AI 博资考：拿到题后的 30 秒识别流程

人工智能博士资格考试–考前速查速记表

0–5 秒：先选方向，不要被题面故事带走

A 机器学习理论：PAC、VC、ERM、稳定性、核、分类器构造。 **B 深度学习与强化学习**：网络、生成模型、Bellman、策略梯度。 **C 优化**：光滑/强凸、prox、随机梯度、递推与复杂度。 **D NLP**：语言模型、embedding、Transformer、概率模型、检索/对齐/系统设计。

模块判断	题目究竟在问“统计保证、表示/生成、优化收敛、语言/知识系统”的哪一个对象？
⇓ 题型判断	定义/选择；计算；证明；算法步骤；复杂度；系统设计。先按分值决定展开深度。
⇓ 写核心对象	风险 $L, \widehat{L}$；目标 $F = f + R$；分布/联合分解；状态动作奖励；张量形状。
⇓ 查假设	i.i.d./可实现？凸/强凸/光滑？无偏/有界方差？核 PSD？可逆？Markov？
⇓ 查维数	每次乘法先写形状；标量目标、梯度同参数形状、Hessian 方阵。
⇓ 收口	最优性：全局/局部、唯一性、KKT 条件；收敛：率 + 噪声底；复杂度：时间 + 空间。

六类题的第一反应

- **定义题**：定义式 + 量词 + 条件 + 一个反例边界。
- **概率界**：单个坏事件 → 指数界 → union bound。
- **VC 题**：下界构造全部标记；上界排除任意更大点集。
- **优化证明链**：fixed point → nonexpansive → 展平方 → 条件期望 → 递推。
- **生成/概率模型**：随机变量 → 联合分解 → 目标 → 训练 → 采样。
- **系统设计**：输入/输出 → 模块 → 目标 → 推理 → 指标/失败模式。

交卷前 20 秒

标签是 $\{0,1\}$ 还是 $\{-1,+1\}$？损失是和还是平均？KL 方向？softmax 轴？占用测度是否归一化？步长条件？复杂度的输入规模？所有公式量纲是否闭合？

材料核对提醒

25S-A-Q3(b)：仅由 VCdim(矩形) = 4、题给 VCdim(三角形) = 7 及并类定理，只能严格推出 $7 \le \mathrm{VCdim}(R \cup T) \le 12$，不能推出精确等于 12。 **25F-A-Q10(c)**：若题面版本写严格 $>$，强凸结论应为 $\ge$，二次函数可取等号。 **25F-B-Q3(a)**：题面把未归一化折扣访问量与概率期望混写；答题先声明 ρ_π 或 $d_\pi = (1-\gamma)\rho_\pi$。

A1–A2 概率、ERM、PAC 与 VC

经验损失方差在 25S、26S 重复；PAC/VC 在三套真题中均为核心。

风险与经验风险

设 $Z_i \sim P$，$S = (Z_1, \dots, Z_m)$：

$$L_P(h) = \mathbb{E}\ell(h, Z), \qquad \widehat{L}_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h, Z_i).$$

固定分类器经验损失方差（高频证明）

【看到】固定 h、零一损失、求 $\mathrm{Var}(\widehat{L})$。
【想到】把错误写成 Bernoulli。令 $U_i = \mathbf{1}[h(X_i) \ne f(X_i)] \sim \mathrm{Bern}(p)$。
【写】$\widehat{L} = m^{-1}\sum_i U_i$，故

$$\mathbb{E}\widehat{L} = p, \qquad \mathrm{Var}(\widehat{L}) = \frac{p(1-p)}{m} \le \frac{1}{4m}.$$

【做】写独立性 $\Rightarrow$ 方差相加；最后说明方差 $O(m^{-1})$、标准差 $O(m^{-1/2})$。
【错】算法输出 $A(S)$ 依赖 S 时，不能把它当固定 h。

Bayes / MLE / MAP

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)}, \quad \hat{\theta}_{\mathrm{MLE}} = \arg\max_\theta p(D \mid \theta),$$

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_\theta \{\log p(D \mid \theta) + \log p(\theta)\}.$$

Bernoulli：$\hat{p} = \frac{1}{n}\sum_i x_i$。 **Gaussian**：$\hat{\mu} = \bar{x}, \hat{\sigma}^2_{\mathrm{MLE}} = n^{-1}\sum_i (x_i - \bar{x})^2$。

概率必错点

MLE $\ne$ MAP；后验均值 $\ne$ MAP。独立 $X \perp Y$ 不等于条件独立 $X \perp Y \mid Z$。$\mathrm{Cov}(X,Y) = 0$ 一般不推出独立（联合 Gaussian 才可）。Gaussian 方差 MLE 分母是 n；$n-1$ 是无偏估计。

Jensen、熵、KL

凸 ϕ：$\phi(\mathbb{E}X) \le \mathbb{E}\phi(X)$；严格凸时等号通常要求 X a.s. 常数。

$$H(p) = -\sum_x p(x)\log p(x), \qquad \mathrm{KL}(p\|q) = \mathbb{E}_p \log\frac{p}{q} \ge 0.$$

KL ≥ 0 证明骨架

$-\log$ 凸：$\mathrm{KL}(p\|q) = \mathbb{E}_p[-\log(q/p)] \ge -\log \mathbb{E}_p(q/p) = -\log 1 = 0$。离散情形须要求 $q > 0$ 于 $p > 0$ 处；等号当且仅当 $p = q$ a.e.

Gaussian 与条件分布速记

若 $\binom{x}{y} \sim \mathcal{N}\left(\binom{\mu_x}{\mu_y}, \begin{smallmatrix}\Sigma_{xx} & \Sigma_{xy}\\ \Sigma_{yx} & \Sigma_{yy}\end{smallmatrix}\right)$，则

$$x \mid y \sim \mathcal{N}(\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}).$$

维数：$x \in \mathbb{R}^p, y \in \mathbb{R}^q$，$\Sigma_{xy} \in \mathbb{R}^{p\times q}$，条件协方差 $p \times p$。

Bias–variance 与数据划分

平方损失下（噪声方差 σ^2）：

$$\mathbb{E}[(Y - \hat{f}(x))^2] = \sigma^2 + \mathrm{Bias}^2[\hat{f}(x)] + \mathrm{Var}[\hat{f}(x)].$$

诊断

训练误差高：容量不足、优化失败、特征/实现错。训练低而验证高：过拟合，增数据/正则/增强/早停。验证集用于选模型；测试集只做最终无偏评估。

【错】regularization 不等于 early stopping；训练损失小不代表 generalization 好。

A2 ERM、PAC、增长函数与 VC 维

三套真题均考；优先掌握定义、样本复杂度、上下界证明量词。

PAC 两种口径

可实现：$f \in \mathcal{H}$，目标 $L_P(\hat{h}) \le \epsilon$。 **Agnostic**：不要求真目标在类中，目标

$$L_P(\hat{h}) \le \inf_{h\in\mathcal{H}} L_P(h) + \epsilon$$

以至少 $1-\delta$ 的训练样本概率成立。

有限类一致学习器

【看到】$\widehat{L}(h) = 0$、$|\mathcal{H}| < \infty$。
对任意坏 h（$L(h) > \epsilon$），$\mathbb{P}(\widehat{L}(h) = 0) = (1 - L(h))^m \le e^{-m\epsilon}$；union bound：

$$\mathbb{P}(\exists \text{坏的一致} h) \le |\mathcal{H}| e^{-m\epsilon}.$$

故 $m \ge \epsilon^{-1}(\log|\mathcal{H}| + \log(1/\delta))$。Agnostic 有界损失通常为 $O((\log|\mathcal{H}| + \log(1/\delta))/\epsilon^2)$。

独立但不同分布（25S）

$X_i \sim D_i$ 独立，$p_i = L_{D_i,f}(h)$，平均风险 $\bar{p} = m^{-1}\sum p_i$。若 $\bar{p} > 1/4$：

$$\mathbb{P}(\widehat{L} = 0) = \prod_i (1 - p_i) \le e^{-\sum p_i} < e^{-m/4}.$$

再对 $h \in \mathcal{H}$ union bound。【错】不得偷换为共同 p，也不得声称样本 i.i.d.

