{R}_S(h)| > \varepsilon\right) \le 2Me^{-2m\varepsilon^2}.$$

证明. 对每个固定 h 应用 Hoeffding，再对 M 个事件作 union bound。这里允许同一数据同时出现在所有经验风险中；union bound 不要求这些事件独立。 □

因此只要

$$m \ge \frac{\log(2M/\delta)}{2\varepsilon^2},$$

便以至少 $1-\delta$ 的概率得到统一偏差不超过 ε。McDiarmid 适合控制“替换一个样本时函数值变化有限”的统计量；Hoeffding 适合固定函数的独立平均。考试中应先写随机对象，再选择工具。

3.2　Ghost sample 与双侧 VC 统一收敛

无限类不能把 M 换成 ∞。需要引入独立 ghost sample $S' \sim \mathcal{D}^m$，条件于 $S \cup S'$ 后，再利用 $\Pi_{\mathcal{H}}(2m)$ 计数固定 trace。

注 3.1 (可测性与近似极大元)*.* 以下假设经验风险过程可分且相关上确界事件可测。若 supremum 不达到，在可数稠密子类中取 η-近似极大元，以阈值减去 η 完成论证，再令 $\eta \downarrow 0$。对一般不可分类可把概率解释为外概率；数值上界不变。

引理 3.2 (双样本两侧对称化)**.** *定义*

$$\Delta_S^+ = \sup_h (R(h) - \widehat{R}_S(h)), \qquad \Delta_S^- = \sup_h (\widehat{R}_S(h) - R(h)).$$

若 $m\alpha^2 \geq 2$*，则*

$$\mathbb{P}(\Delta_S^+ > \alpha) \leq 2\mathbb{P}\left(\sup_h (\widehat{R}_{S'}(h) - \widehat{R}_S(h)) > \alpha/2 \right), \tag{3.1}$$

$$\mathbb{P}(\Delta_S^- > \alpha) \leq 2\mathbb{P}\left(\sup_h (\widehat{R}_S(h) - \widehat{R}_{S'}(h)) > \alpha/2 \right). \tag{3.2}$$

证明. 只证第一式。事件 $\Delta_S^+ > \alpha$ 发生时，按 说明 3.1 选取一个超过阈值的近似极大元 h_S。条件于 S 后它固定。ghost 经验风险的方差至多 $1/(4m)$，故 Chebyshev 给

$$\mathbb{P}\Big(|\widehat{R}_{S'}(h_S) - R(h_S)| \leq \alpha/2 \mid S \Big) \geq 1 - \frac{1}{m\alpha^2} \geq \frac{1}{2}.$$

在该事件上 $\widehat{R}_{S'}(h_S) - \widehat{R}_S(h_S) > \alpha/2$。对 S 积分即得第一式。对第二式，在 $\Delta_S^- > \alpha$ 上选取满足 $\widehat{R}_S(h_S) - R(h_S) > \alpha$ 的近似极大元；仍由同一个 Chebyshev 事件 $|\widehat{R}_{S'}(h_S) - R(h_S)| \leq \alpha/2$，得到 $\widehat{R}_S(h_S) - \widehat{R}_{S'}(h_S) > \alpha/2$。条件化后再对 S 积分，便得所写第二个不等式。 □

定理 3.3 (VC 类的双侧统一收敛)**.** *对任意二分类假设类和* $\alpha > 0$*,*

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} |R(h) - \widehat{R}_S(h)| > \alpha \right) \leq 4\Pi_{\mathcal{H}}(2m) \exp\left(-\frac{m\alpha^2}{8} \right).$$

证明. 若 $m\alpha^2 < 2$，右端大于 1，结论平凡。以下设 $m\alpha^2 \geq 2$。由 引理 3.2 分别控制正、负两侧。固定 $2m$ 个带标签观测。对每个 h，其损失向量只由 h 在这 $2m$ 个输入上的标记决定，因此至多有 $\Pi_{\mathcal{H}}(2m)$ 种。

在每对 (Z_i, Z_i') 内用独立 Rademacher 符号随机交换。对固定损失向量，

$$\widehat{R}_{S'}(h) - \widehat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} \sigma_i \big(\ell_h(Z_i') - \ell_h(Z_i) \big).$$

各项条件均值为零、绝对值至多 $1/m$，Hoeffding 给一侧概率

$$\mathbb{P}_\sigma\Big(\widehat{R}_{S'}(h) - \widehat{R}_S(h) > \alpha/2 \Big) \leq e^{-m\alpha^2/8}.$$

对固定后的至多 $\Pi_{\mathcal{H}}(2m)$ 个向量作 union bound，乘一侧对称化因子 2；最后再合并正负两侧，得到因子 4。 □

若 $\mathrm{VCdim}(\mathcal{H}) = d$ 且 $2m \geq d$，Sauer 界给

$$\mathbb{P}\left(\sup_h |R - \widehat{R}| > \alpha \right) \leq 4 \exp\left[d \log \frac{2em}{d} - \frac{m\alpha^2}{8} \right].$$

3.3 Realizable 与 agnostic PAC

3.3.1 Agnostic ERM：双侧偏差

在事件

$$\sup_h |R(h) - \widehat{R}_S(h)| \le \varepsilon/2$$

上，精确 ERM $\widehat{h}$ 和类内最优 h^* 满足

$$R(\widehat{h}) \le \widehat{R}(\widehat{h}) + \varepsilon/2 \le \widehat{R}(h^*) + \varepsilon/2 \le R(h^*) + \varepsilon.$$

因此

$$\mathbb{P}(R(\widehat{h}) - R(h^*) > \varepsilon) \le 4\Pi_{\mathcal{H}}(2m)e^{-m\varepsilon^2/32}.$$

若最小元不达到，取 η-近似总体最小元和 η-ERM，超额风险变为 $\varepsilon + 2\eta$，再令 $\eta \downarrow 0$。

3.3.2 Realizable：合法的双样本一致性证明

定理 3.4 (一致学习器的失败界)**.** *若存在 $h^* \in \mathcal{H}$ 使 $R(h^*) = 0$，学习器返回任意 $\widehat{R}_S(\widehat{h}) = 0$ 的假设。若 $0 < \varepsilon < 1$ 且 $m\varepsilon \ge 8\log 2$，则*

$$\mathbb{P}(R(\widehat{h}) > \varepsilon) \le 2\Pi_{\mathcal{H}}(2m)\exp\left(-\frac{\log 2}{2}m\varepsilon\right).$$

证明. 令

$$\mathcal{E} = \{\exists h : \widehat{R}_S(h) = 0,\ R(h) > \varepsilon\}.$$

条件于 S 和一个坏的一致假设 h。若 $p = R(h) > \varepsilon$，则 $m\widehat{R}_{S'}(h) \sim \mathrm{Bin}(m, p)$，乘法 Chernoff 界给

$$\mathbb{P}(\widehat{R}_{S'}(h) \le \varepsilon/2 \mid S) \le \mathbb{P}(\widehat{R}_{S'}(h) \le p/2 \mid S) \le e^{-mp/8} \le \frac{1}{2}.$$

所以

$$\mathbb{P}(\mathcal{E}) \le 2\mathbb{P}\left(\exists h : \widehat{R}_S(h) = 0,\ \widehat{R}_{S'}(h) > \varepsilon/2\right).$$

现在条件于合并的 $2m$ 个带标签样本，并在每对位置内随机交换。对固定 trace，设它在 $2m$ 个位置上共有 k 个错误。若训练半样本零错而 ghost 半样本错误率大于 $\varepsilon/2$，则所有错误必须进入 ghost 一侧且 $k > m\varepsilon/2$。一对中若两处都错，所求事件概率为零；否则每个错误位置都必须被公平硬币送入 ghost 一侧，故条件概率至多

$$2^{-k} \le \exp\left(-\frac{\log 2}{2}m\varepsilon\right).$$

条件化后候选 trace 已固定，数量至多 $\Pi_{\mathcal{H}}(2m)$，此时 union bound 合法。再乘前面的因子 2 即得结论。 □

不能改写为“在随机训练输入上共有 $\Pi_{\mathcal{H}}(m)$ 个 trace，所以直接 union bound”：这些候选在抽样前并未固定。ghost sample 和条件化随机分半正是修复量词的步骤。

3.3.3 样本量与量级

一个可检查的充分条件是

$$\frac{m\varepsilon^2}{32} \geq d\log\frac{2em}{d} + \log\frac{4}{\delta}$$

用于 agnostic ERM，而 realizable 一致学习器可用

$$\frac{\log 2}{2} m\varepsilon \geq d\log\frac{2em}{d} + \log\frac{2}{\delta}, \qquad m\varepsilon \geq 8\log 2.$$

吸收对数后分别是

$$m_{\text{agn}} = O\left(\frac{d\log(1/\varepsilon) + \log(1/\delta)}{\varepsilon^2}\right), \quad m_{\text{real}} = O\left(\frac{d\log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}\right).$$

3.4 为什么这两个量级不能整体改进

定理 3.5 (PAC 下界的考试级形式)**.** *存在普适常数 $c > 0$，使 VC 维为 $d \geq 2$ 的某些类满足：*

1. *realizable 学习要使错误不超过 ε、成功概率至少 $1-\delta$，最坏情形需要 $m \geq c(d+\log(1/\delta))/\varepsilon$；*
2. *agnostic 学习要使超额风险不超过 ε，最坏情形需要 $m \geq c(d+\log(1/\delta))/\varepsilon^2$。*

证明路线 取一个被打散的 d 点集。realizable 情形把总概率质量 $\Theta(\varepsilon)$ 均匀放在 $d-1$ 个“稀有点”上，其余质量放在一个固定标签点。每个稀有点标签独立选择。若样本没有看到其中相当比例的点，任何学习器在这些点上平均至少错一半；coupon-collector 的一阶矩与 Paley–Zygmund 界给 $m = \Omega(d/\varepsilon)$。只保留一个概率质量 ε 的稀有点，并在两个 realizable 标签之间二选一，未观察到它的概率为 $(1-\varepsilon)^m$，要把失败概率压到 δ 需要 $m = \Omega(\log(1/\delta)/\varepsilon)$。

agnostic 情形仍用被打散点，但在第 j 点令标签为 Bernoulli$(1/2+\sigma_j\gamma)$，其中 $\sigma_j \in \{-1,1\}$ 未知、$\gamma \asymp \varepsilon$。在每个坐标上判断 σ_j 是两个相距 2γ 的硬币检验；其 n_j 次观测的 KL 散度为 $O(n_j\gamma^2)$。Pinsker 不等式表明当 $n_j\gamma^2$ 为小常数时，判断错误概率仍有正常数下界。对 d 个坐标求和并用 $\sum_j n_j = m$，得到 $m = \Omega(d/\varepsilon^2)$。二点检验再给 $\Omega(\log(1/\delta)/\varepsilon^2)$。

这一定理说明统一收敛上界的某些对数和常数可以改善，但 ε^{-1} 与 ε^{-2} 的根本差异不是证明技巧造成的。

3.5 Rademacher 复杂度：数据依赖的函数类大小

对实值函数类 $\mathcal{F}$ 定义经验 Rademacher 复杂度

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i f(Z_i) \,\middle|\, S\right],$$

其中 σ_i 独立且等概率取 ± 1。

定理 3.6 (期望对称化界)**.** *若 $f(Z) \in [0,1]$，则*

$$\mathbb{E}_S \sup_{f\in\mathcal{F}} \big(R(f) - \widehat{R}_S(f)\big) \leq 2\mathbb{E}_S\widehat{\mathfrak{R}}_S(\mathcal{F}).$$

证明. 引入独立 S'。由 Jensen,

$$\mathbb{E}_S \sup_f \left(\mathbb{E}_{S'} \widehat{R}_{S'}(f) - \widehat{R}_S(f) \right) \le \mathbb{E}_{S,S'} \sup_f \left(\widehat{R}_{S'}(f) - \widehat{R}_S(f) \right).$$

成对随机交换不改变联合分布，所以右端等于

$$\mathbb{E}_{S,S',\sigma} \sup_f \frac{1}{m} \sum_i \sigma_i (f(Z_i') - f(Z_i)).$$

用 $\sup_f(a_f + b_f) \le \sup_f a_f + \sup_f b_f$，再分别对 S' 与 S 取期望，两个项都等于 $\mathbb{E}\widehat{\mathfrak{R}}_S(\mathcal{F})$。 □

引理 3.7 (收缩原理的常用形式)**.** *若每个 $\phi_i : \mathbb{R} \to \mathbb{R}$ 都是 L-Lipschitz 且 $\phi_i(0) = 0$，则*

$$\mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \frac{1}{m} \sum_i \sigma_i \phi_i(f(Z_i)) \le L\, \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \frac{1}{m} \sum_i \sigma_i f(Z_i).$$

它允许先控制预测类，再把界传给 Lipschitz 损失类。例如 $\mathcal{F} = \{x \mapsto w^\top x : \|w\|_2 \le B\}$ 且 $\|x_i\|_2 \le R$ 时，

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) = \frac{B}{m} \mathbb{E}_\sigma \left\| \sum_i \sigma_i x_i \right\|_2 \le \frac{BR}{\sqrt{m}}.$$

因此复杂度直接反映样本几何和范数约束，不只依赖参数个数。

3.6 Uniform stability：控制算法而不是整个类

定义 3.1 (替换一点稳定性)**.** 样本 S, S' 只相差一个观测。若对任意测试点 z,

$$|\ell(A(S), z) - \ell(A(S'), z)| \le \beta_m,$$

称算法具有 β_m-uniform stability。若只要求对替换位置和样本取期望后的单侧差不超过 β_m，称为 on-average replace-one stability。

定理 3.8 (稳定性给期望泛化)**.** *若 A 具有 on-average replace-one 稳定率 β_m，则*

$$\mathbb{E}_S \left[R(A(S)) - \widehat{R}_S(A(S)) \right] \le \beta_m.$$

证明. 取独立 $Z_i' \sim \mathcal{D}$，令 $S^{i \leftarrow Z_i'}$ 替换第 i 个样本。交换 (Z_i, Z_i') 的联合分布不变，所以

$$\mathbb{E} R(A(S)) = \frac{1}{m} \sum_i \mathbb{E}\, \ell(A(S^{i \leftarrow Z_i'}), Z_i).$$

而

$$\mathbb{E} \widehat{R}_S(A(S)) = \frac{1}{m} \sum_i \mathbb{E}\, \ell(A(S), Z_i).$$

逐项相减并使用 on-average stability 即得。 □

定理 3.9 (二次正则 ERM 的稳定性)**.** *设 $\ell(w,z)$ 关于 w 凸且 ρ-Lipschitz,*

$$A(S) \in \arg\min_{w} \left\{ \widehat{R}_S(w) + \lambda \|w\|_2^2 \right\}.$$

则替换一个样本引起的测试损失变化不超过 $2\rho^2/(\lambda m)$；因而

$$\mathbb{E}R(A(S)) \le R(w^*) + \lambda \|w^*\|_2^2 + \frac{2\rho^2}{\lambda m}$$

对任意总体风险比较点 w^* 成立。

证明. 令 $w = A(S)$、$w' = A(S')$，两样本只在第 i 项不同。目标函数含 $\lambda\|w\|^2$，故为 2λ-强凸。分别用 w' 比较 S 上的最优值、用 w 比较 S' 上的最优值，两式相加后共同样本项抵消，得到

$$2\lambda \|w - w'\|^2 \le \frac{1}{m}\left[\ell(w', Z_i) - \ell(w, Z_i) + \ell(w, Z_i') - \ell(w', Z_i')\right] \le \frac{2\rho}{m}\|w - w'\|.$$

所以 $\|w - w'\| \le \rho/(\lambda m)$，进而测试损失差至多 $\rho^2/(\lambda m)$，当然也不超过题目采用的保守常数 $2\rho^2/(\lambda m)$。

另一方面，经验最优性给

$$\widehat{R}_S(A(S)) + \lambda\|A(S)\|^2 \le \widehat{R}_S(w^*) + \lambda\|w^*\|^2.$$

取期望、使用 $\mathbb{E}\widehat{R}_S(w^*) = R(w^*)$，再用 定理 3.8 即得风险界。 □

3.7 PAC–Bayes、在线遗憾与过参数化边界

VC 与 Rademacher 界同时控制一个类中的全部确定性假设；PAC–Bayes 改为控制由数据选出的随机后验 Q，并用它相对一个抽样前固定先验 P 的信息代价约束数据依赖性。对 $h \sim Q$ 记

$$\widehat{L}_S(Q) = \mathbb{E}_{h\sim Q}\widehat{L}_S(h), \qquad L(Q) = \mathbb{E}_{h\sim Q}L(h).$$

定理 3.10 (一个二元相对熵型 PAC–Bayes 界)**.** *损失取值于 $[0,1]$，先验 P 与样本独立。则以至少 $1-\delta$ 的概率，同时对所有满足 $\mathrm{KL}(Q\|P) < \infty$ 的后验 Q 有*

$$\mathrm{kl}\Big(\widehat{L}_S(Q) \,\Big\|\, L(Q)\Big) \le \frac{\mathrm{KL}(Q\|P) + \log(2\sqrt{m}/\delta)}{m},$$

其中 $\mathrm{kl}(a\|b)$ 是 Bernoulli 相对熵。先验若看过同一份样本，结论一般不再成立。

证明. 固定 h 时，二项型指数矩估计给

$$\mathbb{E}_S \exp\Big\{ m\ \mathrm{kl}(\widehat{L}_S(h)\|L(h)) \Big\} \le 2\sqrt{m}.$$

先对 $h \sim P$ 积分，再对 S 用 Markov 不等式，得到一个概率至少 $1-\delta$ 的事件；在该事件上

$$\mathbb{E}_{h\sim P} e^{F_S(h)} \le 2\sqrt{m}/\delta, \qquad F_S(h) = m\,\mathrm{kl}(\widehat{L}_S(h)\|L(h)).$$

Donsker–Varadhan 变分不等式

$$\mathbb{E}_Q F \le \mathrm{KL}(Q\|P) + \log \mathbb{E}_P e^F$$

对每个 Q 同时成立。最后用二元相对熵的联合凸性和 Jensen 不等式，把 $\mathbb{E}_Q F_S(h)$ 下界为

$$m\,\mathrm{kl}\Big(\widehat{L}_S(Q)\,\Big\|\,L(Q)\Big),$$

即得结论。统一量词来自同一个关于 P 的指数矩事件，而不是对不可数个 Q 做 union bound。 □

例题 3.1 (后验复杂度为何不是参数个数)**.** 若 P 在 N 个假设上均匀，而 Q 集中在一个数据选出的假设 $\widehat{h}$，则 $\mathrm{KL}(Q\|P) = \log N$，界退化为有限类复杂度。若 Q 只在少量相近假设间分散，KL 可显著小于 $\log N$；这说明 PAC–Bayes 惩罚的是后验相对先验的移动，而不是裸参数数目。

在线学习没有 i.i.d. 总体风险。第 t 轮先选 $w_t \in K$，再观察凸损失 ℓ_t，遗憾定义为

$$\mathrm{Reg}_T = \sum_{t=1}^{T} \ell_t(w_t) - \min_{w\in K} \sum_{t=1}^{T} \ell_t(w).$$

定理 3.11 (在线梯度下降的基本遗憾界)**.** *设闭凸集 K 的直径不超过 D，且 $\|g_t\|_2 \le G$，其中 $g_t \in \partial \ell_t(w_t)$。更新*

$$w_{t+1} = \Pi_K(w_t - \eta g_t)$$

满足

$$\mathrm{Reg}_T \le \frac{D^2}{2\eta} + \frac{\eta G^2 T}{2}.$$

取 $\eta = D/(G\sqrt{T})$ 得 $\mathrm{Reg}_T \le DG\sqrt{T}$。

证明. 由投影不扩张和次梯度不等式，对任意 $w \in K$，

$$\begin{aligned} \|w_{t+1} - w\|^2 &\le \|w_t - \eta g_t - w\|^2 \\ &= \|w_t - w\|^2 - 2\eta\langle g_t, w_t - w\rangle + \eta^2\|g_t\|^2, \\ \ell_t(w_t) - \ell_t(w) &\le \langle g_t, w_t - w\rangle. \end{aligned}$$

整理后对 t 求和，距离平方望远镜相消，得到所述界；再对 w 取累计损失最优比较点。 □

易错点

遗憾界比较的是同一损失序列上的最佳固定决策，不是 i.i.d. 泛化概率。过参数化模型可以插值训练集，但“插值”本身既不推出泛化，也不推出失败；结果还依赖数据结构、优化隐式偏置、正则化和噪声。双降与神经切线等现象可用于解释实验，却不能替代带假设的定理。No Free Lunch 的考试结论正是：没有任务分布或归纳偏置时，不存在对所有问题都严格优越的学习算法。

3.8 正式真题：平均稳定性与正则化

正式真题：2025 春 MLT Q4（原卷第 2 页，10 分）

1. 设 $S=(z_1,\dots,z_m)$ i.i.d.，算法 A 具有平均 replace-one 稳定率 $\epsilon(m)$。证明

$$\mathbb{E}_S[L_D(A(S))-L_S(A(S))]\le\epsilon(m).$$

2. 设损失凸且 ρ-Lipschitz。证明带 $\lambda\|w\|^2$ 正则的 ERM 满足

$$\mathbb{E}_S L_D(A(S))\le L_D(w^*)+\lambda\|w^*\|^2+\frac{2\rho^2}{\lambda m}.$$

完整解答（独立核验稿） 第一问按 定理 3.8 引入独立替换点，利用交换对称性把总体损失写成替换后算法在原位置上的损失，逐项应用稳定性。

第二问按 定理 3.9：两组正则经验目标的强凸最优性不等式相加，共同样本项消去；Lipschitz 性把右端控制为 $(2\rho/m)\|A(S)-A(S')\|$，得到 $O(\rho^2/(\lambda m))$ 的稳定率。再把经验正则目标与固定 w^* 比较并取期望，即得题目给出的保守常数。本解答是独立核验稿，不代表官方答案。

易错点

VC/Rademacher 控制一整个候选类，稳定性控制具体算法。模型类即使很大，强正则算法仍可能稳定；反之，有限参数模型若选择规则高度不连续，也不能仅凭“参数少”断言泛化。

答题方法

泛化题先识别事件：

- 固定假设：Hoeffding 或 Bernstein；
- 有限候选：固定假设界加 union bound；
- 无限分类类：ghost sample、trace、Sauer；
- 实值/范数类：Rademacher 与收缩；
- 给定正则算法：replace-one stability；
- agnostic ERM：必须使用双侧偏差。

自测

1. 从 定理 3.3 推出置信度 $1-\delta$ 的隐式样本量条件。
2. 在 realizable 双样本证明中，为什么一对位置都错会使事件概率为零？
3. 计算半径 R 输入、范数不超过 B 的线性类经验 Rademacher 复杂度。
4. 若正则项写为 $(\lambda/2)\|w\|^2$，稳定率常数如何改变？

5. 说明统一收敛上界与 minimax 最优样本复杂度为何不是同一命题。

本节结论

组合路线把无限类在 $2m$ 个点上离散化；Rademacher 路线保留样本几何；稳定性路线直接比较相邻样本下的算法输出。三条路线的共同点是：先修复数据依赖选择的量词，再使用集中或确定性敏感度。

第四章　经典监督与无监督方法

学习目标

本章把前面的风险与复杂度语言落到具体方法。读者应能推导感知机更新与收敛界、SVM 的间隔和对偶、核技巧与 representer theorem；能从重构误差推导 PCA，从交替最小化理解 k-means，并从 Jensen 下界推导 EM 的单调性。

4.1　线性分类、间隔与感知机

把标签写成 $y_i \in \{-1, 1\}$，增广后把偏置并入特征，线性分类器为

$$h_w(x) = \operatorname{sign}(w^\top x).$$

样本的函数间隔是 $y_i w^\top x_i$，几何间隔是 $y_i w^\top x_i / \|w\|$。由于 w 的正比例缩放不改变分类器，讨论最大间隔时必须固定尺度。

定理 4.1 (感知机收敛)**.** *假设 $\|x_i\| \le R$，存在单位向量 u 使*

$$y_i u^\top x_i \ge \gamma > 0$$

对所有样本成立。感知机从 $w_0 = 0$ 开始，每次在误分类样本上更新

$$w_{t+1} = w_t + y_i x_i.$$

则误更新次数不超过 $(R/\gamma)^2$。

证明. 若共发生 T 次误更新，则

$$u^\top w_T = \sum_{t=1}^{T} y_t u^\top x_t \ge T\gamma.$$

另一方面，误分类给 $y_t w_t^\top x_t \le 0$，故

$$\|w_{t+1}\|^2 = \|w_t\|^2 + 2y_t w_t^\top x_t + \|x_t\|^2 \le \|w_t\|^2 + R^2,$$

所以 $\|w_T\| \le R\sqrt{T}$。Cauchy–Schwarz 给 $T\gamma \le R\sqrt{T}$，从而 $T \le (R/\gamma)^2$。 □

这个结论是“有限次犯错”而不是随机样本上的直接泛化界。它依赖数据线性可分；不可分时更新可能永不停止，需要 pocket、平均感知机或软间隔方法。

4.2 SVM：从几何间隔到凸优化

4.2.1 硬间隔与规范化

可分样本上，希望最大化 $\min_i y_i(w^\top x_i+b)/\|w\|$。利用缩放自由度规定

$$y_i(w^\top x_i+b)\ge 1,$$

便得到等价问题

$$\min_{w,b}\frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad y_i(w^\top x_i+b)\ge 1.$$

两条支持超平面 $w^\top x+b=\pm 1$ 的距离为 $2/\|w\|$。

4.2.2 软间隔与 hinge

引入松弛变量 $\xi_i\ge 0$：

$$\min_{w,b,\xi}\frac{1}{2}\|w\|^2+C\sum_i \xi_i, \quad y_i(w^\top x_i+b)\ge 1-\xi_i.$$

对固定 w,b，最小可行松弛为 $\xi_i=(1-y_i(w^\top x_i+b))_+$，所以等价于正则 hinge 风险。$C$ 大意味着更重视训练违约；若写成平均损失加 $\lambda\|w\|^2$，C、λ 与样本量之间的换算取决于是否除以 m。

定理 4.2 (硬间隔 SVM 对偶)**.** *硬间隔 SVM 的对偶为*

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

最优解满足

$$w^*=\sum_i \alpha_i^* y_i x_i, \qquad \alpha_i^*\big(y_i((w^*)^\top x_i+b^*)-1\big)=0.$$

证明. Lagrangian 为

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

对 w 求驻点得到 $w=\sum_i\alpha_i y_i x_i$；对 b 求驻点得到 $\sum_i\alpha_i y_i=0$。代回即得对偶目标。可分情形存在严格可行点，Slater 条件给强对偶；KKT 互补松弛给最后一式。 □

4.3 多类间隔、结构化预测与 boosting

多类线性模型使用联合特征 $\Psi(x,y)$ 与分数 $s_w(x,y)=\langle w,\Psi(x,y)\rangle$。给定代价 $\Delta(y,\widetilde{y})$，结构化 hinge 损失为

$$\ell_{\text{str}}(w;x,y)=\max_{\widetilde{y}\in\mathcal{Y}}\{\Delta(y,\widetilde{y})+s_w(x,\widetilde{y})-s_w(x,y)\}.$$

它是仿射函数族的上确界，故关于 w 凸；若损失为零，则所有错误标签都满足所要求的间隔。训练时的核心算法步骤是 loss-augmented inference：寻找上式中的最坏 $\widetilde{y}$，而不是枚举参数坐标。普

通多类 hinge 是取块特征 $\Psi(x,y)$ 的特例。

例题 4.1 (三类块特征)**.** 令 $\Psi(x,y)$ 把 x 放在第 y 个参数块，其余块为零。则

$$\ell(w;x,y)=\max_{j\neq y}[1+w_j^\top x-w_y^\top x]_+.$$

若活动错误类为 j^*，一个次梯度只在两个块非零：$\partial_{w_{j^*}}\ell=x$、$\partial_{w_y}\ell=-x$。这同时给出维度检查和可执行更新。

AdaBoost 在第 t 轮维护样本权重 $D_t(i)$，选择取值于 $\{-1,1\}$ 的弱分类器 h_t。其加权误差为

$$\epsilon_t=\sum_i D_t(i)\mathbf{1}\{h_t(x_i)\neq y_i\},$$

并取

$$\alpha_t=\frac{1}{2}\log\frac{1-\epsilon_t}{\epsilon_t},\qquad D_{t+1}(i)=\frac{D_t(i)e^{-\alpha_t y_i h_t(x_i)}}{Z_t}.$$

最终 $F_T(x)=\sum_t\alpha_t h_t(x)$。

命题 4.3 (AdaBoost 训练误差证书)**.** *若边缘* $\gamma_t=\frac{1}{2}-\epsilon_t>0$，*则*

$$\frac{1}{m}\sum_i\mathbf{1}\{y_iF_T(x_i)\le 0\}\le\prod_{t=1}^T Z_t=\prod_{t=1}^T 2\sqrt{\epsilon_t(1-\epsilon_t)}\le\exp\left(-2\sum_{t=1}^T\gamma_t^2\right).$$

证明. $\mathbf{1}\{u\le 0\}\le e^{-u}$，故训练误差不超过 $\frac{1}{m}\sum_i e^{-y_iF_T(x_i)}$。由权重递推归纳，该平均指数损失正好等于 $\prod_t Z_t$。将正确/错误样本分组得

$$Z_t=(1-\epsilon_t)e^{-\alpha_t}+\epsilon_t e^{\alpha_t}=2\sqrt{\epsilon_t(1-\epsilon_t)}.$$

最后用 $\sqrt{1-4\gamma_t^2}\le e^{-2\gamma_t^2}$。 □

易错点

训练误差指数下降不等于无条件泛化；还要控制基分类器复杂度、轮数或最终间隔分布。structured prediction 的难点常在 loss-augmented inference，而 boosting 的关键条件是每轮相对当前权重分布确有正边缘。

只有 $\alpha_i>0$ 的样本进入 w^*，这些是支持向量。软间隔中约束变为 $0\le\alpha_i\le C$；$\alpha_i=C$ 常对应间隔内或误分类点，但不能仅凭 $\alpha_i=C$ 判断其恰好位于哪一侧。

4.4 核与 RKHS

定义 4.1 (正半定核)**.** 对称函数 $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ 称为正半定核，若对任意 $x_1,\dots,x_m$ 和 $c\in\mathbb{R}^m$，

$$\sum_{i,j}c_ic_jk(x_i,x_j)\ge 0.$$

等价地，任意 Gram 矩阵 $K=(k(x_i,x_j))$ 半正定。

Moore–Aronszajn 定理保证存在 Hilbert 空间 $\mathcal{H}_k$ 和特征映射 ϕ，使

$$k(x,x') = \langle \phi(x), \phi(x') \rangle .$$

特征映射不唯一，也可能无限维；核技巧的价值正是算法只需要内积。

例题 4.2 (二次核分开 XOR)**.** 二维 XOR 点在原空间线性不可分。使用

$$k(x,x') = (1 + x^\top x')^2$$

等价于加入常数、一次项和所有二次单项式，乘积 x_1x_2 成为一个线性坐标，从而可把同号象限与异号象限分开。

定理 4.4 (Representer theorem)**.** *设目标为*

$$\min_{f \in \mathcal{H}_k} \Phi(f(x_1), \dots, f(x_m)) + \Omega(\|f\|_{\mathcal{H}_k}),$$

其中 Ω 严格递增。若最小点存在，则每个最小点都可写成

$$f^*(\cdot) = \sum_{i=1}^{m} \alpha_i k(x_i, \cdot).$$

证明. 令 $\mathcal{M} = \operatorname{span}\{k(x_i,\cdot) : 1 \le i \le m\}$，作正交分解 $f = f_\parallel + f_\perp$。再生性质给

$$f_\perp(x_i) = \langle f_\perp, k(x_i,\cdot) \rangle_{\mathcal{H}_k} = 0,$$

故数据拟合项只依赖 $f_\parallel$。同时

$$\|f\|^2 = \|f_\parallel\|^2 + \|f_\perp\|^2 .$$

若 $f_\perp \neq 0$，删除它保持所有训练预测不变并严格降低正则项，和最优性矛盾。因此 $f_\perp = 0$。 □

易错点

“核矩阵元素非负”既非必要也非充分。正半定要求 $c^\top Kc \ge 0$ 对所有 c 成立。核技巧也不自动降低统计复杂度；复杂度仍由 RKHS 范数约束、核对角和正则化共同决定。

4.5 PCA：最大方差与最小重构误差

设中心化数据矩阵 $X \in \mathbb{R}^{m \times p}$，样本协方差 $\widehat{\Sigma} = X^\top X/m$。

定理 4.5 (第一主成分)**.** *单位向量 v 上投影的样本方差为 $v^\top \widehat{\Sigma} v$。最大方差方向是 $\widehat{\Sigma}$ 最大特征值的特征向量。前 r 个主成分组成的矩阵 V_r 同时最小化*

$$\left\| X - XV_rV_r^\top \right\|_F^2 .$$

证明. Rayleigh 商定理给第一结论。对任意正交列矩阵 V，

$$\left\| X - XVV^\top \right\|_F^2 = \|X\|_F^2 - \operatorname{tr}(V^\top X^\top XV).$$

所以最小化重构误差等价于最大化捕获方差。谱分解或 Courant–Fischer 原理说明最优子空间由前 r 个特征向量张成。 □

若数据未中心化，第一主成分可能主要解释均值而不是变化。PCA 是线性无监督表示，不等同于选择原始变量；loadings 通常是多个变量的线性组合。

例题 4.3 (二维手算)**.** 若

$$\widehat{\Sigma} = \begin{pmatrix} 4 & 3 \\ 3 & 4 \end{pmatrix},$$

特征对为 $(7,(1,1)/\sqrt{2})$ 与 $(1,(1,-1)/\sqrt{2})$。一维 PCA 保留方向 $(1,1)$，解释方差比例为 $7/8$，丢失的平均平方重构误差等于被删特征值 1。

4.6 k-means：交替最小化而非全局保证

k-means 目标为

$$J(c,\mu) = \sum_{i=1}^{m} \|x_i - \mu_{c_i}\|^2, \qquad c_i \in \{1,\dots,k\}.$$

固定中心时，把每个点分配给最近中心使目标最小；固定分配时，对 μ_j 求导得到簇内均值。Lloyd 算法交替执行两步，每一步都不增大 J。有限分配数保证算法在排除平局循环的规则下有限停止，但停止点只保证坐标局部最优，不保证全局最优。

初始化会改变结果。k-means++ 用与已有中心距离平方成比例的概率选新中心，降低极差初始化的风险；它不是“必得全局最优”的证明。

4.7 EM：缺失变量、下界与单调性

设观测为 x，隐变量为 z，参数为 θ。观测对数似然

$$\ell(\theta) = \log p_\theta(x) = \log \sum_z p_\theta(x,z).$$

对任意分布 $q(z)$，Jensen 给

$$\ell(\theta) = \log \mathbb{E}_q \frac{p_\theta(x,Z)}{q(Z)} \geq \mathbb{E}_q \log p_\theta(x,Z) + H(q) =: \mathcal{L}(q,\theta).$$

差距恰为

$$\ell(\theta) - \mathcal{L}(q,\theta) = \mathrm{KL}(q(z) \| p_\theta(z \mid x)).$$

定理 4.6 (EM 单调性)**.** *E 步取*

$$q^{(t)}(z) = p_{\theta^{(t)}}(z \mid x),$$

M 步取任意满足 $\mathcal{L}(q^{(t)},\theta^{(t+1)}) \geq \mathcal{L}(q^{(t)},\theta^{(t)})$ 的参数，则

$$\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)}).$$

证明. E 步使 KL 差距在 $\theta^{(t)}$ 处为零，所以

$$\ell(\theta^{(t)}) = \mathcal{L}(q^{(t)}, \theta^{(t)}).$$

M 步提高下界，而任意参数处下界不超过真实对数似然，因此

$$\ell(\theta^{(t+1)}) \geq \mathcal{L}(q^{(t)}, \theta^{(t+1)}) \geq \mathcal{L}(q^{(t)}, \theta^{(t)}) = \ell(\theta^{(t)}).$$

□

对高斯混合，E 步计算责任度

$$r_{ij} = \frac{\pi_j \varphi(x_i; \mu_j, \Sigma_j)}{\sum_\ell \pi_\ell \varphi(x_i; \mu_\ell, \Sigma_\ell)},$$

