arXiv:2611.29692v1 [cs.CL] 22 Nov 2026

文法稀疏性决定 Transformer 如何学习随机语言

Edgar Andres Hernandez Moreno巴塞
罗那大学ehernamo48@alumnes.ub.edu

Francesco Cagnetta Simplex

摘要

我们研究在随机语言模型上训练的 Transformer 如何习得语言。随机语言模型是一种具有随机产生式规则的合成层次文法，其稀疏度由单一参数控制，条件熵则可以精确获知。对于深度二文法，我们发现，稀疏度决定了学习的顺序和速度。稠密文法是逐层习得的：经过 $t \propto V$ 步后，模型会根据一个词元的兄弟项来预测该词元；只有在 $t \propto V^3$ 之后，才会根据完整上下文进行预测。因此，这两个阶段之间存在一个平台期，且该平台期会随着 V 增大而延长。稀疏文法则会在 $t \propto \sqrt{V}$ 一次性习得，其指数甚至小于稠密文法第一阶段的指数。指数 1 和 $1/2$ 在深度一时便已出现；对于使用 Adam 训练的线性模型，我们推导出了这两个指数。

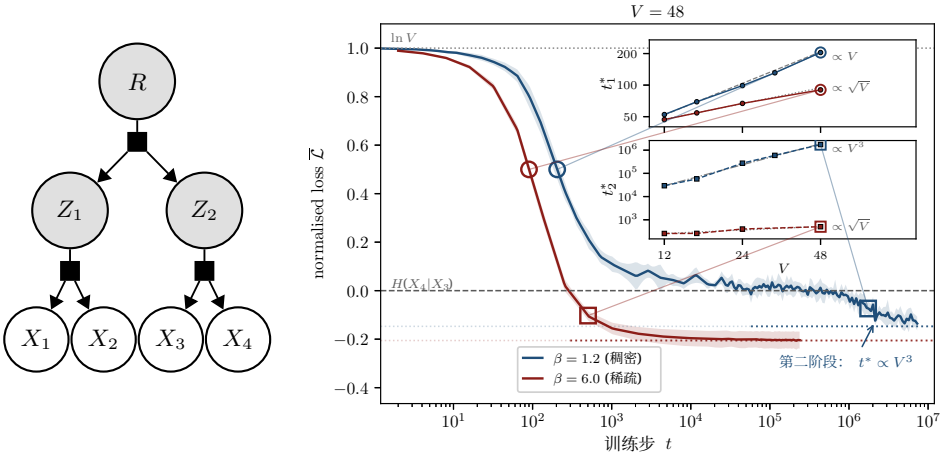

图 1：**文法的稀疏度决定了调度和时间复杂度。左：RLM**。根符号 R 通过式 (1) 的产生式规则 $\mathcal{Q}_a$ 递归展开（黑色方块）；先展开为一对潜变量（Z_1、Z_2），再展开为观测序列（X_1、……、X_4）。**右：深层 Transformer 在稠密**（$\beta = 1.2$、$V = 48$，**蓝色**）**和稀疏**（$\beta = 6.0$、$V = 48$，**红色**）**文法下的学习动力学**。测试损失经过归一化，$\bar{\mathcal{L}} = (\mathcal{L} - H_1)/(\ln V - H_1)$：对于均匀预测器，其值为 1；在兄弟熵 $H_1 = H(X_4 \mid X_3)$ 处为 0（虚线）。点线标示各文法的完整上下文下限 $H_2 = H(X_4 \mid X_1, X_2, X_3)$。稠密文法分两个阶段下降，中间在 H_1 处进入平台期；稀疏文法则一步到位。标记指出两个时间复杂度，t_1^*（圆圈）和 t_2^*（方块），并在插图中以 V 为基准绘出。对 8 次文法实现和 Transformer 初始化取平均；色带表示一个标准差。