增长函数与 Sauer

$$\tau_{\mathcal{H}}(m) = \max_{x_{1:m}} |\{(h(x_1), \dots, h(x_m)) : h \in \mathcal{H}\}|.$$

$\mathrm{VCdim}(\mathcal{H}) = d$：存在 d 点可实现全部 2^d 标记，任何 $d+1$ 点不可。

$$\tau_{\mathcal{H}}(m) \le \sum_{i=0}^{d}\binom{m}{i} \le \left(\frac{em}{d}\right)^d \quad (m \ge d).$$

典型 VC 泛化量级：可实现 $\tilde{O}((d+\log 1/\delta)/\epsilon)$；agnostic $\tilde{O}((d+\log 1/\delta)/\epsilon^2)$。

Sauer 归纳证明骨架

对最后一点 x_m，把限制类分成 A（至少一个延拓）与 B（两种标记均可延拓）。则 $\tau_H(m) \le \tau_A(m-1)+\tau_B(m-1)$，且 $\mathrm{VCdim}(A) \le d$、$\mathrm{VCdim}(B) \le d-1$。归纳后用 Pascal 恒等式收口。

VC 题标准模板

1. **猜 d**：先画最小几何反例。
2. **下界**：给定一组 d 点，对任意标记构造参数。
3. **上界**：取任意 $d+1$ 点，找一种必不能实现的标记。
4. 两边合并写“故 $\mathrm{VCdim}=d$”。

实考值：threshold 1；interval 2；signed interval 3；同心圆 1；二维轴对齐矩形 4；半空间 $\mathbb{R}^d$ 为 $d+1$；26S 的“区间 $\cup$ 右射线”类为 3。

并类上界（25S）

若 $d_i = \mathrm{VCdim}(\mathcal{H}_i)$，则

$$\mathrm{VCdim}(\mathcal{H}_1 \cup \mathcal{H}_2) \le d_1 + d_2 + 1.$$

思路：反设 $m = d_1+d_2+2$ 被并类打散，则 $2^m \le \tau_{H_1}(m)+\tau_{H_2}(m)$；分别用 Sauer 上界导出矛盾。

A3–A4 稳定性、经典模型与表达性

稳定性、回归、SVM/核、压缩感知及网络表达构成第二组机器学习母题。

Rademacher complexity

固定样本 $S=(x_i)_{i=1}^m$：

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

它衡量拟合随机符号的能力。典型有界损失泛化界：

$$\forall f: \quad L(f) \lesssim \widehat{L}(f) + 2\widehat{\mathfrak{R}}_S(\ell\circ\mathcal{F}) + O\left(\sqrt{\frac{\log(1/\delta)}{m}}\right).$$

若 $\|w\| \le B, \|x_i\| \le R$，线性类复杂度 $\le BR/\sqrt{m}$（维数不显式出现）。

Replace-one stability（25S）

令 $S^{i\leftarrow z'}$ 把 z_i 换成独立 z'。On-average replace-one stability 控制

$$\frac{1}{m}\sum_i \mathbb{E}[\ell(A(S^{i\leftarrow z'}), z_i) - \ell(A(S), z_i)] \le \epsilon(m).$$

泛化差证明骨架

写 $\mathbb{E}L(A(S)) = m^{-1}\sum_i \mathbb{E}_{S,z'}\ell(A(S), z')$；利用 (S, z') 的交换对称性，把测试点 z' 与训练点 z_i 对换；差恰好化为 replace-one 稳定项并求和。

若 $\ell(\cdot, z)$ 凸且 ρ-Lipschitz，目标 $\widehat{L}_S(w) + \frac{\lambda}{2}\|w\|^2$ 为 λ-强凸，典型稳定率 $O(\rho^2/(\lambda m))$。比较 $\hat{w}$ 与任意 w^*：

$$\mathbb{E}L(\hat{w}) \le L(w^*) + \frac{\lambda}{2}\|w^*\|^2 + O\left(\frac{\rho^2}{\lambda m}\right).$$

线性回归与 Ridge（维数必写）

$X \in \mathbb{R}^{m\times d}$，$w \in \mathbb{R}^d$，$y \in \mathbb{R}^m$：

$$J(w) = \frac{1}{2m}\|Xw - y\|^2,$$

$$\nabla J = \frac{1}{m}X^\top(Xw-y), \qquad \nabla^2 J = \frac{1}{m}X^\top X.$$

满列秩时 $\hat{w} = (X^\top X)^{-1}X^\top y$。Ridge：

$$J_\lambda = J + \frac{\lambda}{2}\|w\|^2, \quad \hat{w} = (X^\top X + m\lambda I_d)^{-1}X^\top y.$$

【错】$X^\top X$ 秩不足时 OLS 解不唯一；Ridge 的系数取决于“和/平均、1/2”约定。

Binary logistic regression

高频计算模板

【看到】binary logistic regression。
【想到】Bernoulli likelihood + sigmoid。
【写】$p=\sigma(Xw)$，$\sigma(a)=1/(1+e^{-a})$；

$$L(w) = -\sum_i [y_i \log p_i + (1-y_i)\log(1-p_i)],$$

$$\nabla L = X^\top(p-y) \in \mathbb{R}^d, \qquad \nabla^2 L = X^\top W X \succeq 0,$$

$W = \mathrm{diag}(p_i(1-p_i)) \in \mathbb{R}^{m\times m}$。
【做】likelihood → 负 log → gradient/Hessian → PSD 判凸。
【错】完全分离且无正则时，有限 MLE 可能不存在；logistic 不是 least squares。

模型选择

ERM 最小化训练风险；SRM 最小化“经验风险 + 复杂度罚”。交叉验证只在训练/验证层；预处理统计量（均值、词表等）须只在训练折估计。

数据泄漏

测试集反复调参会把测试集变成验证集。先划分，再做标准化/PCA/特征选择；最终只报告一次 test。

A4 SVM、核、树/近邻、压缩感知与表达性

纲内经典模型；26S 直接考有效核，25F/26S 考神经网络表达构造。

SVM primal ↔ dual

数据 $x_i \in \mathbb{R}^d, y_i \in \{\pm 1\}$。硬间隔：

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

软间隔：加 $\xi_i \ge 0$ 与 $C\sum_i \xi_i$。Lagrangian 消去 w, b：

$$\max_\alpha \sum_i \alpha_i - \frac{1}{2}\sum_{ij}\alpha_i\alpha_j y_i y_j K(x_i, x_j),$$

$$0 \le \alpha_i \le C, \qquad \sum_i \alpha_i y_i = 0, qquad w = \sum_i \alpha_i y_i \phi(x_i).$$

KKT 与 support vector

原始可行、对偶可行、stationarity、complementary slackness：

$$\alpha_i[y_i f(x_i) - 1 + \xi_i] = 0, \qquad (C-\alpha_i)\xi_i = 0.$$

$\alpha_i > 0$ 才是支持向量；$0 < \alpha_i < C \Rightarrow y_i f(x_i) = 1$。预测 $\mathrm{sign}(\sum_i \alpha_i y_i K(x_i, x) + b)$。

【错】kernel 替换的是内积；函数间隔 $y_i f_i$，几何间隔还要除以 $\|w\|$；primal/dual 的正负号与 C, λ 缩放勿混。

有效核与 kernel trick

K 对称，且任意点集的 Gram $G_{ij} = K(x_i, x_j)$ 均 PSD：$c^\top G c \ge 0$。等价于存在（可能无限维）Hilbert 空间特征映射 $K(x,x') = \langle\phi(x), \phi(x')\rangle$；映射不唯一。

只检查 $K(x,x) \ge 0$ 不够。常用闭包：非负线性组合、乘积、点态极限（在适当条件下）仍为核。

树与 k-NN

分类树分裂常最大化 impurity decrease：entropy $H = -\sum_c p_c \log p_c$；Gini $1 - \sum_c p_c^2$。深度越大通常偏差降、方差升；剪枝/最小叶样本控制复杂度。

k-NN：训练近似存储，朴素预测 $O(md)$。小 k 低偏差高方差；大 k 相反。先标准化；高维距离集中导致维数灾难。

压缩感知

$y = Ax_0$，x_0 为 s-稀疏。Basis pursuit：

$$\min_x \|x\|_1 \quad \text{s.t. } Ax = y.$$

NSP(s)：对任意 $0 \ne h \in \ker A$、$|S| \le s$，$\|h_S\|_1 < \|h_{S^c}\|_1$，则所有 s-稀疏 x_0 唯一恢复。噪声时用 $\|Ax - y\|_2 \le \eta$ 或 Lasso。

NSP 证明骨架

任意另一可行解 $x_0 + h$：$\|x_0+h\|_1 \ge \|x_0\|_1 - \|h_S\|_1 + \|h_{S^c}\|_1 > \|x_0\|_1$。

神经网络表达性真题

Sigmoid 通用逼近构造（25F）：将 $[-1,1]^n$ 划为直径 $\le \epsilon/(2\rho)$ 的盒；Lipschitz 保证盒内近似常数；sigmoid 逼近阈值/盒指示；输出层加权盒值；分配离散化与门函数误差。

ReLU 与 sigmoid 交（26S）：有限 ReLU 网络是连续分段仿射；标准 sigmoid 网络表示实解析且有界函数。若两者全局相同，解析函数在某个开集等于仿射函数，恒等定理推广到全域；全域有界仿射只能是常数。