M 步用软计数更新 π_j, μ_j, Σ_j。EM 单调提高似然，但可停在局部极值或鞍点；协方差塌缩还可能使无约束混合似然无上界。

本节结论

经典方法看似分散，实则都由“目标—约束/正则—可计算表示—单调或收敛证书”组织。SVM 与核方法连接监督学习和凸优化；PCA、k-means 与 EM 分别代表谱方法、硬交替和软隐变量三种无监督范式。

答题方法

经典方法题先问目标函数，再问每步是否精确最小化某个子问题：

- 感知机：误分类驱动的在线可行性更新；
- SVM：间隔规范化后的凸二次规划；
- PCA：Rayleigh 商或低秩重构；
- k-means：硬分配的交替最小化；
- EM：后验填充的变分下界坐标上升。

不要只写算法步骤而不说明其下降量、条件和失败模式。

自测

1. 感知机证明中哪一步使用了误分类条件？哪一步使用了可分间隔？
2. 从软间隔 primal 推导 $0 \leq \alpha_i \leq C$。
3. 证明线性核、多项式核和高斯核均为正半定核。
4. 用 SVD 证明 PCA 的秩 r 重构误差等于尾部奇异值平方和。
5. 说明 k-means 与等协方差、等先验高斯混合的“小方差极限”联系。

6. EM 的似然单调为何不意味着参数收敛到全局最优？

自测核对要点 第 1 题中, 误分类给 $y_t u^\top x_t \le 0$, 可分间隔给 $y_t u_*^\top x_t \ge \gamma$; 二者分别控制范数增长和正确方向的线性增长。第 2 题对松弛变量的非负约束引入乘子 $\mu_i \ge 0$, 驻点条件 $C-\alpha_i-\mu_i=0$ 与 $\alpha_i,\mu_i \ge 0$ 合并即得 $0 \le \alpha_i \le C$。第 3 题可分别使用显式特征映射、核的非负线性组合与高斯核的幂级数表示。第 4 题由 Eckart–Young 定理, $\|X - X_r\|_F^2 = \sum_{j>r} \sigma_j^2$。第 5 题令各混合分量协方差为 $\tau^2 I$ 并令 $\tau \downarrow 0$, 后验责任度趋于最近中心的硬分配。第 6 题的单调性只说明似然值不下降; 非凸似然仍有局部极值、鞍点、标签交换与协方差塌缩。

4.8 感知机收敛定理：间隔如何控制犯错次数

设 $\|x_i\| \le R$，存在单位向量 w_* 使

$$y_i \langle w_*, x_i \rangle \ge \gamma > 0.$$

感知机在犯错样本上更新 $w_{t+1} = w_t + y_i x_i$。若共犯错 M 次，则

$$\langle w_M, w_* \rangle = \sum_{\text{犯错} i} y_i \langle x_i, w_* \rangle \ge M\gamma.$$

另一方面，犯错意味着 $y_i \langle w_t, x_i \rangle \le 0$，所以

$$\|w_{t+1}\|^2 = \|w_t\|^2 + 2y_i \langle w_t, x_i \rangle + \|x_i\|^2 \le \|w_t\|^2 + R^2,$$

从而 $\|w_M\| \le R\sqrt{M}$。Cauchy–Schwarz 给

$$M\gamma \le \langle w_M, w_* \rangle \le R\sqrt{M}, \qquad M \le (R/\gamma)^2.$$

可分性与正间隔不可删除；数据不可分时算法可能无限循环。

4.9 SVM 对偶、支持向量与核化

硬间隔原问题

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad y_i(w^\top x_i + b) \ge 1$$

的 Lagrangian 为

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

对 w, b 求极小给

$$w = \sum_i \alpha_i y_i x_i, \qquad \sum_i \alpha_i y_i = 0.$$

代回得到对偶

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

互补松弛

$$\alpha_i\left[y_i(w^\top x_i + b) - 1\right] = 0$$

表明 $\alpha_i > 0$ 的样本位于间隔边界，是支持向量。将内积替换为正半定核 $k(x_i, x_j)$，即可在特征空间执行同一对偶。

4.10 正定核与 representer theorem

函数 k 是正定核，指对任意样本 $x_1, \dots, x_m$ 和系数 c，

$$\sum_{i,j} c_i c_j k(x_i, x_j) \ge 0.$$

其 Gram 矩阵因此半正定。和、非负倍数、逐点乘积等运算保持正定性，可据此构造组合核。

定理 4.7 (表示定理的考试级版本)**.** *在 RKHS $\mathcal{H}_k$ 中，若目标为*

$$\min_{f\in\mathcal{H}_k} \Phi(f(x_1), \dots, f(x_m)) + \lambda\|f\|_{\mathcal{H}_k}^2, \qquad \lambda > 0,$$

则存在最优解形如

$$f^*(\cdot) = \sum_{i=1}^m \alpha_i k(x_i, \cdot).$$

证明. 令 $S = \operatorname{span}\{k(x_i,\cdot)\}_{i=1}^m$，将任意 $f = f_S + f_\perp$ 作正交分解。再生性质给

$$f_\perp(x_i) = \langle f_\perp, k(x_i,\cdot)\rangle = 0,$$

故数据拟合项只依赖 f_S。而

$$\|f\|^2 = \|f_S\|^2 + \|f_\perp\|^2 \ge \|f_S\|^2.$$

删去 $f_\perp$ 不增目标，因此至少有一个最优解属于有限维空间 S。 □

4.11 PCA 的两种目标为何等价

中心化数据协方差为

$$S = \frac{1}{m}\sum_{i=1}^m x_i x_i^\top.$$

取正交列矩阵 $U \in \mathbb{R}^{d\times r}$。投影方差为

$$\frac{1}{m}\sum_i \|U^\top x_i\|^2 = \operatorname{tr}(U^\top S U).$$

重构误差为

$$\frac{1}{m}\sum_i \|x_i - UU^\top x_i\|^2 = \operatorname{tr}(S) - \operatorname{tr}(U^\top S U).$$

所以最大方差与最小重构误差等价。Rayleigh–Ritz 原理给出最优 U 由 S 最大的 r 个特征向量组成。若数据未中心化，第一主成分可能主要追踪均值方向。

4.12 k-means 与 EM 的单调性

k-means 目标为

$$J(c,\mu) = \sum_i \|x_i - \mu_{c_i}\|^2.$$

固定中心时，把每个样本分给最近中心逐点最小化 J；固定分配时，组内均值由平方和的一阶条件得到并全局最小化该块。因此每次交替不增目标。由于分配有限，若排除距离并列的循环处理，算法有限进入一个坐标局部最优；这不保证全局最优。

EM 对隐变量 Z 构造

$$Q(\theta \mid \theta^{(t)}) = \mathbb{E}_{\theta^{(t)}}[\log p_\theta(X,Z) \mid X].$$

Jensen 不等式给

$$\log p_\theta(X) \geq Q(\theta \mid \theta^{(t)}) - \mathbb{E}_{q_t} \log q_t(Z),$$

且在 $\theta = \theta^{(t)}$ 处取等。M 步提高右侧，故观测似然不下降。若 M 步只提高而不完全最大化，得到 generalized EM，单调性仍可保留。

易错点

PCA、k-means、EM 都可写成迭代或优化问题，但保证不同：PCA 的标准矩阵问题有全局谱解；k-means 交替最小化通常只到局部解；EM 保证似然不降，不保证参数收敛到全局最大值。

自测

1. 感知机犯错界的两条不等式分别使用了间隔与犯错条件中的哪一条？
2. 推导 SVM 对偶并解释支持向量为何对应非零乘子。
3. 表示定理证明中为何正交分量不影响样本点函数值？
4. 证明 PCA 最大方差与最小重构误差等价。

第五章　三期正式真题与完整解答

本章按年份收入 Machine Learning Theory 的全部正式题面与解答，共保留既有审计口径下的 24 道大题、34 个最细作答单元。除特别注明外，解答为复习用整理稿，不代表官方评分标准。每个题组前的“理论入口”用于回看本 Part 前文。

5.1　2025 春

理论入口：参见章 三、四。
Q1 唯一解答：题面与完整解答（第 5 页）
Q3 唯一解答：题面与完整解答（第 12 页）
Q4 唯一解答：题面与完整解答（第 21 页）

2025 春·MLT·Q2（原卷第 1 页，7 分）

Original. Let $D_1, \dots, D_m$ be distributions over $\mathcal{X}$, let $\mathcal{H}$ be finite, and $f \in \mathcal{H}$. Draw $x_i \sim D_i$ independently and set $y_i = f(x_i)$. Let $\bar{D}_m = (D_1 + \cdots + D_m)/m$. For $\epsilon = 1/4$, show

$$\Pr\{\exists h \in \mathcal{H} : L_{\bar{D}_m, f}(h) > \epsilon, L_{S,f}(h) = 0\} \le |\mathcal{H}| e^{-m/4}.$$

Hint: $\ln(3/4) < -1/4$.

中文译述

样本独立但不必同分布，第 i 个样本来自 D_i。证明在平均分布上错误率超过 $1/4$ 的假设仍与全部训练样本一致的概率至多为 $|\mathcal{H}|e^{-m/4}$。

思路：固定坏假设，乘独立“未犯错”概率，再用 $1-u \le e^{-u}$ 和并合界。

完整解答：Q2

固定 h，记 $p_i = \Pr_{D_i}(h(X) \neq f(X))$。则 $L_{\bar{D}_m,f}(h) = m^{-1}\sum_i p_i$。若它大于 $1/4$，独立性给

$$\Pr(L_{S,f}(h) = 0) = \prod_i (1 - p_i) \le \exp(-\sum_i p_i) < e^{-m/4}.$$

对所有满足真实平均错误率大于 $1/4$ 的 h 作并合界，数量不超过 $|\mathcal{H}|$，得到所求。这里不需要同分布，但需要各 x_i 独立。

易错点：把 D_i 错写成共同分布；把平均错误率条件误变成每个 $p_i > 1/4$。

正文对应：节 3.3中的有限假设类 PAC 证明。

理论入口：参见章 三、四。

5.2 2025 秋

理论入口：参见章 三、四。

2025 秋·MLT·Part I Q1–Q8（原卷第 1–2 页，5 分 +1 bonus）

Original. Select one answer for each question; no justification is required.

1. Which best describes the bias–variance tradeoff? (a) Increasing bias reduces variance but increases approximation error; (b) reduces both; (c) variance is independent; (d) only applies to neural networks.
2. Why does No Free Lunch matter? (a) every algorithm is optimal; (b) all tasks require exponential time; (c) no universally superior algorithm across all tasks; (d) only smooth losses can be optimized.
3. Which is true of Rademacher complexity? (a) it measures fitting random labels; (b) always equals VC dimension; (c) decreases with model complexity; (d) is independent of sample size.
4. In ERM, what is minimized? (a) true risk; (b) average training loss; (c) VC dimension; (d) variance.
5. Which VC statement is true? (a) infinite VC dimension can never be PAC learnable; (b) halfspaces in $\mathbb{R}^d$ have VC dimension $d+1$; (c) VC dimension always equals parameter count; (d) finite VC dimension implies zero generalization error.
6. In agnostic PAC, what changes? (a) zero training error is required; (b) the class contains the target; (c) compete with the best hypothesis in the class despite noise; (d) sample complexity is independent of ϵ.
7. Why can a decreasing SGD learning rate help? (a) eliminates overfitting; (b) ensures convergence under suitable convexity assumptions; (c) increases gradient variance; (d) eliminates backpropagation.

8 (bonus). Sauer's Lemma implies: (a) shatter beyond d; (b) growth is polynomial once $m > d$; (c) ERM has zero risk for $m \le d$; (d) sample complexity independent of d.

中文译述

八道单选依次考查偏差–方差、No Free Lunch、Rademacher、ERM、VC 维、agnostic PAC、递减步长和 Sauer 引理；第 8 题为 1 分附加题。

完整解答：Q1–Q8

答案依次为 **(a), (c), (a), (b), (b), (c), (b), (b)**。Q1 的偏差增大通常降低估计方差但增大逼近误差；Q2 说明学习保证必须依赖归纳偏置；Q3 的经验 Rademacher 直接测量对

随机符号的拟合；Q4 是 ERM 定义；Q5 的齐次/仿射半空间约定需看是否含偏置，本题用含偏置的 $d+1$；Q6 是 agnostic 超额风险；Q7 递减步长可压低噪声平台，但仍需步长序列与凸性条件；Q8 由 Sauer 得 $\Pi_{\mathcal{H}}(m) \le \sum_{i=0}^{d} \binom{m}{i} = O(m^d)$。
正文对应： 章 三、节 12.1中的 PAC/VC/Rademacher 与 SGD。

理论入口： 参见章 三、四。

2025 秋·MLT·Q9（原卷第 2 页，5 分）

Original. Let $\mathcal{H} = \{h_{a,b,s} : a \le b, s \in \{-1, 1\}\}$, where $h_{a,b,s}(x) = s$ for $x \in [a, b]$ and $-s$ outside. Calculate VCdim($\mathcal{H}$).

中文译述

求实线上的“带符号区间”类 VC 维。

完整解答：Q9

三个有序点可被打散：正类集合既可取一个区间（$s=1$），也可取一个区间的补集（$s=-1$），从而包含模式 010 与 101，其余模式更直接。四个有序点不能被打散，因为任一假设沿直线最多改变标签两次，而模式 1010 有三次改变。因此 VC 维为 3。
常见错误： 只考虑 $s=1$ 得到普通区间 VC 维 2。

理论入口： 参见章 三、四。

2025 秋·MLT·Q10（原卷第 2–3 页，6 分）

Original. Prove: (a) $f(w) = \lambda\|w\|^2$ is 2λ-strongly convex; (b) if f is λ-strongly convex and g convex, then $f+g$ is λ-strongly convex; (c) if u minimizes a λ-strongly convex f, then $f(w) - f(u) \ge \lambda\|w-u\|^2/2$.

中文译述

证明二次正则的强凸常数、强凸与凸函数相加保持强凸，以及最优点二次增长。

完整解答：Q10

(a) Hessian 为 $2\lambda I$，或直接展开强凸定义。(b) 将 f 的强凸一阶不等式与 g 的凸一阶不等式相加。(c) 最优性给 $0 \in \partial f(u)$，将 $x = u, y = w, g = 0$ 代入强凸次梯度不等式即得。
关键条件： 若不可微，(c) 使用凸次微分的 Fermat 条件；最小点必须存在。

理论入口： 参见章 三、四。

2025 秋·MLT·Q11（原卷第 3 页，5 分）

Original. Let $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{H} = \{h_r : r \in \mathbb{R}_+\}$, $h_r(x) = \mathbf{1}\{\|x\| \le r\}$. Prove realizable PAC learnability and $m_{\mathcal{H}}(\epsilon, \delta) \le \log(1/\delta)/\epsilon$.

中文译述

对以原点为心的圆盘阈值类，构造一致学习器并证明给定样本复杂度。

完整解答：Q11

令目标半径 r^*。学习器输出训练集中正样本的最大半径 $\hat{r}$（没有正样本则取 0），故 $\hat{r} \le r^*$ 且训练一致；错误只可能出现在环带 $A = \{x : \hat{r} < \|x\| \le r^*\}$。若某个内层半径 r_ϵ 使外侧目标环带质量至少 ϵ，学习失败意味着 m 个样本都未落入该环带，概率至多 $(1-\epsilon)^m \le e^{-m\epsilon}$。令其不超过 δ 得 $m \ge \log(1/\delta)/\epsilon$。等价地，该类 VC 维为 1，但直接一致性证明给出题目常数。

理论入口： 参见章 三、四。

2025 秋·MLT·Q12（原卷第 3 页，7 分）

Original. Let $\hat{g}_t$ be an unbiased estimate of $g_t \in \partial F(w_t)$, $w_{t+1} = \Pi_W(w_t - \eta_t \hat{g}_t)$, and $\mathbb{E}\|\hat{g}_t\|^2 \le G^2$. Show for every $w \in W$,

$$\mathbb{E}\|w_{t+1} - w\|^2 \le \mathbb{E}\|w_t - w\|^2 - 2\eta_t \mathbb{E}\langle g_t, w_t - w\rangle + \eta_t^2 G^2.$$

The stated schedule is $\eta_t = 1/(\lambda t)$.

中文译述

用投影非扩张性和随机次梯度无偏性证明一步距离递推。

完整解答：Q12

因 $w \in W$ 且投影非扩张，

$$\|w_{t+1} - w\|^2 \le \|w_t - w - \eta_t \hat{g}_t\|^2 = \|w_t - w\|^2 - 2\eta_t \langle \hat{g}_t, w_t - w\rangle + \eta_t^2 \|\hat{g}_t\|^2.$$

条件于历史取期望，$\mathbb{E}_t \hat{g}_t = g_t$，再用二阶矩界并取全期望即得。强凸性和具体 η_t 对本一步不等式不是必需，但用于后续求和率。

理论入口： 参见章 三、四。

2025 秋·MLT·Q13（原卷第 3 页，5 分）

Original. Let $f : [-1,1]^n \to [-1,1]$ be ρ-Lipschitz. For any $\epsilon > 0$ construct a sigmoid network N with $|f(x) - N(x)| \le \epsilon$ for all x. Hint: partition into boxes, use Lipschitzness, detect the box, and output its average value.

中文译述

按提示构造 sigmoid 网络一致逼近 Lipschitz 函数。

完整解答：Q13

取网格边长 $a < \epsilon/(3\rho\sqrt{n})$，每个盒 B_j 选中心 c_j 与常数 $v_j = f(c_j)$，则盒内 $|f(x)-v_j| < \epsilon/3$。用大斜率 sigmoid $\sigma(M(x_i-l_{ji}))$ 与 $\sigma(M(u_{ji}-x_i))$ 近似每个坐标的上下界指示；把 $2n$ 个近似指示的和送入第二个大斜率 sigmoid，得到盒门 $G_j(x)$。在盒边界使用相邻盒的 partition-of-unity 归一化（或先取略重叠盒）避免硬阈值不连续，令 $N(x) = \sum_j v_j G_j(x) / \sum_j G_j(x)$。有限盒数使可选统一足够大的 M，门控误差贡献小于 $2\epsilon/3$；加上盒内误差得一致误差不超过 ϵ。最后一层可用仿射组合并将范围裁入 $[-1, 1]$。

易错点： 直接声称 sigmoid 等于不连续盒指示；忽略边界和统一误差。

5.3 2026 春

理论入口： 参见章 三、四。

2026 春·MLT·MQ1–MQ6（原卷第 1–2 页，共 3 分）

Original. Select one answer.

MQ1. In agnostic PAC, compete with (a) Bayes; (b) zero-training-error hypothesis; (c) the best hypothesis in $\mathcal{H}$; (d) validation minimizer.

MQ2. If $\operatorname{VCdim}\mathcal{H} = d < \infty$, which is necessary? (a) $\tau(m) = 2^m$; (b) $O(m^d)$ for all m; (c) $\tau(m) \le \sum_{i=0}^{d} \binom{m}{i}$ for all m; (d) constant after d.

MQ3. Which ReLU statement is correct? (a) smooth functions; (b) depth always lowers VC; (c) piecewise linear with complexity depending on depth/width; (d) universal approximation implies PAC learnability.

MQ4. Which ERM statement is correct? (a) consistent for any class; (b) minimizes true risk; (c) finite VC implies realizable PAC learnability by ERM; (d) only finite classes.

MQ5. A consequence of No Free Lunch is (a) restricting always improves; (b) universal optimum exists; (c) inductive bias is needed for nontrivial guarantees; (d) random guessing always optimal.

MQ6. For symmetric K, which is correct? (a) $K(x,x) \ge 0$ suffices; (b) valid iff a finite-dimensional feature map exists; (c) PSD implies a possibly infinite-dimensional Hilbert feature map; (d) feature map is unique.

中文译述

六道单选依次考 agnostic PAC、Sauer、ReLU 表达、ERM、No Free Lunch 和核的 Hilbert 表示。

完整解答：MQ1–MQ6

答案均为 **(c)**。MQ2 中 Sauer 上界对所有 m 可写该求和（$m \le d$ 时等于 2^m）；MQ6 是 Moore–Aronszajn 表示，特征映射不唯一且可无限维。

理论入口： 参见章 三、四。

2026 春·MLT·Q1（原卷第 2 页，6 分）

Original. For fixed binary h, $p = L_{D,f}(h)$ and $L_S(h) = m^{-1}\sum_i \mathbf{1}[h(x_i) \neq f(x_i)]$: (a) show $\mathbb{E}(L_S(h) - p)^2 = p(1-p)/m$; (b) deduce it is at most $1/(4m)$ and hence $O(1/m)$.

中文译述

计算固定分类器经验错误的均方误差并用 Bernoulli 方差最大值给上界。

完整解答：Q1

完整推导见第 1.6 节的固定假设经验风险方差题（第 5 页）。这里令 $Z_i = \mathbf{1}\{h(X_i) \neq f(X_i)\}$，则 Z_i 独立同分布为 Bernoulli(p)，所以 L_S 无偏且方差为 $p(1-p)/m$，均方误差等于该方差。因 $p(1-p) = 1/4 - (p-1/2)^2 \leq 1/4$，得 $1/(4m)$。

理论入口： 参见章 三、四。

2026 春·MLT·Q2（原卷第 2 页，4 分）

Original. Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable and λ-strongly convex. Prove it has a unique global minimizer.

中文译述

证明有限维强凸光滑函数的全局最小点存在且唯一。

完整解答：Q2

强凸性在 0 点给 $f(x) \geq f(0) + \nabla f(0)^T x + \lambda\|x\|^2/2 \to \infty$ 当 $\|x\| \to \infty$，故 f coercive。连续性使某个下水平集非空紧，Weierstrass 给存在性。若 $x \neq y$ 都最优，强凸中点不等式给严格低于最优值，矛盾。

理论入口： 参见章 三、四。

2026 春·MLT·Q3（原卷第 3 页，8 分）

Original. On $\mathbb{R}$, let $\mathcal{H} = \{h_{a,b,c}(x) = \mathbf{1}_{[a,b]}(x) \vee \mathbf{1}_{[c,\infty)}(x) : a \leq b < c\}$. (a) for n distinct points give an upper bound on $s(\mathcal{H}, n)$; (b) determine VC dimension.

中文译述

假设的正类是一个有界区间与其右侧射线之并，求增长函数上界和 VC 维。

完整解答：Q3

在 n 个有序点上，a, b, c 各只需落在相邻点形成的 $n+1$ 个间隙，因此粗略上界 $s(\mathcal{H}, n) \leq (n+1)^3$（也可进一步去重）。三个点可打散：唯一非普通区间模式 101 可由左侧单点区间加右侧射线实现。四点标记 1010 不可实现：一个有界正块加一个正后缀不能产生两个分离

正点而最后一点为负。因此 VC 维为 3，Sauer 又给更紧上界 $s(\mathcal{H},n)\le\sum_{i=0}^{3}\binom{n}{i}$。

理论入口： 参见章 三、四。

2026 春·MLT·Q4（原卷第 3 页，7 分）

Original. Examples $(x,y)\in\mathbb{R}^n\times[k]$ lie in radius-r balls around μ_y, $r\ge 1$, and $\|\mu_i-\mu_j\|\ge 4r$. With the multiclass feature map $\Psi(x,y)$ placing $(x,1)$ in block y, show a vector w gives zero multiclass loss. Hint: use blocks $w_i=(\mu_i,-\|\mu_i\|^2/2)$.

中文译述

用带偏置的多向量构造证明间隔分离的多类样本可被零损失分类。

完整解答：Q4

类别 i 分数 $s_i(x)=\mu_i^Tx-\|\mu_i\|^2/2$。写 $x=\mu_y+v$、$\|v\|\le r$，则对 $i\ne y$，

$$s_y-s_i=\frac{1}{2}\|\mu_y-\mu_i\|^2+(\mu_y-\mu_i)^Tv\ge\frac{1}{2}(4r)^2-4r^2=4r^2\ge 1.$$

因此正确类分数至少高出 1，标准 multiclass hinge $\max_{i\ne y}[1+s_i-s_y]_+$ 为零；所有样本均如此。

关键条件： 最后一维常数 1 产生 $-\|\mu_i\|^2/2$ 偏置；没有它无法得到距离差恒等式。

理论入口： 参见章 三、四。

2026 春·MLT·Q5（原卷第 3 页，5 分）

Original. A finite ReLU feedforward network and a finite sigmoid feedforward network with the same input dimension and scalar output represent the same function on all of $\mathbb{R}^d$. Show the function must be constant.

中文译述

证明全空间上同时可由有限 ReLU 网络和有限 sigmoid 网络表示的标量函数只能是常数。

完整解答：Q5

有限 ReLU 网络是有限多胞形分区上的分段仿射函数。有限 sigmoid 网络由实解析函数复合而成，故输出实解析。在每个 ReLU 线性区域内部，所有二阶偏导为零；至少有一个非空开区域，解析函数的二阶偏导在开集为零，由解析延拓在连通的 $\mathbb{R}^d$ 上处处为零，所以公共函数是全局仿射 a^Tx+b。sigmoid 隐层输出有界，有限线性组合仍有界（若输出也 sigmoid 更明显）；全局仿射函数有界只能 $a=0$，故为常数。

易错点： 只说 ReLU 分段线性、sigmoid 光滑，未用解析延拓和有界性闭合证明。

5.4 高频错误、作答模板与复习检查

固定假设集中界不能直接套在数据依赖输出上；realizable 与 agnostic 的 $1/\epsilon$、$1/\epsilon^2$ 量级不可混写；VC 界必须声明常数/对数版本；Rademacher 对称化不可交换期望与上确界；SVM 要写偏置对应的对偶等式约束；online regret 是序列比较，不是 i.i.d. 泛化概率。

复习检查

考场模板：先写学习设定和比较对象，再选择集中/复杂度工具；证明中标出概率至少 $1-\delta$ 的事件；模型题写目标、约束、对偶、KKT 和预测规则。完成后检查是否能证明有限类界、Sauer/VC 路线、对称化和 Representer 定理，并手算 Bayes 风险、SVM 对偶和核岭系统。

第二部分

Deep Learning and Reinforcement Learning

深度学习与强化学习

第六章　深度网络的计算、反向传播与尺度

学习目标

本章把深度网络还原成带维度的计算图：读者应能逐层写出张量形状，手算反向传播，解释初始化、归一化、正则化和训练/推理状态的差异，并判断梯度消失或爆炸来自哪一段 Jacobian 链。

6.1　计算图、前向传播与反向传播

对 L 层前馈网络，令 $a^{(0)} = x$，

$$z^{(\ell)} = W^{(\ell)} a^{(\ell-1)} + b^{(\ell)}, \qquad a^{(\ell)} = \phi_\ell(z^{(\ell)}), \quad 1 \le \ell \le L. \tag{6.1}$$

分类常以 $p = \mathrm{softmax}(z^{(L)})$，交叉熵 $\mathcal{L} = -\sum_c y_c \log p_c$。单样本的 softmax–交叉熵梯度是

$$\frac{\partial \mathcal{L}}{\partial z^{(L)}} = p - y. \tag{6.2}$$

这条抵消关系是手算反向传播的起点。

完整证明. 写 $p_i = e^{z_i} / \sum_j e^{z_j}$，则

$$\frac{\partial p_i}{\partial z_j} = p_i(\mathbf{1}\{i = j\} - p_j).$$

交叉熵为 $\mathcal{L} = -\sum_i y_i \log p_i$，且 one-hot 或一般概率标签满足 $\sum_i y_i = 1$。因此

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial z_j} &= -\sum_i \frac{y_i}{p_i} p_i(\mathbf{1}\{i = j\} - p_j) \\ &= -y_j(1 - p_j) + p_j \sum_{i \ne j} y_i = p_j - y_j. \end{aligned}$$

若标签未归一化，最后一步不能使用；若损失还取 batch 均值，梯度需再除以 batch 大小。　□

例题 6.1 (三分类 softmax–交叉熵). 令 logits $z = (\log 2, 0, 0)$，真实类别为第一类。则 $p = (1/2, 1/4, 1/4)$，$\mathcal{L} = -\log(1/2) = \log 2$，且 $\nabla_z \mathcal{L} = (-1/2, 1/4, 1/4)$。梯度分量之和为零，反映 softmax 对所有 logits 同加常数的不变性；若求出不为零的和，通常是 Jacobian 或标签处理有误。

定理 6.1 (反向传播递推)**.** *设损失为标量，逐元素激活可微。记 $\delta^{(\ell)}=\partial\mathcal{L}/\partial z^{(\ell)}$，则*

$$\delta^{(L)}=\nabla_{a^{(L)}}\mathcal{L}\odot\phi_L'(z^{(L)}),\quad \delta^{(\ell)}=(W^{(\ell+1)})^\top\delta^{(\ell+1)}\odot\phi_\ell'(z^{(\ell)}),$$

并且

$$\nabla_{W^{(\ell)}}\mathcal{L}=\delta^{(\ell)}(a^{(\ell-1)})^\top,\qquad \nabla_{b^{(\ell)}}\mathcal{L}=\delta^{(\ell)}.$$

完整证明. 由式(6.1)，$z^{(\ell+1)}$ 对 $a^{(\ell)}$ 的 Jacobian 为 $W^{(\ell+1)}$。反向模式链式法则给出

$$\nabla_{a^{(\ell)}}\mathcal{L}=(W^{(\ell+1)})^\top\nabla_{z^{(\ell+1)}}\mathcal{L}.$$

逐元素激活的 Jacobian 为 $\mathrm{diag}(\phi_\ell'(z^{(\ell)}))$，故左乘它得到 $\delta^{(\ell)}$ 的递推。又 $z_i^{(\ell)}=\sum_j W_{ij}^{(\ell)}a_j^{(\ell-1)}+b_i^{(\ell)}$，所以 $\partial\mathcal{L}/\partial W_{ij}^{(\ell)}=\delta_i^{(\ell)}a_j^{(\ell-1)}$，写成矩阵即结论。证明的维度检查为 $\delta^{(\ell)}\in\mathbb{R}^{n_\ell}$、$a^{(\ell-1)}\in\mathbb{R}^{n_{\ell-1}}$，外积恰为 $n_\ell\times n_{\ell-1}$。 □

例题 6.2 (两层网络手算)**.** 令 $x=(1,-1)^\top$，$W^{(1)}=I$，$b^{(1)}=0$，激活为 ReLU，$W^{(2)}=(2,-1)$，目标 $y=1$，损失 $\frac12(a^{(2)}-y)^2$。前向得到 $z^{(1)}=(1,-1)^\top$、$a^{(1)}=(1,0)^\top$、$a^{(2)}=2$。于是 $\delta^{(2)}=1$，$\delta^{(1)}=(2,-1)^\top\odot(1,0)^\top=(2,0)^\top$，$\nabla W^{(2)}=(1,0)$，$\nabla W^{(1)}=(2,0)^\top(1,-1)=\begin{pmatrix}2&-2\\0&0\end{pmatrix}$。易错点是把 ReLU 在负预激活处的导数写成 1。

6.2 激活、初始化与梯度传播

激活	公式	特点	主要风险
sigmoid	$(1+e^{-z})^{-1}$	输出可作概率门控	饱和区梯度小，非零中心
tanh	$\tanh z$	零中心、RNN 常用	仍会饱和
ReLU	$\max(0,z)$	正区梯度稳定、稀疏激活	dying ReLU，零点次梯度约定
GELU	$z\Phi(z)$	平滑门控，Transformer 常用	计算较复杂

必须能同时写出导数:

$$\sigma'(z)=\sigma(z)(1-\sigma(z)),\qquad(\tanh z)'=1-\tanh^2 z,\qquad \mathrm{ReLU}'(z)=\mathbf{1}\{z>0\},$$

以及 $\mathrm{GELU}'(z)=\Phi(z)+z\varphi(z)$（精确高斯 CDF 版本）。ReLU 在 $z=0$ 不可微，软件通常选定某个次梯度；这不影响几乎处处反传，但答题应说明约定。

命题 6.2 (通用逼近的考试级表述)**.** *设 $K\subset\mathbb{R}^d$ 为紧集，激活函数 ϕ 连续且不是多项式（例如 sigmoid、tanh 或 ReLU）。则单隐层网络的有限线性组合 $\sum_{j=1}^m a_j\phi(w_j^\top x+b_j)$ 在 $C(K)$ 中稠密：对任意连续 f 与 $\epsilon>0$，存在有限宽网络使 $\sup_{x\in K}|f(x)-f_m(x)|<\epsilon$。*

证明思路

标准证明用 Hahn–Banach：若网络张成空间不稠密，则存在非零有限符号测度 μ 消去所有 ridge 函数 $\int \phi(w^\top x + b)\, d\mu(x) = 0$；对 w, b 变化并利用非多项式激活的判别性质可推出 $\mu = 0$，矛盾。完整泛函分析细节省略。结论只说明“存在足够宽的近似网络”，不提供所需宽度、训练算法或泛化保证，也不意味着梯度下降一定找到它。

命题 6.3 (ReLU 深度分离的一个可手证模型)**.** *一维单隐层、宽度 m 的 ReLU 网络至多有 m 个折点，因而至多 $m+1$ 个线性片段。另一方面，固定宽度的 ReLU “帐篷映射”复合 L 次可产生 2^L 个交替线性片段。因此要精确表示该复合函数，单隐层网络必须有 $m \geq 2^L - 1$。*

完整证明. 单个 $\mathrm{ReLU}(a_j x + b_j)$ 只在 $a_j x + b_j = 0$ 处改变斜率；m 个此类函数的线性组合的折点集合包含在这 m 个位置中，所以线性片段不超过 $m+1$。取可由常数个 ReLU 实现的三角形帐篷函数 $T : [0,1] \to [0,1]$，它在左右半区分别线性上升、下降。复合 $T^{\circ L}$ 时，每个既有单调片段都被再分成两个，归纳得 2^L 个片段。若单隐层网络精确等于它，必须至少拥有 $2^L - 1$ 个折点。该下界针对特定函数、ReLU、精确表示和一维域；不能泛化成“任何深网络都指数优于浅网络”。 □

若 $z_j = \sum_{i=1}^{n_{\text{in}}} W_{ji} a_i$，假设独立、零均值，则 $\mathrm{Var}(z_j) \approx n_{\text{in}} \mathrm{Var}(W_{ji}) \mathrm{Var}(a_i)$。Xavier 初始化对近线性/对称激活取 $\mathrm{Var}(W) \approx 2/(n_{\text{in}} + n_{\text{out}})$；He 初始化针对 ReLU 约一半单元被截断，取 $\mathrm{Var}(W) \approx 2/n_{\text{in}}$。这些推导依赖独立与方差稳定近似，不是训练全程的精确定理。

深链梯度包含 Jacobian 乘积。若奇异值长期小于 1，梯度消失；长期大于 1，梯度爆炸。残差连接 $h_{\ell+1} = h_\ell + F_\ell(h_\ell)$ 使 Jacobian 为 $I + J_{F_\ell}$，给梯度提供恒等路径，但不保证任意深度都稳定。

6.3 正则化、归一化与泛化

- **权重衰减**：$J(w) + \frac{\lambda}{2}\|w\|^2$ 的梯度更新为 $w^+ = (1 - \eta\lambda)w - \eta \nabla J(w)$。对 Adam 等自适应法，解耦式 weight decay 与把 L_2 项混入梯度并不完全等价。
- **Dropout**：训练时用掩码 $m_j \sim \mathrm{Bernoulli}(q)$，$\widetilde{h}_j = m_j h_j / q$，故 $\mathbb{E}\widetilde{h}_j = h_j$；测试时关闭随机掩码。它近似集成并抑制共适应。
- **Early stopping**：以验证集选择停止点，是算法层正则化；反复查看测试集会造成测试泄漏。
- **数据增强**：应保持标签语义。旋转对自然图像可能合理，对数字 6/9 未必合理。

6.3.1 训练、验证与测试：三个集合不能互换

训练集用于计算梯度；验证集用于选架构、学习率、正则强度与停止时刻；测试集只在这些选择冻结后评估一次。若根据测试结果反复改模型，测试集事实上已参与模型选择，报告的测试误差不再是未见数据上的诚实估计。设经验训练风险与验证风险分别为

$$\widehat{R}_{\text{tr}}(\theta), \qquad \widehat{R}_{\text{val}}(\theta).$$

训练风险持续下降而验证风险先降后升，是过拟合的常见信号，但不是逻辑上的充要条件：训练、验证分布不一致、标签噪声或验证集过小也会产生类似曲线。Early stopping 应保存验证指标最优时的参数，而不是最后一次迭代的参数。