B1–B2 网络训练、视觉与 Attention

25S 考完整分类流水线；25F/26S 考 CNN、backprop、attention 与复杂度。

网络与尺寸

批 $X \in \mathbb{R}^{B\times d_0}$，第 ℓ 层

$$Z^\ell = H^{\ell-1}W^\ell + \mathbf{1}(b^\ell)^\top, \quad H^\ell = \phi_\ell(Z^\ell),$$

$W^\ell \in \mathbb{R}^{d_{\ell-1}\times d_\ell}, b^\ell \in \mathbb{R}^{d_\ell}, H^\ell \in \mathbb{R}^{B\times d_\ell}$。分类输出 logits $Z^L \in \mathbb{R}^{B\times C}$，$P = \mathrm{softmax}(Z^L)$。

Softmax + cross entropy

对单样本 $z \in \mathbb{R}^C$、one-hot y：

$$p_c = \frac{e^{z_c}}{\sum_j e^{z_j}}, \quad \ell = -\sum_c y_c \log p_c, \quad \frac{\partial \ell}{\partial z} = p - y.$$

数值实现用 log-sum-exp：$\log\sum_j e^{z_j} = a + \log\sum_j e^{z_j - a}$，$a = \max_j z_j$。

Backprop 维数模板

令 $\Delta^\ell = \partial J/\partial Z^\ell \in \mathbb{R}^{B\times d_\ell}$：

$$\Delta^L = (P-Y)/B,$$

$$\frac{\partial J}{\partial W^\ell} = (H^{\ell-1})^\top\Delta^\ell \in \mathbb{R}^{d_{\ell-1}\times d_\ell}, \quad \frac{\partial J}{\partial b^\ell} = (\Delta^\ell)^\top\mathbf{1},$$

$$\Delta^{\ell-1} = (\Delta^\ell(W^\ell)^\top)\odot\phi'_{\ell-1}(Z^{\ell-1}).$$

反向传播就是共享中间量的链式法则；时间/空间与一次前向同阶（常数倍）。

激活与梯度

$\sigma'(z) = \sigma(z)(1-\sigma(z)) \le 1/4$；$\tanh'(z) = 1-\tanh^2 z \le 1$；$\mathrm{ReLU}'(z) = \mathbf{1}[z>0]$（$z=0$ 取次梯度约定）。深层 Jacobian 连乘，小于 1 的因子导致 vanishing gradient；ReLU 正区导数 1，但会 dead ReLU。

SGD、动量与训练流程

mini-batch 平均梯度：$g_k = B^{-1}\sum_{i\in\mathcal{B}_k}\nabla\ell_i(\theta_k)$。

$$v_{k+1} = \beta v_k + g_k, \qquad \theta_{k+1} = \theta_k - \eta v_{k+1}.$$

动量在梯度持续方向累积、振荡方向相消。流程：划分数据 → 定义网络/损失 → 前向 → 反传 → 更新 → 验证选超参 → 一次测试。

Batch Normalization

训练时对 mini-batch：

$$\hat{x} = (x-\mu_B)/\sqrt{\sigma_B^2+\varepsilon}, \qquad y = \gamma\hat{x}+\beta.$$

同时更新 running mean/variance；测试时用 running statistics，不用当前测试批统计。参数 γ,β 可训练。

25S 训练诊断满分话术

训练误差高（表达/优化）：加宽/加深、残差连接、更合适激活；调学习率/优化器/初始化；检查梯度与数据。
训练低、测试高（泛化）：数据增强/更多数据、weight decay/dropout、早停、减小容量、交叉验证。至少区分“容量不足”和“过拟合”。

深度学习易错

softmax+CE 梯度是 $p-y$；activation derivative 不得漏；dropout/BN 训练与测试行为不同；regularization 不等于 early stopping；低训练误差不等于低测试误差。

B2 CNN、视觉任务、Attention 与长序列

25F 考 CNN 参数、ViT；26S 考 attention scaling；NLP 真题考长上下文复杂度。

卷积尺寸与参数量

输入 $X\in\mathbb{R}^{B\times H\times W\times C_{in}}$，核 $K\in\mathbb{R}^{k_h\times k_w\times C_{in}\times C_{out}}$：

$$H_{out} = \left\lfloor \frac{H+2p_h-d_h(k_h-1)-1}{s_h} \right\rfloor + 1$$

（宽同理），参数量 $k_hk_wC_{in}C_{out}+C_{out}$。Dense 若把整图映到 $H_{out}W_{out}C_{out}$，参数量为 $(HWC_{in})(H_{out}W_{out}C_{out})$+bias，远大于卷积。

卷积归纳偏置：局部连接、参数共享、平移等变；池化/全局聚合带来近似不变性。

视觉题统一答法

先写输入/输出：分类 C logits；分割 $H\times W\times C$；光流 $H\times W\times 2$；深度 $H\times W$；三维重构为深度/点云/mesh/辐射场。再写监督信号、损失（CE/ℓ_1/photometric/geometric）、多尺度特征、评价指标与遮挡/标定/域移失败模式。

Scaled dot-product attention（维数）

单批写法：$X\in\mathbb{R}^{n\times d}$，

$$Q=XW_Q\in\mathbb{R}^{n\times d_k},\ K=XW_K\in\mathbb{R}^{n\times d_k},\ V=XW_V\in\mathbb{R}^{n\times d_v},$$

$$S = QK^\top/\sqrt{d_k}\in\mathbb{R}^{n\times n},$$

$$\mathrm{Attn}(Q,K,V) = \mathrm{softmax}_{\mathrm{row}}(S)V\in\mathbb{R}^{n\times d_v}.$$

若 q_i,k_j 分量方差约 1，则 $q_i^\top k_j$ 方差 d_k；除 $\sqrt{d_k}$ 保持 logits $O(1)$，防 softmax 饱和和梯度过小。

多头：每头 $d_k=d_v=d_{model}/h$；拼接各头后乘 W_O。必须加位置编码，否则 self-attention 对 token 排列等变，无法区分顺序。复杂度主项时间 $O(n^2d)$、注意力内存 $O(n^2)$。

Minimal Conv–ViT（25F）

图像切 $P\times P$ patch；$N=HW/P^2$。Conv stem 或 stride-P 卷积把 patch 映为 d 维 token；加位置编码/CLS；若干 Transformer blocks；CLS/平均池化后线性分类。卷积负责局部纹理，注意力负责全局交互。

RNN vs Transformer

RNN：$h_t=f(h_{t-1},x_t)$，长度维必须串行，时间 $O(nd^2)$、关键路径 $O(n)$。Transformer 训练时所有 token 并行，关键路径每层 $O(1)$，但 attention $O(n^2d)$。推理自回归仍逐 token；KV cache 避免重复计算历史 K/V，空间随层数和长度线性增。

长序列设计模板（26S-D）

滑窗 + g 个 global tokens：

$$\mathcal{N}(i) = \begin{cases} \{j:|i-j|\le w\}\cup G, & i\notin G,\\ \{1,\dots,L\}, & i\in G. \end{cases}$$

时间/注意力内存 $O(L(w+g)d)$ / $O(L(w+g))$；任意两个局部 token 经 global token 两层可交互，图直径 ≤ 2（需说明 global 更新与局部读写）。

Attention 易错

$QK^\top$ 是 query 行、key 列；softmax 通常沿 key 维；causal mask 禁止看未来；输出不是 $n\times n$ 而是 $n\times d_v$；长序列方案必须同时写精确连接、复杂度和跨远距离路径。

B3–B4 生成模型、Flow 与强化学习

三套 B 卷连续考生成模型：25S/26S 分别考 score SDE 与 flow matching。

统一识别

显式精确密度：autoregressive、normalizing flow。 **下界**：VAE、diffusion（层次 VAE）。 **隐式**：GAN。先写训练目标，再写采样过程和失败模式。

MLE ⇔ forward KL

$$\mathrm{KL}(p_{data}\|p_\theta) = \mathbb{E}_{p_{data}}\log p_{data}(x) - \mathbb{E}_{p_{data}}\log p_\theta(x).$$

第一项与 θ 无关，故最大化 likelihood 等价于最小化 $\mathrm{KL}(p_{data}\|p_\theta)$。【错】方向不能反。

GAN

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

固定 G 时 $D^*(x) = p_{data}/(p_{data}+p_g)$；代回得到常数加 $2\,\mathrm{JS}(p_{data}\|p_g)$。失败：mode collapse、梯度饱和、博弈振荡。生成器常用 non-saturating loss $-\mathbb{E}_z\log D(G(z))$。

VAE / ELBO

$p_\theta(x,z)=p(z)p_\theta(x\mid z)$，引入 $q_\phi(z\mid x)$：

$$\log p_\theta(x) = \mathrm{ELBO}(x) + \mathrm{KL}(q_\phi(z\mid x)\|p_\theta(z\mid x)),$$

$$\mathrm{ELBO}(x) = \mathbb{E}_q\log p_\theta(x\mid z) - \mathrm{KL}(q_\phi(z\mid x)\|p(z)).$$

重构项保真；prior KL 规范潜空间；ELBO gap 是后验近似误差。重参数化：$z=\mu_\phi(x)+\sigma_\phi(x)\odot\epsilon$，$\epsilon\sim\mathcal{N}(0,I)$。

Normalizing flow

$z_0\sim p_0$，$z_\ell=f_\ell(z_{\ell-1})$，$x=z_L$，各 f_ℓ 双射同维：

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

训练最小化数据平均 NLL；推断密度用逆变换。【错】Jacobian 符号取决于写 forward 还是 inverse，必须自洽。

DDPM 高频闭式

$\beta_t=1-\alpha_t$，$\bar\alpha_t=\prod_{s=1}^t\alpha_s$：

$$q(x_t\mid x_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}x_{t-1},\beta_tI),$$

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

所以可一步采样 $x_t=\sqrt{\bar\alpha_t}x_0+\sqrt{1-\bar\alpha_t}\epsilon$。

后验 $q(x_{t-1}\mid x_t,x_0)=\mathcal{N}(\tilde\mu_t,\tilde\beta_tI)$：

$$\tilde\beta_t = \frac{1-\bar\alpha_{t-1}}{1-\bar\alpha_t}\beta_t,$$

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

反向模型 $p_\theta(x_{t-1}\mid x_t)=\mathcal{N}(\mu_\theta,\sigma_t^2I)$。噪声参数化

$$\mu_\theta = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar\alpha_t}}\epsilon_\theta(x_t,t)\right),$$

简化损失 $\mathbb{E}\left\|\epsilon-\epsilon_\theta(\sqrt{\bar\alpha_t}x_0+\sqrt{1-\bar\alpha_t}\epsilon,t)\right\|^2$。

生成模型易错

VAE 最大化 ELBO，训练 loss 常取负；$\mathrm{KL}(q\|p)$ 不可换向；flow 要可逆且同维；扩散的 α_t 与 $\bar\alpha_t$ 勿混；训练从真实 x_0 加噪，采样从 $x_T\sim\mathcal{N}(0,I)$ 逐步反推。

B4 Score/Flow Matching 与强化学习

25S 考 VP-SDE；26S 考 flow matching、Fokker–Planck、Bellman/actor–critic。

Reverse-time SDE 与 score

前向 $dx=f(x,t)dt+g(t)dW_t$，反向时间（$dt<0$ 的约定）为

$$dx = [f(x,t)-g(t)^2\nabla_x\log p_t(x)]dt + g(t)d\bar W_t.$$

未知 score $\nabla_x\log p_t(x)$ 用 $s_\theta(x,t)$ 拟合。

VP-SDE：$dx=-\frac{1}{2}\beta(t)xdt+\sqrt{\beta(t)}dW_t$。令 $a(t)=\exp[-\frac{1}{2}\int_0^t\beta(s)ds]$，则

$$p_{t|0}(x_t\mid x_0) = \mathcal{N}(a(t)x_0,(1-a(t)^2)I),$$

$$\nabla_{x_t}\log p_{t|0} = -\frac{x_t-a(t)x_0}{1-a(t)^2} = -\frac{\epsilon}{\sqrt{1-a(t)^2}}.$$

Denoising score matching 等价证明

展开 $\mathbb{E}\|s_\theta-\nabla\log p_t\|^2$；用 Fisher identity $\mathbb{E}[\nabla\log p_{t|0}(x_t\mid x_0)\mid x_t]=\nabla\log p_t(x_t)$；交叉项相同，两个目标仅差与 θ 无关常数。最终回归 conditional score。

Flow matching（26S）

给 coupling $(x_0,x_1)\sim\gamma$，线性插值 $x_t=(1-t)x_0+tx_1$，条件速度 $u_t=x_1-x_0$。人口速度

$$v_t(x) = \mathbb{E}[u_t\mid x_t=x].$$

回归目标 $\mathbb{E}_{t,\gamma}\|v_\theta(t,x_t)-u_t\|^2$ 的平方损失最优解是条件期望 v_t。其密度满足连续性方程

$$\partial_t \pi_t + \nabla\cdot(\pi_t v_t) = 0.$$

Fokker–Planck：SDE $dx = b_t dt + \sqrt{2\beta_t} dW_t$ 满足 $\partial_t\pi = -\nabla\cdot(\pi b) + \beta_t\Delta\pi$。取 $b = v + \beta\nabla\log\pi$，扩散项与 score drift 抵消，边缘仍满足同一连续性方程。

Bellman：先分 expectation / optimality

$$V^\pi(s) = \mathbb{E}_{a\sim\pi,s'\sim P}[r(s,a)+\gamma V^\pi(s')],$$
$$Q^\pi(s,a) = r(s,a)+\gamma\mathbb{E}_{s'}\mathbb{E}_{a'\sim\pi}Q^\pi(s',a'),$$
$$V^*(s) = \max_a\{r(s,a)+\gamma\mathbb{E}_{s'}V^*(s')\}.$$

Bellman optimality operator 在 $\|\cdot\|_\infty$ 下是 γ-压缩，故唯一不动点、value iteration 收敛。

PI / VI / TD 控制

Policy iteration：精确/迭代评估 V^{π_k}，再 $\pi_{k+1}(s)\in\arg\max_a[r+\gamma PV^{\pi_k}]$。Value iteration：每次直接做 optimality backup。

$$Q\leftarrow Q+\alpha[r+\gamma\max_{a'}Q(s',a')-Q(s,a)]\quad\text{(off-policy)},$$

$$\text{SARSA: } Q\leftarrow Q+\alpha[r+\gamma Q(s',a')-Q(s,a)]\quad\text{on-policy}.$$

Policy gradient / actor–critic

$$\nabla_\theta J = \mathbb{E}_\pi\sum_t \gamma^t\nabla_\theta\log\pi_\theta(a_t\mid s_t)Q^\pi(s_t,a_t).$$

任意 state-only baseline 不改期望，因为 $\mathbb{E}_{a\sim\pi}\nabla\log\pi(a\mid s)b(s) = b(s)\nabla\sum_a\pi(a\mid s) = 0$。取 $b=V^\pi$ 得 advantage、降方差。Actor 更新 policy，critic 用 TD error $\delta = r+\gamma V_\phi(s')-V_\phi(s)$ 拟合价值。

RL 易错

Bellman expectation ≠ optimality；Q-learning off-policy，SARSA on-policy；value iteration ≠ policy iteration；baseline 不能依赖当前动作（除非另做零均值修正）；声明折扣访问分布是否归一化。

C1–C2 凸分析与 Proximal SGD

三套优化卷均从定义进入 fixed point、非扩张、条件期望、线性收缩与噪声底。

凸集 / 凸函数

C 凸：$\theta x+(1-\theta)y\in C$，$\theta\in[0,1]$。$f$ 凸：

$$f(\theta x+(1-\theta)y)\le\theta f(x)+(1-\theta)f(y).$$

可微等价一阶支撑：$f(y)\ge f(x)+\langle\nabla f(x),y-x\rangle$；二阶连续可微时，凸 $\Leftrightarrow\nabla^2 f(x)\succeq 0$ 于凸开域。

严格凸：不同点、$\theta\in(0,1)$ 取严格不等式，保证至多一个极小点。**强凸**：

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

$f(w)=\lambda\|w\|^2$ 是 2λ-强凸；$\frac{\lambda}{2}\|w\|^2$ 是 λ-强凸。

L-smooth 与核心不等式

$$\|\nabla f(x)-\nabla f(y)\|\le L\|x-y\|.$$

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

凸且 L-smooth：cocoercivity

$$\langle\nabla f(x)-\nabla f(y),x-y\rangle\ge\frac{1}{L}\|\nabla f(x)-\nabla f(y)\|^2.$$

μ-强凸且 L-smooth 的联合插值：

$$\langle\Delta g,\Delta x\rangle\ge\frac{\mu L}{\mu+L}\|\Delta x\|^2+\frac{1}{\mu+L}\|\Delta g\|^2.$$

25S 证明骨架

令 $g(x)=f(x)-\frac{\mu}{2}\|x\|^2$。由强凸，$g$ 凸；若 Hessian 口径，$0\preceq\nabla^2 g\preceq(L-\mu)I$，故 $(L-\mu)$-smooth。对 g 用 cocoercivity，代入 $\nabla g=\nabla f-\mu x$，展开并整理即得联合插值。

存在、唯一与驻点

连续 coercive（$\|x\|\to\infty\Rightarrow f(x)\to\infty$）给存在；强凸给至多一个最优解，二者合并得唯一。无约束可微凸：$\nabla f(x^*)=0\Leftrightarrow x^*$ 全局最优。