数据增强可写成变换分布下的目标

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{T \sim \mathcal{A}} \ell(f_\theta(Tx_i), y_i).$$

该式隐含 T 保持任务标签；若变换改变语义，增强会系统性引入错标。Dropout 则改变网络内部随机计算图，不能与输入增强混为同一机制。分类评估还必须先看类别不平衡与错误代价：accuracy、macro-F1、AUROC 与校准误差回答的问题不同。

答题方法

训练流程题按“数据划分 → 训练目标 → 验证选择 → 一次测试 → 失效诊断”作答。若调了 M 组超参数，应承认验证集承担了 M 次比较；数据很少时用嵌套交叉验证，而不是让同一测试集同时选模型和报分。

BatchNorm 对一个 mini-batch 和每个通道计算均值方差，

$$\widehat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}}, \qquad y_i = \gamma \widehat{x}_i + \beta.$$

训练与推断统计不同。LayerNorm 对单样本的特征维归一化，推断不依赖 batch，因而适合 Transformer 和小批量。归一化有助于优化，但不能替代泛化控制。

答题方法

网络推导先列每层输入/输出形状，再写局部 Jacobian；反向传播只做向量–Jacobian 乘积。初始化题检查前向方差与反向方差，归一化题必须区分训练统计量和推理统计量。

自测

1. 对两层 MLP 写出每个参数梯度的形状，并说明批量维度如何求和。
2. 从独立零均值权重假设推导 Xavier 与 He 方差标度。
3. 比较 batch normalization 与 layer normalization 在训练和推理阶段使用的统计量。
4. 构造一个饱和激活导致梯度消失的数值例，并给出至少两种修复策略。

本节结论

深度网络的难点不是记住模块名，而是同时保持维度、概率尺度和训练状态一致。反向传播、初始化和归一化共同决定信号能否跨层传播。

6.4　张量维度与参数量：先写形状，再写公式

深度网络题最常见的失分不是不会求导，而是没有先固定批量、样本与特征轴。以下统一采用“样本在行”的约定。令

$$X \in \mathbb{R}^{B\times d_0}, \qquad W_\ell \in \mathbb{R}^{d_{\ell-1}\times d_\ell}, \qquad b_\ell \in \mathbb{R}^{d_\ell}.$$

广播后的仿射层与激活为

$$Z_\ell = H_{\ell-1}W_\ell + \mathbf{1}_B b_\ell^\top, \qquad H_\ell = \phi_\ell(Z_\ell).$$

因此一个全连接网络的可训练参数总数是

$$\sum_{\ell=1}^{L}(d_{\ell-1}d_\ell + d_\ell).$$

这个数字只描述参数化规模；它既不是计算复杂度，也不是 VC 维。一次稠密前向传播的主复杂度为 $\sum_\ell O(Bd_{\ell-1}d_\ell)$，而 VC 维还依赖激活形式、网络深度和允许的实数运算。

例题 6.3. 输入维数 64，两层隐藏宽度分别为 $128, 32$，十分类输出。参数量为

$$64\cdot 128 + 128 + 128\cdot 32 + 32 + 32\cdot 10 + 10 = 12778.$$

若只写三个权重矩阵的元素数，会漏掉 170 个偏置参数。若批量大小从 32 改成 256，参数量不变，但前向乘加量近似增加八倍。

6.5　向量–Jacobian 反向传播的完整维度链

设标量损失为

$$\mathcal{L} = \frac{1}{B}\sum_{i=1}^{B}\ell\big((H_L)_{i:}, y_i\big).$$

记 $G_\ell = \partial\mathcal{L}/\partial Z_\ell \in \mathbb{R}^{B\times d_\ell}$。链式法则给出

$$\frac{\partial\mathcal{L}}{\partial W_\ell} = H_{\ell-1}^\top G_\ell \in \mathbb{R}^{d_{\ell-1}\times d_\ell}, \qquad \frac{\partial\mathcal{L}}{\partial b_\ell} = G_\ell^\top \mathbf{1}_B \in \mathbb{R}^{d_\ell},$$

以及

$$\frac{\partial\mathcal{L}}{\partial H_{\ell-1}} = G_\ell W_\ell^\top, \qquad G_{\ell-1} = \left(G_\ell W_\ell^\top\right)\odot \phi'_{\ell-1}(Z_{\ell-1}).$$

每一步都可由微分恒等式验证。例如

$$\mathrm{d}\mathcal{L} = \langle G_\ell,\, \mathrm{d}H_{\ell-1}W_\ell + H_{\ell-1}\,\mathrm{d}W_\ell + \mathbf{1}_B\,\mathrm{d}b_\ell^\top\rangle_F,$$

再利用 $\langle A, BC\rangle_F = \langle AC^\top, B\rangle_F$ 读取三个梯度。这个推导避免把 Jacobian 显式展开成四阶张量。

例题 6.4. 二分类输出 $z = Xw + b\mathbf{1}$，损失

$$\mathcal{L} = -\frac{1}{B}\sum_i \{y_i \log \sigma(z_i) + (1-y_i)\log[1-\sigma(z_i)]\}.$$

先对单点求导得 $\partial \ell_i/\partial z_i = \sigma(z_i) - y_i$，于是

$$\nabla_w \mathcal{L} = \frac{1}{B} X^\top(\sigma(z) - y), \qquad \partial_b \mathcal{L} = \frac{1}{B}\mathbf{1}^\top(\sigma(z) - y).$$

“sigmoid 加交叉熵”的简化发生在对数似然层面，不是把 sigmoid 的导数任意删去。

6.6 残差网络的恒等路径与梯度

普通深层堆叠把每层写成 $x_{\ell+1} = F_\ell(x_\ell)$，反向梯度包含一长串 Jacobian 乘积。残差块改写为

$$x_{\ell+1} = x_\ell + F_\ell(x_\ell; \theta_\ell),$$

于是

$$\frac{\partial x_{\ell+1}}{\partial x_\ell} = I + J_{F_\ell}(x_\ell), \qquad \frac{\partial L}{\partial x_\ell} = \frac{\partial L}{\partial x_{\ell+1}} + J_{F_\ell}(x_\ell)^\top \frac{\partial L}{\partial x_{\ell+1}}.$$

第一项提供不经过残差分支的恒等梯度路径；这解释了为什么把残差分支初始化在较小尺度时，深层网络更接近恒等映射且更易优化。

命题 6.4 (残差链的梯度乘积)**.** *对 L 个残差块,*

$$\frac{\partial x_L}{\partial x_0} = \prod_{\ell=0}^{L-1} (I + J_{F_\ell}(x_\ell)),$$

乘积按前向复合的逆序作用于反向向量。若 $\|J_{F_\ell}\| \le \varepsilon_\ell$，则

$$\left\| \frac{\partial x_L}{\partial x_0} \right\| \le \prod_{\ell=0}^{L-1} (1+\varepsilon_\ell) \le \exp\left(\sum_{\ell=0}^{L-1} \varepsilon_\ell \right).$$

证明. 链式法则逐层给出 Jacobian 乘积。再用诱导范数的次乘性与 $\|I + J\| \le 1 + \|J\|$，最后使用 $1 + u \le e^u$。该上界只控制爆炸；它不保证最小奇异值远离零，因此不能把“有恒等路径”误写成“梯度永不消失”。 □

例题 6.5 (标量残差块与普通块)**.** 若普通块为 $x_{\ell+1} = a_\ell x_\ell$，梯度因子为 $\prod_\ell a_\ell$。残差块 $x_{\ell+1} = x_\ell + a_\ell x_\ell$ 的因子为 $\prod_\ell (1 + a_\ell)$。当 a_ℓ 都接近零时，后者接近一；但若某层 $a_\ell = -1$，梯度仍会归零。这给出残差结构作用与边界的同时说明。

6.7 初始化的前向与反向方差推导

考虑 $z_j = \sum_{i=1}^{d} W_{ij} h_i$，假设 W_{ij} 独立、均值为零，并与 h_i 独立。忽略相关性后，

$$\mathrm{Var}(z_j) = d\ \mathrm{Var}(W_{ij})\,\mathrm{Var}(h_i).$$

若激活近似线性且希望层间方差不变，应取 $\mathrm{Var}(W_{ij}) \approx 1/d$，即 Xavier 尺度。对 ReLU，若 z 关于零对称，

$$\mathbb{E}[\mathrm{ReLU}(z)^2] = \frac{1}{2}\mathbb{E}z^2,$$

故前向保方差要求

$$\mathrm{Var}(W_{ij}) \approx \frac{2}{d},$$

即 He 尺度。

反向传播中

$$g_i = \sum_{j=1}^{d_{\text{out}}} W_{ij}\phi'(z_j)g'_j.$$

对 ReLU，$\mathbb{E}[\phi'(z_j)^2] \approx 1/2$，于是

$$\mathrm{Var}(g_i) \approx \frac{d_{\text{out}}}{2}\,\mathrm{Var}(W_{ij})\,\mathrm{Var}(g'_j).$$

若输入、输出宽度差异很大，只满足前向或只满足反向都可能造成尺度漂移；Xavier 的 fan-in/fan-out 平均形式正是在二者之间折中。

反例 6.1. 把所有权重初始化为同一个非零常数并不能“保证每个神经元都收到信号”。同一层神经元具有相同输入、输出与梯度，更新后仍保持相同，网络无法打破置换对称性，等价于只有一个有效隐藏单元。

6.8 归一化：归一化哪些轴，推理时使用什么量

Batch normalization 对每个通道使用批量统计量

$$\mu_c = \frac{1}{M}\sum_{r=1}^{M} x_{rc}, \qquad v_c = \frac{1}{M}\sum_{r=1}^{M}(x_{rc}-\mu_c)^2,$$

$$\widehat{x}_{rc} = \frac{x_{rc}-\mu_c}{\sqrt{v_c+\epsilon}}, \qquad y_{rc} = \gamma_c\widehat{x}_{rc} + \beta_c.$$

训练时 μ_c, v_c 依赖同一批量，因此样本间耦合；推理时通常改用运行均值和运行方差。忘记切换会使单样本预测依赖偶然拼入的其他样本。若动量参数记为 $\alpha \in (0,1]$，一种常见运行统计更新是

$$\mu_{\text{run}} \leftarrow (1-\alpha)\mu_{\text{run}} + \alpha\mu_B, \qquad v_{\text{run}} \leftarrow (1-\alpha)v_{\text{run}} + \alpha v_B.$$

软件对批内方差是否作无偏修正可能不同，答题应采用题目给定约定。以 $\varepsilon = 0$ 手算：batch 为 $(0,2)$、$\gamma = 2, \beta = -1$ 时，$\mu_B = 1, v_B = 1$，训练输出为 $(-3,1)$；若推理运行统计为 $\mu_{\text{run}} = 1/2, v_{\text{run}} = 4$，输入 $x = 2$ 的输出为 $1/2$，不能继续使用单样本自己的均值方差。

Layer normalization 对单个 token 的特征轴归一化。若 $x_{bt:} \in \mathbb{R}^d$，则

$$\mu_{bt} = \frac{1}{d}\sum_{j=1}^{d} x_{btj}, \qquad v_{bt} = \frac{1}{d}\sum_{j=1}^{d}(x_{btj}-\mu_{bt})^2.$$

它不依赖批量统计量，更适合长度可变的序列模型。两种归一化都不能从数学上消除所有梯度爆

炸，也不能替代合适的初始化和学习率。

易错点

考试中必须同时写清四件事：归一化轴、可学习参数 γ, β、训练统计量、推理统计量。只写“使均值为零、方差为一”通常不足以区分 BN 与 LN。

6.9 计算、存储与数值检查

反向传播需要保存足以重构局部 Jacobian 的中间激活。若网络深度为 L，朴素存储量约为各层激活大小之和；gradient checkpointing 只保存若干检查点，反向时重算其间前向值，以额外计算换内存。混合精度训练则必须区分：

- 前向/反向张量的低精度表示；
- 累加与主权重的较高精度表示；
- loss scaling 用于避免小梯度下溢。

这些技术改变数值实现，不改变链式法则本身。

答题方法

维度题的稳定作答顺序是：先列每个张量形状；再写每层前向式；从标量损失逆序定义伴随量；最后用 Frobenius 内积读取参数梯度。每得到一个梯度，都核对它与对应参数同形。

自测

1. 为什么批量大小不计入网络参数量，却计入激活存储量？
2. 推导带偏置仿射层对 W, b, X 的三个梯度，并逐项核对维度。
3. tanh 网络若采用过大方差初始化，前向与反向分别会发生什么？
4. BN 在批量大小为一时为什么可能不稳定？LN 为什么没有同样的批量依赖？

第七章　卷积、序列架构与生成模型

学习目标

本章建立从结构归纳偏置到概率训练目标的完整链：会计算 CNN 尺寸、感受野和反向梯度，会展开 RNN/LSTM 的时间反传，会从隐变量模型推导 VAE 的 ELBO，从最优判别器解释 GAN，并理解 diffusion/score 模型的前向加噪与反向生成。Transformer 的完整推导统一放在章 十五。

7.1　卷积网络与视觉任务

二维卷积（深度学习库实际常为互相关）可写为

$$Y_{c_{\text{out}},i,j} = b_{c_{\text{out}}} + \sum_{c_{\text{in}},u,v} K_{c_{\text{out}},c_{\text{in}},u,v} X_{c_{\text{in}},i+u,j+v}.$$

若输入宽高为 H, W，核 K_h, K_w，padding P_h, P_w，stride S_h, S_w，dilation D_h, D_w，则

$$H_{\text{out}} = \left\lfloor \frac{H + 2P_h - D_h(K_h - 1) - 1}{S_h} + 1 \right\rfloor,$$

宽度同理。参数数目为 $C_{\text{out}}(C_{\text{in}}K_hK_w + 1)$，与空间尺寸无关。局部连接和权重共享带来平移等变性；池化或全局平均才增强一定的平移不变性。

例题 7.1 (卷积尺寸与参数)**.** 输入 $32 \times 32 \times 3$，用 64 个 3×3 核、padding 1、stride 2，输出为 $16 \times 16 \times 64$，参数数为 $64(3 \cdot 3 \cdot 3 + 1) = 1792$。若忘记 bias，则少算 64。

7.1.1　卷积层反向传播：共享参数意味着梯度求和

为突出索引，先取步长 1、无填充的二维互相关形式。令上游梯度 $\Delta_{oij} = \partial \mathcal{L}/\partial Y_{oij}$，则

$$\frac{\partial \mathcal{L}}{\partial K_{ocuv}} = \sum_{i,j} \Delta_{oij} X_{c,i+u,j+v}, \tag{7.1}$$

$$\frac{\partial \mathcal{L}}{\partial b_o} = \sum_{i,j} \Delta_{oij}, \tag{7.2}$$

$$\frac{\partial \mathcal{L}}{\partial X_{crs}} = \sum_{\substack{o,i,j,u,v \\ r=i+u,\ s=j+v}} \Delta_{oij} K_{ocuv}. \tag{7.3}$$

式(7.1)对所有空间位置求和, 因为同一个核参数在每个位置复用; 若把每个位置当成独立权重, 就丢失了 CNN 的参数共享。加入 stride、padding 与 dilation 后公式结构不变，只需把求和限制为前向时真实对齐的输出位置，并把填充区视为常数零而不对其求梯度。

完整证明. 对前向式取微分:

$$\mathrm{d}Y_{oij} = \mathrm{d}b_o + \sum_{c,u,v} \big(\mathrm{d}K_{ocuv} X_{c,i+u,j+v} + K_{ocuv}\, \mathrm{d}X_{c,i+u,j+v} \big).$$

代入 $\mathrm{d}\mathcal{L} = \sum_{o,i,j} \Delta_{oij}\, \mathrm{d}Y_{oij}$，再分别收集 $\mathrm{d}K_{ocuv}$、$\mathrm{d}b_o$ 与 $\mathrm{d}X_{crs}$ 的系数，得到三条梯度式。输入梯度常被口头称为“用翻转核卷积”，但是否显式翻转取决于前向采用卷积还是互相关以及张量布局；索引式(7.3)不含歧义。 □

例题 7.2 (一维共享核反传手算)**.** 取输入 $x = (1, 2, 0)$、核 $k = (2, -1)$，步长 1、无填充，则互相关输出 $y = (0, 4)$。令 $\mathcal{L} = \frac{1}{2}\|y\|^2$，上游梯度为 $\Delta = (0, 4)$。由共享位置求和,

$$\nabla_k \mathcal{L} = (0 \cdot 1 + 4 \cdot 2,\ 0 \cdot 2 + 4 \cdot 0) = (8, 0),$$

$$\nabla_x \mathcal{L} = (0 \cdot 2,\ 0 \cdot (-1) + 4 \cdot 2,\ 4 \cdot (-1)) = (0, 8, -4).$$

用有限差分扰动任一分量可核对符号。这个例子同时说明：输入梯度和核梯度的形状、求和轴与“翻转”口号都不能凭记忆决定。

官方视觉考纲可按下表复习:

任务	输出	常见损失	关键结构/指标
分类	单标签/多标签	交叉熵/BCE	CNN/ViT, accuracy/AUROC
检测	类别与框	分类 +box 回归	anchor 或 set prediction, mAP
语义分割	每像素类别	像素 CE/Dice	encoder–decoder, mIoU
图像恢复	清晰图像	L_1/L_2/感知损失	U-Net/残差, PSNR/SSIM
光流	每像素二维位移	端点误差/重建一致性	多尺度、warping
三维重建	深度/点云/辐射场	几何或渲染损失	多视图约束

7.2 序列模型：RNN、LSTM 与门控

基本 RNN:

$$h_t = \phi(W_x x_t + W_h h_{t-1} + b), \qquad p(y_t \mid h_t) = \mathrm{softmax}(W_o h_t).$$

BPTT 将网络沿时间展开，梯度含 $\prod_{s=k+1}^{t} \partial h_s/\partial h_{s-1}$，因此同样有梯度消失/爆炸。梯度裁剪只限制爆炸，不解决长期记忆本身。

LSTM 的一套统一记号为

$$\begin{aligned} i_t &= \sigma(W_i[x_t; h_{t-1}] + b_i), & f_t &= \sigma(W_f[x_t; h_{t-1}] + b_f), \\ o_t &= \sigma(W_o[x_t; h_{t-1}] + b_o), & \widetilde{c}_t &= \tanh(W_c[x_t; h_{t-1}] + b_c), \\ c_t &= f_t \odot c_{t-1} + i_t \odot \widetilde{c}_t, & h_t &= o_t \odot \tanh(c_t). \end{aligned}$$

加性 cell path 允许 $\partial c_t/\partial c_{t-1} = f_t$，若遗忘门接近 1，长距离梯度较易保留。LSTM 缓解而非数学上消除梯度消失。

若 $x_t \in \mathbb{R}^{d_x}$、$h_t, c_t \in \mathbb{R}^{d_h}$，把四个门合并成一个仿射变换时，

$$W_{\text{gate}} \in \mathbb{R}^{4d_h \times (d_x + d_h)}, \qquad b_{\text{gate}} \in \mathbb{R}^{4d_h},$$

故标准 LSTM 核心共有 $4d_h(d_x + d_h + 1)$ 个参数，不含输出层。标量手算时，若 $c_{t-1} = 2, f_t = 0.75, i_t = 0.25, \widetilde{c}_t = -1, o_t = 0.5$，则

$$c_t = 0.75 \cdot 2 + 0.25(-1) = 1.25, \qquad h_t = 0.5 \tanh(1.25).$$

固定当前门值的直接记忆通道导数是 $\partial c_t/\partial c_{t-1} = 0.75$；完整 BPTT 还包含门对 h_{t-1} 的间接依赖，不能把这条直接通道当成完整 Jacobian。

7.2.1 GRU：两门结构与候选状态

GRU 用更新门和重置门压缩 LSTM 的三门结构。一套常用记号为

$$\begin{aligned} z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z), \\ r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r), \\ \widetilde{h}_t &= \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h), \\ h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h}_t. \end{aligned}$$

不同教材可能交换 z_t 与 $1 - z_t$ 的角色，只要定义与更新一致即可。GRU 参数少、并行性仍受递归限制；它和 LSTM 都缓解而不保证消除长期梯度问题。

7.3 自监督、表示学习与迁移

自编码器最小化重建损失 $\mathcal{L}_{\text{AE}} = \|x - g_\theta(f_\phi(x))\|^2$；若瓶颈过宽且无约束，可能只学恒等映射。去噪自编码器从扰动输入恢复原样本，迫使模型利用数据结构。

对比学习的 InfoNCE 损失（anchor i、正样本 i^+）为

$$\ell_i = -\log \frac{\exp(\text{sim}(z_i, z_{i^+})/\tau)}{\sum_{j \in \mathcal{A}_i} \exp(\text{sim}(z_i, z_j)/\tau)}.$$

温度 τ 控制分布尖锐程度；负样本集合与数据增强决定学到的不变性。它不是简单“把相似样本

拉近”，而是相对正负对比。

迁移学习先预训练再微调；冻结特征适合小数据或域接近，完全微调表达力强但更易遗忘。元学习以任务分布为单位优化“快速适应”，例如 MAML 的双层目标

$$\min_{\theta} \mathbb{E}_{\mathcal{T}} L_{\mathcal{T}}^{\text{query}}\big(\theta - \alpha \nabla_{\theta} L_{\mathcal{T}}^{\text{support}}(\theta)\big).$$

其外层梯度包含 Hessian 项；一阶 MAML 忽略该项，是近似而非完全相同算法。

7.4 深度网络的优化与泛化辨析

非凸网络中，训练损失低不代表测试风险低；宽网络可能存在大量全局最小点。隐式正则、数据增强、架构先验和优化轨迹共同影响泛化。考题若问“更深是否一定更好”“更宽是否一定过拟合”，正确回答应是否定无条件断言，并说明优化、数据规模和显式/隐式正则条件。

历年题专题 7.1: 正式卷中的深度学习稳定题型

三期正式卷持续考查前向/反向传播、架构比较、attention/序列建模及生成/强化学习接口。没有可靠官方答案时，本书把解答标为独立推导。作答顺序建议固定为：张量形状 → 前向公式 → 损失 → 局部梯度 → 链式法则 → 复杂度与失败模式。

易错点

易错点：softmax 后又把 cross-entropy 导数多乘一次 sigmoid；矩阵梯度转置错误；混淆 BatchNorm 的训练/推断统计；把 dropout 测试时也随机开启；把卷积等变性说成完全不变性；漏掉因果掩码；声称位置编码“可有可无”；把 Transformer 的 $O(n^2)$ 只写成时间而忽略存储；把 LSTM 写成保证不消失；把经验性的过参数化现象写成无条件定理。

复习检查

1. 能否在不跳步的情况下完成一个两层网络的前向、softmax 交叉熵和反向传播？
2. 能否从方差传播解释 Xavier/He 初始化，并说明近似假设？
3. 能否计算卷积输出尺寸、参数量和感受野？
4. 能否写出 RNN/LSTM 全部门并解释 cell path？
5. 能否推导缩放点积注意力、因果掩码和 $O(n^2)$ 瓶颈？
6. 能否比较 BatchNorm/LayerNorm、CNN/Transformer、冻结/微调？
7. 能否为分类、检测、分割、恢复选择合理损失与指标？

7.5 对比学习与 InfoNCE

对 anchor 的候选相似度 $s_1, \dots, s_M$，设正样本索引为 $+$、温度为 $\tau > 0$。InfoNCE 为

$$\ell = -\log \frac{e^{s_+/\tau}}{\sum_{j=1}^{M} e^{s_j/\tau}}, \qquad p_j = \frac{e^{s_j/\tau}}{\sum_k e^{s_k/\tau}}.$$

命题 7.1 (InfoNCE 对相似度的梯度)**.**

$$\frac{\partial \ell}{\partial s_j} = \frac{1}{\tau}\big(p_j - \mathbf{1}\{j = +\}\big).$$

完整证明. 把损失写成 $\ell = -s_+/\tau + \log \sum_k e^{s_k/\tau}$。第一项求导为 $-\mathbf{1}\{j = +\}/\tau$，第二项为 p_j/τ，相加即得。因此正样本相似度被提高，负样本相似度按当前 softmax 概率降低；温度同时改变概率尖锐度和梯度尺度。 □

例题 7.3 (InfoNCE 梯度手算)**.** 若 $\tau = 1$、相似度为 $(\log 2, 0, 0)$ 且第一个为正样本，则 $p = (1/2, 1/4, 1/4)$、$\ell = \log 2$，梯度为 $(-1/2, 1/4, 1/4)$。分母若漏掉某个负样本，损失和全部梯度都会改变；把温度只放在正样本项也不是同一个目标。

7.6 概率生成建模的统一视角

生成模型试图学习数据分布 $p_{\text{data}}(x)$ 或可采样近似。考试应按四问组织：是否有显式似然？潜变量如何推断？训练目标是什么？采样要多少步？

7.6.1 自回归模型

链式法则给出

$$p_\theta(x) = \prod_{t=1}^{T} p_\theta(x_t \mid x_{<t}), \qquad \mathcal{L}_{\text{NLL}} = -\sum_t \log p_\theta(x_t \mid x_{<t}).$$

教师强制训练可并行计算所有条件概率；推断必须逐 token 生成，存在暴露偏差（exposure bias）和串行瓶颈。因果 Transformer 是当前常见参数化。精确似然是优点，但生成速度通常为 $O(T)$ 次网络调用。

7.7 变分自编码器

潜变量模型 $p_\theta(x, z) = p(z)p_\theta(x \mid z)$ 的边缘似然 $\log p_\theta(x) = \log \int p_\theta(x, z)\, dz$ 通常难算，引入近似后验 $q_\phi(z \mid x)$。

定理 7.2 (证据下界 ELBO 分解)**.** *若 $q_\phi(z \mid x)$ 对真实后验绝对连续，则*

$$\log p_\theta(x) = \underbrace{\mathbb{E}_{q_\phi} \log p_\theta(x \mid z) - \mathrm{KL}(q_\phi(z \mid x) \| p(z))}_{\mathcal{L}_{\text{ELBO}}(x)} + \mathrm{KL}(q_\phi(z \mid x) \| p_\theta(z \mid x)). \tag{7.4}$$

故 $\mathcal{L}_{\mathrm{ELBO}} \le \log p_\theta(x)$，等号当且仅当两个后验几乎处处相同。

完整证明. 从非负 KL 出发，也可直接代数分解：

$$\begin{aligned}\mathrm{KL}(q_\phi(z \mid x) \| p_\theta(z \mid x)) &= \mathbb{E}_q \log \frac{q_\phi(z \mid x)}{p_\theta(z \mid x)} \\ &= \mathbb{E}_q \log \frac{q_\phi(z \mid x) p_\theta(x)}{p_\theta(x, z)} \\ &= \log p_\theta(x) - \mathbb{E}_q \log p_\theta(x, z) + \mathbb{E}_q \log q_\phi(z \mid x).\end{aligned}$$

移项并用 $p_\theta(x, z) = p(z) p_\theta(x \mid z)$，得到式(7.4)。绝对连续保证比值与 KL 有定义；若 KL 无穷，下界仍可用扩展实数解释，但不能作有限优化目标。 □

高斯编码器 $q_\phi(z \mid x) = \mathcal{N}(\mu_\phi(x), \mathrm{diag}\, \sigma_\phi^2(x))$ 用重参数化（reparameterization trick）$z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon$、$\epsilon \sim \mathcal{N}(0, I)$，把随机性移到与 ϕ 无关的噪声上，从而低方差反传。标准正态先验下

$$\mathrm{KL}(q \| p) = \frac{1}{2} \sum_j (\mu_j^2 + \sigma_j^2 - \log \sigma_j^2 - 1).$$

重建项与输出分布匹配：Bernoulli 像素对应 BCE，高斯观测对应带常数的平方误差。

重参数化允许把 $\nabla_\phi \mathbb{E}_{z \sim q_\phi(z|x)} f_\theta(z)$ 写成路径导数

$$\mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)} \nabla_z f_\theta(z) \frac{\partial (\mu_\phi + \sigma_\phi \odot \epsilon)}{\partial \phi},$$

从而对固定噪声样本用普通反向传播；若把随机采样节点视为与 ϕ 无关，编码器将收不到重建项梯度。

例题 7.4 (一维 VAE 的 KL 与重参数化)**.** 若 $q(z \mid x) = \mathcal{N}(\mu = 1, \sigma^2 = 4)$、$p(z) = \mathcal{N}(0, 1)$，则

$$\mathrm{KL}(q \| p) = \tfrac{1}{2}(1^2 + 4 - \log 4 - 1) = 2 - \tfrac{1}{2} \log 4.$$

采样时写 $z = 1 + 2\epsilon$。若题目给的是 $\log \sigma^2$，必须先指数化得到方差，不能把 log-variance 直接代入 σ。

7.8 GAN 与对抗学习

标准 GAN 解

$$\min_G \max_D \; \mathbb{E}_{x \sim p_{\mathrm{data}}} \log D(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D(G(z))).$$

假设 p_{data} 与 p_g 对同一支配测度有密度。固定生成器分布 p_g，在 $p_{\mathrm{data}}(x) + p_g(x) > 0$ 的点逐点最大化可得

$$D^\star(x) = \frac{p_{\mathrm{data}}(x)}{p_{\mathrm{data}}(x) + p_g(x)}.$$

在两个密度都为零的点，D 的值不影响目标。代回后目标为 $-\log 4 + 2\,\mathrm{JS}(p_{\mathrm{data}} \| p_g)$，全局最优在 $p_g = p_{\mathrm{data}}$。实际训练常用非饱和生成器损失 $-\mathbb{E}_z \log D(G(z))$ 改善早期梯度，但它与原 minimax 的生成器更新不同。

完整证明. 对固定 x, 最大化 $a\log D+b\log(1-D)$, 一阶条件 $a/D-b/(1-D)=0$ 给 $D=a/(a+b)$; 二阶导负，故为最大值。令混合分布 $m=(p_{\text{data}}+p_g)/2$，代回积分:

$$\int p_{\text{data}}\log\frac{p_{\text{data}}}{2m}+p_g\log\frac{p_g}{2m}=-\log 4+\mathrm{KL}(p_{\text{data}}\|m)+\mathrm{KL}(p_g\|m),$$

即所述 JS 形式。 □

GAN 不给易计算的显式似然，采样一次前向即可，但训练可能 mode collapse、梯度不稳。WGAN 用 Wasserstein-1 距离及 1-Lipschitz critic；如何强制 Lipschitz（裁剪/梯度惩罚）决定实现细节。考试若未额外给出假设，不应声称通用全局收敛。

7.9 扩散模型与流模型

离散前向扩散令

$$q(x_t\mid x_{t-1})=\mathcal{N}(\sqrt{1-\beta_t}x_{t-1},\beta_t I),\quad \alpha_t=1-\beta_t,\quad \bar\alpha_t=\prod_{s=1}^{t}\alpha_s.$$

高斯闭包给

$$x_t=\sqrt{\bar\alpha_t}x_0+\sqrt{1-\bar\alpha_t}\epsilon,\qquad \epsilon\sim\mathcal{N}(0,I). \tag{7.5}$$

常用简化目标随机采样 t,x_0,ϵ:

$$\mathbb{E}\,\|\epsilon-\epsilon_\theta(x_t,t)\|_2^2.$$

由于前向链为线性高斯，给定 x_0,x_t 的一步后验仍为高斯:

$$q(x_{t-1}\mid x_t,x_0)=\mathcal{N}(\widetilde{\mu}_t(x_t,x_0),\widetilde{\beta}_t I),\quad \widetilde{\beta}_t=\frac{1-\bar\alpha_{t-1}}{1-\bar\alpha_t}\beta_t,$$

$$\widetilde{\mu}_t=\frac{\sqrt{\bar\alpha_{t-1}}\beta_t}{1-\bar\alpha_t}x_0+\frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}x_t.$$

生成模型用 $p_\theta(x_{t-1}\mid x_t)=\mathcal{N}(\mu_\theta(x_t,t),\sigma_t^2 I)$ 近似未知的反向条件。把式(7.5)解出 $x_0=(x_t-\sqrt{1-\bar\alpha_t}\epsilon)/\sqrt{\bar\alpha_t}$，再以 ϵ_θ 替代真实噪声，可将均值参数化为噪声预测；加权变分项由此化为上述 MSE。采样质量高但通常需要多步网络调用。Classifier-free guidance 用条件/无条件噪声预测组合 $\widehat{\epsilon}=(1+w)\epsilon_\theta(x_t,c)-w\epsilon_\theta(x_t,\varnothing)$，增大条件一致性但可能降低多样性。

Flow matching 直接学习概率路径 p_t 的速度场。若 ODE

$$\frac{dX_t}{dt}=v_\theta(X_t,t)$$

推动密度，则它满足连续性方程

$$\partial_t p_t(x)+\nabla\cdot(p_t(x)v_\theta(x,t))=0.$$

选定可采样的条件路径 X_t 和目标条件速度 u_t 后，训练

$$\min_\theta\mathbb{E}_{t,X_t}\,\|v_\theta(X_t,t)-u_t(X_t,t)\|^2.$$

例如线性插值 $X_t = (1-t)X_0 + tX_1$ 的条件速度为 $X_1 - X_0$。这与 normalizing flow 的区别是：flow matching 回归速度并通过 ODE 采样，normalizing flow 依赖显式可逆映射及 Jacobian 行列式。

连续时间 SDE、score matching 与 Fokker–Planck 只作联系：score 是 $\nabla_x \log p_t(x)$，反向时间动力学依赖它。离散前向条件的 score 为

$$\nabla_{x_t} \log q(x_t \mid x_0) = -\frac{\epsilon}{\sqrt{1-\bar{\alpha}_t}},$$

故噪声预测器对应 $s_\theta(x_t, t) = -\epsilon_\theta(x_t, t)/\sqrt{1-\bar{\alpha}_t}$。连续时间的 Fokker–Planck 方程及其边界条件只作模型联系；考试主线是离散前向过程、后验高斯形式和噪声预测目标，不要把连续时间推导当作必背完整证明。

可逆 normalizing flow 用 $x = f_\theta(z)$ 和换元公式

$$\log p_X(x) = \log p_Z(f_\theta^{-1}(x)) + \log \left| \det J_{f_\theta^{-1}}(x) \right|.$$

它有精确似然与单次可逆采样，但架构受 Jacobian 行列式可计算性约束。

例题 7.5 (一维换元似然)**.** 若 $Z \sim \mathcal{N}(0,1)$、$X = 2Z + 1$，观测 $x = 3$ 对应 $z = 1$，故

$$\log p_X(3) = \log p_Z(1) - \log 2 = -\tfrac{1}{2}\log(2\pi) - \tfrac{1}{2} - \log 2.$$

若误加 $\log 2$，就是把正向 Jacobian 与逆向 Jacobian 的方向写反。

答题方法

架构题先算形状、参数量和依赖范围；生成模型题先写联合分布，再指出训练目标是似然、变分下界、对抗散度还是去噪/score 匹配。任何“生成质量”结论都要说明它来自目标函数还是经验现象。

自测

1. 给定卷积核、步幅、填充和膨胀率，计算输出尺寸、参数量与两层感受野。
2. 从 LSTM 门方程解释加性 cell path 为什么缓解但不保证消除梯度消失。
3. 由 $\log p_\theta(x)$ 插入任意 $q_\phi(z \mid x)$，完整推导 ELBO 与 KL 缺口。
4. 求固定生成器时 GAN 的最优判别器，并说明模式崩溃为何不与该闭式解矛盾。
5. 比较 VAE、GAN、diffusion 在归一化似然、采样速度和训练稳定性上的边界。

本节结论

CNN、RNN 和生成模型都在利用结构降低无约束函数学习的难度；考试中应把“结构、目标、梯度、失败模式”四件事连成一条可检查的推导链。

7.10　卷积尺寸、感受野与等变性

一维卷积输入长度 n，核宽 k，左右总填充 $2p$，步长 s，膨胀率 d 时，有效核宽

$$k_{\mathrm{eff}} = 1 + d(k-1), \qquad n_{\mathrm{out}} = \left\lfloor \frac{n + 2p - k_{\mathrm{eff}}}{s} \right\rfloor + 1.$$

二维情形对高、宽分别应用此式。若输入通道 c_{in}、输出通道 c_{out}，普通卷积参数量为

$$k_h k_w c_{\mathrm{in}} c_{\mathrm{out}} + c_{\mathrm{out}}.$$

空间尺寸不进入参数量，正是参数共享的结果。

令第 ℓ 层在原输入上的跳距为 j_ℓ，感受野宽度为 r_ℓ。初始化 $j_0 = r_0 = 1$，递推

$$j_\ell = j_{\ell-1} s_\ell, \qquad r_\ell = r_{\ell-1} + (k_{\mathrm{eff},\ell} - 1) j_{\ell-1}.$$

这个公式说明步长层虽然本身核不一定很大，却会放大后续每层对原输入的覆盖间隔。

命题 7.3. *忽略边界并令 T_a 表示整数平移，则共享核卷积 $K * \cdot$ 满足*

$$K * (T_a x) = T_a (K * x).$$

证明.

$$[K * (T_a x)]_u = \sum_v K_v (T_a x)_{u-v} = \sum_v K_v x_{u-v-a} = [T_a (K * x)]_u.$$

若加入步长抽样，等变性只对与总步长相容的平移成立；零填充还会在边界破坏严格等变性。 □

7.11　循环网络的梯度乘积与门控记忆

简单 RNN 写成

$$h_t = \phi(W h_{t-1} + U x_t + b), \qquad o_t = V h_t.$$

从时刻 T 向 t 传播的 Jacobian 包含

$$\frac{\partial h_T}{\partial h_t} = \prod_{r=t+1}^{T} D_r W, \qquad D_r = \operatorname{diag}\big(\phi'(W h_{r-1} + U x_r + b)\big).$$

因此

$$\left\| \frac{\partial h_T}{\partial h_t} \right\| \le \prod_{r=t+1}^{T} \|D_r W\|.$$

若这些因子的典型范数小于一，梯度指数衰减；若大于一，则可能爆炸。这是乘积结构造成的，不只是“sigmoid 导数小”一句话。

LSTM 的细胞状态更新

$$c_t = f_t \odot c_{t-1} + i_t \odot \widetilde{c}_t$$

给出直接通道

$$\frac{\partial c_t}{\partial c_{t-1}} = \operatorname{diag}(f_t)$$

（完整导数还包括门值间接依赖）。当遗忘门接近一时，直接通道能减轻长乘积衰减；但若门长期饱和或状态间接通道不稳定，LSTM 仍不保证任意长依赖都可学习。

7.12 VAE：从边缘似然到可训练下界

生成模型规定先验 $p_\theta(z)$ 和条件分布 $p_\theta(x \mid z)$，目标是

$$\log p_\theta(x) = \log \int p_\theta(x, z)\, \mathrm{d}z.$$

引入任意近似后验 $q_\phi(z \mid x)$，有

$$\begin{aligned}\log p_\theta(x) &= \mathbb{E}_{q_\phi} \log \frac{p_\theta(x, z)}{q_\phi(z \mid x)} + \mathbb{E}_{q_\phi} \log \frac{q_\phi(z \mid x)}{p_\theta(z \mid x)} \\ &= \mathcal{L}_{\mathrm{ELBO}}(x; \theta, \phi) + \mathrm{KL}(q_\phi(z \mid x) \,\|\, p_\theta(z \mid x)).\end{aligned}$$

KL 非负，故

$$\mathcal{L}_{\mathrm{ELBO}} = \mathbb{E}_{q_\phi(z|x)} \log p_\theta(x \mid z) - \mathrm{KL}(q_\phi(z \mid x) \| p_\theta(z)) \le \log p_\theta(x).$$

第一项要求重构，第二项防止每个样本的潜变量分布任意偏离共同先验。

当

$$q_\phi(z \mid x) = N(\mu_\phi(x), \operatorname{diag} \sigma_\phi^2(x))$$

时，令 $\epsilon \sim N(0, I)$，

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon.$$

随机性被移到与参数无关的 ϵ，从而可对 θ, ϕ 反向传播。对标准正态先验，

$$\mathrm{KL}(q_\phi \| p) = \frac{1}{2} \sum_j \left(\mu_j^2 + \sigma_j^2 - \log \sigma_j^2 - 1\right).$$

易错点

ELBO 与 $\log p_\theta(x)$ 的差是后验 KL，不是一个未知常数。最大化 ELBO 同时改变生成参数与推断参数；只有近似后验等于真实后验时下界才取等。

7.13 GAN：最优判别器与 JS 散度

原始目标为

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\mathrm{data}}} \log D(x) + \mathbb{E}_{x \sim p_g} \log(1 - D(x)).$$

固定 G 后，对每个 x 最大化

$$a\log d + b\log(1-d), \qquad a = p_{\text{data}}(x),\ b = p_g(x).$$

一阶条件 $a/d - b/(1-d) = 0$ 给

$$D_G^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_g(x)}$$

（在两密度都为零处任取）。代回并记 $m = (p_{\text{data}} + p_g)/2$，得到

$$V(D_G^*, G) = -\log 4 + \mathrm{KL}(p_{\text{data}} \| m) + \mathrm{KL}(p_g \| m) = -\log 4 + 2\ \mathrm{JS}(p_{\text{data}}, p_g).$$

故理想无限容量博弈在 $p_g = p_{\text{data}}$ 处达到全局最优。

这个结论不等于实际梯度训练必然收敛。若两分布支撑几乎不交，最优判别器过强，原始饱和生成器损失会给出很小梯度；实践中常改用非饱和损失 $-\mathbb{E}_z \log D(G(z))$。Mode collapse 则表现为多个潜变量区域映到少量模式，判别器目标本身没有直接保证有限训练过程覆盖所有模式。

7.14 扩散模型：闭式加噪、噪声预测与反向一步

离散前向过程取

$$q(x_t \mid x_{t-1}) = N(\sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I), \qquad \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s.$$

高斯递推可归纳得到

$$q(x_t \mid x_0) = N(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I),$$

即

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \qquad \epsilon \sim N(0, I).$$

因此训练时可直接从任意时刻 t 构造噪声样本，无须真的执行前面 $t-1$ 步。

真实后验 $q(x_{t-1} \mid x_t, x_0)$ 仍为高斯，其均值是 x_t, x_0 的线性组合。把 x_0 用噪声参数化

$$\widehat{x}_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}$$

代入后验均值，可定义反向模型 $p_\theta(x_{t-1} \mid x_t)$。在常用固定方差参数化下，变分目标的主要可训练部分化为加权噪声回归；简化目标为

$$\mathbb{E}_{t,x_0,\epsilon} \left\| \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t) \right\|_2^2.$$

“预测噪声”并非脱离概率模型的经验技巧，而是高斯后验均值的一种重参数化。

答题方法

生成模型题先回答四层：随机变量与联合分布；训练目标如何从似然或散度得到；梯度如何计算；训练与采样分别执行什么流程。VAE、GAN、扩散的目标不能只按网络结构比较。

自测

1. 推导膨胀卷积的有效核宽，并计算两层步长卷积的感受野。
2. ELBO 中哪一项衡量近似后验与先验，哪一项衡量近似后验与真实后验？
3. 为什么 D_G^* 的公式不能直接当作有限神经网络判别器的训练结果？
4. 从 $q(x_t \mid x_{t-1})$ 归纳证明 $q(x_t \mid x_0)$ 的闭式。

第八章　强化学习：从 Bellman 方程到策略梯度

学习目标

本章以 MDP 为统一对象，从 Bellman 压缩映射建立动态规划，再区分 Monte Carlo、TD、SARSA 与 Q-learning 的目标和采样分布，最后推导 policy gradient、baseline 与 actor–critic。

8.1　MDP 与 Bellman 方程

定义 8.1 (折扣 MDP)**.** MDP 为 $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$，其中 $P(s' \mid s, a)$ 是转移核，$r(s,a)$ 为期望即时回报，$0 \le \gamma < 1$。策略 $\pi(a \mid s)$ 的价值为

$$V^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t) \mid S_0 = s\right].$$

动作价值 $Q^\pi(s,a)$ 固定第一动作后再遵循 π。

Bellman 期望方程:

$$V^\pi(s) = \sum_a \pi(a \mid s)\left[r(s,a) + \gamma \sum_{s'} P(s' \mid s, a) V^\pi(s')\right].$$

最优性方程:

$$V^\star(s) = \max_a \left[r(s,a) + \gamma \sum_{s'} P(s' \mid s, a) V^\star(s')\right].$$

定理 8.1 (Bellman 最优算子的压缩性)**.** *在有限或有界价值函数空间的上确界范数下, Bellman 最优算子*

$$(TV)(s) = \max_a \{r(s,a) + \gamma \mathbb{E}[V(S') \mid s, a]\}$$

满足 $\|TV - TW\|_\infty \le \gamma \|V - W\|_\infty$。因此它有唯一不动点 $V^\star$，价值迭代 $V_{k+1} = TV_k$ 以 γ^k 几何收敛。

完整证明. 对任意 s，使用 $|\max_a u_a - \max_a v_a| \le \max_a |u_a - v_a|$:

$$\begin{aligned}|(TV)(s) - (TW)(s)| &\le \gamma \max_a \left|\mathbb{E}[V(S') - W(S') \mid s, a]\right| \\ &\le \gamma \|V - W\|_\infty .\end{aligned}$$

对 s 取上确界即得压缩性。完备度量空间上的 Banach 不动点定理给唯一不动点和 $\|V_k - V^\star\|_\infty \le \gamma^k \|V_0 - V^\star\|_\infty$。$\gamma < 1$ 是此证明的关键；平均回报或无折扣问题需其他条件。 □

8.2 动态规划、Monte Carlo 与 TD

策略评估（policy evaluation）与策略改进构成策略迭代；其更新交替为：

$$V^{\pi_k} = T^{\pi_k} V^{\pi_k}, \qquad \pi_{k+1}(s) \in \arg\max_a [r(s,a) + \gamma \mathbb{E} V^{\pi_k}(S')].$$

价值迭代（value iteration）直接应用 T。两者都需要已知模型或能够精确计算期望。

定理 8.2 (策略改进定理)**.** *若对所有状态*

$$\pi'(s) \in \arg\max_a Q^\pi(s,a),$$

则 $V^{\pi'}(s) \ge V^\pi(s)$ 对所有 s 成立；若某个可达状态严格改进，在相应可达区域价值严格提高。

完整证明. 由贪心定义，

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

Bellman 策略算子 $T^{\pi'}$ 单调且为 γ-压缩。反复应用得到

$$V^\pi \le T^{\pi'} V^\pi \le (T^{\pi'})^2 V^\pi \le \cdots,$$

而该序列收敛到 $T^{\pi'}$ 的唯一不动点 $V^{\pi'}$，故 $V^{\pi'} \ge V^\pi$。严格性还需新动作在可达状态确实严格优于旧策略混合。 □

例题 8.1 (Bellman 方程求解)**.** 在固定策略下，s_1 以奖励 1 确定转到 s_2，s_2 以奖励 2 自环，$\gamma = 1/2$。方程为 $V_1 = 1 + \frac{1}{2}V_2$、$V_2 = 2 + \frac{1}{2}V_2$，故 $V_2 = 4, V_1 = 3$。矩阵形式是 $(I - \gamma P^\pi)V = r^\pi$；只要 $\gamma < 1$，有限状态下该矩阵可逆。

例题 8.2 (值迭代手算)**.** 单状态 MDP 有两个自环动作，奖励分别为 1 与 2，$\gamma = 1/2$。从 $V_0 = 0$ 开始，$V_1 = 2$、$V_2 = 2 + \frac{1}{2} \cdot 2 = 3$、$V_3 = 3.5$，极限为 4，每步贪心动作都是奖励为 2 的动作。误差满足 $|V_k - 4| = 2^{-k}|V_0 - 4|$。

Monte Carlo 在 episode 结束后用完整回报 $G_t = \sum_{k \ge 0} \gamma^k R_{t+k+1}$ 更新，目标无 bootstrap、方差较大。TD(0) 一步更新

$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)],$$

有 bootstrap、偏差较大但可在线更新。n-step 与 TD(λ) 在二者之间折中。

8.3 Q-learning、策略梯度与 Actor–Critic

表格型 Q-learning：

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha_t[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)].$$

它是 off-policy：行为策略可探索，目标使用贪心最大值。经典几乎必然收敛需要有限状态动作、所有对被无限访问、奖励有界以及 Robbins–Monro 步长 $\sum_t \alpha_t = \infty, \sum_t \alpha_t^2 < \infty$。深度函数逼近下这些条件不再直接成立。

SARSA 的 on-policy 更新为

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha_t[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)].$$

目标动作 A_{t+1} 来自当前行为策略，因此探索动作的风险进入学习目标。Q-learning 的目标使用 $\max_a Q(S_{t+1}, a)$，可由另一个探索策略收集数据。ϵ-greedy 以概率 $1-\epsilon$ 选贪心动作、以概率 ϵ 探索；表格型收敛常用 GLIE 条件：所有状态动作被无限访问且策略渐近贪心。

例题 8.3 (Q-learning 一步更新)**.** 若当前 $Q(s,a) = 1$，观察奖励 2，下一状态最大动作价值为 3，$\gamma = 0.9, \alpha = 0.5$，则 TD 目标为 $2 + 0.9 \cdot 3 = 4.7$，新值为 $1 + 0.5(4.7 - 1) = 2.85$。SARSA 若实际下一动作价值仅为 2，同样参数下更新为 $1 + 0.5(2 + 1.8 - 1) = 2.4$。

8.3.1 DQN：回放、延迟目标与过估计的不同机制

用神经网络 $Q_\theta(s,a)$ 逼近动作价值时，DQN 从回放缓冲区 $\mathcal{D}$ 抽取转移 (s, a, r, s', d)，其中终止标志 $d \in \{0, 1\}$，并最小化

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathcal{D}}\left[(y - Q_\theta(s,a))^2\right], \qquad y = r + \gamma(1-d)\max_{a'} Q_{\bar{\theta}}(s', a').$$

$\bar{\theta}$ 是若干步后硬复制或缓慢 Polyak 更新的目标参数。经验回放复用数据并减弱连续样本相关性；延迟目标使 bootstrap 目标在若干更新内近似固定。二者改善数值稳定性，但都不能把“函数逼近 + off-policy + bootstrap”的 deadly triad 变成一般收敛定理。

Algorithm 1 DQN 的一次训练更新

1: 从缓冲区均匀或按给定优先级抽取 mini-batch
2: 对每条转移计算 $y = r + \gamma(1-d)\max_{a'} Q_{\bar{\theta}}(s', a')$
3: 对 $\frac{1}{B}\sum(y - Q_\theta(s,a))^2$ 反向传播并更新 θ
4: 按固定周期令 $\bar{\theta} \leftarrow \theta$，或作慢速平均

最大化含噪估计会产生过估计偏差。Double DQN 把动作选择和动作评价拆开：

$$y_{\text{DDQN}} = r + \gamma(1-d)Q_{\bar{\theta}}\left(s', \arg\max_{a'} Q_\theta(s', a')\right).$$

目标网络的首要作用是稳定目标；Double DQN 才专门处理选择与评价使用同一噪声所致的过估计，二者不能混写。

例题 8.4 (DQN 与 Double DQN 的一步目标)**.** 设 $r = 1, \gamma = 0.9, d = 0$，在线网络在下一状态的

两个动作值为 $(4,3)$，目标网络为 $(2,2.5)$。DQN 目标为 $1+0.9\max(2,2.5)=3.25$；Double DQN 用在线网络选第一个动作，再由目标网络评价，得到 $1+0.9\cdot 2=2.8$。该例说明“选择—评价分离”，不表示 Double DQN 每个样本的目标都必然更小。

策略目标 $J(\theta)=\mathbb{E}_{\tau\sim\pi_\theta}[\sum_t \gamma^t R_t]$。

定理 8.3 (策略梯度：常用形式)**.** *在可微随机策略和可交换微分/期望的正则条件下，*

$$\nabla_\theta J(\theta)=\mathbb{E}_{\pi_\theta}\left[\sum_t \gamma^t \nabla_\theta \log \pi_\theta(A_t\mid S_t)\, G_t\right].$$

任意只依赖状态的 baseline $b(S_t)$ 可从 G_t 中减去而不改变期望。

完整证明. 对轨迹概率 $p_\theta(\tau)=p(s_0)\prod_t \pi_\theta(a_t\mid s_t)P(s_{t+1}\mid s_t,a_t)$ 使用 log-derivative：$\nabla p = p\nabla\log p$。环境转移不依赖 θ，故 $\nabla\log p_\theta(\tau)=\sum_t \nabla\log\pi_\theta(a_t\mid s_t)$。展开总折扣回报后，时刻 t 的 score 与更早奖励的乘积条件期望为零：给定 S_t 前的历史，$\mathbb{E}_{A_t\sim\pi_\theta}\nabla_\theta\log\pi_\theta(A_t\mid S_t)=0$。因此只保留从 t 开始、相对于 t 折扣的 G_t，而原目标中的时间位置贡献 γ^t，得到式(8.3)。baseline 项期望为 $\mathbb{E}_{a\sim\pi}\nabla\log\pi(a\mid s)b(s)=b(s)\nabla\sum_a\pi(a\mid s)=0$。无限时域下还需有界回报或支配条件，以交换求和、期望与求导。 □

Actor–Critic 用 actor 更新策略，用 critic 估计 V^π 或 Q^π。TD 误差 $\delta_t=R_{t+1}+\gamma V_\phi(S_{t+1})-V_\phi(S_t)$ 可作为 advantage 的低方差近似，actor 更新 $\theta\leftarrow\theta+\alpha\delta_t\nabla_\theta\log\pi_\theta(A_t\mid S_t)$。若 critic 偏差大，策略梯度也会被带偏。

8.4 方法比较、历年题与检查

方法	训练目标	采样/交互	典型问题
自回归	精确 NLL	串行生成	暴露偏差、速度
VAE	ELBO	一次解码	后验坍塌、样本偏平滑
GAN	对抗目标	一次生成	mode collapse、不稳定
扩散	去噪/score	多步反演	采样慢、schedule 敏感
Flow	精确换元似然	可逆一次/数层	可逆与 Jacobian 约束
DP	Bellman 精确期望	需要模型	状态空间爆炸
MC	回报均值	完整 episode	高方差
TD/Q	bootstrap	在线	偏差；off-policy 不稳
Policy gradient	轨迹对数导数	on-policy 常见	高方差、样本效率

历年题专题 8.1: 生成与强化学习的正式卷答题结构

正式卷常把“目标函数—采样/更新—优缺点”作为同一题的三个层次。VAE 应从边缘似然到 ELBO；扩散应从前向闭式到噪声预测；RL 应从 MDP 到 Bellman，再说明 model-based/off-policy/on-policy。

易错点

易错点：把 ELBO 写成上界；KL 方向颠倒；重参数化仍让噪声分布依赖 ϕ；GAN 生成器/判别器极值方向写反；扩散把 α_t 与 $\bar{\alpha}_t$ 混淆；Bellman 期望与最优方程混用；Q-learning 漏掉 off-policy 与探索条件；REINFORCE 漏写 log-policy；baseline 依赖动作时仍声称无偏。

复习检查

1. 能否完整推导 ELBO，并计算对角高斯 KL？
2. 能否推出最优判别器和 JS 目标，同时解释非饱和损失？
3. 能否由前向扩散递推写出式(7.5)和噪声预测目标？
4. 能否完整证明 Bellman 压缩性并写出误差界？
5. 能否比较 DP/MC/TD、SARSA/Q-learning、value-based/policy-based？
6. 能否用 log-derivative 推导策略梯度并证明 state baseline 无偏？

例题 8.5 (生成模型与优化综合题)**. 题目**：比较 VAE、GAN、diffusion 在高质量条件图像生成中的训练目标和部署成本。

思路：按似然/目标、条件注入、采样步数、覆盖与稳定性比较。

完整解答：VAE 优化 ELBO，条件进入编码/解码器，采样一次但重建分布假设可能使图像偏平滑；GAN 优化对抗目标，条件进入 G, D，推断一次且锐利，但 mode collapse 和训练博弈不稳；diffusion 优化条件噪声预测，classifier-free guidance 调整条件强度，质量/覆盖通常好但多步采样慢。部署应报告网络调用次数、显存、延迟和安全过滤，而非只比较 FID。

易错点：声称 GAN 有显式精确似然，或扩散前向过程就是生成过程。

变式：要求为移动端选择方法并说明蒸馏/少步采样。

答题方法

强化学习题固定先写状态、动作、转移、奖励、折扣和策略；再辨认要估计的是 V^π、Q^π 还是最优值。证明收敛先找 Bellman 压缩性，算法比较再看 on/off-policy、bootstrap 与函数逼近。

自测

1. 证明 T^π 与 T^* 在无穷范数下都是 γ-压缩。
2. 从一步 TD 误差推导 SARSA 与 Q-learning 更新，并指出两者的策略分布差异。
3. 用 log-derivative trick 推导 REINFORCE，并证明与状态有关的 baseline 不引入偏差。
4. 解释 deadly triad 的三个组成部分，并给出删除其中一个因素的稳定方案。

本节结论

Bellman 方程给出递归结构，采样算法给出随机近似，策略梯度给出直接优化。三条路线的共同核验点是目标算子、采样分布和误差是否有偏。

8.5 Bellman 算子：压缩性、唯一性与误差证书

对有限折扣 MDP，固定策略的 Bellman 算子为

$$(T^\pi V)(s) = \sum_a \pi(a \mid s) \left[r(s,a) + \gamma \sum_{s'} P(s' \mid s,a) V(s') \right],$$

最优算子为

$$(T^* V)(s) = \max_a \left[r(s,a) + \gamma \sum_{s'} P(s' \mid s,a) V(s') \right].$$

对任意 V, W，概率加权平均与最大值都不放大上确界范数，因此

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

Banach 不动点定理给出唯一不动点 V^π, V^*，并给出值迭代误差

$$\|V_k - V^*\|_\infty \le \gamma^k \|V_0 - V^*\|_\infty.$$

若只知道残量 $\|T^* V - V\|_\infty \le \varepsilon$，则

$$\|V - V^*\|_\infty \le \|V - T^* V\|_\infty + \|T^* V - T^* V^*\|_\infty \le \varepsilon + \gamma \|V - V^*\|_\infty,$$

故

$$\|V - V^*\|_\infty \le \frac{\varepsilon}{1-\gamma}.$$

这个残量证书比只报告相邻两次迭代差更接近考试所需的可验证误差。

8.6 策略迭代为何单调改进

设 V^π 已精确评价，并令新策略 π' 在每个状态贪心选择：

$$T^{\pi'} V^\pi = T^* V^\pi \ge T^\pi V^\pi = V^\pi.$$

因为 $T^{\pi'}$ 单调，

$$V^\pi \le T^{\pi'} V^\pi \le (T^{\pi'})^2 V^\pi \le \cdots.$$

右侧收敛到 $T^{\pi'}$ 的唯一不动点 $V^{\pi'}$，故

$$V^{\pi'} \ge V^\pi.$$

若某状态严格改进且可从有关初始状态以正概率到达，则相应价值严格提高。有限确定策略数有限，所以精确策略迭代在不重复策略的情况下有限终止；实际近似策略迭代因评价误差可能振荡，需要额外误差控制。

8.7 Monte Carlo 与 TD(0)：目标、偏差与随机近似

首次访问 MC 使用完整回报

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots$$

更新 $V(S_t)$。给定 $S_t = s$，G_t 对 $V^\pi(s)$ 无偏，但方差可能很大，而且必须等到回合结束。

TD(0) 使用一步自举目标

$$Y_t = R_{t+1} + \gamma V_t(S_{t+1}), \qquad \delta_t = Y_t - V_t(S_t),$$

$$V_{t+1}(S_t) = V_t(S_t) + \alpha_t \delta_t.$$

由于目标包含当前估计，有限时刻的 Y_t 一般对真实价值有偏；但条件期望漂移指向 Bellman 误差：

$$\mathbb{E}[\delta_t \mid S_t = s, \mathcal{F}_t] = (T^\pi V_t)(s) - V_t(s).$$

在有限表格、遍历访问、有界奖励和 Robbins–Monro 步长

$$\sum_t \alpha_t(s) = \infty, \qquad \sum_t \alpha_t(s)^2 < \infty$$

下，噪声平方可和而每个状态仍获得无限总学习量，TD(0) 收敛到 V^π。

8.8 SARSA 与 Q-learning：同策略与异策略

SARSA 更新

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha_t \left[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t) \right].$$

它评价实际生成下一动作的行为策略，是 on-policy 方法。Q-learning 改成

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha_t \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \right],$$

目标对应最优 Bellman 算子，即使样本由含探索的另一策略生成，仍是 off-policy。

表格 Q-learning 的标准收敛条件包括：有限状态动作集、奖励有界、$\gamma < 1$、每个 (s, a) 被访问无穷多次，以及逐状态动作步长满足 Robbins–Monro 条件。只写“ϵ-greedy”不够；固定 $\epsilon > 0$ 保证持续探索，却不保证行为策略最终变成贪心策略。若要求 GLIE，可令探索逐渐消失同时保持所有动作无限访问。

反例 8.1. 把表格收敛定理直接搬到“函数逼近 + off-policy + bootstrapping”是错误的。这三项组合被称为 deadly triad：投影、采样分布与 Bellman 更新可能形成不再压缩的线性算子，即使

奖励有界，参数也可能发散。

8.9 策略梯度定理的占用测度推导

令

$$J(\theta) = \mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\sum_{t=0}^{\infty}\gamma^t R_t\right].$$

轨迹密度分解为

$$p_\theta(\tau) = p(s_0)\prod_{t\geq 0}\pi_\theta(a_t \mid s_t)P(s_{t+1} \mid s_t, a_t).$$

环境转移不依赖 θ，故对数导数为

$$\nabla_\theta \log p_\theta(\tau) = \sum_{t\geq 0}\nabla_\theta \log \pi_\theta(a_t \mid s_t).$$

似然比分解给

$$\nabla J(\theta) = \mathbb{E}_\tau\left[\left(\sum_{t\geq 0}\nabla \log \pi_\theta(A_t \mid S_t)\right)\left(\sum_{k\geq 0}\gamma^k R_k\right)\right].$$

当 $k < t$ 时，给定 S_t 后，过去奖励不依赖当前动作，而

$$\mathbb{E}[\nabla \log \pi_\theta(A_t \mid S_t) \mid S_t] = \sum_a \pi_\theta(a \mid S_t)\nabla \log \pi_\theta(a \mid S_t) = \nabla \sum_a \pi_\theta(a \mid S_t) = 0.$$

因此可删除每个 score 前的过去奖励，只保留 reward-to-go，得到

$$\nabla J(\theta) = \mathbb{E}\sum_{t\geq 0}\gamma^t \nabla \log \pi_\theta(A_t \mid S_t) Q^{\pi_\theta}(S_t, A_t).$$

把折扣状态访问次数归一化为 d_γ^π，等价写成

$$\nabla J(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{\substack{S\sim d_\gamma^\pi \\ A\sim\pi_\theta(\cdot|S)}}\left[\nabla \log \pi_\theta(A \mid S) Q^\pi(S, A)\right].$$

8.10 Baseline 为什么不引入偏差

对任意只依赖状态的 $b(s)$，

$$\mathbb{E}_{A\sim\pi_\theta(\cdot|s)}\left[\nabla \log \pi_\theta(A \mid s) b(s)\right] = b(s)\nabla \sum_a \pi_\theta(a \mid s) = 0.$$

故可用优势函数

$$A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$$

替代 Q^π 而不改变期望梯度。它减少的不是所有噪声，而是与动作无关的回报波动。若 baseline 依赖动作，则一般不再自动为零，除非加入相应校正。

8.11 Actor–critic 的两时间尺度解释

Critic 以 TD 误差学习 V_w 或 Q_w，actor 使用

$$\theta_{t+1} = \theta_t + \beta_t \nabla_\theta \log \pi_{\theta_t}(A_t \mid S_t)\, \widehat{A}_t.$$

典型的一步优势估计为

$$\widehat{A}_t = R_{t+1} + \gamma V_w(S_{t+1}) - V_w(S_t).$$

理论分析常让 critic 的步长 α_t 比 actor 的 β_t 更快，即 $\beta_t/\alpha_t \to 0$：在 actor 看来，critic 近似已追踪当前策略的价值；在 critic 看来，策略参数缓慢变化。若两个网络共享表示，这种理想分离只是一种分析近似，实践仍需监控目标漂移。

例题 8.6. 两臂 bandit 的策略为 $\pi_\theta(1) = \sigma(\theta)$，奖励均值 μ_1, μ_0。则

$$J(\theta) = \sigma(\theta)\mu_1 + [1 - \sigma(\theta)]\mu_0,$$

$$J'(\theta) = \sigma(\theta)[1 - \sigma(\theta)](\mu_1 - \mu_0).$$

用 score-function 公式计算也得到同一结果。若 $\mu_1 > \mu_0$，梯度为正，但当策略已接近确定性时 $\sigma(1-\sigma)$ 很小，说明过早饱和会减慢探索修正。

答题方法

强化学习证明题按“对象—算子—范数—不动点—采样近似”组织。先判断是在解固定策略方程还是最优方程，再区分精确动态规划、表格随机近似和函数逼近；三类结论的条件不能混用。

自测

1. 由 Bellman 残量推导价值函数误差上界。
2. 策略改进定理中，哪里使用了 Bellman 算子的单调性？
3. 比较 MC、TD、SARSA、Q-learning 的目标是否自举、是否 on-policy。
4. 证明状态 baseline 不改变策略梯度，并说明动作相关 baseline 的风险。
5. 函数逼近 Q-learning 为什么不能只引用表格收敛定理？

第九章　三期正式真题与完整解答

本章按年份收入 Deep Learning and Reinforcement Learning 的全部正式题面与解答，共保留既有审计口径下的 8 道大题、31 个最细作答单元。除特别注明外，解答为复习用整理稿，不代表官方评分标准。每个题组前的“理论入口”用于回看本 Part 前文。

9.1　2025 春

理论入口：参见章 七、八。

2025 春·DL/RL·Q1（原卷第 2–3 页，18 分）

Original. Consider a classification problem: the training dataset is given as $\{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^d$ represents the input features and $y_i \in \{1, 2, \ldots, C\}$ represents the class labels. A supervised deep-learning pipeline typically includes preparing training data, defining a hypothesis space, designing a training scheme, and optimizing the network. Answer the following questions: (a) Define the hypothesis space consisting of all neural networks structured with a feature extractor—an MLP mapping the inputs to a learned feature representation—and a classifier—an MLP mapping that representation to class probabilities. Specify the input, output, and training parameters. (b) Specify an appropriate loss and explain how SGD optimizes the network. (c) Describe batch normalization in the training and testing phases. (d) If training error is unsatisfactory, describe at least two adjustments that improve expressivity. (e) If training error is low but testing error is high, discuss at least two strategies that reduce overfitting.

中文译述

围绕一个多类监督深度学习流程，分别说明网络假设空间、损失与 SGD、BatchNorm 的训练/测试差异、欠拟合修正和过拟合修正。

完整解答：(a)–(b)

设特征器 $\phi_{\theta_f} : \mathbb{R}^d \to \mathbb{R}^p$，分类器产生 logits $a_{\theta_c} : \mathbb{R}^p \to \mathbb{R}^C$，$p_\theta(y = c \mid x) = \mathrm{softmax}(a_{\theta_c}(\phi_{\theta_f}(x)))_c$，$\theta = (\theta_f, \theta_c)$ 是全部权重和偏置。交叉熵 $J(\theta) = -N^{-1} \sum_i \log p_\theta(y_i \mid x_i)$。minibatch B_k 上令 $g_k = |B_k|^{-1} \sum_{i \in B_k} \nabla_\theta[-\log p_\theta(y_i \mid x_i)]$，更新 $\theta_{k+1} = \theta_k - \eta_k g_k$；若均匀抽样，$g_k$ 对全梯度无偏。

完整解答：(c)

训练时对每个通道/特征计算 batch 均值 μ_B 和方差 σ_B^2，$\hat{z}=(z-\mu_B)/\sqrt{\sigma_B^2+\varepsilon}$，输出 $\gamma\hat{z}+\beta$，同时更新 running mean/variance。测试时不能依赖当前小 batch，使用训练期累计统计量；γ,β 始终可训练。

完整解答：(d)–(e)

训练误差高表示优化或表达不足：增宽/加深、改用合适激活/残差、延长训练并调学习率、检查归一化与数据质量。至少列两项并说明原因。训练误差低而测试误差高表示过拟合：更多数据/增强、weight decay、dropout、早停、减小模型、交叉验证；不能用“继续增大模型”作为唯一方案。

易错点： 把 BN 测试统计量写成当前 batch；不区分欠拟合与过拟合。

正文对应： 节 6.1、章 七中的前向/反向、BN 与正则化。

理论入口： 参见章 七、八。

2025 春·DL/RL·Q2（原卷第 3 页，15 分）

Original. For the VP SDE $dX_t=-\frac{1}{2}\beta(t)X_t dt+\sqrt{\beta(t)}dW_t$: (a) write the reverse-time SDE and explain score matching; (b) prove

$$\mathbb{E}_{x_t}\|s_\theta-\nabla\log p_t(x_t)\|^2=\mathbb{E}_{x_0,x_t}\|s_\theta-\nabla\log p_{t|0}(x_t|x_0)\|^2+C;$$

(c) find $p_{t|0}(x_t|x_0)$ and derive the denoising score-matching loss.

中文译述

对 variance-preserving SDE，写反向 SDE，证明边缘 score matching 与条件 denoising score matching 只差常数，并推出可训练损失。

完整解答：(a)

一般 $dX=f dt+g,dW$ 的反向时间动力学（沿 $t:T\to 0$ 积分）为

$$dX_t=[f(X_t,t)-g(t)^2\nabla_x\log p_t(X_t)]dt+g(t)d\bar{W}_t.$$

故本题漂移为 $-\frac{1}{2}\beta(t)x-\beta(t)\nabla\log p_t(x)$。边缘 score 不可直接计算，故训练网络 s_θ 逼近它。

完整解答：(b)

记 $a=s_\theta(x_t,t)$，$u=\nabla\log p_t(x_t)$，$v=\nabla\log p_{t|0}(x_t|x_0)$。Fisher identity 给 $\mathbb{E}[v\mid x_t]=u$。展开平方并条件于 x_t：

$$\mathbb{E}\|a-v\|^2=\mathbb{E}\|a-u\|^2+\mathbb{E}\|v-u\|^2,$$

交叉项因 $\mathbb{E}[v-u\mid x_t]=0$ 消失。故题中 $C=-\mathbb{E}\|v-u\|^2$ 若按左等于右加 C 的写法；它与 θ 无关。

完整解答：(c)

令 $B(t)=\int_0^t \beta(s)ds$、$\alpha(t)=e^{-B(t)/2}$，线性 SDE 解给

$$X_t \mid X_0=x_0 \sim \mathcal{N}(\alpha(t)x_0,(1-\alpha(t)^2)I).$$

写 $X_t=\alpha x_0+\sqrt{1-\alpha^2}\epsilon$，则条件 score $-\epsilon/\sqrt{1-\alpha^2}$。因此可最小化

$$\mathbb{E}_{t,x_0,\epsilon}\lambda(t)\left\|s_\theta(\alpha x_0+\sqrt{1-\alpha^2}\epsilon,t)+\frac{\epsilon}{\sqrt{1-\alpha^2}}\right\|^2.$$

等价噪声参数化令 $\epsilon_\theta=-\sqrt{1-\alpha^2}s_\theta$，得到加权噪声 MSE。

正文对应： 节 7.9中的扩散与 score matching。

9.2 2025 秋

理论入口： 参见章 七、八。

2025 秋·DL/RL·Q1（原卷第 4 页，9 分）

Original. (a) Derive the parameters of a $k\times k$ convolution from C_{in} to C_{out} (with bias) and compare with a dense layer of the same input/output size. (b) Write scaled dot-product self-attention, define Q,K,V, explain positional information and one injection method. (c) State one benefit each of convolution and self-attention, then design a minimal vision Transformer using both.