凸性必错点

stationary point 一般不一定全局最优；凸不等于严格凸；二阶可微时是 Hessian PSD 对应凸，不是“determinant ≥ 0”；强凸常数受 1/2 约定影响。

次微分与 prox

$g\in\partial f(x)$ 若 $f(y)\ge f(x)+\langle g,y-x\rangle$ 对所有 y。凸 f：$0\in\partial f(x^*)\Leftrightarrow x^*$ 最优。单调性：$g_x\in\partial f(x),g_y\in\partial f(y)$ 时 $\langle g_x-g_y,x-y\rangle\ge 0$。

$$\mathrm{prox}_{\gamma R}(u)=\arg\min_x\left\{R(x)+\frac{1}{2\gamma}\|x-u\|^2\right\},$$

$$x^+=\mathrm{prox}_{\gamma R}(u)\Longleftrightarrow(u-x^+)/\gamma\in\partial R(x^+).$$

KKT（低频但必须会条件）

问题 $\min f(x)$ s.t. $g_i(x)\le 0, h_j(x)=0$。Lagrangian $\mathcal{L}=f+\sum_i\lambda_i g_i+\sum_j\nu_j h_j$。KKT：primal feasibility；$\lambda_i\ge 0$；stationarity $\nabla_x\mathcal{L}=0$；complementary slackness $\lambda_i g_i(x)=0$。凸问题 + Slater 时 KKT 对最优性充要（适当正则性）。

KKT 不是无条件充要；必须说明凸性/约束资格。非凸时通常只是一阶必要条件。

C2 GD、Proximal SGD 与噪声底证明链

25S、26S 几乎同一母题；这是优化卷最高优先级。

GD 三种结论

$x_{k+1}=x_k-\gamma\nabla f(x_k)$。$L$-smooth 且 $\gamma\le 1/L$：

$$f(x_{k+1})\le f(x_k)-\frac{\gamma}{2}\|\nabla f(x_k)\|^2.$$

非凸：平均梯度范数平方 $O(1/K)$；凸：函数值 $O(1/K)$；μ-强凸：线性率，条件数 $\kappa=L/\mu$ 决定难度。

Newton：$x^+=x-[\nabla^2 f(x)]^{-1}\nabla f(x)$，局部快但每步解线性系统，Hessian 非正定时未必下降。SGD 用无偏随机梯度降低每步成本，但有方差。

Prox fixed point 与 nonexpansive

复合 $F=f+R$，f 光滑强凸，R proper closed convex。最优性

$$0\in\nabla f(x^*)+\partial R(x^*)\Longleftrightarrow x^*=\mathrm{prox}_{\gamma R}(x^*-\gamma\nabla f(x^*)).$$

由 prox 最优性取 $r_u,r_v\in\partial R$，次微分单调：

$$\begin{aligned}\|\mathrm{prox}(u)-\mathrm{prox}(v)\|^2&\le\langle\mathrm{prox}(u)-\mathrm{prox}(v),u-v\rangle\\&\le\|\mathrm{prox}(u)-\mathrm{prox}(v)\|\,\|u-v\|,\end{aligned}$$

故 nonexpansive。

Proximal SGD 一步递推

$g^k=\nabla f(x^k)+\xi^k$，$\mathbb{E}_k\xi^k=0$，$\mathbb{E}_k\|\xi^k\|^2\le\sigma^2$：

$$x^{k+1}=\mathrm{prox}_{\gamma R}(x^k-\gamma g^k).$$

与 fixed point 比较并用 nonexpansive：

$$\|x^{k+1}-x^*\|^2\le\|x^k-x^*-\gamma(\Delta\nabla f+\xi^k)\|^2.$$

条件期望，交叉项因零均值消失：

$$\mathbb{E}_k\|x^{k+1}-x^*\|^2\le\|x^k-x^*-\gamma\Delta\nabla f\|^2+\gamma^2\sigma^2.$$

联合插值 → 线性收缩

取 $\gamma=2/(L+\mu)$，展开平方并代入上一页联合插值：

$$\mathbb{E}\|x^{k+1}-x^*\|^2\le(1-\rho)\mathbb{E}\|x^k-x^*\|^2+\gamma^2\sigma^2,$$

$$\rho=\frac{4\mu L}{(L+\mu)^2}.$$

递推展开：

$$e_k\le(1-\rho)^k e_0+\frac{\gamma^2\sigma^2}{\rho}=(1-\rho)^k e_0+\frac{\sigma^2}{\mu L}.$$

常步长线性靠近最优点附近后停在 noise floor；减小 γ 降噪声底但收敛慢。

方差分解

对随机向量 Z 与定值 a：

$$\mathbb{E}\|Z-a\|^2=\|\mathbb{E}Z-a\|^2+\mathbb{E}\|Z-\mathbb{E}Z\|^2.$$

本题把 $Z=\Delta\nabla f+\xi$，条件于 x^k；无偏保证混合项为零。

迭代平均（26S）

$\bar{x}^k=k^{-1}\sum_{t=1}^k x^t$。若误差 $e_t=x^t-\mathbb{E}x^t$ 两两不相关且 $\mathbb{E}\|e_t\|^2\le V$：

$$\mathbb{E}\|\bar{e}_k\|^2=\frac{1}{k^2}\sum_t\mathbb{E}\|e_t\|^2\le\frac{V}{k}.$$

对 $R=0$、$\gamma_t=1/(\mu t)$，结合一步函数值界、求和与 Jensen，典型 $\mathbb{E}[f(\bar{x}^k)-f^*]=O(\log k/k)$。

优化证明链易错

条件期望应以当前 iterate 为条件；无偏不等于零方差；不要在取期望前擅自删随机交叉项；noise floor 要写来源与步长依赖；prox 的 fixed point 用 $\nabla f(x^*)$，不是假设它为 0。

C3 Mini-batch、Expected Smoothness、坐标/动量/自适应

25F 整卷围绕有限和采样；纲内还要求坐标下降、动量与自适应学习率。

有限和与 importance sampling

$f(x)=n^{-1}\sum_{i=1}^n f_i(x)$，采样 s 满足 $\mathbb{P}(s=i)=q_i$：

$$g_1(x)=\frac{1}{nq_s}\nabla f_s(x),\qquad\mathbb{E}g_1(x)=\nabla f(x).$$

独立 mini-batch $s_1,\dots,s_\tau$：$g_\tau=\tau^{-1}\sum_t(nq_{s_t})^{-1}\nabla f_{s_t}$。**平均，不是求和。**

若每个 f_i 凸且 L_i-smooth，则真题常数

$$A''(\tau,q)=\frac{1}{\tau}\max_i\frac{L_i}{nq_i}+\left(1-\frac{1}{\tau}\right)L,$$

$$\mathbb{E}\|g(x)-g(y)\|^2\le 2A''D_f(x,y).$$

证明：展开平方；对角项用 $\|\nabla f_i(x)-\nabla f_i(y)\|^2 \le 2L_i D_{f_i}(x,y)$；交叉项因独立化为 $\|\nabla f(x)-\nabla f(y)\|^2 \le 2LD_f(x,y)$。

采样与方差

最小化 $\max_i L_i/(nq_i)$ 得 $q_i^* = L_i/\sum_j L_j$，值 $n^{-1}\sum_i L_i$；uniform 的值 $\max_i L_i$。最优点噪声

$$\mathrm{Var}[g_\tau(x^*)] = \frac{1}{\tau}\,\mathrm{Var}\left[\frac{1}{nq_s}\nabla f_s(x^*)\right].$$

增加 τ 使 expected-smoothness 常数从单样本值插值到 L、方差按 $1/\tau$ 降，但每步算力近似乘 τ。

AC inequality 识别

若 $\mathbb{E}[\|g^k-\nabla f(x^*)\|^2 \mid x^k] \le 2AD_f(x^k,x^*)+C$，则步长常要求 $\gamma \le 1/A$，递推形如 $\mathbb{E}\|x^k-x^*\|^2 \le (1-\gamma\mu)^k\|x^0-x^*\|^2+\gamma C/\mu$。

随机坐标下降

选坐标 i_k，$x^{k+1} = x^k - \eta_i e_i \nabla_i f(x^k)$。若坐标光滑常数 L_i，常取 $\eta_i = 1/L_i$，采样可与 L_i 成比例。每步只算一列/一个 block；比较算法必须同时报告“迭代数 × 单步代价”。

动量、Nesterov、自适应

Heavy-ball：$v_{k+1} = \beta v_k + \nabla f(x_k)$，$x_{k+1} = x_k - \eta v_{k+1}$。Nesterov 在 look-ahead 点求梯度；凸光滑可达 $O(1/k^2)$（需相应参数/证明条件），强凸可加速到 $O((1-1/\sqrt{\kappa})^k)$ 量级。