中文译述

计算卷积参数、写注意力和位置编码，并设计卷积与自注意结合的最小视觉 Transformer。

完整解答：Q1

(a) 卷积参数 $k^2C_{in}C_{out}+C_{out}$，与 HWC_{in} 到 $H'W'C_{out}$ 的全连接参数 $HWC_{in}H'W'C_{out}+H'W'C_{out}$ 相比不随空间位置复制，体现权共享。(b) $Q=XW_Q,K=XW_K,V=XW_V$，$\mathrm{Attn}(X)=\mathrm{softmax}(QK^T/\sqrt{d_k})V$。无位置项时对 token 置换等变，故加正弦位置编码、可学习位置向量或相对位置偏置。(c) 卷积有局部性/平移等变偏置，自注意可直接建模全局关系。最小系统：卷积 stem 或 patch 卷积把图像变成 token，叠加位置编码，经若干 Transformer block，在 block 内或 tokenization 前用 depthwise/普通卷积注入局部性，最后 CLS/池化分类。

维度检查： $Q,K\in\mathbb{R}^{N\times d_k}$、$V\in\mathbb{R}^{N\times d_v}$，注意力矩阵 $N\times N$。

理论入口： 参见章 七、八。

2025 秋·DL/RL·Q2（原卷第 4–5 页，14 分）

Original. (a) Show maximum likelihood for p_θ is equivalent to minimizing $\mathrm{KL}(p_{data}\|p_\theta)$. (b) For a flow $z_0\sim p_0$, $z_\ell=f_\ell(z_{\ell-1})$, $x=z_L$, write the dataset loss. (c) Explain the VAE

ELBO and its reconstruction/regularization terms. (d) For $q(x_{1:T}|x_0) = \prod_t q(x_t|x_{t-1})$ with Gaussian noise, give the backward model and derive the diffusion ELBO decomposition shown in the paper.

中文译述

统一考查最大似然、normalizing flow、VAE ELBO 和扩散模型的分层 VAE 解释。

完整解答：Q2

(a) $\mathrm{KL}(p_d\|p_\theta) = \mathbb{E}_{p_d}\log p_d - \mathbb{E}_{p_d}\log p_\theta$，第一项与 θ 无关。(b) 对数据 x 反演 $z_{\ell-1} = f_\ell^{-1}(z_\ell)$，

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

数据集取平均。(c) $\log p_\theta(x) = \mathrm{ELBO} + \mathrm{KL}(q_\phi(z|x)\|p_\theta(z|x))$，$\mathrm{ELBO} = \mathbb{E}_q \log p_\theta(x|z) - \mathrm{KL}(q_\phi(z|x)\|p(z))$；前者重建，后者使潜变量接近先验。(d) 取 $p(x_T) = N(0, I)$、$p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$。将 q 插入 Jensen 后用 Markov 分解重排，得到

$$\begin{aligned}\log p_\theta(x_0) \geq\ & -\mathrm{KL}(q(x_T|x_0)\|p(x_T)) \\ & -\sum_{t=2}^{T}\mathbb{E}_q\mathrm{KL}\big(q(x_{t-1}|x_t,x_0)\|p_\theta(x_{t-1}|x_t)\big) \\ & +\mathbb{E}_q\log p_\theta(x_0|x_1).\end{aligned}$$

理论入口：参见章 七、八。

2025 秋·DL/RL·Q3（原卷第 5–6 页，10 分）

Original. In a discounted MDP with differentiable policy π_θ, fixed s_0, return G_0, objective $J = V^{\pi_\theta}(s_0)$ and discounted visitation d^π: (a) prove

$$\nabla J = \mathbb{E}_\pi \sum_{t\geq 0}\gamma^t \nabla\log\pi_\theta(a_t|s_t)Q^\pi(s_t,a_t) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\pi, a\sim\pi}[\nabla\log\pi(a|s)Q^\pi(s,a)];$$

(b) prove any state baseline has zero expectation and explain the advantage baseline.

中文译述

证明策略梯度定理、状态基线无偏和 advantage 降方差。

完整解答：Q3

轨迹概率 $p_\theta(\tau) = p(s_0)\prod_t \pi_\theta(a_t|s_t)P(s_{t+1}|s_t,a_t)$，环境转移与 θ 无关，故 $\nabla\log p_\theta(\tau) = \sum_t \nabla\log\pi_\theta(a_t|s_t)$。似然比给 $\nabla J = \mathbb{E}[G_0\sum_t\nabla\log\pi_t]$。对第 t 项，早于 t 的奖励在条件于历史时与 a_t 的 score 均值零，故只保留从 t 起的回报；条件期望即 $Q^\pi(s_t,a_t)$，得到第一式。按 $d^\pi(s) = \sum_t \gamma^t \Pr(s_t = s)$ 重排求和得第二式（若把 d^π 定义为未归一分布，则不应再除

$1-\gamma$；原卷同时写“unnormalized”与 $1/(1-\gamma)$，这里采用归一化采样约定并明确差异）。对任意 $b(s)$，$\mathbb{E}_{a\sim\pi}\nabla\log\pi(a|s)b(s)=b(s)\nabla\sum_a\pi(a|s)=0$。取 $b(s)=V^\pi(s)$ 得 advantage，保持无偏并通常降低方差。

题目歧义： 原卷 d^π 的“unnormalized”文字与第二等式的 $1/(1-\gamma)$ 存在归一化约定冲突；显式区分。

9.3 2026 春

理论入口： 参见章 七、八。

2026 春·DL/RL·Q1（原卷第 3–4 页，10 分）

Original. (a) For $x^\ell=f_\ell(x^{\ell-1},\theta_\ell)$ and $J=\mathcal{L}(x^L,y)$, derive $\partial J/\partial\theta_\ell$ and explain reverse-mode cost. (b) explain sigmoid vanishing gradients vs. ReLU; (c) write momentum SGD and explain oscillation reduction; (d) explain $1/\sqrt{d_k}$ in attention and what happens without it; (e) compare RNN and Transformer computation and hardware efficiency.

中文译述

考查反向传播、激活梯度、动量、注意力缩放和 RNN/Transformer 并行性。

完整解答：Q1

(a) 令 $\delta^L=\partial J/\partial x^L$，反推 $\delta^\ell=\delta^{\ell+1}(\partial x^{\ell+1}/\partial x^\ell)$，$\partial J/\partial\theta_\ell=\delta^\ell(\partial x^\ell/\partial\theta_\ell)$；每条计算图边只做常数次局部 Jacobian–vector product，故总复杂度与 forward 同阶、常数倍。(b) $\sigma'(z)=\sigma(1-\sigma)\le 1/4$ 且饱和区趋零，多层乘积指数衰减；ReLU 正区导数 1，缓解但不能完全消除消失/死亡神经元。(c) $v_{k+1}=\beta v_k+g_k$、$x_{k+1}=x_k-\eta v_{k+1}$（等价 heavy-ball 形式）；累积一致方向并平滑高曲率方向符号翻转。(d) 独立零均值分量使 q^Tk 方差约 d_k，除以 $\sqrt{d_k}$ 保持 logits $O(1)$；否则 softmax 饱和近 one-hot，梯度变小。(e) RNN 在时间上串行、路径 $O(L)$；Transformer 同层 token 并行并用矩阵乘法，现代硬件吞吐高，但注意力时间/存储 $O(L^2)$。

理论入口： 参见章 七、八。

2026 春·DL/RL·Q2（原卷第 4–5 页，15 分）

Original. Let $(x_0,x_1)\sim\gamma$, $x_t=(1-t)x_0+tx_1$. (a) compute conditional velocity; (b) define $v_t(x)=\mathbb{E}[v_t(x_t|x_0,x_1)|x_t=x]$ and prove the continuity equation and transport; (c) give a squared regression loss and prove its minimizer is the marginal velocity; (d) for $dX_t=(v_t(X_t)+\beta_t\nabla\log\pi_t(X_t))dt+\sqrt{2\beta_t}dW_t$, derive Fokker–Planck, show the same marginals, and explain score matching.

中文译述

完整推导 flow matching 的条件/边缘速度、回归目标，以及加入 score 后保持同一边缘分布的随机动力学。

完整解答：Q2(a)–(c)

$\dot{x}_t = x_1 - x_0$。对任意光滑测试函数 φ，$\frac{d}{dt}\mathbb{E}\varphi(x_t) = \mathbb{E}\nabla\varphi(x_t)^T(x_1 - x_0) = \mathbb{E}\nabla\varphi(x_t)^T v_t(x_t)$；分部积分的弱形式即 $\partial_t \pi_t + \nabla \cdot (\pi_t v_t) = 0$，端点边缘为 π_0, π_1。训练损失 $\mathcal{L}(\theta) = \mathbb{E}_{t,(x_0,x_1),x_t}\|v_\theta(t, x_t) - (x_1 - x_0)\|^2$。条件平方损失的唯一 L^2 最优预测是条件均值，因此最优 $v_\theta = v_t$。

完整解答：Q2(d)

Fokker–Planck 为

$$\partial_t p = -\nabla \cdot [p(v_t + \beta_t \nabla \log \pi_t)] + \beta_t \Delta p.$$

代入 $p = \pi_t$，后两项 $-\nabla\cdot(\beta_t\pi_t\nabla\log\pi_t) + \beta_t\Delta\pi_t = -\beta_t\Delta\pi_t + \beta_t\Delta\pi_t = 0$，余下原 continuity equation，故边缘相同。未知 score 可用 denoising score matching，从可采样的条件扰动 score 回归得到。

理论入口：参见章 七、八。

2026 春·DL/RL·Q3（原卷第 5–6 页，8 分）

Original. In a discounted MDP: (a) derive the Bellman equation for V^π; (b) give policy iteration evaluation/improvement; (c) adapt to model-free actor–critic; (d) specify either reward-model-based RLHF with KL regularization or a DPO-style objective for preference triples (x, y^+, y^-).

中文译述

从 Bellman 到 policy iteration、actor–critic，再写一个完整 RLHF/DPO 目标。

完整解答：Q3

(a) 对第一步条件化：$V^\pi(s) = \sum_a \pi(a|s)[r(s,a) + \gamma \sum_{s'} P(s'|s,a)V^\pi(s')]$。(b) 固定 π_k 解 $V^{\pi_k} = T^{\pi_k}V^{\pi_k}$，再 $\pi_{k+1}(s) \in \arg\max_a [r + \gamma P V^{\pi_k}]$，直至稳定。(c) critic 用采样 TD 误差 $\delta_t = r_t + \gamma V_w(s_{t+1}) - V_w(s_t)$ 更新 w；actor 用 $\theta \leftarrow \theta + \eta\delta_t \nabla \log \pi_\theta(a_t|s_t)$，无需已知 P。(d) DPO 可写

$$-\mathbb{E}\log\sigma\left(\beta\left[\log\frac{\pi_\theta(y^+|x)}{\pi_{ref}(y^+|x)} - \log\frac{\pi_\theta(y^-|x)}{\pi_{ref}(y^-|x)}\right]\right).$$

它直接用偏好对；奖励模型路线则先做 Bradley–Terry，再最大化奖励减 KL。

9.4 高频错误、作答模板与复习检查

反向传播先核对维度；attention 不能漏 $\sqrt{d_k}$ 和 mask；Transformer 需同时报告 attention 与 FFN 的时间/空间；ELBO 是下界；GAN 最优判别器需处理共同支配测度；diffusion 的 forward 与生成 reverse 不可颠倒；Bellman expectation/optimality、SARSA/Q-learning、on/off-policy 必须区分；baseline 只有在不依赖当前动作等条件下保持无偏。

复习检查

考场模板：网络题列形状和计算图，生成题列联合分布与目标，RL 题列 MDP、价值、Bellman、更新及收敛边界。完成后应能手算一层反向传播、卷积尺寸、attention、ELBO、Q-learning，并完整证明 Bellman 压缩性、策略改进和 state baseline 零期望。

第三部分

Optimization Methods in Artificial Intelligence

人工智能中的优化方法

第十章　凸分析、次梯度与对偶证书

学习目标

本章建立优化 Part 的语言：凸集/函数、光滑与强凸、次梯度、共轭、法锥、KKT 和 Slater。目标是能从几何条件写出可验证的最优性证书，而不是只背算法名称。

10.1　凸性、光滑性与最优性条件

10.1.1　凸集、分离与 Farkas 引理

集合 $C \subseteq \mathbb{R}^d$ 为凸集，若任意 $x, y \in C$ 和 $\theta \in [0,1]$ 都有 $\theta x + (1-\theta)y \in C$。仿射集要求该条件对所有 $\theta \in \mathbb{R}$ 成立；凸锥还要求 $\alpha x + \beta y \in C$ 对任意 $\alpha, \beta \geq 0$ 成立。凸组合、仿射组合和锥组合是后续分离定理的基本语言。

定理 10.1 (点与闭凸集的严格分离)**.** *设 $C \subset \mathbb{R}^d$ 非空、闭、凸，$b \notin C$。则存在 $a \neq 0$ 和 $\alpha \in \mathbb{R}$，使*

$$a^T x \leq \alpha < a^T b, \qquad \forall x \in C.$$

完整证明. 令 $\gamma = \inf_{x \in C} \|b - x\| > 0$。取极小序列 $x_k \in C$。由凸性，$(x_k + x_m)/2 \in C$，平行四边形恒等式给出

$$\|x_k - x_m\|^2 = 2\|x_k - b\|^2 + 2\|x_m - b\|^2 - 4\left\|\frac{x_k + x_m}{2} - b\right\|^2 \to 0.$$

故 x_k 为 Cauchy 列；闭性给 $\bar{x} = \lim x_k \in C$，且 $\|b - \bar{x}\| = \gamma$。对任意 $x \in C$ 与 $t \in [0,1]$，$\bar{x} + t(x - \bar{x}) \in C$，函数 $\phi(t) = \|b - \bar{x} - t(x - \bar{x})\|^2$ 在 $t = 0$ 取最小，故 $\phi'(0) = -2(b - \bar{x})^T(x - \bar{x}) \geq 0$。令 $a = b - \bar{x} \neq 0$，得到 $a^T x \leq a^T \bar{x}$；而 $a^T b - a^T \bar{x} = \|b - \bar{x}\|^2 > 0$。取 $\alpha = a^T \bar{x}$ 即得。 □

分离定理还有边界点支撑超平面、两个不交闭凸集的分离版本。条件必须逐一核对：没有闭性时最近点未必存在；没有凸性时最近点的一阶方向论证不能控制整个集合。

引理 10.2 (Farkas 二择一)**.** *给定 $A \in \mathbb{R}^{m \times n}$、$b \in \mathbb{R}^m$，下列系统恰有一个可行：*

$$Ax = b,\ x \geq 0; \qquad A^T y \geq 0,\ b^T y < 0.$$

两者不能同时可行，因为 $b^T y = x^T A^T y \geq 0$。至少一个可行由锥 $\{Ax : x \geq 0\}$ 与点 b 的分离得到。Gordan 定理是齐次对应版本；它们是对偶证书与 KKT 乘子的几何前置。

10.1.2 proper、闭凸、epigraph 与存在唯一性

扩展实值函数 $f:\mathbb{R}^d\to(-\infty,+\infty]$ 为 proper，若 $f(x)>-\infty$ 且 $\operatorname{dom} f=\{x:f(x)<+\infty\}\neq\varnothing$。其 epigraph 为 $\operatorname{epi} f=\{(x,t):f(x)\le t\}$；$f$ 凸当且仅当 $\operatorname{epi} f$ 凸，f 闭当且仅当 epigraph 闭。indicator $\iota_C(x)=0$（$x\in C$）而在 C 外为 $+\infty$，把约束并入目标。

定理 10.3 (Weierstrass 存在性)**.** *若 proper 闭函数 f 的某个非空下水平集 $\{x:f(x)\le\alpha\}$ 有界，则 f 在该下水平集上取得全局最小值。若 f 还为 μ-强凸（$\mu>0$），最小点唯一。*

完整证明. 取极小序列 x_k，从充分大的 k 起它落在某个非空有界下水平集内。闭下水平集在有限维空间中紧，故有收敛子列 $x_{k_j}\to x^*$；下半连续性给 $f(x^*)\le\liminf_j f(x_{k_j})=\inf f$，所以取得最小值。若存在不同最小点 x^*,y^*，强凸性在中点给 $f((x^*+y^*)/2)\le\frac12 f(x^*)+\frac12 f(y^*)-\mu\|x^*-y^*\|^2/8<\inf f$，矛盾。 □

10.1.3 凸、严格凸、强凸、光滑和联合不等式

凸函数在任意不同 x,y 与 $\theta\in(0,1)$ 上若总满足严格的 Jensen 不等式，则称为严格凸（strictly convex）。μ-强凸要求减去 $\mu\|x\|^2/2$ 后仍凸，因而强凸蕴含严格凸，严格凸蕴含凸；反向一般不成立。例如 x^4 在 $\mathbb{R}$ 上严格凸但不是全局强凸。

一阶凸性判据：可微函数在凸域上凸，当且仅当 $f(y)\ge f(x)+\langle\nabla f(x),y-x\rangle$；**二阶凸性判据：**二阶可微时等价于 $\nabla^2 f(x)\succeq 0$。L-smooth 指 $\|\nabla f(x)-\nabla f(y)\|\le L\|x-y\|$，从而有下降引理

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

μ-强凸指

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

定理 10.4 (光滑强凸联合不等式)**.** *若 f 可微、L-smooth、μ-强凸，$0<\mu\le L$，则*

$$\langle\nabla f(x)-\nabla f(y),x-y\rangle\ge\frac{\mu L}{\mu+L}\|x-y\|^2+\frac{1}{\mu+L}\|\nabla f(x)-\nabla f(y)\|^2.$$

完整证明. 令 $g(x)=f(x)-\frac{\mu}{2}\|x\|^2$。则 g 凸且 $(L-\mu)$-smooth。对凸光滑函数应用 cocoercivity：$\langle\nabla g(x)-\nabla g(y),x-y\rangle\ge\|\nabla g(x)-\nabla g(y)\|^2/(L-\mu)$。代入 $\nabla g(x)-\nabla g(y)=\Delta\nabla f-\mu\Delta x$，展开平方并整理即得。$L=\mu$ 时，对任意 $\varepsilon>0$ 把光滑常数放宽为 $L+\varepsilon$，应用上述证明后令 $\varepsilon\downarrow 0$ 即得。 □

10.1.4 次微分（subdifferential）、法锥（normal cone）、投影与 prox

对 proper 凸函数，次梯度（subgradient）$g\in\partial f(x)$ 当且仅当 $f(y)\ge f(x)+\langle g,y-x\rangle$ 对所有 y 成立。闭凸集 C 在 $x\in C$ 的法锥

$$N_C(x)=\{z:\langle z,y-x\rangle\le 0,\forall y\in C\}=\partial\iota_C(x).$$

故 x^* 解 $\min_{x\in C}f(x)$ 的一阶条件是 $0\in\partial f(x^*)+N_C(x^*)$。投影 $\Pi_C=\operatorname{prox}_{\iota_C}$。对闭凸集上的光滑问题 $\min_{x\in C}f(x)$，投影梯度（projected gradient）更新为

$$x_{k+1}=\Pi_C\left(x_k-t\nabla f(x_k)\right).$$

若 f 凸且 L-smooth、$0 < t \le 1/L$，可由投影的 firm nonexpansiveness 与下降引理得到标准 $O(1/k)$ 函数值率；若只讨论非凸问题，则通常改以投影梯度映射趋零作为驻点度量。

命题 10.5 (prox 最优性、单调性与 firm nonexpansiveness)**.** *对 proper 闭凸 h 和 $t > 0$,*

$$p = \mathrm{prox}_{th}(u) \iff \frac{u-p}{t} \in \partial h(p).$$

若 $p = \mathrm{prox}_{th}(u)$、$q = \mathrm{prox}_{th}(v)$，则 $\|p-q\|^2 \le \langle p-q, u-v\rangle$，因而 prox 非扩张。

完整证明. prox 的目标 $h(z) + \|z-u\|^2/(2t)$ 强凸，最优性条件是 $0 \in \partial h(p) + (p-u)/t$。对 p, q 分别写该条件；凸次微分的单调性给 $\langle (u-p)/t - (v-q)/t, p-q\rangle \ge 0$，整理得到 firm 不扩张。 □

例题 10.1 (软阈值、投影和维度检查)**.** 若 $h(x) = \lambda\|x\|_1$，则 $\mathrm{prox}_{th}(u)_i = \mathrm{sign}(u_i)(|u_i| - t\lambda)_+$。取 $u = (-3, 0.5, 2)^T, t\lambda = 1$，结果为 $(-2, 0, 1)^T$。若 $h = \iota_{\{x:\|x\|\le r\}}$，则 prox 为欧氏投影；两者输入输出均在 $\mathbb{R}^d$，不能把标量阈值误作向量范数。

10.2 对偶、KKT 与约束优化

10.2.1 Fenchel 共轭、双共轭和 Moreau

proper 函数的 Fenchel 共轭 $f^*(y) = \sup_x\{\langle y, x\rangle - f(x)\}$，Fenchel–Young 不等式 $f(x) + f^*(y) \ge \langle x, y\rangle$，等号当且仅当 $y \in \partial f(x)$。proper 闭凸函数满足 $f^{**} = f$；证明用 epigraph 分离。

易错点

Fenchel 强对偶需要正则性条件，常用的有效形式是相关函数定义域的相对内部交非空。不能把该条件误写为“交为空集”，也不能只凭凸性就无条件声称强对偶。

Moreau 分解（单位缩放）为 $x = \mathrm{prox}_f(x) + \mathrm{prox}_{f^*}(x)$；一般缩放为

$$x = \mathrm{prox}_{tf}(x) + t\,\mathrm{prox}_{f^*/t}(x/t).$$

Moreau envelope $f^{(t)}(x) = \min_u\{f(u) + \|u-x\|^2/(2t)\}$ 满足 $\nabla f^{(t)}(x) = (x - \mathrm{prox}_{tf}(x))/t$。

10.2.2 Lagrange 对偶、Slater 与 KKT（考纲保留）

对 $\min f_0(x)$ s.t. $f_i(x) \le 0$、$h_j(x) = 0$，$\mathcal{L} = f_0 + \sum_i \lambda_i f_i + \sum_j \nu_j h_j$，$\lambda \ge 0$，$g(\lambda, \nu) = \inf_x \mathcal{L}$ 总给弱对偶。凸问题若存在严格可行点 $f_i(x) < 0$ 且仿射等式成立（Slater），则强对偶成立并可用 KKT：原始可行、对偶可行、互补松弛和驻点。作答时应能从 Lagrangian 依次写出对偶函数、对偶问题和 KKT 四组条件，并单独核对 Slater 条件是否成立。

答题方法

先判断可行集和目标是否凸，再写一阶条件；含约束时依次检查约束资格、原始可行、对偶可行、互补松弛和驻点。只有凸性与资格条件齐全时，KKT 才能作为全局最优充要证书。

自测

1. 证明光滑凸函数的 cocoercivity，并说明它与下降引理的关系。
2. 计算 ℓ_1 范数、最大函数和集合指示函数的次微分。
3. 从 Lagrangian 推导弱对偶，并在 Slater 条件下解释强对偶的作用。
4. 给出一个非凸问题满足 KKT 但不是全局最优的反例。

本节结论

凸分析把“最优”转换为分离、次梯度和对偶等可检查条件；算法收敛证明将在这些不等式上反复迭代。

10.3 光滑与强凸的可用不等式

若 f 可微且梯度 L-Lipschitz，则沿线段积分：

$$f(y)-f(x)=\int_0^1 \langle \nabla f(x+t(y-x)),y-x\rangle \,\mathrm{d}t.$$

减去 $\langle \nabla f(x),y-x\rangle$，用 Lipschitz 条件得下降引理

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

若 f 又为凸函数，则梯度是 $1/L$-cocoercive：

$$\langle \nabla f(x)-\nabla f(y),x-y\rangle \ge \frac{1}{L}\|\nabla f(x)-\nabla f(y)\|^2.$$

这比普通单调性更强，是证明梯度映射非扩张的关键。

μ-强凸性要求

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

在最优点 x^* 处 $\nabla f(x^*)=0$，故

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

若还 L-光滑，则条件数 $\kappa=L/\mu$ 控制一阶法的最坏速率。

10.4 Fenchel 共轭与 Fenchel–Young 等号条件

共轭定义为

$$f^*(y)=\sup_x\{\langle y,x\rangle-f(x)\}.$$

直接由上确界定义，

$$f(x)+f^*(y)\ge \langle x,y\rangle.$$

对闭 proper 凸函数，等号成立当且仅当

$$y \in \partial f(x) \quad \Longleftrightarrow \quad x \in \partial f^*(y).$$

例如 $f(x) = \frac{1}{2}\|x\|^2$ 时 $f^*(y) = \frac{1}{2}\|y\|^2$，Fenchel–Young 就是

$$\frac{1}{2}\|x - y\|^2 \geq 0.$$

若 $f = \mathbf{1}_C$ 是闭凸集 C 的指示函数，则

$$f^*(y) = \sup_{x \in C} \langle y, x \rangle = \sigma_C(y),$$

即支撑函数。

10.5　投影、prox 与 Moreau 分解

闭凸集投影 $p = P_C(x)$ 的最优性条件是

$$\langle x - p, z - p \rangle \leq 0, \qquad \forall z \in C.$$

取两个点的投影并相加相应不等式，可得 firm nonexpansiveness：

$$\|P_C x - P_C y\|^2 \leq \langle P_C x - P_C y, x - y \rangle.$$

对 proper 闭凸 g，

$$\operatorname{prox}_{\lambda g}(x) = \arg\min_z \left\{ g(z) + \frac{1}{2\lambda}\|z - x\|^2 \right\}.$$

强凸二次项保证唯一解，最优性条件为

$$\frac{x - z}{\lambda} \in \partial g(z).$$

令 $u = (x - z)/\lambda$，利用共轭次梯度互逆，得到 Moreau 分解

$$x = \operatorname{prox}_{\lambda g}(x) + \lambda \operatorname{prox}_{g^*/\lambda}(x/\lambda).$$

例题 10.2. $g(x) = \|x\|_1$ 时，问题按坐标分离。标量最优性给

$$\operatorname{prox}_{\lambda|\cdot|}(v) = \operatorname{sgn}(v)(|v| - \lambda)_+,$$

即软阈值。$v = \pm\lambda$ 处虽然 $|x|$ 不可微，次梯度条件仍给唯一解零。

10.6　KKT 是何时充分、何时只是必要

考虑

$$\min_x f_0(x) \quad \text{s.t. } f_i(x) \le 0,\ i=1,\dots,m, \quad Ax=b.$$

Lagrangian

$$L(x,\lambda,\nu)=f_0(x)+\sum_i \lambda_i f_i(x)+\nu^\top(Ax-b).$$

KKT 条件包括原始可行、对偶可行 $\lambda\ge 0$、驻点和互补松弛。在凸问题且满足 Slater 条件时，强对偶成立，KKT 对最优性充要。非凸问题中，即便约束资格成立，KKT 通常只是局部极小的必要条件。

例题 10.3 (二次约束投影)**.**

$$\min_x \frac{1}{2}\|x-a\|^2 \quad \text{s.t. } \|x\|\le r.$$

若 $\|a\|\le r$，解为 a。否则约束活跃，驻点

$$x-a+\lambda x=0$$

给 $x=a/(1+\lambda)$。由 $\|x\|=r$ 得 $\lambda=\|a\|/r-1$，所以

$$x^*=r\frac{a}{\|a\|}.$$

该问题凸且有严格可行点，KKT 结论充分。

反例 10.1. 无约束函数 $f(x)=x^3$ 在 $x=0$ 满足 $\nabla f=0$，但不是局部极小。驻点条件从不等价于非凸最优性。

答题方法

凸分析题先判断集合与函数是否凸、是否闭、是否 proper；写对偶前检查 Slater 或其他资格条件；使用 KKT 时明确“必要”还是“充要”。次梯度、法锥与 prox 都应从同一个最优性包含式 $0\in\partial f(x)+N_C(x)$ 出发。

第十一章　确定性一阶、近端与二阶方法

学习目标

本章从下降引理推导梯度下降速率，解释加速、投影、镜像和近端更新的最优性条件，并把 Newton/拟 Newton 放在局部模型与全局化框架中比较。

11.1　梯度下降及其收敛率

GD 为 $x_{k+1}=x_k-\gamma\nabla f(x_k)$。考试必须同时写“目标类别、步长、度量和速率”。

定理 11.1 (三类 GD 结论)**.** *设 f 下有界且 L-smooth。*

(i) 若 $0<\gamma\le 1/L$，则 $\frac{1}{K+1}\sum_{k=0}^{K}\|\nabla f(x_k)\|^2\le 2(f(x_0)-f^*)/[\gamma(K+1)]$。

(ii) 若再假设凸，取 $\gamma=1/(2L)$，则 $f(x_K)-f^*\le 2L\|x_0-x^*\|^2/(K+1)$。

(iii) 若再假设 μ-强凸，取 $\gamma=2/(L+\mu)$，则 $\|x_K-x^*\|\le((L-\mu)/(L+\mu))^K\|x_0-x^*\|$。

完整证明. 下降引理和更新式给 $f(x_{k+1})\le f(x_k)-\gamma(1-L\gamma/2)\|\nabla f(x_k)\|^2$；$\gamma\le 1/L$ 时求和得 (i)。凸情形展开距离并用 $\|\nabla f(x_k)\|^2\le 2L(f(x_k)-f^*)$、$\langle\nabla f(x_k),x_k-x^*\rangle\ge f(x_k)-f^*$，在 $\gamma=1/(2L)$ 下得到 $f(x_k)-f^*\le 2L(\|x_k-x^*\|^2-\|x_{k+1}-x^*\|^2)$；函数值单调后求和得 (ii)。强凸情形将上一节联合不等式代入

$$\|x_{k+1}-x^*\|^2=\|x_k-x^*\|^2-2\gamma\langle\nabla f(x_k),x_k-x^*\rangle+\gamma^2\|\nabla f(x_k)\|^2,$$

取 $\gamma=2/(L+\mu)$ 后梯度平方项抵消，收缩因子为 $((L-\mu)/(L+\mu))^2$，迭代即得 (iii)。 □

另一常用版本取 $\gamma=1/L$，得 $f(x_k)-f^*\le L\|x_0-x^*\|^2/(2k)$。本书以 $1/(2L)$ 的证明为主模板；作答时必须让步长、常数和证明链彼此一致。

例题 11.1 (二维二次型的最优固定步长)**.** 令 $f(x)=\frac12 x^T\operatorname{diag}(1,9)x$，则 $\mu=1,L=9$。步长 $\gamma=2/(L+\mu)=0.2$，迭代矩阵 $I-\gamma Q=\operatorname{diag}(0.8,-0.8)$，距离每步至多乘 0.8；若错取 $\gamma=1$，第二坐标乘 -8 发散。该题同时检查 Hessian、步长和谱半径。

11.2　坐标、镜像与近端方法

近端梯度（proximal gradient, PG）是复合凸优化的主线算法。对 $\psi=f+h$，假设 f 凸、L-smooth，h proper 闭凸，最小值可取，

$$x_{k+1}=\operatorname{prox}_{th}(x_k-t\nabla f(x_k)),\qquad G_t(x)=\frac{1}{t}\big(x-\operatorname{prox}_{th}(x-t\nabla f(x))\big).$$

$G_t(x)=0$ 当且仅当 x 最优。

定理 11.2 (PG 的凸函数值率)**.** *若 $0<t\le 1/L$，则*

$$\psi(x_k)-\psi^\ast \le \frac{\|x_0-x^\ast\|^2}{2kt}.$$

完整证明. 令 $x^+=x-tG_t(x)$。由光滑上界、f,h 的凸性和 prox 最优性，对任意 $z\in\operatorname{dom}\psi$ 有

$$\psi(x^+)\le \psi(z)+\langle G_t(x),x-z\rangle-\frac{t}{2}\|G_t(x)\|^2.$$

取 $z=x^\ast$、$x=x_{i-1}$，并用 $x_i=x_{i-1}-tG_t(x_{i-1})$，右端等于 $[\|x_{i-1}-x^\ast\|^2-\|x_i-x^\ast\|^2]/(2t)$。求和得到平均界；同一不等式取 $z=x$ 可知 $\psi(x_i)$ 单调不增，因此最后迭代点不超过平均，结论成立。 □

Douglas–Rachford splitting 的一种写法为：

$$T(z)=z+\operatorname{prox}_{tf}(2\operatorname{prox}_{th}(z)-z)-\operatorname{prox}_{th}(z),\quad z_{k+1}=z_k+\rho_k(Tz_k-z_k).$$

其不动点经 prox_{th} 映射到 $\arg\min(f+h)$。这是教师讲授内容但三期卷未出现，掌握要求为“会写更新、了解不动点联系”。

坐标下降每步只更新一坐标，成本低但率依赖坐标光滑常数；mirror descent 以 Bregman 散度替代欧氏平方距离，更新 $x_{k+1}=\arg\min_{x\in C}\{\eta\langle g_k,x\rangle+D_\psi(x,x_k)\}$。除非题目给出对应几何和常数，不要求背尖锐率。

11.3 Newton、拟 Newton 与非凸边界

二阶法可从预条件视角统一写成：$x_{k+1}=x_k-\gamma P_k\nabla f(x_k)$；取 $P_k=[\nabla^2 f(x_k)]^{-1}$ 得 Newton。若 Hessian 正定且 Lipschitz、初值在解附近，Newton 具有局部二次收敛；非凸时 Hessian 可不定，必须阻尼或正则化。BFGS 用割线条件更新正定近似，Armijo/Wolfe 用函数值/曲率控制步长。

这些方法在三期正式卷中未被直接考查，掌握要求低于 GD、PG 和 SGD：会写更新、会说明局部收敛所需条件，不必背高级全局化定理。

11.3.1 线搜索与 trust region 的边界

线搜索沿方向 p_k 选择步长。Armijo 条件要求

$$f(x_k+\alpha p_k)\le f(x_k)+c_1\alpha\nabla f(x_k)^Tp_k,$$

Wolfe 条件再以曲率条件避免步长过小。trust region 则直接求局部二次模型

$$\min_{\|p\|\le\Delta_k} m_k(p)=f(x_k)+g_k^Tp+\tfrac12 p^TB_kp,$$

并用实际下降与预测下降之比 $\rho_k=[f(x_k)-f(x_k+p_k)]/[m_k(0)-m_k(p_k)]$ 决定接受步长及调整半径。它在 Hessian 不定时仍可通过子问题边界获得下降；考试主线是模型、比值和半径更新逻

辑，不要求高级子问题求解器。

答题方法

收敛率题先写函数类与步长，再找单步势函数下降。复合问题把不可微项放进 prox；约束问题把指示函数放进 prox。二阶方法必须同时说明 Hessian 可逆性、局部模型误差和线搜索/信赖域。

自测

1. 由下降引理证明凸光滑 GD 的 $O(1/k)$ 函数值率。
2. 证明强凸光滑情形的线性收敛，并写出条件数依赖。
3. 推导软阈值是 ℓ_1 的 proximal map，并由此写出 ISTA。
4. 说明 Nesterov 加速的势函数中为何需要外推点。
5. 比较 Newton、BFGS 与梯度下降的每步成本和局部保证。

本节结论

确定性优化的核心是“单步不等式 + 势函数 + 条件”。加速和近端不是经验修饰，而是改变势函数或精确处理非光滑结构。

11.4 梯度下降的两个核心速率

令 $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$。下降引理给

$$f(x_{k+1}) \le f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2.$$

若 f 凸，以 x^* 为最优点，则

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^*\|^2 - \frac{2}{L}\langle \nabla f(x_k), x_k - x^*\rangle + \frac{1}{L^2}\|\nabla f(x_k)\|^2 \\ &\le \|x_k - x^*\|^2 - \frac{2}{L}[f(x_k) - f^*] + \frac{1}{L^2}\|\nabla f(x_k)\|^2.\end{aligned}$$

再用下降量吸收最后一项，可得

$$f(x_{k+1}) - f^* \le \frac{L}{2}\big(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\big).$$

求和并利用函数值单调，得到

$$f(x_k) - f^* \le \frac{L\|x_0 - x^*\|^2}{2k}.$$

若 f 还 μ-强凸，则

$$\|\nabla f(x)\|^2 \ge 2\mu[f(x) - f^*].$$

代入下降式:

$$f(x_{k+1}) - f^* \le \left(1 - \frac{\mu}{L}\right)[f(x_k) - f^*].$$

因此迭代复杂度分别为 $O(LR^2/\varepsilon)$ 与 $O(\kappa \log(1/\varepsilon))$。

11.5 加速为何不是简单增加动量

Nesterov 方法的一种形式为

$$y_k = x_k + \beta_k(x_k - x_{k-1}), \qquad x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k).$$