AdaGrad：$G_k = \sum_{t\le k} g_t \odot g_t$，$x_{k+1} = x_k - \eta g_k/(\sqrt{G_k}+\epsilon)$。RMSProp 用指数二阶矩；Adam：

$$m_k = \beta_1 m_{k-1} + (1-\beta_1)g_k, \quad v_k = \beta_2 v_{k-1} + (1-\beta_2)g_k^2,$$
$$\hat m_k = m_k/(1-\beta_1^k), \quad \hat v_k = v_k/(1-\beta_2^k),$$
$$x_{k+1} = x_k - \eta \hat m_k/(\sqrt{\hat v_k}+\epsilon).$$

AdamW 将 weight decay 与梯度中的 ℓ_2 正则解耦。

复杂度速记

GD 强凸光滑：$O(\kappa\log(1/\epsilon))$ 次梯度；Nesterov：$O(\sqrt{\kappa}\log(1/\epsilon))$；凸光滑 GD：$O(1/\epsilon)$；加速：$O(1/\sqrt{\epsilon})$。随机法结论必须写是梯度范数、函数值还是距离，以及期望/高概率。

材料与结论边界

Adam 的常用更新式不等于无条件收敛定理；动量也可能在步长过大时发散。Expected smoothness 的常数依赖采样和 batch；“方差降 $1/\tau$”需要独立/不相关等条件。

D1–D2 语言模型、Embedding、Transformer 与对齐

25S/25F/26S 覆盖统计建模、表示、架构复杂度、Scaling Law 与 RLHF。

N-gram MLE 与平滑

Markov 假设：$p(w_t \mid w_{<t}) \approx p(w_t \mid w_{t-n+1:t-1})$。

$$\hat p_{\mathrm{MLE}}(w_3 \mid w_1,w_2) = \frac{C(w_1,w_2,w_3)}{C(w_1,w_2)}.$$

Add-k：$(C(h,w)+k)/(C(h)+k|V|)$；更好答案写 backoff/interpolation：

$$p(w \mid h) = \lambda_3 p_{ML}(w \mid w_1,w_2) + \lambda_2 p_{ML}(w \mid w_2) + \lambda_1 p_{ML}(w),$$

$\lambda_i \ge 0, \sum\lambda_i = 1$，并提 Kneser–Ney/held-out 估权。OOV 用训练阶段确定的 UNK/subword；不能在 test 重新建词表。

Perplexity：$\mathrm{PPL} = \exp[-N^{-1}\sum_t \log p(w_t \mid w_{<t})]$；只在同 tokenization/词表/数据上可比。

表示学习第一句话

把离散对象映到连续向量，使关系/相似性由几何结构表达；训练信号来自上下文共现、随机游走、图边或知识图谱三元组。

四种 embedding 三句模板

GloVe：全局词–上下文共现矩阵；加权最小二乘拟合 $w_i^\top \tilde w_j + b_i + \tilde b_j \approx \log X_{ij}$；偏置是共现比率编码语义。
Skip-gram：局部中心词预测上下文；softmax/negative sampling；分布式假设“上下文相似则语义相似”。
Node2Vec：图上带偏的随机游走生成 node sequences，再做 Skip-gram；p,q 控制 BFS-like 同质性与 DFS-like 结构角色。
TransE：知识图谱正负三元组 ranking/contrastive loss；关系为平移 $e_h + e_r \approx e_t$，对一对多/多对多有限制。

Skip-gram negative sampling

正对 (w,c) 与噪声 c_j^-：

$$\ell = -\log\sigma(v_c^\top u_w) - \sum_{j=1}^{K}\log\sigma(-v_{c_j^-}^\top u_w).$$

【想到】正样本拉近，负样本推远；输入/输出 embedding 是两套参数，使用时可选其一或组合。

InfoNCE（26S-D）

令 $a_0 = z^+$、$a_i = z_i^-$，$p_j = \exp(z^\top a_j/\tau)/\sum_k \exp(z^\top a_k/\tau)$：

$$\nabla_z \mathcal{L} = \frac{1}{\tau}\left(\sum_{j=0}^{n} p_j a_j - z^+\right).$$

单位球上还需考虑归一化诱导的切空间投影。各向同性负样本在高维中 $z^\top z_i^- \approx \mathcal{N}(0,1/d)$，最大值约 $\sqrt{2\log n/d}$。False negative 相似度高、softmax 权重大，会把语义同类强行推开。

LDA / supervised topic model

LDA：$\theta_d \sim \mathrm{Dir}(\alpha)$，$\phi_k \sim \mathrm{Dir}(\eta)$；每 token $z_{dn} \sim \mathrm{Cat}(\theta_d)$，$w_{dn} \sim \mathrm{Cat}(\phi_{z_{dn}})$。监督版再加

$$p(y_d \mid \bar z_d, \beta) = \sigma(\beta^\top \bar z_d)^{y_d}[1-\sigma(\beta^\top \bar z_d)]^{1-y_d}.$$

联合分解必须包含 priors、所有 token 项和 label 项；监督使 topics 同时解释词并预测标签，可避免高频但与任务无关的主题占主导。

概率模型易错

先画/写随机变量再分解联合；条件独立来自图结构，不是常识。Dirichlet 生成 multinomial 参数，不直接生成词。$\bar z_d$ 是 topic frequency，维数 K。

D2 Transformer、解码、Scaling Law 与对齐

大纲明确；三套真题均从原理、复杂度或系统角度考。

Transformer block 答题顺序

tokens → embedding + position → (masked) multi-head attention → residual + norm → FFN → residual + norm → logits。Pre-LN 常写

$$H' = H + \mathrm{MHA}(\mathrm{LN}(H)), \quad H'' = H' + \mathrm{FFN}(\mathrm{LN}(H')).$$

FFN 逐 token：$\mathrm{FFN}(h) = W_2\phi(W_1 h + b_1) + b_2$。Decoder causal mask：$M_{ij} = -\infty$ 若 $j > i$。

解码：质量/多样性/代价

Greedy 每步 argmax，快但易局部最优/重复；beam 保留 B 条高分序列，适合确定性任务但可能偏短、缺多样性；top-k 固定候选数；top-p 取最小集合 S 使 $\sum_{w\in S} p(w) \ge p_0$ 后重归一化，候选数随分布尖锐度自适应，开放式生成通常更自然。temperature：$p_i \propto e^{z_i/T}$；$T\downarrow$ 更尖锐。

解码易错

top-p 不是“概率大于 p 的 token”；它是累计质量阈值。开放式生成偏好 sampling 不代表所有任务都优于 beam；翻译等低熵目标可能仍用 beam。

Scaling law

经验形式 $L(P,T) = L_\infty + AP^{-\alpha} + BT^{-\beta}$。Log–log 上单项主导时斜率为 $-\alpha$ 或 $-\beta$；另一项/不可约损失会形成平台。统计学习解释：数据增多降低估计误差（复杂度项常约 $O(m^{-1/2})$，具体指数依分布/模型/优化），不能为方便假设 Gaussian 后宣称普适定律。

比较 LSTM/Transformer 时只在“相同 tokenization、上下文、优化、算力口径”下谈 A,B,α,β；说明长依赖、并行吞吐、归纳偏置/样本效率。经验拟合参数不是由架构名称必然决定，结论应带实验条件。

SFT、Reward Model、RLHF

SFT：教师强制最小化 $-\sum_t \log\pi_\theta(y_t \mid x, y_{<t})$。偏好数据 (x,y^+,y^-)，Bradley–Terry reward loss：

$$\mathcal{L}_{RM} = -\log\sigma(r_\psi(x,y^+) - r_\psi(x,y^-)).$$

Reward-model-based RLHF：采样 $y\sim\pi_\theta$，优化

$$J(\theta) = \mathbb{E}[r_\psi(x,y)] - \beta\mathbb{E}_x\,\mathrm{KL}(\pi_\theta(\cdot\mid x)\|\pi_{ref}(\cdot\mid x)).$$

score-function 梯度（把序列视为动作）：$\mathbb{E}[(r - \beta\,\mathrm{KL\ signal} - b)\nabla_\theta \log\pi_\theta(y\mid x)]$；PPO 用 clipped ratio 限制更新。

DPO

令 $\Delta_\theta = \log\frac{\pi_\theta(y^+|x)}{\pi_{ref}(y^+|x)} - \log\frac{\pi_\theta(y^-|x)}{\pi_{ref}(y^-|x)}$：

$$\mathcal{L}_{DPO} = -\log\sigma(\beta\Delta_\theta).$$

直接提升 preferred 相对 dispreferred 的参考归一化 log-ratio，无显式 reward model/online RL。

对齐局限与改进

Reward hacking、偏好偏差/标注噪声、分布外失真、KL 过强欠优化/过弱漂移。回答改进：数据审计与多样偏好、uncertainty/ensemble reward、约束/拒答、在线评估、红队、KL/长度偏置校准。不要把“对齐”写成事实正确性的保证。