对一般光滑凸函数，合适变化的 β_k 可给

$$f(x_k) - f^* = O\left(\frac{L\|x_0 - x^*\|^2}{k^2}\right).$$

对强凸函数，固定

$$\beta = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$$

给 $O((1 - 1/\sqrt{\kappa})^k)$ 量级。证明需要同时控制函数误差与一个辅助点的距离势能；仅检查每一步函数值下降并不足够，因为加速序列可能非单调。

易错点

“momentum”是一族递推，不等于任何参数都加速。步长或动量过大时，哪怕一维二次函数也会使特征根越出单位圆而发散。

11.6 近端梯度与 ISTA 的下降证书

对 $F = f + g$，其中 f 凸且 L-光滑，g proper 闭凸，定义

$$x^+ = \mathrm{prox}_{\alpha g}(x - \alpha \nabla f(x)).$$

它等价于最小化局部上界

$$g(z) + f(x) + \langle \nabla f(x), z - x\rangle + \frac{1}{2\alpha}\|z - x\|^2.$$

当 $\alpha \le 1/L$ 时，此二次模型上界真实 $F(z)$，由 $z = x$ 与 $z = x^+$ 比较得到

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

对 Lasso,

$$\min_x \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1,$$

$\nabla f = A^\top(Ax - b)$，可取 $\alpha \le 1/\|A\|_2^2$，再施加逐坐标软阈值。

11.7 投影梯度与约束最优性

闭凸集 C 上的投影梯度

$$x^+ = P_C(x - \alpha \nabla f(x))$$

满足投影最优性

$$\langle x - \alpha \nabla f(x) - x^*, z - x^+ \rangle \le 0, \quad z \in C.$$

在最优点，固定点条件

$$x^* = P_C(x^* - \alpha \nabla f(x^*))$$

等价于

$$\langle \nabla f(x^*), z - x^* \rangle \ge 0, \quad \forall z \in C.$$

因此“投影后梯度为零”不是正确说法；正确对象是 gradient mapping

$$G_\alpha(x) = \frac{1}{\alpha}\left[x - P_C(x - \alpha \nabla f(x))\right].$$

11.8 非凸光滑问题保证什么

若 f 仅下有界且 L-光滑，步长 $1/L$ 仍给

$$f(x_{k+1}) \le f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2.$$

从 $k = 0, \dots, K-1$ 求和:

$$\min_{0 \le k < K} \|\nabla f(x_k)\|^2 \le \frac{2L[f(x_0) - f_{\inf}]}{K}.$$

故得到一阶驻点的梯度范数保证，而不是函数值到全局最优的保证。要逃离严格鞍点通常需要随机扰动、负曲率步或更强几何假设。

11.9 Newton 法、阻尼与局部二次率

Newton 步解

$$\nabla^2 f(x_k) p_k = -\nabla f(x_k), \qquad x_{k+1} = x_k + p_k.$$

在最优点附近，若 Hessian Lipschitz 且 $\nabla^2 f(x^*)$ 非奇异，则

$$\|x_{k+1} - x^*\| \le C\|x_k - x^*\|^2.$$

远离最优点,Hessian 可能不正定,Newton 方向甚至不是下降方向。阻尼线搜索用 $x_{k+1} = x_k + \alpha_k p_k$，信赖域则限制 $\|p\| \le \Delta_k$ 并比较模型下降与真实下降；二者都是全局化机制，不改变局部 Newton 模型。

答题方法

优化算法题先写目标的凸性、光滑性、强凸性与约束；再选择势函数或下降证书；最后说清保证对象是函数值、距离、gradient mapping 还是梯度范数。不同保证不能用同一个“收敛”概括。

第十二章　随机优化与现代训练算法

学习目标

本章把 SGD 看成 Robbins–Monro 型随机递推，区分无偏性、方差、批量和步长的作用；随后分析 momentum、AdaGrad、Adam 的更新、偏差修正和可能失败的条件。

12.1　随机梯度、动量（momentum）与自适应方法

令过滤 $\mathcal{F}_k$ 包含迭代 k 前全部历史，标准随机梯度假设是

$$\mathbb{E}[g_k \mid \mathcal{F}_k] = \nabla f(x_k), \qquad \mathbb{E}[\|g_k - \nabla f(x_k)\|^2 \mid \mathcal{F}_k] \le \sigma^2.$$

“无偏”必须是条件无偏；只写无条件期望不足以消去一步递推中的交叉项。

定理 12.1 (非凸 SGD)**.** *若 f L-smooth、下有界，以上条件成立且 $0 < \gamma \le 1/L$，则*

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\nabla f(x_k)\|^2 \le \frac{2\Delta_0}{\gamma(K+1)} + \gamma L\sigma^2, \quad \Delta_0 = f(x_0) - f^*.$$

完整证明. 条件于 $\mathcal{F}_k$ 使用下降引理:

$$\mathbb{E}_k f(x_{k+1}) \le f(x_k) - \gamma\|\nabla f(x_k)\|^2 + \frac{L\gamma^2}{2}\mathbb{E}_k\|g_k\|^2.$$

方差分解给 $\mathbb{E}_k\|g_k\|^2 \le \|\nabla f(x_k)\|^2 + \sigma^2$；$\gamma L \le 1$ 后得到 $\mathbb{E}_k f(x_{k+1}) \le f(x_k) - \gamma\|\nabla f(x_k)\|^2/2 + L\gamma^2\sigma^2/2$。取全期望并从 0 到 K 求和，利用 $f(x_{K+1}) \ge f^*$ 即得。 □

凸光滑时，用 $\gamma \le 1/(2L)$ 可得平均函数值的 $O(1/K) + O(\gamma\sigma^2)$；强凸光滑时有

$$\mathbb{E}[f(x_K) - f^*] \le \Delta_0 e^{-\gamma\mu K} + \frac{L\gamma\sigma^2}{\mu}, \qquad \gamma \le 1/L.$$

第一项是优化偏差，第二项是常步长噪声平台；递减步长才可继续压低平台。

12.1.1　正式卷的 minibatch 与重要性抽样

对 $f = \frac{1}{n}\sum_i f_i$，以 $q_i > 0$ 抽取 s_t，

$$g(x) = \frac{1}{\tau}\sum_{t=1}^{\tau}\frac{1}{nq_{s_t}}\nabla f_{s_t}(x), \qquad \mathbb{E}g(x) = \nabla f(x).$$

独立性给

$$\mathbb{E}\|g(x)-g(y)\|^2 \le 2\left[\frac{1}{\tau}\max_i \frac{L_i}{nq_i} + (1-\tfrac{1}{\tau})L\right] D_f(x,y).$$

为最小化最大项，令 $q_i^* = L_i/\sum_j L_j$，此时 $\max_i L_i/(nq_i^*) = (\sum_i L_i)/n$；均匀抽样则为 $\max_i L_i$。方差在 i.i.d. minibatch 下按 $1/\tau$ 缩放。完整逐问推导见 节 13.2 Q3–Q9。

12.1.2 动量、NAG 与自适应预条件

Heavy-ball：$x_{k+1} = x_k - \gamma g_k + \beta(x_k - x_{k-1})$；NAG 先 look-ahead：$y_k = x_k + \beta_k(x_k - x_{k-1})$，再在 y_k 取梯度。

定理 12.2 (凸光滑 NAG)**.** *若 f 凸且 L-smooth，$x_{-1} = x_0$，$\gamma = 1/L$，取等价参数 $\theta_k = 2/(k+1)$ (对应动量 $\beta_k = (k-2)/(k+1)$)，则*

$$f(x_k) - f^* \le \frac{2L\|x_0 - x^*\|^2}{(k+1)^2}.$$

证明思路

定义辅助点 $v_k = x_{k-1} + (x_k - x_{k-1})/\theta_k$。光滑下降与凸性给

$$\frac{f(x_k)-f^*}{L\theta_k^2} + \frac{1}{2}\|v_k - x^*\|^2 \le \frac{f(x_{k-1})-f^*}{L\theta_{k-1}^2} + \frac{1}{2}\|v_{k-1}-x^*\|^2.$$

望远镜求和并代入 $\theta_k = 2/(k+1)$ 得结论。完整代数展开的关键是写出势函数和 $\theta_k^2 \ge (1-\theta_k)\theta_{k-1}^2$，不能只背 $O(1/k^2)$。

AdaGrad、RMSProp、Adam、AdamW 的常用更新为：

$$s_k = s_{k-1} + g_k \odot g_k \quad \text{(AdaGrad)}, \qquad s_k = \beta_2 s_{k-1} + (1-\beta_2) g_k \odot g_k,$$

$$m_k = \beta_1 m_{k-1} + (1-\beta_1) g_k, \qquad x_{k+1} = x_k - \gamma\, m_k \oslash (\sqrt{s_k} + \epsilon).$$

AdamW 另作 $-\gamma\lambda x_k$ 的解耦权重衰减。上述简式不自动导出一般非凸问题的通用收敛保证，因而不能声称“默认参数必收敛”。

12.2 方法比较、历年题与易错点

方法	假设/步长	度量与率	每步主要成本	掌握要求
GD 非凸	L-smooth，$\gamma \le 1/L$	平均 $\|\nabla f\|^2 = O(1/K)$	一次全梯度	会证明
GD 凸	convex+L-smooth，$1/(2L)$	$f - f^* = O(1/K)$	一次全梯度	会证明

GD 强凸	μ-SC+L-smooth, $2/(L+\mu)$	距离几何收缩	一次全梯度	会算收缩因子
PG	convex composite, $t \le 1/L$	$\psi - \psi^* = O(1/k)$	梯度 + 一次 prox	会证明、会算 prox
SGD	条件无偏、方差有界	偏差项 + 噪声项	一次随机梯度	会推期望递推
NAG	convex+L-smooth, $1/L$	$O(1/k^2)$	一次梯度 + 动量	会写势函数
DRS	两个 proper 闭凸函数	不动点收敛	两次 prox	了解
AdamW	随机梯度、经验参数	无默认通用率	向量二阶矩	会写更新

知识点	核心条件	历年题与正文位置	掌握要求
凸集、分离、Farkas	非空、闭性、凸性与线性系统可行性	三期 MLT/Opt 定义题；节 10.1	定义、会证明
闭凸、存在唯一	proper、下半连续、有界下水平集；唯一性需强凸	2025 秋 Opt Q2；2026 春 MLT Q2/Opt Q3	会证明
光滑强凸联合不等式	$0 < \mu \le L$，梯度 L-Lipschitz	2025 春 Opt Q2–Q6；2026 春 Opt Q8	会推导
次微分、prox、法锥	proper 闭凸；prox 缩放约定一致	2025 春 Opt Q3；2026 春 Opt Q2–Q6	会证明、会计算
Fenchel/Moreau	proper 闭凸；强对偶需正则性	三期卷未直接出现；节 10.2	定义、会比较
GD 三类率	L-光滑；凸/强凸结论需相应额外假设	多期定义/递推前置；节 11.1	会证明、会计算
PG/DRS	复合凸目标，$t \le 1/L$	2025 春、2025 秋、2026 春 Opt	会推导；DRS 了解
SGD 三类率	条件无偏、条件二阶矩/方差界	三期 Optimization 全部	会证明、会计算
小批量/重要性抽样	$q_i > 0$，重加权保证无偏	节 13.2 Q3–Q9	会推导
NAG/动量	凸光滑率需正确的参数序列	2026 春 DL Q1(c)；节 12.1	会写、会证明主率
预条件/自适应	预条件矩阵与目标几何相容	正式卷概念接口；节 11.3、12.1	会比较

易错点

高频错误：把无条件无偏当条件无偏；把常步长 SGD 的噪声平台漏掉；混用 $1/L$、$1/(2L)$ 与 $2/(L+\mu)$；未说明 prox 的缩放约定；把 AdamW 的解耦衰减等同于所有自适应方法中的 L_2 正则；把 Newton/BFGS 等低频扩展与 GD/PG/SGD 当作等权主线。三期 Optimization 完整题面与逐问解答分别见 节 13.1 to 13.3。

答题方法

随机递推题先写条件期望下的无偏/有偏结构，再展开平方距离或函数值；交叉项用条件期望消失，噪声项由方差界控制。自适应方法必须写清坐标缩放、偏差修正和步长下界/上界。

自测

1. 证明强凸目标上递减步长 SGD 的均方误差递推。
2. 计算 mini-batch 独立采样下梯度方差随批量的变化，并说明相关采样为何不同。
3. 从指数滑动平均推导 Adam 的一阶、二阶矩偏差修正。
4. 给出“训练损失下降”不能推出迭代点收敛的例子。

本节结论

随机优化用可控噪声换取廉价迭代。收敛结论必须同时声明目标几何、梯度估计、步长序列和输出方式。

12.3 SGD 的基本递推与步长权衡

设

$$x_{k+1} = x_k - \alpha_k g_k, \qquad \mathbb{E}[g_k \mid \mathcal{F}_k] = \nabla f(x_k), \qquad \mathbb{E}\|g_k - \nabla f(x_k)\|^2 \le \sigma^2.$$

若 f 为 μ-强凸且梯度 L-Lipschitz，则

$$\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k] = \|x_k - x^*\|^2 - 2\alpha_k \langle \nabla f(x_k), x_k - x^* \rangle + \alpha_k^2 \mathbb{E}[\|g_k\|^2 \mid \mathcal{F}_k].$$

用强凸与光滑界，可整理为

$$\mathbb{E}\|x_{k+1} - x^*\|^2 \le (1 - c\mu\alpha_k)\mathbb{E}\|x_k - x^*\|^2 + C\alpha_k^2\sigma^2$$

（当 α_k 足够小）。固定步长使误差下降到 $O(\alpha\sigma^2/\mu)$ 的噪声邻域；递减步长 $\alpha_k \asymp 1/(\mu k)$ 可给 $O(1/k)$ 均方量级。

Robbins–Monro 条件

$$\sum_k \alpha_k = \infty, \qquad \sum_k \alpha_k^2 < \infty$$

分别保证总移动量不会过早耗尽、平方噪声累积有限。典型 $\alpha_k = k^{-p}$ 要求 $1/2 < p \le 1$。

12.4 Mini-batch 与重要性抽样

若单样本梯度独立且方差为 σ^2，大小 B 的均值梯度方差约为 σ^2/B。但墙钟时间不一定线性改善，因为数据读取、同步与硬件饱和会限制收益。若有限和目标

$$f(x)=\frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

按概率 p_i 抽样，使用

$$g=\frac{\nabla f_i(x)}{n p_i}$$

保持无偏。其二阶矩

$$\mathbb{E}\|g\|^2=\frac{1}{n^2}\sum_{i}\frac{\|\nabla f_i(x)\|^2}{p_i}$$

在已知当前梯度范数时由 $p_i\propto\|\nabla f_i(x)\|$ 最小化。实际中常用 Lipschitz 常数或历史梯度作近似，必须保留权重校正。

12.5 AdaGrad、Adam 与偏差修正

坐标 AdaGrad 累积

$$v_{k,j}=\sum_{t=1}^{k} g_{t,j}^2,\qquad x_{k+1,j}=x_{k,j}-\frac{\alpha}{\sqrt{v_{k,j}}+\epsilon}g_{k,j}.$$

稀疏坐标若很少出现，其累计平方梯度小，得到较大有效步长。在线凸分析中的遗憾界自然含 $\sum_j\sqrt{\sum_t g_{t,j}^2}$。

Adam 使用指数滑动平均

$$m_k=\beta_1 m_{k-1}+(1-\beta_1)g_k,\qquad v_k=\beta_2 v_{k-1}+(1-\beta_2)g_k^{\odot 2}.$$

若初值为零，$\mathbb{E}m_k$ 含因子 $1-\beta_1^k$，$\mathbb{E}v_k$ 含因子 $1-\beta_2^k$，故偏差修正为

$$\widehat{m}_k=\frac{m_k}{1-\beta_1^k},\qquad \widehat{v}_k=\frac{v_k}{1-\beta_2^k}.$$

更新为

$$x_{k+1}=x_k-\alpha\frac{\widehat{m}_k}{\sqrt{\widehat{v}_k}+\epsilon}.$$

这不意味着任意目标上 Adam 都比 SGD 收敛；自适应分母的历史权重若造成有效步长不受控，在线凸反例中也可能不收敛。

12.6 方差缩减的控制变量思想

SVRG 类方法在快照 $\widetilde{x}$ 计算全梯度 $\nabla f(\widetilde{x})$，内循环使用

$$g_k=\nabla f_i(x_k)-\nabla f_i(\widetilde{x})+\nabla f(\widetilde{x}).$$

对均匀随机 i,

$$\mathbb{E}_i g_k = \nabla f(x_k),$$

且当 x_k 接近 $\widetilde{x}$ 时，两单样本梯度之差变小。控制变量在不引入偏差的同时降低方差。代价是周期性全梯度与快照存储，因此更适合有限和、可重复访问的数据。

12.7 随机非凸优化的保证对象

若 f 为 L-光滑、下有界，g_k 无偏且二阶矩有界，常数步长下可得到

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(x_k)\|^2 \le \frac{2[f(x_0)-f_{\inf}]}{\alpha K} + L\alpha\sigma^2.$$

右端第一项随 αK 减小，第二项随 α 增大；取 $\alpha \asymp K^{-1/2}$ 得平均梯度平方 $O(K^{-1/2})$。输出通常取均匀随机迭代点，这与最后迭代点全局最优是完全不同的陈述。

易错点

报告 SGD 结论必须列出：无偏或有偏；方差/二阶矩假设；目标的凸性等级；步长；输出方式；保证对象。少任何一项，都可能把只对平均迭代成立的界误写成最后一点的确定性结论。

自测

1. 解释固定步长 SGD 为什么通常只能到噪声邻域。
2. 重要性抽样时若不除以 np_i，梯度目标发生什么变化？
3. 从零初始化推导 Adam 两个偏差修正因子。
4. 验证 SVRG 梯度估计无偏，并说明方差何时降低。

第十三章　三期正式真题与完整解答

本章按年份收入 Optimization Methods in Artificial Intelligence 的全部正式题面与解答，共保留既有审计口径下的 25 道大题、28 个最细作答单元。除特别注明外，解答为复习用整理稿，不代表官方评分标准。每个题组前的“理论入口”用于回看本 Part 前文。

13.1　2025 春

理论入口： 参见章 十。

2025 春·Optimization·Q1–Q2（原卷第 4 页，11 分）

Original. Let f be L-smooth and μ-strongly convex. Q1: give both definitions. Q2: show $g(x)=f(x)-\frac{\mu}{2}\|x\|^2$ is convex and $(L-\mu)$-smooth, and using cocoercivity prove

$$\mu\|x-y\|^2+\frac{1}{L}\|\nabla f(x)-\nabla f(y)\|^2\le(1+\mu/L)\langle\nabla f(x)-\nabla f(y),x-y\rangle.$$

中文译述

给出光滑与强凸定义；对去除二次项后的函数应用凸光滑函数的 cocoercivity，推出联合不等式。

完整解答：Q1–Q2

定义见节 10.1。强凸等价于 $g=f-\mu\|\cdot\|^2/2$ 凸；若二阶可微，$\mu I\preceq\nabla^2 f\preceq LI$，所以 $0\preceq\nabla^2 g\preceq(L-\mu)I$。一般可微情形可由一阶强凸与光滑不等式相减得到。对 g 用 $\|\nabla g(x)-\nabla g(y)\|^2/(L-\mu)\le\langle\nabla g(x)-\nabla g(y),x-y\rangle$，代入 $\nabla g(x)-\nabla g(y)=\Delta\nabla f-\mu\Delta x$，展开并消去 $\mu^2\|\Delta x\|^2$，整理为

$$\frac{\mu L}{\mu+L}\|\Delta x\|^2+\frac{1}{\mu+L}\|\Delta\nabla f\|^2\le\langle\Delta\nabla f,\Delta x\rangle,$$

两边乘 $(\mu+L)/L$ 即为原式。

正文对应： 节 10.1中的光滑强凸联合不等式。

理论入口： 参见章 十。

2025 春·Optimization·Q3–Q4（原卷第 4 页，12 分）

Original. For $x_{k+1}=\mathrm{prox}_{\gamma R}(x_k-\gamma(\nabla f(x_k)+\xi_k))$, derive by prox nonexpansiveness and the optimality condition

$$\|x_{k+1}-x^*\|^2\le\|x_k-x^*-\gamma(\nabla f(x_k)-\nabla f(x^*)+\xi_k)\|^2,$$

then use variance decomposition to prove

$$\mathbb{E}_k\|x_{k+1}-x^*\|^2\le\|x_k-x^*-\gamma(\nabla f(x_k)-\nabla f(x^*))\|^2+\gamma^2\sigma^2.$$

中文译述

先利用最优点的 prox 固定点和非扩张性得到一步距离界，再条件期望并分解随机噪声。

完整解答：Q3–Q4

最优性 $0\in\nabla f(x^*)+\partial R(x^*)$ 等价于 $x^*=\mathrm{prox}_{\gamma R}(x^*-\gamma\nabla f(x^*))$。将它与迭代式代入 prox 非扩张性即得 Q3。令 $a=x_k-x^*-\gamma(\nabla f(x_k)-\nabla f(x^*))$，右端为 $\|a-\gamma\xi_k\|^2$。条件于历史，$\mathbb{E}_k\xi_k=0$，故交叉项消失：$\mathbb{E}_k\|a-\gamma\xi_k\|^2=\|a\|^2+\gamma^2\mathbb{E}_k\|\xi_k\|^2\le\|a\|^2+\gamma^2\sigma^2$。

关键条件：噪声必须条件零均值；R proper 闭凸确保 prox 单值。

正文对应：节 11.2中的 prox；2026 春 Opt Q2–Q7。

理论入口：参见章 十。

2025 春·Optimization·Q5–Q6（原卷第 4 页，10 分）

Original. With $\gamma=2/(\mu+L)$ prove

$$\mathbb{E}\|x_{k+1}-x^*\|^2\le(1-\rho)\mathbb{E}\|x_k-x^*\|^2+\gamma^2\sigma^2,\quad \rho=\frac{4\mu L}{(\mu+L)^2},$$

and prove that $k\ge(L/\mu+3)\log(1/\epsilon)/4$ implies $\mathbb{E}\|x_k-x^*\|^2\le\epsilon\|x_0-x^*\|^2+\gamma\sigma^2/\mu$.

中文译述

选择联合不等式对应的最优固定步长，得到带噪声平台的线性递推，并转化为迭代复杂度。

完整解答：Q5–Q6

展开 Q4 的确定性平方，代入联合不等式。$\gamma=2/(L+\mu)$ 时梯度差平方系数抵消，留下 $1-4\mu L/(L+\mu)^2=1-\rho$。取全期望并迭代：

$$a_k\le(1-\rho)^k a_0+\gamma^2\sigma^2\sum_{j=0}^{k-1}(1-\rho)^j\le e^{-\rho k}a_0+\frac{\gamma^2\sigma^2}{\rho}.$$

又 $1/\rho = (L+\mu)^2/(4\mu L) \le (L/\mu+3)/4$，故题给 k 使 $e^{-\rho k} \le \epsilon$。并且 $\rho \ge \gamma\mu$，所以 $\gamma^2/\rho \le \gamma/\mu$，得到题中较松的噪声平台。

易错点： 把 $1-\rho$ 误写为 ρ；漏掉几何级数；把噪声平台写成随 k 消失。

正文对应： 节 11.1、12.1中的强凸梯度收缩与 SGD 噪声平台。

13.2 2025 秋

理论入口： 参见章 十。

2025 秋·Optimization·Q1–Q9（原卷第 6–8 页，共 33 分）

Original setting. Minimize $F(x) = f(x) + R(x)$, $f = n^{-1}\sum_i f_i$, where f_i is convex and L_i-smooth, f is L-smooth and μ-strongly convex, and R is proper closed convex. Draw s_t i.i.d. with $\Pr(s=i) = q_i > 0$ and define

$$g(x) = \frac{1}{\tau}\sum_{t=1}^{\tau} \frac{1}{nq_{s_t}} \nabla f_{s_t}(x), \qquad x_{k+1} = \mathrm{prox}_{\gamma R}(x_k - \gamma g^k),$$

$D_f(x,y) = f(x) - f(y) - \langle \nabla f(y), x-y\rangle$. Q1 defines smoothness/strong convexity; Q2 proves a unique minimizer; Q3 proves unbiasedness; Q4 proves

$$\mathbb{E}\|g(x) - g(y)\|^2 \le 2A''(\tau,q) D_f(x,y), \quad A'' = \frac{1}{\tau}\max_i \frac{L_i}{nq_i} + (1-\tfrac{1}{\tau})L;$$

Q5 evaluates limits and monotonicity; Q6 designs optimal q; Q7 derives variance at x^* and $1/\tau$ scaling; Q8 specializes the AC constants; Q9 gives the stepsize and explicit convergence/noise-floor bound.

中文译述

该 section 从复合有限和问题出发，要求逐步建立：定义与唯一性、multisampling 无偏性、expected smoothness、batch 插值、重要性抽样、最优点方差、AC 不等式以及线性收敛加噪声平台。

完整解答：Q1–Q3

定义见节 10.1。f 强凸且有限连续，$F = f + R$ proper 闭强凸并强制增长，故最小点存在且唯一。对单样本 $Y = (nq_s)^{-1}\nabla f_s(x)$，$\mathbb{E}Y = \sum_i q_i (nq_i)^{-1}\nabla f_i = \nabla f$；batch 平均仍无偏。

完整解答：Q4

令 $Z_t = (nq_{s_t})^{-1}(\nabla f_{s_t}(x) - \nabla f_{s_t}(y))$。独立同分布平均的二阶矩分解为

$$\mathbb{E}\left\|\frac{1}{\tau}\sum_t Z_t\right\|^2 = \frac{1}{\tau}\mathbb{E}\|Z\|^2 + (1-\tfrac{1}{\tau})\|\mathbb{E}Z\|^2.$$

由每个 f_i 的 cocoercivity，

$$\mathbb{E}\|Z\|^2 = \sum_i \frac{\|\nabla f_i(x) - \nabla f_i(y)\|^2}{n^2 q_i} \le 2 \max_i \frac{L_i}{n q_i} D_f(x, y),$$

而 $\|\nabla f(x) - \nabla f(y)\|^2 \le 2L D_f(x, y)$，合并即得。

完整解答：Q5–Q6

$\tau = 1$ 时 $A'' = \max_i L_i/(nq_i)$，对应单样本 SGD；$\tau \to \infty$ 时 $A'' \to L$，对应 GD。写 $M = \max_i L_i/(nq_i) \ge n^{-1}\sum_i L_i \ge L$，则 $A'' = L + (M - L)/\tau$ 随 τ 不增。最小化 M 需使 L_i/q_i 等高：$q_i^* = L_i/\sum_j L_j$，最优值 $(\sum_i L_i)/n$；均匀抽样为 $\max_i L_i$。

完整解答：Q7–Q8

令 $Y = (nq_s)^{-1}\nabla f_s(x^*)$。独立平均给

$$\mathrm{Var}[g(x^*)] = \frac{1}{\tau}\mathrm{Var}(Y) = \frac{1}{\tau}\left[\sum_i \frac{\|\nabla f_i(x^*)\|^2}{n^2 q_i} - \|\nabla f(x^*)\|^2\right].$$

注意复合最优点一般 $\nabla f(x^*) \ne 0$。Q4 中 $C'' = 0$，原题给出的 implication 因此得到 $A = 2A''$、$C = 2\,\mathrm{Var}[g(x^*)]$。

完整解答：Q9

题给定理要求 $0 < \gamma \le 1/A = 1/(2A'')$，所以

$$\mathbb{E}\|x_k - x^*\|^2 \le (1 - \gamma\mu)^k \|x_0 - x^*\|^2 + \frac{2\gamma}{\mu\tau}\left[\sum_i \frac{\|\nabla f_i(x^*)\|^2}{n^2 q_i} - \|\nabla f(x^*)\|^2\right].$$

第一项线性收敛，第二项为 batch、sampling 和常步长共同决定的噪声平台。

易错点：忘记 cross terms 产生 $(1-1/\tau)L$；把 $q_i \propto 1/L_i$；在复合最优点错误设 $\nabla f(x^*) = 0$。

正文对应：节 12.1中的 minibatch、重要性抽样与噪声平台。

13.3 2026 春

理论入口：参见章 十。

2026 春·Optimization·Q1–Q10（原卷第 6–8 页，共 33 分）

Original setting. Minimize $F = f + R$ with f differentiable, L-smooth and μ-strongly convex, R proper closed convex, $g(x, \xi) = \nabla f(x) + \xi$, $\mathbb{E}\xi = 0$, $\mathbb{E}\|\xi\|^2 \le \sigma^2$, and $x_{k+1} = \mathrm{prox}_{\gamma R}(x_k - \gamma g_k)$. Q1 asks definitions; Q2 prox optimality; Q3 uniqueness and fixed point; Q4 monotonicity of ∂R; Q5 prox nonexpansiveness; Q6 one-step inequality; Q7 conditional expectation and variance decomposition; Q8 contraction for $\gamma = 2/(L + \mu)$ and the noise floor; Q9 proves averaging uncorrelated errors reduces variance by $1/k$; Q10, with $R = 0$, derives

$$\mathbb{E}_t\|x^{t+1} - x^*\|^2 \le \|x^t - x^*\|^2 - 2\gamma_t(1 - L\gamma_t)\mathbb{E}_t[f(x^t) - f^*] + \gamma_t^2\sigma^2,$$

then for $\gamma_t \le 1/(2L)$ the simplified bound and, choosing $\gamma_t = 1/(\mu t)$, the averaged rate $O(\log k/k)$.

中文译述

该 section 从 prox 次微分基础开始，逐步证明 composite SGD 一步递推、强凸线性收缩、平均降方差与递减步长平均率。

完整解答：Q1–Q5

定义见节 10.1。F proper 闭且 μ-强凸，故存在唯一最优点。prox 最优性为 $(u - x^+)/\gamma \in \partial R(x^+)$；与 $0 \in \nabla f(x^*) + \partial R(x^*)$ 合并得到固定点。若 $r_x \in \partial R(x)$、$r_y \in \partial R(y)$，两条次梯度不等式相加得 $\langle r_x - r_y, x - y\rangle \ge 0$；将 prox 的两个最优性条件代入即得 firm nonexpansiveness 和非扩张。

完整解答：Q6–Q8

由最优点固定点和非扩张，$\|x_{k+1} - x^*\|^2 \le \|x_k - x^* - \gamma(g_k - \nabla f(x^*))\|^2$。条件于 x_k，交叉噪声项因 $\mathbb{E}_k \xi_k = 0$ 消失，增加 $\gamma^2\sigma^2$。再用联合不等式并取 $\gamma = 2/(L+\mu)$，得到

$$\mathbb{E}\|x_{k+1} - x^*\|^2 \le (1-\rho)\mathbb{E}\|x_k - x^*\|^2 + \gamma^2\sigma^2, \quad \rho = 4\mu L/(L+\mu)^2.$$

稳态上界 $\gamma^2\sigma^2/\rho$ 是常步长噪声平台。

完整解答：Q9

$\bar{e}_k = k^{-1}\sum_{t=1}^k e_t$，故

$$\mathbb{E}\|\bar{e}_k\|^2 = k^{-2}\sum_t \mathbb{E}\|e_t\|^2 + 2k^{-2}\sum_{s<t}\mathbb{E}\langle e_s, e_t\rangle \le V/k.$$

无关/不相关假设是消去交叉项的关键；真实 SGD 误差通常相关，因此该结论是题设简化模型。

完整解答：Q10

对 $R = 0$ 展开距离，条件期望后得到 $a_{t+1} \le a_t - 2\gamma_t\langle \nabla f(x^t), x^t - x^*\rangle + \gamma_t^2(\|\nabla f(x^t)\|^2 + \sigma^2)$。凸性与 $\|\nabla f(x)\|^2 \le 2L(f(x) - f^*)$ 给题中系数 $-2\gamma_t(1 - L\gamma_t)$；若 $\gamma_t \le 1/(2L)$，它不大于 $-\gamma_t(f - f^*)$。再用强凸递推控制 $a_t = O(1/t)$，取 $\gamma_t = 1/(\mu t)$ 并从满足步长条件的 $t_0 = \lceil 2L/\mu\rceil$ 起求和，得到 $\sum_{t=t_0}^k \mathbb{E}[f(x^t) - f^*] = O(\log k)$；Jensen 给 $\mathbb{E}[f(\bar{x}_k) - f^*] = O(\log k/k)$。原卷直接从 $t = 1$ 同时要求 $1/(\mu t) \le 1/(2L)$ 一般不成立，严格证明需 burn-in 或改用 $\gamma_t = 1/(\mu(t + t_0))$；保留这一条件修正。

易错点： 忽略条件期望；把真实相关误差当不相关；不检查 $1/(\mu t)$ 与 $1/(2L)$ 的兼容性。

正文对应： 节 10.1、11.2、12.1中的次梯度、prox 不动点与随机递推。

13.4 高频错误、作答模板与复习检查

严格凸不等于强凸；L-smooth 与 Hessian 上界需说明可微性；GD 的 $O(1/k)$、线性率和非凸驻点率属于不同假设；SGD 必须写条件期望和噪声模型；prox 的参数缩放要统一；Slater 是强

对偶/KKT 的资格条件而非任意约束问题的自动事实；Newton 二次收敛只在局部成立。

复习检查

考场模板：先列假设和步长，再写一步下降或距离递推；随机项先条件化；对偶题写 Lagrangian、dual function、可行域、Slater/KKT。完成后应能证明下降引理、凸 GD 率、强凸收缩、prox 非扩张，并推导三期卷中的 minibatch、importance sampling 和噪声平台。

第四部分

Natural Language Processing

自然语言处理

第十四章　统计语言模型、表示与序列标注

学习目标

本章从计数概率模型进入 NLP：n-gram、平滑、perplexity、词表示、HMM、前向后向、Viterbi 和 CRF。要求读者能区分生成式归一化与条件式归一化，并手算动态规划。

14.1　稀疏文本表示与 Naive Bayes

词袋（BoW）把文档表示为词频向量，忽略词序。常见 TF–IDF 约定为

$$\mathrm{tfidf}(t,d)=\mathrm{tf}(t,d)\,\log\frac{N}{\mathrm{df}(t)},$$

其中 N 是文档数，$\mathrm{df}(t)$ 是包含词 t 的文档数。实践中 TF 可用原始次数、$1+\log c$ 或频率归一化，IDF 也可平滑为 $\log((N+1)/(\mathrm{df}+1))+1$；答题必须声明约定。

例题 14.1 (TF–IDF 手算)**.** 语料有 $N=10$ 篇文档，词 t 出现在其中 2 篇，在目标文档出现 3 次。按原始 TF 和未平滑自然对数 IDF，$\mathrm{tfidf}=3\log(10/2)=3\log 5$。若另一答案使用对数 TF 或平滑 IDF，数值不同不一定错误，关键是公式和约定一致。

多项式 Naive Bayes 假设给定类别后 token 条件独立：

$$\hat{y}=\arg\max_{y}\left[\log p(y)+\sum_{t\in\mathcal{V}}c_t(d)\log p(t\mid y)\right],\qquad \hat{p}(t\mid y)=\frac{C_{t,y}+\alpha}{\sum_{v}C_{v,y}+\alpha|\mathcal{V}|}.$$

这是对词计数的生成模型；实际计算应在对数域，避免许多小概率相乘下溢。条件独立是假设，不是从文本事实推出的结论。

例题 14.2 (Naive Bayes 分类)**.** 两类先验均为 $1/2$，词表为 {good,bad}。正类计数 $(3,1)$，负类计数 $(1,3)$，取加一平滑。文档为“good good”时，未归一化后验分数为

$$\text{正类}:\frac{1}{2}(4/6)^2=\frac{2}{9},\qquad \text{负类}:\frac{1}{2}(2/6)^2=\frac{1}{18},$$