Embedding 知识库推理

Dense retriever：$s(q,d) = f(q)^\top g(d)$；训练用对比损失/难负样本。知识图谱可用 TransE/双线性/图神经网络；局限：近似几何关系、组合推理弱、实体更新与可解释性不足。改进：结构约束、多跳检索、cross-encoder rerank、符号规则/MLN 与校准。

D3 MLN、RAG、长记忆与系统设计模板

25S/25F/26S 的 NLP 大题稳定要求“模型 + 目标 + 推理 + 失败模式”。

Markov Logic Network

加权一阶逻辑公式 (ϕ_j, w_j)：

$$P(X \mid E) = \frac{1}{Z(E)} \exp\left(\sum_j w_j n_j(X, E)\right).$$

每个 ground atom 是节点；同一 ground formula 中共同出现的 atoms 形成 clique/potential。$w_j > 0$ 鼓励满足，$w_j < 0$ 惩罚满足；硬约束可视为无限权重/直接排除不可行世界。MAP：最大化 $\sum_j w_j n_j$，不必计算 Z。

实体链接 + 关系抽取规则模板

局部证据：$HighLink(m, e) \Rightarrow Link(m, e)$；若 $HighRel(m_1, m_2, r) \wedge Link(m_1, e_1) \wedge Link(m_2, e_2)$，则 $Rel(e_1, e_2, r)$。
唯一性硬约束：$e_1 \neq e_2 \Rightarrow \neg(Link(m, e_1) \wedge Link(m, e_2))$。
类型一致：$Rel(e_1, e_2, worksFor) \Rightarrow Type(e_1, Person) \wedge Type(e_2, Org)$。
同代约束：关系证据 $\Rightarrow Contemporary(e_1, e_2)$；若 $|Birth(e_1) - Birth(e_2)| > \Delta$，强烈惩罚 Contemporary。硬规则留给逻辑不可能，模型/KB 噪声相关规则设软权重。

上游 $HighLink$：bi-encoder 召回 + cross-encoder 打分，CE/ranking loss，阈值/top-k+margin。$HighRel$：实体对上下文 encoder 输出 relation softmax/multilabel sigmoid，以阈值和校准概率转证据。

RAG 无标注领域问答（25S）

1. 文档清洗、chunk + metadata；离线 embedding，建 ANN/BM25 索引。
2. Query rewrite；hybrid retrieve；cross-encoder rerank；保留出处。
3. 将 top-k context + 问题输入 LLM，提示仅据证据回答并引用；证据不足则拒答。
4. 无 QA 标签时自监督：相邻段/标题生成 query、inverse cloze、对比学习；可用人工小集校准。
5. 评估 retrieval recall@k/MRR 与 answer EM/F1/faithfulness，分开定位召回和生成错误。

【错】只“把领域语料继续预训练”不保证可追溯事实；chunk 太小断上下文，太大稀释检索；必须防 prompt injection 与过期知识。

对话外部记忆（26S）

收益 $\Delta_t = s(f(H_t, R_t), y_t^*) - s(f(H_t, \emptyset), y_t^*)$。若 $\mathbb{P}(C_t = 0) = \epsilon$：

$$\Delta = (1 - \epsilon) g_c + \epsilon g_{\neg c} > 0.$$

若 $g_c > g_{\neg c}$，等价于 $\epsilon < g_c/(g_c - g_{\neg c})$（分母正；还要结合 $g_{\neg c}$ 符号解释是否为有效 $[0, 1]$ 阈值）。
MDP：state $s_t = (H_t, M_t$，预算、检索质量信号 $)$；action 同时含 write/merge/evict 与 query rewrite、hybrid weights、k、rerank；reward $r_t = \Delta_t - \lambda Cost(a_t)$；目标 $\max_\pi \mathbb{E} \sum_t \gamma^t r_t$。

概率/图模型系统题答法

1 随机变量：observed/latent、尺寸、support。 **2 生成过程**：按拓扑顺序逐步写 distribution。 **3 联合分解**：products over documents/-topics/tokens。 **4 学习**：MLE/MAP/variational/Gibbs，写目标。 **5 推理**：posterior/MAP/采样。 **6 失败**：identifiability、稀疏数据、近似偏差、复杂度。

系统设计评分点

输入输出要可测；每个模块给模型和 objective；训练/推理分开；复杂度给主导项；消融验证每个部件；评价含质量、延迟、内存、鲁棒/安全；最后给两个具体 failure modes 与 fallback。

NLP 易错

MLN 权重作用于满足 grounding 数，负权重不是“规则取反”；MAP 不需 partition function。RAG 的 retrieval recall 是回答上限之一，但召回成功仍可能 hallucinate。系统题不要只列模型名，必须写数据流和训练目标。

终极公式速查 1/3：概率、学习理论与经典 ML

最后三页可独立复习：先认对象，再抄公式，再核条件。

概率 / 信息论

$$p(\theta|D) \propto p(D|\theta)p(\theta), \quad \hat{\theta}_{MAP} = \arg\max(\log p(D|\theta) + \log p(\theta)).$$

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n,$$

$$\mathbb{E}\|Z - a\|^2 = \|\mathbb{E}Z - a\|^2 + \mathbb{E}\|Z - \mathbb{E}Z\|^2.$$

$$H(p) = -\mathbb{E}_p \log p, \quad H(p, q) = -\mathbb{E}_p \log q = H(p) + \mathrm{KL}(p\|q).$$

Jensen：凸 ϕ，$\phi(\mathbb{E}X) \le \mathbb{E}\phi(X)$。$\mathrm{KL} \ge 0$；等号 $p = q$ a.e.

ERM / PAC / VC

$$L(h) = \mathbb{E}\ell(h, Z), \qquad \hat{L}(h) = m^{-1} \sum_i \ell(h, Z_i),$$

$$\hat{h}_{ERM} \in \arg\min_{h \in \mathcal{H}} \hat{L}(h).$$

固定零一损失：$\mathbb{E}\hat{L} = p$，$\mathrm{Var}(\hat{L}) = p(1-p)/m \le 1/(4m)$。

$$\mathbb{P}(\exists \text{坏的一致} h) \le |H| e^{-m\epsilon}, \quad m \ge \frac{\log|H| + \log(1/\delta)}{\epsilon}.$$

$$\tau_H(m) \le \sum_{i=0}^{d} \binom{m}{i} \le (em/d)^d, \quad d = \mathrm{VCdim}(H).$$

可实现 VC 样本量 $\tilde{O}((d + \log 1/\delta)/\epsilon)$；agnostic $\tilde{O}((d + \log 1/\delta)/\epsilon^2)$。

线性 / logistic

$X : m \times d, w : d, y : m$：

$$\nabla \frac{1}{2m} \|Xw - y\|^2 = \frac{1}{m} X^\top (Xw - y),$$

$$w_{ridge} = (X^\top X + m\lambda I)^{-1} X^\top y.$$

Logistic：$p = \sigma(Xw)$；NLL gradient/Hessian（损失为和）：

$$\nabla L = X^\top (p - y), \qquad \nabla^2 L = X^\top \mathrm{diag}(p_i(1 - p_i)) X \succeq 0.$$

SVM / Kernel

$$\min \tfrac{1}{2}\|w\|^2 + C \sum_i \xi_i, \quad y_i(w^\top \phi_i + b) \ge 1 - \xi_i, \ \xi_i \ge 0.$$

$$\max_{0 \le \alpha \le C} \mathbf{1}^\top \alpha - \tfrac{1}{2} \alpha^\top (YKY) \alpha, \qquad y^\top \alpha = 0.$$

$w = \sum_i \alpha_i y_i \phi_i$；$\alpha_i[y_i f_i - 1 + \xi_i] = 0$；$K$ 有效 $\Leftrightarrow$ 任意 Gram PSD。

经典小抄

Tree：entropy $-\sum p_c \log p_c$；Gini $1 - \sum p_c^2$。 **k-NN**：标准化；$k \uparrow$ 通常 bias↑、variance↓。 **压缩感知**：$\min \|x\|_1$ s.t. $Ax = y$；NSP：$\|h_S\|_1 < \|h_{S^c}\|_1$。

证明关键词 → 骨架

KL ≥ 0	$-\log$ Jensen，检查 support，等号条件。
VC 值	指定点集全标记下界；任意更大点集排除上界。
泛化界	单假设集中/一致概率 → growth/union bound。
稳定性	测试点与 replace-one 训练点交换对称。
logistic 凸	Hessian $X^\top W X \succeq 0$。
SVM 对偶	Lagrangian → stationarity → 代回 → KKT。