故预测正类。若漏掉分母中的 $\alpha|\mathcal{V}|$，平滑概率不再归一化。

14.2 统计语言模型与评测

语言模型给序列概率 $p(w_{1:T}) = \prod_{t=1}^{T} p(w_t \mid w_{<t})$。$n$-gram 用 Markov 假设近似 $p(w_t \mid w_{<t}) \approx p(w_t \mid w_{t-n+1:t-1})$。最大似然估计为

$$\widehat{p}(w \mid h) = \frac{C(h, w)}{C(h)}.$$

未见组合概率为零，因此需要平滑。加 α 平滑为

$$p_\alpha(w \mid h) = \frac{C(h, w) + \alpha}{C(h) + \alpha|\mathcal{V}|},$$

易算但会给大量未见词过多质量；更成熟方法使用 backoff/interpolation 与折扣。

定义 14.1 (交叉熵与困惑度)**.** 测试序列长度为 T，平均负对数似然 $H = -T^{-1} \sum_t \log p(w_t \mid w_{<t})$；若用自然对数，perplexity 为

$$\mathrm{PPL} = \exp(H) = p(w_{1:T})^{-1/T}.$$

只有在相同 tokenization、词表与测试集上，PPL 才可直接比较。更小 PPL 表示模型给真实序列更高平均概率，不自动等价于事实性或任务质量更高。

例题 14.3 (trigram 与平滑)**.** 若上下文 (a, b) 出现 10 次，其中后接 c 三次，词表大小 100，则 MLE 为 0.3；加一平滑为 $(3+1)/(10+100) = 4/110$。平滑显著降低已见事件概率，显示加一法在大词表中过强。

例题 14.4 (语言模型概率、交叉熵与困惑度)**.** 若两 token 测试序列的条件概率依次为 1/2、1/4，则序列概率为 1/8，平均 NLL 为 $H = -\frac{1}{2}(\log \frac{1}{2} + \log \frac{1}{4}) = \frac{3}{2} \log 2$，故 $\mathrm{PPL} = e^H = 2^{3/2} = \sqrt{8}$。不要把序列概率 1/8 直接取倒数得到 8；PPL 还要取每 token 几何平均。

14.3 分词与词表示

词级模型有 OOV，字符级序列过长，subword 在二者间折中。BPE 从字符/字节开始，反复合并最高频相邻 token 对；WordPiece 以似然改进准则选择合并。算法依赖训练语料与预处理，同一文本在不同 tokenizer 下的 token 数不可直接比较。考试主要掌握 BPE/WordPiece 的目标与更新；Unigram LM 的动态规划实现细节只需了解。

one-hot 无法表达相似性。分布式表示基于“上下文相似的词语义相近”。Skip-gram with negative sampling 对中心词 w、上下文 c 最大化

$$\log \sigma(u_c^\top v_w) + \sum_{j=1}^{K} \mathbb{E}_{n_j \sim P_n} \log \sigma(-u_{n_j}^\top v_w).$$

正对提高内积，负样本降低内积；负采样分布常对频率作 3/4 次幂平滑。CBOW 从上下文预测中心词。GloVe 拟合全局共现对数：

$$\sum_{i,j} f(X_{ij})(u_i^\top v_j + b_i + \widetilde{b}_j - \log X_{ij})^2.$$

Word2vec 强调局部预测，GloVe 显式使用全局共现；二者都可能编码语料偏见。

点互信息

$$\mathrm{PMI}(w,c)=\log\frac{p(w,c)}{p(w)p(c)}$$

衡量共现超出独立基线的程度。理想化大样本和独立参数条件下，带 K 个负样本的 SGNS 最优内积满足 $u_c^\top v_w\approx\mathrm{PMI}(w,c)-\log K$；这是目标的矩阵分解解释，不表示有限维训练一定精确等于 PMI。

14.4 HMM、前向后向（forward–backward）与 Viterbi

HMM 假设隐藏状态一阶 Markov 且观测在给定当前状态后条件独立：

$$p(z_{1:T},x_{1:T})=p(z_1)p(x_1\mid z_1)\prod_{t=2}^{T}p(z_t\mid z_{t-1})p(x_t\mid z_t).$$

前向量 $\alpha_t(j)=p(x_{1:t},z_t=j)$ 递推

$$\alpha_t(j)=p(x_t\mid z_t=j)\sum_i\alpha_{t-1}(i)p(z_t=j\mid z_{t-1}=i).$$

总似然为 $\sum_j\alpha_T(j)$。Viterbi 把求和换成最大并保存回溯指针，求 MAP 状态序列。前向算法计算“所有路径概率之和”，Viterbi 计算“概率最大的一条路径”，二者不能互换。

例题 14.5 (两状态 Viterbi 一步)**.** 前一时刻最优分数为 $(0.3,0.2)$，转移矩阵 $A=\left(\begin{smallmatrix}0.8&0.2\\0.4&0.6\end{smallmatrix}\right)$，当前观测在两状态的发射概率为 $(0.5,0.1)$。新分数为 $\delta_t(1)=0.5\max(0.3\cdot0.8,0.2\cdot0.4)=0.12$，$\delta_t(2)=0.1\max(0.3\cdot0.2,0.2\cdot0.6)=0.012$，第二状态的回溯来自前一第二状态。

14.5 CRF 与判别式序列标注（sequence labeling）

线性链 CRF 直接建模条件分布

$$p_\theta(y\mid x)=\frac{\exp s_\theta(x,y)}{Z_\theta(x)},\qquad Z_\theta(x)=\sum_{y'}\exp s_\theta(x,y').$$

若得分按相邻标签分解，配分函数由 log-sum-exp 前向算法在 $O(TK^2)$ 计算，MAP 解码由 Viterbi 完成。负对数似然梯度是“模型期望特征减经验特征”。CRF 不需像 HMM 那样建模 $p(x)$，可使用丰富输入特征；代价是规范化与解码。

命题 14.1 (CRF 梯度结构)**.** *若 $s_\theta(x,y)=\theta^\top F(x,y)$，单样本负对数似然 $\ell=-\theta^\top F(x,y)+\log Z_\theta(x)$ 的梯度为*

$$\nabla_\theta\ell=-F(x,y)+\mathbb{E}_{Y\sim p_\theta(\cdot\mid x)}F(x,Y).$$

完整证明. 第一项直接求导。对第二项，

$$\nabla\log Z=\frac{1}{Z}\sum_{y'}\exp(\theta^\top F(x,y'))F(x,y')=\sum_{y'}p_\theta(y'\mid x)F(x,y').$$

有限状态时交换求和与求导合法；大结构空间由动态规划（dynamic programming）计算边缘期望而非枚举。 □

14.6 神经语言模型、RNN 与注意力

神经 LM 把离散 token 映射为 embedding，经 RNN/LSTM/Transformer 得隐藏状态，再用 softmax 预测下一个 token。RNN 参数跨时间共享但难并行；Transformer 训练并行且长距离路径短，但全注意力为 $O(T^2)$。结构公式见第七章。

encoder-only（仅编码器）模型双向看上下文，适合理解；decoder-only（仅解码器）模型以 causal language modeling 因果预测，适合开放生成；encoder–decoder 用交叉注意力连接输入与输出，适合翻译/摘要等条件生成。分类不是绝对用途限制，而是预训练掩码与信息流的主要区别。

答题方法

序列题先画状态/观测依赖，再决定求总概率、后验边缘还是最优路径；三者分别对应 forward、forward–backward 和 Viterbi。评价语言模型用 token 级负对数似然，比较时统一词表与分词。

自测

1. 推导加一平滑为何改变条件概率的归一化，并说明其偏差。
2. 手算一个两状态 HMM 的 forward、backward 和 Viterbi 表。
3. 解释 CRF 的全局归一化如何避免局部归一化的 label bias。
4. 从 skip-gram softmax 写到 negative sampling，并说明后者不是精确似然。

本节结论

统计 NLP 的统一主线是“表示—概率分解—动态规划—规范化—评价”。这些概念仍是理解神经序列模型和预训练目标的基础。

14.7 n-gram 的计数、平滑与困惑度

链式法则给

$$p(w_{1:T}) = \prod_{t=1}^{T} p(w_t \mid w_{<t}).$$

n-gram 假设把历史截断为最近 $n-1$ 个词。最大似然估计为

$$\widehat{p}(w \mid h) = \frac{C(h, w)}{C(h)}.$$

未见组合得到零概率，会使整个句子对数似然为负无穷。Add-α 平滑为

$$\widehat{p}_\alpha(w \mid h) = \frac{C(h, w) + \alpha}{C(h) + \alpha |V|},$$

但大词表下它可能把过多质量分给未见词。插值法

$$p(w \mid h) = \lambda_h p_{\mathrm{MLE}}(w \mid h) + (1-\lambda_h) p(w \mid h')$$

通过低阶模型回退，且 λ_h 可随上下文计数调整。

测试集 token 数为 N 时，交叉熵与困惑度为

$$H = -\frac{1}{N}\sum_{t=1}^{N} \log p(w_t \mid w_{<t}), \qquad \mathrm{PPL} = e^{H}.$$

若使用以 2 为底的对数，则 $\mathrm{PPL} = 2^{H_2}$。不同 tokenization 下的 token 数与事件空间不同，困惑度不能直接横向比较。

14.8 HMM 的三种动态规划

HMM 参数为初始分布 π_i、转移 a_{ij} 与发射 $b_j(o_t)$。前向量

$$\alpha_t(j) = p(o_{1:t}, z_t = j)$$

递推为

$$\alpha_1(j) = \pi_j b_j(o_1), \qquad \alpha_t(j) = b_j(o_t)\sum_i \alpha_{t-1}(i) a_{ij}.$$

似然 $p(o_{1:T}) = \sum_j \alpha_T(j)$。后向量

$$\beta_t(i) = p(o_{t+1:T} \mid z_t = i)$$

满足

$$\beta_T(i) = 1, \qquad \beta_t(i) = \sum_j a_{ij} b_j(o_{t+1}) \beta_{t+1}(j).$$

于是后验

$$p(z_t = i \mid o_{1:T}) = \frac{\alpha_t(i)\beta_t(i)}{p(o_{1:T})}.$$

Viterbi 把“求和”替换为“最大”：

$$\delta_t(j) = b_j(o_t) \max_i \delta_{t-1}(i) a_{ij},$$

并保存取得最大值的前驱。前向算法求所有路径概率之和，Viterbi 求单条最可能路径；两者不可互换。

例题 14.6. 若状态数 K、序列长 T，三种算法均为 $O(TK^2)$。采用对数域时，Viterbi 的乘法变加法；前向算法则需使用 log-sum-exp，不能也替换成 max，除非只做近似。

14.9 CRF 的归一化与梯度

线性链 CRF 条件分布写为

$$p_\theta(y \mid x) = \frac{\exp\{\theta^\top F(x,y)\}}{Z_\theta(x)}, \qquad Z_\theta(x) = \sum_{y'} \exp\{\theta^\top F(x,y')\}.$$

单样本对数似然梯度为

$$\nabla_\theta \log p_\theta(y \mid x) = F(x,y) - \mathbb{E}_{Y \sim p_\theta(\cdot \mid x)} F(x,Y).$$

经验特征计数与模型期望计数的差由前向–后向算法计算。HMM 建模联合分布 $p(x,y)$，需要发射模型；CRF 直接建模 $p(y \mid x)$，可容纳重叠、非独立输入特征，但训练需要全局归一化。

14.10 Word2vec 与负采样目标

Skip-gram softmax 建模

$$p(c \mid w) = \frac{\exp(u_c^\top v_w)}{\sum_{c'} \exp(u_{c'}^\top v_w)}.$$

词表很大时分母昂贵。负采样对观测对 (w,c) 和噪声词 $\tilde{c}_j \sim q$ 最大化

$$\log \sigma(u_c^\top v_w) + \sum_{j=1}^{K} \log \sigma(-u_{\tilde{c}_j}^\top v_w).$$

这是一个真实/噪声二分类目标，不等于完整 softmax 似然。对正样本分数 $s = u_c^\top v_w$，导数为

$$\frac{\partial}{\partial s} \log \sigma(s) = 1 - \sigma(s),$$

推动正对内积上升；负样本项导数为 $-\sigma(s)$，推动其内积下降。

答题方法

序列模型题先确定是在“对所有路径求和”“求单条最优路径”还是“条件归一化训练”。对应运算分别是 sum-product、max-product 和带配分函数的条件似然。

自测

1. 为什么未见 n-gram 会使句子 MLE 概率为零？Add-α 如何修复？
2. 写出 HMM 前向、后向、Viterbi 的状态量与递推。
3. 推导 CRF 对数似然梯度的“观测减期望”形式。
4. 负采样目标为什么不能冒充完整 softmax 的精确计算？

第十五章　Transformer：注意力、位置、掩码与缓存

学习目标

Transformer 只在本章完整讲解。读者应能从张量形状推导 self/cross-attention、多头拼接、掩码、残差与归一化，解释位置编码、复杂度和 KV cache，并把模型结构映射到训练与推理。

15.1　注意力与 Transformer

缩放点积注意力为

$$\mathrm{Attn}(Q,K,V)=\mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}+M\right)V. \tag{15.1}$$

M 可编码 padding 或因果掩码。除以 $\sqrt{d_k}$ 的动机是：若查询、键各分量独立、零均值、方差约 1，其点积方差约为 d_k；缩放使 logits 维持 $O(1)$，避免 softmax 过早饱和。

多头注意力（multi-head attention）令 $\mathrm{head}_r=\mathrm{Attn}(QW_r^Q,KW_r^K,VW_r^V)$，拼接后乘 W^O。每层通常还含前馈网络 $\mathrm{FFN}(x)=W_2\phi(W_1x+b_1)+b_2$、残差连接和 LayerNorm。Pre-LN 把归一化放在子层前，深层训练往往更稳；Post-LN 是原始 Transformer 的常见形式。

没有递归或卷积时，模型还需显式位置。原始正弦位置编码为

$$\mathrm{PE}(p,2i)=\sin\left(p/10000^{2i/d}\right),\qquad \mathrm{PE}(p,2i+1)=\cos\left(p/10000^{2i/d}\right).$$

可学习绝对位置、相对位置偏置和旋转位置编码是替代方案；它们改变外推和相对距离建模方式，但都不能在完全没有位置信息的置换等变自注意力中凭空恢复词序。

在 encoder–decoder 序列到序列模型中，encoder 产生 $H\in\mathbb{R}^{n\times d}$；decoder 的 masked self-attention 只能看已生成前缀，cross-attention 则取 Q 自 decoder、$K=HW^K,V=HW^V$ 自 encoder。训练常用 teacher forcing 的条件 NLL，推断时自回归解码；二者输入分布不同会产生 exposure bias。

命题 15.1 (因果自注意力的不可见性)**.** *若式(15.1)中 $M_{ij}=-\infty$ 对所有 $j>i$，其余为 0，则第 i 个输出不依赖任何未来值 v_j $(j>i)$。*

完整证明. 第 i 行 softmax 中，所有 $j>i$ 的 logit 加上 $-\infty$ 后指数权重为 0；归一化只在 $j\le i$ 上进行，因此输出 $o_i=\sum_{j\le i}a_{ij}v_j$。该结论要求掩码在每一层都正确应用；若输入预处理已泄露未来信息，矩阵掩码本身不能挽救因果性。 □

长度 n、宽度 d 时，全注意力的主要时间/存储为 $O(n^2d)/O(n^2)$，FFN 为 $O(nd^2)$。当 $n \gg d$，注意力矩阵是长上下文瓶颈。局部/滑窗注意力降到 $O(nwd)$，低秩或核化注意力近似全局交互，但可能损失精确检索能力。

例题 15.1 (单头注意力手算)**.** 设一个查询 $q=(1,0)$，键 $k_1=(1,0), k_2=(0,1)$，值 $v_1=(2,0), v_2=(0,4)$，忽略缩放，则权重为 $(e/(e+1), 1/(e+1))$，输出为 $(2e/(e+1), 4/(e+1))$。若加因果掩码并且当前位置只能看第一个 token，则输出正好为 v_1。

例题 15.2 (完整注意力矩阵与形状)**.** 取 $Q=K=I_2$、$V=\left(\begin{smallmatrix}1&0\\0&2\end{smallmatrix}\right)$，忽略 $1/\sqrt{2}$ 仅为方便手算。则

$$QK^\top = I_2, \qquad A = \mathrm{softmax}_{\mathrm{row}}(I_2) = \frac{1}{e+1}\begin{pmatrix} e & 1 \\ 1 & e \end{pmatrix},$$

$$AV = \frac{1}{e+1}\begin{pmatrix} e & 2 \\ 1 & 2e \end{pmatrix}.$$

softmax 必须逐行做；逐列归一化会改变“每个 query 对所有 key 分配权重”的含义。若第一行加因果掩码 $(0,-\infty)$，其权重变为 $(1,0)$。

15.2 注意力的维度、掩码与复杂度闭环

设输入 $X \in \mathbb{R}^{n\times d_{\mathrm{model}}}$，单头投影

$$Q = XW_Q \in \mathbb{R}^{n\times d_k}, \quad K = XW_K \in \mathbb{R}^{n\times d_k}, \quad V = XW_V \in \mathbb{R}^{n\times d_v}.$$

缩放点积注意力为

$$\mathrm{Attn}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + M\right)V.$$

softmax 沿 key 维逐行归一化。掩码 M_{ij} 对允许位置为 0，禁止位置为 $-\infty$；causal mask 禁止 $j>i$，padding mask 禁止读取填充 token。缩放因子来自独立单位方差坐标下 $\mathrm{Var}(q^\top k)=d_k$：若不除以 $\sqrt{d_k}$，logit 随维度增大而饱和。

定理 15.2 (置换等变性与位置编码的必要性)**.** *若不加入位置编码且不使用破坏置换的掩码，则 self-attention 对 token 行置换 Π 满足*

$$\mathrm{Attn}(\Pi X) = \Pi\,\mathrm{Attn}(X).$$

证明. 行置换给 Q,K,V 同时左乘 Π，于是分数矩阵变成 $\Pi QK^\top\Pi^\top$。逐行 softmax 与同时行列置换可交换，故权重矩阵为 $\Pi A\Pi^\top$，再乘 ΠV 得 ΠAV。因此模型本身不知道绝对或相对次序，必须由位置编码或结构掩码注入。 □

多头注意力把 h 组结果拼接后乘 W_O。当 $d_k=d_v=d_{\mathrm{model}}/h$ 时，投影参数量仍为 $O(d_{\mathrm{model}}^2)$；注意力矩阵的时间和显存为 $O(n^2d_{\mathrm{model}})$ 与 $O(n^2)$。自回归推理用 KV cache 保存历史 K,V，把每个新 token 的重复投影降掉，但总注意力计算仍随上下文长度线性增长、整段生成累计为二次量级。

反例 15.1 (掩码方向写反). 若在生成位置 i 允许读取 $j > i$，训练损失会因未来 token 泄漏而异常低，但推理时未来信息不存在，形成训练–推理不一致。数值上把禁止位置加 0 而不是大负数同样不能实现掩码。

答题方法

系统题先列 $n, d_{\text{model}}, h, d_k, d_v$，逐项写形状；再说明 mask、位置、残差、norm 和 FFN 的作用。复杂度必须区分训练整段、prefill 和逐 token decode。

自测

1. 给定 n, d_{model}, h，写出 Q/K/V、注意力权重、拼接和输出的全部形状。
2. 证明无位置编码的 self-attention 对输入排列等变。
3. 比较 pre-norm 与 post-norm 的梯度路径。
4. 计算带 KV cache 的单 token 解码时间与缓存空间。
5. 说明 causal mask 与 padding mask 的广播维度和语义差别。

本节结论

Transformer 的考试闭环是“维度—归一化—位置—掩码—复杂度—训练/推理状态”。任何只画模块图而不写张量和目标的答案都不完整。

15.3 Scaled dot-product attention 的逐轴推导

设批量 B、序列长 T、模型宽度 d，输入 $X \in \mathbb{R}^{B\times T\times d}$。单头投影

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V,$$

其中 $W_Q, W_K \in \mathbb{R}^{d\times d_k}$，$W_V \in \mathbb{R}^{d\times d_v}$。对每个批量，

$$S = \frac{QK^\top}{\sqrt{d_k}} \in \mathbb{R}^{T\times T}, \qquad A = \operatorname{softmax}_{\text{row}}(S + M), \qquad O = AV \in \mathbb{R}^{T\times d_v}.$$

第 i 行 $A_{i:}$ 是 query i 对所有 key 的概率权重，因此 softmax 必须沿 key 轴。沿 query 轴归一化会改变语义，且每个输出不再是 value 的凸组合。

若 q_j, k_j 独立、均值零、方差一，则

$$\operatorname{Var}(q^\top k) = \sum_{j=1}^{d_k} \operatorname{Var}(q_j k_j) \approx d_k.$$

除以 $\sqrt{d_k}$ 使 logit 尺度在宽度增大时保持 $O(1)$，避免 softmax 过早饱和。

15.4 Multi-head attention 的拼接与参数量

取 H 个头，常用 $d_k = d_v = d/H$。每头输出

$$O_h = \mathrm{Attn}(XW_Q^{(h)}, XW_K^{(h)}, XW_V^{(h)}) \in \mathbb{R}^{B\times T\times d_v}.$$

沿最后一轴拼接

$$O_{\mathrm{cat}} = \mathrm{Concat}(O_1, \dots, O_H) \in \mathbb{R}^{B\times T\times Hd_v},$$

再乘 $W_O \in \mathbb{R}^{Hd_v\times d}$。若不计偏置且 $Hd_k = Hd_v = d$，四组投影共约 $4d^2$ 参数；头数改变各头宽度，但在这一约定下不改变主参数量。

例题 15.3. $B = 16, T = 128, d = 768, H = 12$，则每头 $d_k = d_v = 64$。每头注意力矩阵形状为 $16 \times 128 \times 128$，十二头合计注意力权重元素数为

$$16 \cdot 12 \cdot 128^2,$$

这解释了长序列训练的显存瓶颈。输出仍为 $16 \times 128 \times 768$。

15.5 Causal、padding 与组合 mask

自回归 causal mask 令

$$M_{ij}^{\mathrm{causal}} = \begin{cases} 0, & j \le i, \\ -\infty, & j > i, \end{cases}$$

使第 i 个位置不能读取未来 token。Padding mask 则对所有 query 屏蔽补齐 key。二者组合通常按广播相加：

$$M = M^{\mathrm{causal}} + M^{\mathrm{pad}}.$$

实现中用足够大的负数近似 $-\infty$，必须避免低精度溢出或整行全被屏蔽导致 softmax 的 0/0。

交叉注意力中，query 来自解码器，key/value 来自编码器。causal mask 只作用于解码器自注意力；交叉注意力通常只需要编码器 padding mask。

15.6 位置编码为何必要

没有位置项时，自注意力对输入 token 的同步置换等变。设置换矩阵 P，则

$$Q(PX) = PQ(X), \quad K(PX) = PK(X), \quad V(PX) = PV(X),$$

$$\mathrm{softmax}\left(\frac{PQK^\top P^\top}{\sqrt{d_k}}\right) = P\,\mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)P^\top,$$

从而

$$\mathrm{Attn}(PX) = P\,\mathrm{Attn}(X).$$

模型本身不知道“第一个”与“最后一个”。绝对位置向量、相对位置偏置或旋转位置编码都在注意力计算中注入顺序信息，但外推性质不同。

正弦位置编码

$$\mathrm{PE}_{p,2i} = \sin(p/\omega_i), \qquad \mathrm{PE}_{p,2i+1} = \cos(p/\omega_i)$$

具有相对位移可由同频率二维旋转表示的性质；它不等于自动保证无限长度外推。

15.7 残差、归一化与前馈层

标准 Transformer block 包含注意力子层和逐 token 前馈网络

$$\mathrm{FFN}(x) = W_2\phi(W_1x + b_1) + b_2.$$

FFN 对每个位置使用同一参数，不在 token 之间直接混合；token 混合由注意力完成。若中间宽度 d_{ff}，FFN 主参数量约为 $2dd_{\mathrm{ff}}$，常超过注意力投影参数。

Post-norm 形式

$$x_{l+1} = \mathrm{LN}(x_l + F_l(x_l))$$

与 pre-norm 形式

$$x_{l+1} = x_l + F_l(\mathrm{LN}(x_l))$$

的梯度通道不同。Pre-norm 显式保留恒等残差路径，通常更利于深网络优化；两者不能在不重新训练的情况下直接替换。

15.8 复杂度、KV cache 与自回归生成

完整序列自注意力的主要复杂度为

$$O(BT^2d + BTd^2),$$

前一项来自 $QK^\top$ 与 AV，后一项来自线性投影。注意力矩阵存储为 $O(BHT^2)$。

自回归解码若每一步重新计算全部历史，生成 T 个 token 会重复构造旧 key/value。KV cache 保存每层历史

$$K_{1:t}, V_{1:t},$$

新一步只计算当前 token 的 query/key/value，再让单 query 与缓存 key 做注意力。单步注意力从对全部历史的二次重算降为随当前长度线性增长，但缓存空间为

$$O(LBTd)$$

量级，并随批量、层数和上下文长度增长。

15.9 注意力反向传播的稳定检查

对单行 softmax $a = \mathrm{softmax}(s)$，Jacobian 为

$$\frac{\partial a}{\partial s} = \mathrm{diag}(a) - aa^\top.$$

若上游梯度为 g_a，则

$$g_s = a \odot \big(g_a - \langle g_a, a\rangle \mathbf{1}\big).$$

再由

$$S = QK^\top/\sqrt{d_k}, \qquad O = AV$$

得到

$$g_V = A^\top g_O, \quad g_A = g_O V^\top, \quad g_Q = g_S K/\sqrt{d_k}, \quad g_K = g_S^\top Q/\sqrt{d_k}.$$

这组公式同时给出正确维度和 $1/\sqrt{d_k}$ 因子的位置。

易错点

常见四错：softmax 轴写错；padding mask 只屏蔽 query 而未屏蔽 key；交叉注意力把 Q/K/V 来源写反；有 KV cache 后声称总生成复杂度变成 $O(T)$。缓存消除重复投影，不消除每个新 query 对全部历史 key 的读取。

答题方法

Transformer 题固定六步：列 B, T, d, H, d_k, d_v；写 Q/K/V 来源；写 score 形状；写 mask 广播；写 softmax 轴；写拼接和输出投影。最后分别报告参数量、计算量和缓存量。

自测

1. 证明无位置自注意力的置换等变性。
2. 计算给定 B, T, d, H 时 score、每头输出和拼接输出的形状。
3. 为什么 causal mask 加在 softmax 前，而不是直接把输出 token 清零？
4. 推导 softmax 的向量–Jacobian 乘积。
5. KV cache 节省什么计算，增加什么存储？

第十六章　预训练、LLM、知识表示与 RAG 系统

学习目标

本章从 autoregressive/masked 预训练目标进入解码、扩展规律、SFT、偏好学习、知识图谱与 RAG；重点是把模型概率、检索证据、评价指标、安全与延迟组成可执行系统闭环。

16.1　预训练目标：BERT、GPT 与序列到序列

BERT 类模型用 masked language modeling：随机选择位置集合 M，最小化

$$\mathcal{L}_{\mathrm{MLM}} = -\sum_{t \in M} \log p_\theta(x_t \mid x_{\setminus M}).$$

它利用双向上下文，但预训练中的 mask 与下游无 mask 输入存在差异。GPT 类模型使用因果 NLL $-\sum_t \log p_\theta(x_t \mid x_{<t})$，训练目标与自回归生成一致。T5 等把任务统一为 text-to-text 的条件生成。

预训练后可：

1. **全量微调**：更新全部参数，能力强、显存与遗忘风险大；
2. **参数高效微调**：如 adapter/LoRA，只训练少量参数。LoRA 令 $W' = W + BA$，秩 $r \ll d$；
3. **提示学习**：参数冻结，通过离散/连续提示诱导任务；
4. **上下文学习**：只在输入中给示例，不执行梯度更新。

术语上本书采用常用三阶段区分：大规模自监督训练称为预训练，instruction tuning 属于后训练中的监督微调，偏好优化属于后续对齐阶段，并在第十六章继续说明。

16.2　解码与生成控制

给定 next-token 分布：

- greedy 每步取最大概率，快但可能陷入局部模式；
- beam search 保留 B 个部分序列，近似最大序列概率，易偏短且缺多样性；
- top-k 只在最高 k 个 token 中重归一化采样；
- top-p 选择累计概率至少 p 的最小集合，候选数随分布自适应；
- temperature 用 $p_i \propto e^{z_i/\tau}$，$\tau < 1$ 更尖锐，$\tau > 1$ 更平坦。

采样参数不能修复模型不知道的事实；解码质量需与任务约束、事实验证和安全策略共同评估。

例题 16.1 (greedy 与 beam search)**.** 第一步 $p(A)=0.6, p(B)=0.4$；第二步 $p(\text{EOS}\mid A)=0.5$、$p(\text{EOS}\mid B)=0.9$。greedy 先选 A，得到序列概率 $0.6\cdot 0.5=0.30$；宽度 2 的 beam 同时保留 A, B，会发现 B, EOS 的概率 $0.4\cdot 0.9=0.36$ 更高。beam 仍不是全局最优保证，长序列还常需长度归一化。

16.3 评测（evaluation）、历年题与易错点

分类使用 precision/recall/F1；宏平均先按类平均，重视小类；微平均先汇总计数，受大类支配。BLEU 的 N-gram 修改精确率为 p_n，候选长度 c、参考长度 r 时

$$\text{BP}=\min\{1, e^{1-r/c}\}, \qquad \text{BLEU}_N=\text{BP}\exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

ROUGE-N 更强调参考 N-gram 的召回；ROUGE-L 使用最长公共子序列产生 precision/recall/F 值。BLEU 偏 precision、ROUGE 偏 recall 是复习方向，具体实现仍取决于多参考、分词、平滑和聚合约定。

例题 16.2 (BLEU-1 与 ROUGE-L)**.** 候选为“a b”，参考为“a c”。二者等长，故 BP 为 1；匹配 unigram 只有“a”，所以 BLEU-1 为 1/2。最长公共子序列长度为 1，故 ROUGE-L precision 与 recall 都为 1/2，其 F 值也为 1/2。真实多句评测还需说明句级/语料级聚合，不能把此简例的计算规则无条件外推。

BLEU/ROUGE 侧重表面重叠，语义指标依赖评估模型；事实性、鲁棒性和安全性需要单独评测。

历年题专题 16.1: 2025 春 NLP：语言模型、解码与架构辨析

正式卷要求处理 trigram/平滑、表示学习、top-p 与 top-k/greedy、Transformer 对 RNN 的优势、scaling、RLHF 与领域 RAG。答题应分别给数学定义、结构原因与系统边界，不能写成技术史。

历年题专题 16.2: 2025 秋 NLP：表示、图谱与生成模型

正式卷覆盖 GloVe/skip-gram/node2vec/TransE、LSTM 与 Transformer 的扩展性、Markov logic network、topic–ontology LDA 以及文本到图像扩散。复习时应围绕“数据信号—建模目标—归纳偏置—计算代价”进行比较。

易错点

易错点：PPL 跨 tokenizer 比较；把平滑理解为只给未见事件加概率而不重新归一化；Skip-gram/CBOW 方向颠倒；HMM 条件独立假设遗漏；前向算法与 Viterbi 混用；CRF 配分函数漏求和；BERT 说成自回归；fine-tuning 与 in-context learning 混同；beam search 说成保证全局最优；把 BLEU 当事实性指标。

复习检查

1. 能否计算 n-gram MLE、平滑概率、交叉熵和 PPL？
2. 能否写出 skip-gram negative sampling 与 GloVe 目标并比较？
3. 能否分别完成 HMM forward、Viterbi 和 CRF 梯度推导？
4. 能否从注意力掩码解释 BERT/GPT/encoder–decoder 的信息流？
5. 能否比较全量微调、LoRA、提示与上下文学习？
6. 能否为确定性任务和开放生成选择解码/评测策略？

16.4 扩展规律、涌现与计算分配

经验 scaling law 常用幂律加不可约项描述:

$$L(N, D, C) \approx L_\infty + aN^{-\alpha} + bD^{-\beta} + cC^{-\gamma},$$

其中参数量 N、数据量 D、计算量 C 并非相互独立。固定计算预算时，应在模型和 token 数之间分配，而非只增大参数。幂律是特定数据/架构/区间上的经验拟合，不是任意规模都成立的数学定理。

“涌现”可能指某能力在指标上突然跨阈值，也可能由离散评测或非线性刻度造成。答题应区分连续损失改善、下游阈值现象和真正的新算法能力，避免“规模必然带来推理”这类无条件陈述。

16.5 后训练：SFT、偏好学习（preference learning）与 RLHF

监督微调（SFT）在指令–响应对上最小化自回归 NLL。奖励模型从偏好对 $y^+ \succ y^-$ 学习，例如 Bradley–Terry 损失

$$\mathcal{L}_{\mathrm{RM}} = -\log \sigma(r_\phi(x, y^+) - r_\phi(x, y^-)).$$

RLHF 的策略阶段最大化奖励并用参考策略约束漂移:

$$\max_\theta \mathbb{E}_{x, y\sim\pi_\theta}\left[r_\phi(x, y) - \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)}\right].$$

KL 惩罚同时限制 reward hacking 和语言能力退化，但 β 过大会抑制适应。

PPO 使用旧策略重要性比 $r_t(\theta) = \pi_\theta(a_t \mid s_t)/\pi_{\theta_{\mathrm{old}}}(a_t \mid s_t)$，剪切目标

$$\mathbb{E}\min\{r_t(\theta)\widehat{A}_t, \mathrm{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\widehat{A}_t\}.$$

剪切是保守更新的启发式代理，并非严格保证 KL 上界。DPO 直接对偏好数据优化

$$-\log \sigma\left(\beta\left[\log \frac{\pi_\theta(y^+ \mid x)}{\pi_{\mathrm{ref}}(y^+ \mid x)} - \log \frac{\pi_\theta(y^- \mid x)}{\pi_{\mathrm{ref}}(y^- \mid x)}\right]\right),$$

省去显式奖励模型与在线 RL。GRPO 以同一提示的一组采样响应构造组内相对 advantage，减少独立 critic；具体归一化与裁剪实现有多种变体，考试应先掌握“组内基线 + 策略比更新”的结构，若题目给定具体版本再代入推导。

易错点

术语提醒：本书统一采用“自监督预训练—SFT/instruction tuning—偏好对齐”的三阶段术语。若试题把 instruction tuning 广义归入预训练流程，应先声明口径再作比较。

16.6 知识图谱（knowledge graph）表示与推理

知识图谱由三元组 (h, r, t) 构成。TransE 令 $e_h + e_r \approx e_t$，评分 $s(h, r, t) = -\|e_h + e_r - e_t\|$，用负采样和 margin ranking：

$$\sum_{(h,r,t)} \sum_{(h',r,t')} \left[\gamma + s(h', r, t') - s(h, r, t)\right]_+.$$

它简单可扩展，但对一对多、多对一和对称关系表达受限。Node2vec 在图上以带偏随机游走产生“句子”，再用 skip-gram 学节点表示；参数 p, q 控制回退和向外探索。

Markov Logic Network（MLN）把一阶逻辑公式 F_k 与权重 w_k 结合：

$$p(X = x) = \frac{1}{Z} \exp\left(\sum_k w_k n_k(x)\right),$$

$n_k(x)$ 是世界 x 中公式成立的 grounding 数。权重大表示违反该软规则代价高，无穷权重才近似硬约束。推断需处理巨大 grounding 空间和配分函数，常用近似采样/变分。

命题 16.1 (MLN 的 MAP 目标)**.** *给定证据变量 $X_E = e$，未知变量 X_U 的 MAP 推断等价于*

$$\arg\max_{x_U} p(x_U \mid e) = \arg\max_{x_U} \sum_k w_k n_k(x_U, e).$$

证明. 条件分布的分母 $p(e)$ 与 x_U 无关；联合分布中的配分函数 Z 也与具体世界无关。对正的指数函数取对数不改变极大点，因此只需最大化加权 grounding 计数。若把证据谓词也当未知量优化，就改变了条件问题本身。 □

例题 16.3 (TransE 负样本与限制)**.** 若正三元组 (Paris, capitalOf, France)，负样本可替换头或尾，如 (Paris, capitalOf, Germany)，但必须避免把真实三元组误作负例。对一对多关系“country–hasCity–city”，同一 $e_h + e_r$ 被迫接近多个城市向量，解释了 TransE 的表达瓶颈。变式可要求比较 DistMult 的对称性限制。

16.7 信息抽取（information extraction）、概率逻辑（probabilistic logic）与主题模型

信息抽取包括实体识别、关系抽取、事件抽取和实体链接。NER 常用 BIO/BIOES 标签：可由 HMM 生成建模、CRF 全局条件归一化，或 Transformer token classifier 加 CRF 转移层实现。

独立 token softmax 解码快但可能产生非法标签转移；CRF 用 $O(TK^2)$ 动态规划保持全局一致。Pipeline 易调试但误差传播；联合模型共享表示并保持一致性，但训练和解码更复杂。

LDA 对文档 d 先采样主题混合 $\theta_d \sim \text{Dirichlet}(\alpha)$,再对每个 token 采样 $z_{dn} \sim \text{Categorical}(\theta_d)$、$w_{dn} \sim \text{Categorical}(\beta_{z_{dn}})$，联合分布为

$$p(\theta_d, z_d, w_d \mid \alpha, \beta) = p(\theta_d \mid \alpha) \prod_n p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta).$$

证明. 生成图中，给定 θ_d 后各主题指示 z_{dn} 条件独立；给定 z_{dn} 后词 w_{dn} 只依赖对应主题的词分布。按有向图的条件概率链式分解逐节点相乘，就得到上述联合分布。若主题词分布 β_k 也设 Dirichlet 先验，还需在全语料联合分布前乘 $\prod_k p(\beta_k)$。 □

监督主题模型再乘标签因子 $p_\eta(y_d \mid \theta_d)$，使主题兼顾文本解释与预测。该因子改变后验，不能只在训练完 LDA 后口头附加分类器而仍称同一联合模型。若结合 ontology，可通过先验或层次约束把主题与概念关联，但必须说明是硬约束还是软正则。知识图谱增强语言模型可采用检索、图编码器或联合训练；“加入图谱”本身不保证正确推理，实体链接与证据路径是关键误差源。

16.8 RAG：检索、生成与证据闭环

RAG 的基本链为：

$$\text{query 理解} \to \text{检索} \to \text{重排} \to \text{上下文构造} \to \text{生成} \to \text{引用/验证}.$$

双编码器用内积独立编码 query/document，检索快；cross-encoder 联合编码，相关性高但成本大，通常用于重排。训练可用对比损失；hard negatives 能提升判别力，也可能引入假阴性。

若把外部记忆选择建模为 MDP，状态可包含 query、已检索证据和生成进度，动作是检索/阅读/回答，奖励结合答案正确性、证据性和成本。此建模只在动作影响后续信息与决策时有意义；一次静态 top-k 检索无需强行包装成 RL。

例题 16.4 (领域 RAG 系统题)**. 题目：**为医学指南问答设计 RAG，要求降低幻觉并处理指南版本更新。

思路：建立版本化语料和元数据过滤，混合稀疏/稠密召回，cross-encoder 重排，生成器必须引用证据，回答后执行 entailment/规则核验。

完整解答：离线按章节切分并保留标题、发布日期、适用人群；query 先识别疾病与人群，过滤过期版本；BM25 保留精确术语，dense retrieval 覆盖语义变体；重排后控制上下文去重。提示要求“证据不足则拒答”，输出句级引用。评测分 retrieval recall、reranking NDCG、答案正确性、引用支持率和过期信息率，并以医生审核高风险失败。

易错点：只写“向量库 +LLM”，没有版本控制、证据评测或拒答。

变式：若检索延迟受限，比较预计算摘要、量化索引和级联重排的质量–成本权衡。

16.9 长上下文（long context）、记忆与推理成本

全注意力随长度二次增长，且“放入上下文”不等于“有效使用”。可选：

- sliding window：成本线性于窗口，但远距离信息需层间传播；

- sparse/global token：保留少量全局通道，依赖稀疏模式；
- recurrence/compressive memory：跨块传递状态，可能累积误差；
- retrieval memory：只取相关片段，质量受检索召回限制；
- KV cache：降低自回归时重复计算，但存储随层数和上下文增长。

系统题应同时算训练、prefill、decode 和缓存。batch 增大提高吞吐但可能增加单请求延迟；量化降低存储/带宽但可能损害稀有能力。

16.10 安全（safety）、评测与失效分析

LLM 评测至少分：任务正确性、事实/引用、鲁棒性、安全、公平、校准、成本与延迟。LLM-as-a-judge 可能有位置、长度、自偏好偏差，应随机顺序、使用标尺和人类抽查。数据污染使测试分数高估泛化；对动态知识要记录 cutoff 与检索时间。

常见失效包括 hallucination、prompt injection、检索投毒、reward hacking、jailbreak 和隐私记忆。防护应分层：输入隔离、工具权限最小化、内容/策略过滤、证据验证、审计日志和红队测试。单一系统提示不是安全边界。

例题 16.5 (强化学习与对齐综合题)**. 题目：**为什么 RLHF 中需要 KL 约束？DPO 与 PPO 的关键差异是什么？

完整解答：奖励模型只在有限偏好数据上可靠，直接最大化会把策略推向分布外并利用奖励漏洞；相对参考策略的 KL 惩罚限制漂移。PPO 通过采样轨迹、advantage 和剪切策略比近似优化奖励，通常需奖励模型/critic；DPO 把 KL 正则最优策略关系代入偏好概率，直接在偏好对上做分类式优化，不进行在线 RL。DPO 更简单稳定但受离线偏好覆盖限制；PPO 可在线探索但复杂、方差高。

易错点：把 PPO clipping 说成严格 KL 约束；把 DPO 说成完全不需要参考策略。

变式：比较组相对基线方法为何可减少 critic。

答题方法

系统设计题按“需求—数据/索引—召回—重排—上下文构造—生成—引用—评价—监控”回答。每个组件写输入输出、离线/在线指标、主要失败模式和成本；不得把 RAG 等同于简单拼接搜索结果。

自测

1. 比较 causal LM、masked LM 与 encoder–decoder denoising 的条件分解和适用任务。
2. 推导 temperature 对 softmax 分布熵的定性影响，并比较 top-k 与 nucleus sampling。
3. 写出偏好模型和 KL 约束策略优化的基本目标，说明 reference policy 的作用。
4. 设计一个带可追溯引用的 RAG 系统，并给出检索、忠实性、答案质量和延迟指标。
5. 列出长上下文不能自动替代检索的三个原因。

本节结论

LLM 系统不是单一模型题。高质量答案必须把概率目标、数据、检索、解码、评价、安全和工程约束放在同一因果链中。

16.11 自回归似然、teacher forcing 与曝光偏差

自回归模型分解

$$p_\theta(x_{1:T}) = \prod_{t=1}^{T} p_\theta(x_t \mid x_{<t}).$$

最大似然训练最小化

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

Teacher forcing 指训练条件历史使用真实前缀 $x_{<t}$，因此所有位置可并行计算带 causal mask 的损失。生成时历史来自模型自己的采样，早期错误会改变后续条件分布；这种训练与生成历史分布不匹配称为 exposure bias。它不是把 teacher forcing 简单改成“训练时也完全自由生成”就一定解决，因为后者会失去稳定的逐 token 监督并引入离散采样梯度问题。

Masked LM 目标对被遮位置的条件概率求和，不给出从左到右联合概率的同一分解；encoder–decoder 模型则令编码器双向读取源序列，解码器因果地产生目标序列。

16.12 解码是对同一条件分布的不同决策

Greedy 每步选局部最大概率 token；beam search 保留累计对数概率最高的 K 个前缀；随机采样按模型分布抽取。温度 $T > 0$ 修改 logits：

$$p_T(i) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}.$$

$T < 1$ 使分布更尖，$T > 1$ 更平；$T \to 0$ 接近 argmax。Top-k 只保留概率最高的 k 个 token，top-p 保留累计概率达到阈值的最小集合，再重新归一化。

反例 16.1. Greedy 不是全序列 MAP。若第一步 token a 概率 0.51，b 概率 0.49，但最佳后续条件概率分别为 0.5 与 0.99，则最佳两步序列从 b 开始，概率 0.4851，高于 greedy 路径的 0.255。

长度归一化必须显式说明。直接累计对数概率偏好较短序列，因为每增加一个 token 都加非正数；常用

$$\frac{\log p(y \mid x)}{(5+|y|)^\alpha/(5+1)^\alpha}$$

等启发式校正，但它改变了解码评分，不是原模型联合概率。

16.13 偏好学习的概率模型

给同一提示 x 的偏好对 $y^+ \succ y^-$，Bradley–Terry 奖励模型写为

$$\mathbb{P}(y^+ \succ y^- \mid x) = \sigma(r_\phi(x, y^+) - r_\phi(x, y^-)).$$

最大似然训练奖励差。RLHF 随后优化策略的期望奖励并用 KL 约束限制其偏离参考策略：

$$\max_{\pi} \ \mathbb{E}_{y\sim\pi(\cdot|x)} r_\phi(x,y) - \beta\,\mathrm{KL}(\pi(\cdot\mid x)\|\pi_{\mathrm{ref}}(\cdot\mid x)).$$

KL 项防止策略利用奖励模型的盲区无限漂移。偏好数据只说明比较关系，不把奖励尺度绝对标定；给奖励整体加常数不改变成对概率。

16.14 RAG 的检索—重排—生成闭环

RAG 将条件生成写成潜在文档混合：

$$p(y\mid x) \approx \sum_{d\in\mathcal{D}_K(x)} p_\eta(d\mid x)\, p_\theta(y\mid x,d).$$

系统流程至少包括：

1. 文档切分、元数据与索引；
2. query 构造和候选召回；
3. reranker 精排；
4. 上下文装配、去重和 token 预算；
5. 带证据生成；
6. 引用核验、拒答与日志。

若正确证据未被召回，生成器无从引用；若召回正确但排序靠后或被截断，则是上下文装配失败；若证据已在上下文而答案仍错误，则更可能是生成/归因失败。三类错误需用不同指标诊断。

检索指标包括 recall@K、MRR、nDCG；生成指标包括 exact match、任务评分、事实一致性；系统指标包括延迟、吞吐、索引新鲜度与成本。只报告最终答案准确率无法定位瓶颈。

例题 16.6. 设 100 个问题中 80 个至少有一个正确文档进入 top-5，且这 80 个中生成器答对 60 个。则检索 recall@5 为 0.8，条件生成准确率为 $60/80 = 0.75$，端到端准确率为 0.60。即使把条件生成提高到 0.9，若召回不变，端到端上限仍为 0.72。

16.15 评测、数据泄漏与可复现实验

离线评测集必须与训练、指令微调和检索索引的时间/内容边界明确区分。若测试答案原文存在索引中，RAG 的高分可能只反映直接匹配；若测试题进入预训练语料，模型闭卷高分也不能解释为稳定推理能力。

对随机解码，应报告随机种子、温度、top-p、最大长度与重复次数。对自动裁判，应抽查位置偏差、长度偏差和自偏好，并用人工或规则指标校准。安全评测还要区分拒答率、误拒率与越狱成功率，不能用单一“安全分数”掩盖权衡。

16.16 长上下文并不等于可靠记忆

理论上下文窗口只是允许输入的 token 上限。信息位于中段时可能出现 lost-in-the-middle；重复或冲突证据会竞争注意力；KV cache 随长度线性增长。系统设计应在“把全部资料塞入窗口”与“外部检索/压缩”之间比较：

$$\text{质量} = \text{召回} \times \text{可读上下文} \times \text{生成利用率},$$

这不是严格概率恒等式，而是诊断乘法瓶颈的工程框架：任一环节接近零都会限制端到端表现。

答题方法

LLM 系统题按“目标—数据—模型—推理—指标—失败模式—资源约束”作答。任何经验性规律都标明适用范围；任何概率目标都写清训练分布与推理分布；RAG 必须把检索错误和生成错误分开。

自测

1. 自回归 LM 与 masked LM 的概率分解有什么根本区别？
2. 构造一个 greedy 不是序列 MAP 的两步例子。
3. KL 正则在偏好优化中控制什么？奖励加常数为何不影响成对偏好概率？
4. 为 RAG 分别列出检索、生成和系统指标。
5. 为什么“窗口足够长”不能推出模型能可靠使用窗口内任意证据？

第十七章　三期正式真题与完整解答

本章按年份收入 Natural Language Processing 的全部正式题面与解答，共保留既有审计口径下的 18 道大题、39 个最细作答单元。除特别注明外，解答为复习用整理稿，不代表官方评分标准。每个题组前的“理论入口”用于回看本 Part 前文。

17.1　2025 春

理论入口：参见章 十四、十六。

2025 春·NLP·Q1–Q5（原卷第 5 页，每题 3 分）

Original.

1. In a trigram language model, how is $p(w_3 \mid w_1, w_2)$ learned from a training corpus? In real-world applications, (w_1, w_2), or even w_1 or w_2, may not appear in the training corpus. How can $p(w_3 \mid w_1, w_2)$ be estimated in such cases?
2. What is the central idea behind representation learning? Which algorithms implement this concept?
3. Why do current large language models use top-P sampling during inference instead of greedy or beam search? Why is top-P considered superior to other methods such as top-K sampling?
4. List several key points that contribute to the success of current LLM-based approaches to Artificial General Intelligence and explain their importance.
5. Identify at least two advantages of the Transformer architecture over recurrent models (e.g. LSTM and GRU) for sequence-to-sequence tasks.

中文译述

依次回答 trigram 与平滑、表示学习、nucleus sampling、当前大模型成功因素、Transformer 相对 RNN 的优势。

完整解答：Q1

MLE 为 $\hat{p}(w_3 \mid w_1, w_2) = C(w_1, w_2, w_3)/C(w_1, w_2)$。未见上下文时必须平滑/回退：如插值 $\lambda_3 p_{\text{tri}} + \lambda_2 p_{\text{bi}} + \lambda_1 p_{\text{uni}}$，或 Katz/Kneser–Ney；不能令整句概率为零。

完整解答：Q2

核心是从数据关系中学习对任务有用的连续表示，使相似对象在表示空间接近并可迁移。例：PCA/自编码器、word2vec/GloVe、对比学习、Transformer 预训练；每个算法由重建、上下文预测或对比目标实现。

完整解答：Q3

greedy/beam 偏向高概率、低多样性的序列，开放式生成易重复；top-p 取累计概率达到 p 的最小动态候选集后重归一采样。top-k 固定候选数，分布尖锐时纳入低质词、分布平坦时又排除合理词；top-p 随不确定性调整。确定性任务仍可能适合 greedy/beam，不能绝对化。

完整解答：Q4

可列：大规模自监督数据提供覆盖；Transformer 与并行计算提高可扩展性；scaling law 指导算力/数据/参数配置；instruction tuning 与偏好对齐改善可用性；工具/RAG 增加外部知识；高质量评测和数据治理抑制失败。每点需说明机制而非只列名词。

完整解答：Q5

self-attention 可在常数层路径上连接远距离 token，RNN 路径长度为 $O(L)$；Transformer 在序列维可并行，硬件吞吐高。另可答多头关系建模和残差/归一化改善优化。代价是全注意力 $O(L^2)$ 内存，不能只写优点。

正文对应： 节 14.2、15.1、16.2、章 十六。

理论入口： 参见章 十四、十六。

2025 春·NLP·Q6（原卷第 5 页，5 分）

Original. Explain why a data scaling law relating performance and sample size holds from statistical learning, with a detailed analysis and without assuming Gaussian data.

中文译述

从统计学习角度说明数据规模律的来源，不能靠“数据服从高斯”之类过强假设。

完整解答：Q6

对固定假设类，经验风险围绕真实风险集中；有界损失下 Hoeffding/复杂度界典型给 $L_D(\hat{h}) - \inf_{h\in\mathcal{H}} L_D(h) = O(\mathfrak{R}_m(\mathcal{H})) + O(\sqrt{\log(1/\delta)/m})$。随 m 增大，估计误差下降；若同时增大模型，逼近误差下降但复杂度项上升，最佳包络可表现为幂律。真实神经 scaling law 还混合优化误差、数据重复/噪声、分布尾部和容量瓶颈，因此指数是经验量，不由单一集中不等式普适决定。结论只需 i.i.d./弱依赖、有界或次指数尾与容量控制，不需高斯。

正文对应： 章 三、十六中的泛化界与 scaling law。

理论入口： 参见章 十四、十六。

2025 春·NLP·Q7（原卷第 5 页，5 分）

Original. Given a trained reward model, explain how RL is used in RLHF, including data usage, the RL loss and its gradient.

中文译述

奖励模型已训练好，说明如何采样、构造 KL 正则奖励、写策略目标及梯度。

完整解答：Q7

从提示分布采样 x，用当前策略生成 y，奖励 $\tilde{r}(x,y) = r_\psi(x,y) - \beta \log[\pi_\theta(y \mid x)/\pi_{\text{ref}}(y \mid x)]$。最大化 $J(\theta) = \mathbb{E}_{x,y\sim\pi_\theta}\tilde{r}$；序列级 REINFORCE 梯度为

$$\nabla J = \mathbb{E}\sum_t \nabla \log \pi_\theta(y_t \mid x, y_{<t})\, \hat{A}_t,$$

$\hat{A}_t$ 由回报减 critic/基线估计。PPO 用截断概率比代理目标稳定更新，并监控 KL。偏好数据训练奖励模型；在线 rollout 数据训练策略/critic，二者不可混称。

易错点： 漏参考策略 KL；把奖励模型损失写成策略损失；不说明采样分布。

正文对应： 节 16.5中的 RLHF/PPO/DPO。

理论入口： 参见章 十四、十六。

2025 春·NLP·Q8（原卷第 5–6 页，8 分）

Original. A general LLM lacks domain knowledge. You have a large unlabeled domain corpus but no QA pairs. Design a system that accurately answers domain questions; embedding models may be used.

中文译述

只有领域文档、没有问答标注，设计领域问答系统。

完整解答：Q8

构建 RAG：清洗文档并按语义段落切分，保留标题、版本、权限元数据；同时建立 BM25 与 dense embedding 索引。查询先做意图/实体识别和必要改写，混合召回 top-K，cross-encoder 重排、去重并选入上下文。生成提示要求仅依据证据、给出片段级引用、证据不足拒答。无 QA 对时可用文档标题/段落构造弱监督检索对，或由 LLM 生成合成 QA 后人工抽查；不能用合成答案作为唯一测试集。评测分检索 recall@K、重排 NDCG、答案正确性、引用支持率、拒答和延迟；对高风险领域加入人工审核和版本过滤。

常见错误： 只写“向量数据库 +LLM”，没有切分、重排、证据约束和评测。

正文对应： 节 16.8中的 RAG 与系统设计。

17.2 2025 秋

理论入口： 参见章 十四、十六。

2025 秋·NLP·Q1（原卷第 8 页，3 分）

Original. For GloVe, Skip-gram, Node2Vec and TransE, describe in 2–3 sentences each the data signal, objective type and inductive bias.

中文译述

比较四类文本/图嵌入方法的数据、目标和归纳偏置。

完整解答：Q1

GloVe 用全局词共现计数，拟合内积与对数共现，偏置是全局比率可线性编码；Skip-gram 用局部窗口正样本与负采样/softmax，偏置是分布相似性；Node2Vec 用带 p, q 偏置的随机游走生成“上下文”并做 skip-gram，偏置在同质/结构等价间可调；TransE 用知识图谱三元组并优化 $\|e_h + e_r - e_t\|$ 的 margin ranking，偏置是关系近似平移。

理论入口：参见章 十四、十六。

2025 秋·NLP·Q2（原卷第 8–9 页，5 分）

Original. Assume $L_{arch}(P, T) = L_\infty + A_{arch} P^{-\alpha_{arch}} + B_{arch} T^{-\beta_{arch}}$ for LSTM and Transformer. Compare fixed-data P scaling, fixed-parameter T scaling, and justify by long-range dependency, parallelizability/optimization, and inductive bias/sample efficiency.

中文译述

在给定经验 scaling 形式下比较 LSTM 与 Transformer 的参数/数据扩展曲线，并解释结构原因。

完整解答：Q2

固定 T_0 时，严格的 log–log 斜率只在参数项主导区近似 $-\alpha_{arch}$；固定 P_0 时数据项主导区近似 $-\beta_{arch}$。在通常大规模经验下，Transformer 往往有较低有效系数和较好的参数/数据扩展，曲线更低或下降更快；但题给公式本身并不能推出 A, α, B, β 的必然次序，必须把它视为经验比较。原因是 attention 的短路径和并行吞吐改善长依赖及优化，而 LSTM 的递归/局部顺序偏置在小数据上可能更具样本效率。答题应画出“低曲线/更陡斜率”的条件式结论，不能宣称架构名称数学上决定指数。

理论入口：参见章 十四、十六。

2025 秋·NLP·Q3（原卷第 9–10 页，5 分）

Original. An MLN has formulas $Cites(p_1, p_2) \wedge AI(p_1) \Rightarrow AI(p_2)$ with $w_1 = 1.5$, and $AI(p)$ with $w_2 = -1$ over papers $P1, P2$. (a) describe nodes/cliques; (b) for $P1 \to P2$, $AI(P1) = 1, AI(P2) = 0$, compute true grounding counts and unnormalized probability; (c) interpret changing w_1 to -1.5 and its effect when both are AI; (d) with $P1 \to P2$ observed, find MAP assignments for the two AI atoms.

中文译述

分析一个两论文 MLN 的图结构、grounding 计数、权重影响和 MAP。

完整解答：Q3

节点是 ground atoms（两类 AI 节点和 citation evidence）；每个规则 grounding 连接其涉及的 atoms 成 clique。若只允许两个不同论文间的有向 citation，$P1 \to P2$ 的蕴含为假、$P2 \to P1$ 因前件假为真，故 $n_1 = 1$，$n_2 = 1$，未归一权重 $e^{1.5-1} = e^{0.5}$。若语义把自引用的两组 grounding 也计入且其 citation 为假，则 $n_1 = 3$；原卷未明确自引用域，需声明约定。$w_1 = -1.5$ 奖励违反“AI 论文引用的论文也是 AI”，因而降低双方均 AI 且有引用的世界权重。只考虑观察到的 $P1 \to P2$，四个 $(AI(P1), AI(P2))$ 的 log-score 分别为 $1.5, 0.5, -1, -0.5$（顺序 $00, 01, 10, 11$），MAP 为 00。

易错点：把蕴含的前件为假当作整条公式为假；不说明 grounding 域。

理论入口：参见章 十四、十六。

2025 秋·NLP·Q4（原卷第 10 页，10 分）

Original. A document is a multiset of fact triples (s, r, o) with latent subject/object types $c_s, c_o \in \mathcal{C}$, relation type $t \in \mathcal{R}$, and K topics. (a) design an LDA-style generative process with θ_d, topic-specific ontology distributions and type-specific surface distributions; (b) factorize the full joint; (c) design automatic names for discovered categories.

中文译述

为事实三元组设计同时学习主题和本体类型的 LDA 生成模型、联合分布与自动命名。

完整解答：Q4

取 $\theta_d \sim Dir(\alpha)$；每个主题 k 取 $\pi_k^{(s)}, \pi_k^{(o)} \sim Dir(\eta_C)$、$\pi_k^{(r)} \sim Dir(\eta_R)$；每个实体类型 c 取 $\phi_c^{(s)}, \phi_c^{(o)}$，每个关系类型 t 取 $\phi_t^{(r)}$ 的 Dirichlet 先验。对每个 triple f：$z_f \sim Cat(\theta_d)$，$c_{s,f} \sim Cat(\pi_{z_f}^{(s)})$、$c_{o,f} \sim Cat(\pi_{z_f}^{(o)})$、$t_f \sim Cat(\pi_{z_f}^{(r)})$，再 $s_f \sim Cat(\phi_{c_{s,f}}^{(s)})$、$o_f \sim Cat(\phi_{c_{o,f}}^{(o)})$、$r_f \sim Cat(\phi_{t_f}^{(r)})$。联合分布是全部 Dirichlet 先验乘以各 triple 的上述七个 categorical 因子。命名可取每个潜类后验最高的 surface phrase/实体链接名称，结合 TF–IDF 区分度、知识库类型映射和 LLM 候选摘要，再由 held-out coherence/人工检查去重；不能只选全局最高频词。

理论入口：参见章 十四、十六。

2025 秋·NLP·Q5（原卷第 10–11 页，10 分）

Original. Design a text-to-image diffusion system. (a) For the Gaussian forward process, (i) show $q(x_t|x_0) = N(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I)$, (ii) give the ELBO decomposition, (iii) derive Gaussian $q(x_{t-1}|x_t, x_0)$ with $\tilde{\beta}_t = (1-\bar{\alpha}_{t-1})\beta_t/(1-\bar{\alpha}_t)$ and its mean. (b) Parameterize the

reverse mean by ϵ_θ and prove the weighted noise-MSE objective. (c) specify a Transformer image denoiser and text conditioning.

中文译述

推导离散扩散前向闭式、后验、ELBO、噪声预测目标，并设计文本条件 Transformer。

完整解答：Q5(a)

递推 $x_t=\sqrt{\alpha_t}x_{t-1}+\sqrt{\beta_t}\epsilon_t$，归纳得 $x_t=\sqrt{\bar\alpha_t}x_0+\sqrt{1-\bar\alpha_t}\epsilon$。ELBO 与本章 DL/RL Q2(d) 相同。高斯条件公式给

$$\tilde\mu_t(x_t,x_0)=\frac{\sqrt{\bar\alpha_{t-1}}\beta_t}{1-\bar\alpha_t}x_0+\frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}x_t,\quad q=N(\tilde\mu_t,\tilde\beta_tI).$$

完整解答：Q5(b)–(c)

用 $x_0=(x_t-\sqrt{1-\bar\alpha_t}\epsilon)/\sqrt{\bar\alpha_t}$ 代入后验均值，得到

$$\mu_\theta(x_t,t)=\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{\beta_t}{\sqrt{1-\bar\alpha_t}}\epsilon_\theta(x_t,t)\right).$$

固定方差时两个高斯的 KL 是均值平方误差；代入上式后仅差与 t 有关权重，故训练为 $\mathbb{E}\|\epsilon-\epsilon_\theta(\sqrt{\bar\alpha_t}x_0+\sqrt{1-\bar\alpha_t}\epsilon,t,T)\|^2$。图像用 patch embedding/latent patch token，加入 time embedding 和位置编码，经 Transformer denoiser；文本由 encoder 产生 token states，图像 token 通过 cross-attention 条件化，也可加 classifier-free guidance 的空文本分支。

17.3 2026 春

理论入口：参见章 十四、十六。

2026 春·NLP·Q1（原卷第 8 页，6 分）

Original. For unit embeddings, InfoNCE with one positive and n negatives is $L=-\log\frac{e^{\langle z,z^+\rangle/\tau}}{e^{\langle z,z^+\rangle/\tau}+\sum_i e^{\langle z,z_i^-\rangle/\tau}}$. (a) compute $\nabla_z L$ and interpret pull/push on the sphere; (b) in high dimension give the scale/concentration of negative inner products and their maximum, and explain false negatives.

中文译述

计算 InfoNCE 对 anchor 的梯度，并用高维球面几何解释负样本与 false negative。

完整解答：Q1

令候选 $u_0 = z^+$、$u_i = z_i^-$，$p_j = e^{z^T u_j/\tau} / \sum_r e^{z^T u_r/\tau}$，则

$$\nabla_z L = \frac{1}{\tau}\left(\sum_{j=0}^{n} p_j u_j - z^+\right).$$

梯度下降把 z 拉向正样本并按 softmax 权重推离负样本；单位归一化反传等价再投影到切空间。若负样本均匀在 S^{d-1}，固定 z 时内积均值 0、方差 $1/d$，集中尺度 $O(d^{-1/2})$；近独立时最大值约 $\sqrt{2\log n/d}$（未饱和区）。false negative 与 z 高相似却被当负样本，获得大 p_i 和强排斥梯度，破坏类内聚集。

理论入口： 参见章 十四、十六。

2026 春·NLP·Q2（原卷第 8–9 页，6 分）

Original. For tickets with tokens w_{dn} and binary priority y_d, design a supervised topic model with $\theta_d \sim Dir(\alpha)$, $\phi_k \sim Dir(\eta)$ and empirical topic frequency $\bar{z}_d$. (a) explain how labels change topics and one vanilla-LDA failure; (b) give $p(y_d|\bar{z}_d, \beta)$ via logistic regression and interpret β; (c) specify sampling and factorize $p(w, z, \Theta, \Phi, y|\alpha, \eta, \beta)$.

中文译述

为带紧急标签的工单建立 supervised LDA，并写完整联合分布。

完整解答：Q2

标签项使后验偏向能预测紧急性的主题；vanilla LDA 可能把高频但与优先级无关的词组织成主题，监督可缓解。$\sigma(u) = 1/(1+e^{-u})$，$p(y_d|\bar{z}_d, \beta) = \sigma(\beta^T \bar{z}_d)^{y_d}[1-\sigma(\beta^T \bar{z}_d)]^{1-y_d}$，$\beta_k$ 表示主题 k 对紧急 log-odds 的影响。生成：取 θ_d；每个 token 取 $z_{dn} \sim Cat(\theta_d)$、$w_{dn} \sim Cat(\phi_{z_{dn}})$；最后按上式取 y_d。联合为

$$\prod_k p(\phi_k|\eta) \prod_d p(\theta_d|\alpha) \left[\prod_n p(z_{dn}|\theta_d) p(w_{dn}|\phi_{z_{dn}})\right] p(y_d|\bar{z}_d, \beta).$$

理论入口： 参见章 十四、十六。

2026 春·NLP·Q3（原卷第 9 页，7 分）

Original. Design a Transformer for $L \gg 10^4$ using one of block-sparse, sliding-window+global tokens, linear attention or recurrent memory. Precisely define computation/pattern and hyperparameters; analyze time/memory; argue long-range expressivity by graph connectivity/diameter or memory propagation.

中文译述

选择一种长上下文机制，给精确定义、复杂度和长距离可达性证明。

完整解答：Q3

选 sliding window + g 个 global tokens。普通位置 i 的邻域 $N(i)=\{j:|i-j|\le w\}\cup G$；global token 对所有 $L+g$ 个位置注意。掩码外 logits 设 $-\infty$。普通查询成本 $O(L(w+g)d)$，global 查询 $O(gLd)$，注意力存储 $O(L(w+g)+gL)$；忽略 FFN。任意两个普通 token 均可经 $i\to$ global $\to j$ 在两层内通信，注意力图直径至多 2；若 $g=0$，传播距离需约 L/w 层。

理论入口：参见章 十四、十六。

2026 春·NLP·Q4（原卷第 9–10 页，7 分）

Original. Build an MLN information-extraction system with unknown $Link(m,e)$ and $Rel(e_1,e_2,r)$, evidence $Cand, Type, HighLink, HighRel$. (a) give weighted rules for propagating local evidence, at-most-one linking and relation-type consistency, with hard/soft weights; (b) describe upstream architectures/objectives and threshold/top-k construction; (c) add predicates/rules using relation evidence and birth years to disambiguate same-name people 500 years apart, with a minimal example.

中文译述

把局部实体链接/关系抽取分数放入 MLN，全局施加唯一性、类型和时代一致性。

完整解答：Q4

(a) 软规则：$HighLink(m,e)\Rightarrow Link(m,e)$；$HighRel(m_1,m_2,r)\wedge Link(m_1,e_1)\wedge Link(m_2,e_2)\Rightarrow Rel(e_1,e_2,r)$。硬约束：$Link(m,e)\wedge Link(m,e')\Rightarrow e=e'$；类型规则如 $Rel(e_1,e_2,WorksFor)\Rightarrow Type(e_1,Person)\wedge Type(e_2,Org)$，若 KB 类型可靠可硬，否则高权软。(b) 实体链接器输入 mention/context 与候选描述，用 bi-encoder 召回、cross-encoder softmax/ranking；关系模型输入两 mention span/context，用 Transformer 分类交叉熵。High 信号由校准概率阈值或 top-k 加 margin 产生。(c) 加 $Birth(e,y)$、$Contemp(e_1,e_2)$；$HighRel(m_1,m_2,r)\wedge Link(m_1,e_1)\wedge Link(m_2,e_2)\Rightarrow Contemp(e_1,e_2)$（权重依赖关系 r）；$Birth(e_1,y_1)\wedge Birth(e_2,y_2)\wedge|y_1-y_2|>\Delta\Rightarrow\neg Contemp(e_1,e_2)$ 为高权/硬规则。例：mention “Li”与同文人物有师生关系，候选分别生于 1200/1700 年，对方生于 1210 年；联合推理选择 1200 年候选。

理论入口：参见章 十四、十六。

2026 春·NLP·Q5（原卷第 10–11 页，7 分）

Original. A dialog QA system writes candidate memory c_t, retrieves R_t, and answers $\hat{y}_t$. (a) With $C_t=1$ if key memory is retrieved, $\Pr(C_t=0)=\epsilon$, $g_c=\mathbb{E}[\Delta_t|C_t=1]$, $g_{\neg c}=\mathbb{E}[\Delta_t|C_t=0]$, derive the necessary and sufficient condition for total gain $\Delta>0$, and when $g_c>g_{\neg c}$ show $\epsilon<g_c/(g_c-g_{\neg c})$; explain large ϵ. (b) formulate memory as an MDP whose action includes write/merge/evict and retrieval choices, with reward trading answer quality and cost.

中文译述

推导检索命中率与记忆收益条件，并把记忆管理/检索联合建模为 MDP。

完整解答：Q5

全期望给 $\Delta = (1-\epsilon)g_c + \epsilon g_{\neg c} = g_c - \epsilon(g_c - g_{\neg c})$。故必要充分条件是该量 > 0；若 $g_c > g_{\neg c}$，等价于题给阈值。大漏检率把权重移向 $g_{\neg c}$，若漏检时记忆干扰为负，稳定正收益很快消失。MDP 状态含对话历史、压缩后的存储、预算和检索统计；动作含写入/合并/压缩/淘汰以及 query rewrite、稀疏/稠密权重、k、rerank。转移由用户下一轮和存储更新决定；奖励可取 $r_t = s(\hat{y}_t, y_t^*) - s(f_\theta(H_t, \varnothing), y_t^*) - \lambda Cost(a_t)$，目标最大化 $\mathbb{E}\sum_t \gamma^t r_t$。离线可用 logged bandit/off-policy 校正，在线需探索约束。

17.4 高频错误、作答模板与复习检查

n-gram 条件方向、平滑归一化和 perplexity 指数不可写错；forward 求总概率、Viterbi 求最大路径；HMM 生成归一化与 CRF 全局条件归一化不同；BERT 与 GPT 目标不同；top-k/top-p/temperature 不可只写名称；系统题必须满足检索、复杂度、证据、时效和安全约束。

复习检查

考场模板：概率模型写分解与动态规划，预训练写 mask/因果目标，解码写候选保留规则，系统题写检索—生成—证据—评测闭环。完成后应能手算 n-gram、PPL、HMM forward/Viterbi、InfoNCE 与解码概率，并解释 RLHF/DPO、知识图谱、RAG 和长上下文设计。

附录 A　按年份反查真题索引

本附录只提供反查入口，不重复题面与解答。页码由交叉引用自动生成。

年份	Machine Learning Theory	Deep Learning and Reinforcement Learning	Optimization Methods in AI	Natural Language Processing
2025 春	第32页	第70页	第96页	第124页
2025 秋	第33页	第72页	第98页	第126页
2026 春	第36页	第74页	第99页	第129页

A.1　原题歧义提示

2025 春 MLT 第 3 题 (b) 由题给信息只能推出 VC 维区间；2025 秋 DL/RL 第 3 题对折扣占用测度的 ‘unnormalized’ 措辞与公式中的归一化因子不一致；2025 秋 NLP 第 3 题 (b) 没有给出自引证据的生成机制；2026 春 Optimization 第 10 题 (b) 的早期步长未必同时满足题设上界。各题解答均先声明数学解释再推导。

附录 B　四 Section 联合索引

交叉主题	主归属	次归属如何调用
SVM 与 KKT	MLT 的 SVM 原始–对偶–KKT	Optimization 提供一般 KKT 与约束资格条件
Transformer	DL/RL 完整结构、缩放、复杂度	NLP 引用并讨论语言建模、预训练和解码
对比学习	DL/RL 的 InfoNCE 与表示学习	NLP 只讨论文本/检索应用
RLHF	DL/RL 的策略梯度与 actor–critic	NLP 的 SFT、偏好、PPO/DPO 流程
随机梯度	Optimization 的条件期望和收敛率	深度学习只调用训练更新, 不重复证明
RAG 与知识图谱	NLP 的检索、生成、证据和推理	其他 Part 仅作为应用引用

索引