本页红线

MLE $\neq$ MAP；independence $\neq$ conditional independence；uncorrelated $\not\Rightarrow$ independent；PCA 若出现先 center；train/validation/test 不混；SVM 标签/符号/C-λ 先声明。

终极公式速查 2/3：深度、生成、强化学习

维数、KL 方向、时间方向、策略分布是四个最高风险点。

Softmax / Backprop

$$p = \mathrm{softmax}(z), \quad \ell = -y^\top \log p, \quad \nabla_z \ell = p - y.$$

$$\nabla_{W^\ell} J = (H^{\ell-1})^\top \Delta^\ell, \quad \Delta^{\ell-1} = (\Delta^\ell (W^\ell)^\top) \odot \phi'(Z^{\ell-1}).$$

$\sigma' = \sigma(1-\sigma)$；$\mathrm{ReLU}' = \mathbf{1}[z > 0]$。动量 $v^+ = \beta v + g$，$\theta^+ = \theta - \eta v^+$。

CNN / Attention

Conv 参数 $k_h k_w C_{in} C_{out} + C_{out}$。

$$Q : n \times d_k, \ K : n \times d_k, \ V : n \times d_v,$$

$$\mathrm{Attn} = \mathrm{softmax}(QK^\top/\sqrt{d_k})V : n \times d_v.$$

Full attention 时间 $O(n^2 d)$、矩阵内存 $O(n^2)$；滑窗 +global 为 $O(n(w + g)d)$。

VAE / Flow

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

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

MLE 对应 $\min \mathrm{KL}(p_{data}\|p_\theta)$；GAN 理想目标对应 JS，非显式 likelihood。

DDPM / Score / Flow matching

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t.$$

$$\mu_\theta = \alpha_t^{-1/2} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta\right), \quad L_{simple} = \mathbb{E}\|\epsilon - \epsilon_\theta(x_t, t)\|^2.$$

前向 SDE $dx = f\,dt + g\,dW$；反向 drift $f - g^2\nabla \log p_t$（反向时间约定）。

$$v_t(x) = \mathbb{E}[\dot{x}_t \mid x_t = x],$$
$$\min_\theta \mathbb{E}\,\|v_\theta(t, x_t) - \dot{x}_t\|^2,$$
$$\partial_t \pi + \nabla \cdot (\pi v) = 0.$$

Fokker–Planck：$\partial_t \pi = -\nabla \cdot (\pi b) + \frac{1}{2}\nabla\nabla : (aa^\top \pi)$。

Bellman / TD

$$V^\pi = \mathbb{E}_\pi[r + \gamma V^\pi(s')], \qquad V^* = \max_a \mathbb{E}[r + \gamma V^*(s')].$$

$$Q\text{-learn} : \delta = r + \gamma \max_{a'} Q(s', a') - Q(s, a),$$

$$SARSA : \delta = r + \gamma Q(s', a') - Q(s, a).$$

Bellman optimality operator γ-contraction in $\|\cdot\|_\infty$。

Policy gradient / RLHF / DPO

$$\nabla J = \mathbb{E}\sum_t \gamma^t \nabla \log \pi_\theta(a_t|s_t)(Q^\pi - b(s_t)), \quad b = V^\pi.$$

$$J_{RLHF} = \mathbb{E}[r_\psi(x, y)] - \beta \mathbb{E}_x\,\mathrm{KL}(\pi_\theta \| \pi_{ref}).$$

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

证明关键词 → 骨架

Score 等价	conditional score 的条件期望等于 marginal score；展开平方差常数。 平方损失最优预测是条件期望。
Flow 回归	连续性/Fokker–Planck 两边抵消。
边缘一致	
PG theorem	trajectory log-derivative；去掉过去 reward；换占用测度。
Baseline	$\sum_a \pi \nabla \log \pi = \nabla 1 = 0$。
Bellman 收敛	max/expectation 非扩张，再乘 γ。

本页红线

Backprop 就是 chain rule；softmax+CE 为 $p - y$；attention 检查 $QK^\top$ 与 softmax 轴；VAE KL 方向不换；反向 SDE 先声明时间方向；Bellman expectation $\neq$ optimality；Q-learning off-policy，SARSA on-policy。

终极公式速查 3/3：优化、NLP 与交卷核查

看到长证明先找 fixed point/条件期望；看到系统题先写数据流与目标。

凸优化核心

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

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

$$\langle \Delta g, \Delta x\rangle \geq \frac{\mu L}{\mu + L}\|\Delta x\|^2 + \frac{1}{\mu + L}\|\Delta g\|^2.$$

$\nabla^2 f \succeq 0 \Leftrightarrow$ 二阶可微凸；$\nabla^2 f \succeq \mu I \Rightarrow \mu$-强凸。

Prox / SGD 证明链

$$\mathrm{prox}_{\gamma R}(u) = \arg\min_x R(x) + \frac{1}{2\gamma}\|x - u\|^2,$$

$$(u - x^+)/\gamma \in \partial R(x^+), \quad x^* = \mathrm{prox}_{\gamma R}(x^* - \gamma \nabla f(x^*)).$$

$$\mathbb{E}_k\left\|x^{k+1} - x^*\right\|^2 \leq \left\|x^k - x^* - \gamma \Delta \nabla f\right\|^2 + \gamma^2\sigma^2.$$

$\gamma = 2/(L + \mu)$：

$$e_{k+1} \leq (1 - \rho)e_k + \gamma^2\sigma^2, \quad \rho = 4\mu L/(L + \mu)^2,$$

$$e_k \leq (1 - \rho)^k e_0 + \gamma^2\sigma^2/\rho.$$

Mini-batch / Adam

$$g = \frac{1}{\tau}\sum_{t=1}^{\tau} \frac{1}{n q_{s_t}} \nabla f_{s_t}, \quad \mathbb{E}g = \nabla f,$$

$$A'' = \tau^{-1} \max_i \frac{L_i}{n q_i} + (1 - \tau^{-1})L, \quad q_i^* \propto L_i.$$

Adam：$m \leftarrow \beta_1 m + (1 - \beta_1)g$，$v \leftarrow \beta_2 v + (1 - \beta_2)g^2$；做 bias correction 后 $x \leftarrow x - \eta \hat{m}/(\sqrt{\hat{v}} + \epsilon)$。

NLP 核心

$$p_{ML}(w|h) = C(h, w)/C(h),$$

$$PPL = \exp[-N^{-1}\sum_t \log p(w_t|w_{<t})].$$

Skip-gram NS：$-\log \sigma(v_c^\top u_w) - \sum_j \log \sigma(-v_{c_j}^\top u_w)$。

$$\nabla_z L_{InfoNCE} = \tau^{-1}\left(\sum_j p_j a_j - z^+\right).$$

LDA：$\theta_d \sim Dir(\alpha)$，$z_{dn} \sim Cat(\theta_d)$，$w_{dn} \sim Cat(\phi_{z_{dn}})$。MLN：$P(X|E) \propto \exp(\sum_j w_j n_j)$。RAG：retrieve → rerank → grounded generate → cite/abstain。

题面识别终表

经验损失	Bernoulli indicators；均值、方差、$1/m$。
打散/VC	下界构造 + 任意点集上界；量词写全。
prox 噪声	fixed point + nonexpansive + 条件期望 + noise floor。
diffusion	$\alpha, \bar{\alpha}$；后验；ϵ 或 score 回归。
flow	conditional velocity → 条件期望 → 连续性方程。
长上下文	精确连接图 + time/memory + 图直径。
概率模型	变量/分布/生成顺序/联合/推理。
system design	数据、模块、objective、复杂度、指标、失败/回退。

最后 10 项核查

1. 先声明标签、损失缩放、随机性和所有维数。
2. 每个定理写适用条件；KKT 写约束资格。
3. “凸/严格凸/强凸”“驻点/全局最优”不混。
4. 梯度与参数同形，Hessian 为参数维方阵。
5. KL 方向、softmax 轴、attention 输出尺寸正确。
6. α_t 与 $\bar{\alpha}_t$、forward/reverse 时间不混。
7. RL 写清 expectation/optimality、on/off policy。
8. 概率图联合分解不漏 prior/label/normalizer 角色。
9. 复杂度同时写时间、内存和输入规模。
10. 结尾回到题目：界、率、唯一性、最优性或设计取舍。

材料核对提醒（答卷可直接纠偏）

并类 VC 题只据给定信息写严格可得的上下界；强凸最优差通常是非严格 $\geq$；策略梯度若使用未归一化 ρ_π，不要再无故乘/除 $1 - \gamma$。遇到题面符号冲突，先声明自洽约定再推导。

最短收口句

“由上述条件，所有矩阵乘法维数闭合；所用最优性/收敛结论的假设已满足；因此得到题目要求的界（或算法），并注明其复杂度与失败条件。”