预印本。

1 引言

句子以层级方式构成：词语组成短语，短语组成从句 [1–3], 。仅凭句子的统计特征能否习得这种结构，长期以来一直备受争议 [4–8]。大语言模型已经证明这是可能的：仅通过下一词元预测进行训练，它们便能形成追踪句子句法树的表征 [9, 10] ，并实现类似解析的计算 [11, 12]。然而，这些表征在训练过程中是如何习得的，仍远不明朗：文本中的哪些统计特征推动了它们的涌现？习得每种表征需要多少数据或训练时间？由于自然文本的结构无法改变，研究语言习得的理论需要采用具有可控层级结构的合成数据 [13–15], ，以便精确测量学习者的进展 [16–18]。

在此，我们考察随机语言模型（RLM） [19–21]: 这是一种类似语法的数据生成过程，其中序列通过一组预先设定的规则展开一棵潜变量树而生成。RLM 数据的结构由稀疏参数 β 控制：随着 β 增大，文法从<风格 id='5'>稠密</风格>转为<风格 id='8'>稀疏</风格>；在稠密情况下，每个符号的有效规则数随词汇表大小 V 按幂次增长，而在稀疏情况下，每个符号使用 $O(1)$ 条规则。我们探讨文法稀疏度如何影响 Transformer 的学习动力学：层级结构的各层按什么顺序习得，以及每一层的训练时间如何随词汇表大小变化。

我们的贡献。 我们在由深度为 2 的 RLM 生成的 4 词元序列上训练深层 Transformer，根据前面的上下文预测最后一个词元。我们测量测试损失达到最后一个词元在给定其兄弟项时的熵 $H(X_4|X_3)$，继而达到其在给定完整上下文时的熵 $H(X_4|X_1, X_2, X_3)$ 所需的训练时间，并考察这些时间随词汇表大小 V 的变化（图1）。我们发现，规则的稀疏度决定了学习的调度和速度：

- **稠密语法是分阶段习得的，**每次习得层级结构中的一层。兄弟节点统计量在 $t_1^* \propto V$（图1，上方插图）习得，而前两个词元所携带的信息则在 $t_2^* \propto V^3$ （下方插图）习得。
- **稀疏语法一步即可习得，**在 $t^* \propto \sqrt{V}$（图1，插图）。
- **在深度一时就能观察到不同的机制：**即使训练 Transformer 根据第一个词元预测深度为1的 RLM中的第二个词元，它们在稠密语法上仍需 $t^* \propto V$，在稀疏语法上则需 $t^* \propto \sqrt{V}$（图 3）。我们针对使用 Adam 训练的线性模型，对这两个指数进行了解析推导（第4节，附录C）。

2 随机语言模型（RLM）

数据生成过程。我们使用随机语言模型（RLM）的分层变体 [19–21] ，以迭代方式生成深度为L的二叉树的叶节点（图1，左）：根节点 R 扩展为两个潜变量（Z_1， Z_2），每个潜变量独立生成自己的两个子节点。扩展遵循产生式规则 $a \to bc$，其中 a 是父符号， bc 是有序子节点对，所有符号均取自同一个大小为 V的词汇表。给定父节点 a，子节点对上的分布 $\mathcal{Q}_a$ 定义为

$$\mathcal{Q}_a(b,c) \coloneqq \frac{e^{\beta \mathcal{E}_{a,b,c}}}{\sum_{b',c'=1}^{V} e^{\beta \mathcal{E}_{a,b',c'}}}, \qquad \mathcal{E}_{a,b,c} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \ln V), \tag{1}$$

其中随机耦合 $\mathcal{E}$ 只抽取一次，之后便保持不变。

Spa稀疏性。 β 决定有多少条规则具有显著概率，这可以用 $\mathcal{Q}$ 的熵来量化 $_a$，

$$S_a(\beta) \coloneqq -\sum_{b,c=1}^{V} \mathcal{Q}_a(b,c) \ln \mathcal{Q}_a(b,c), \qquad S(\beta) \coloneqq \frac{1}{V}\sum_{a=1}^{V} S_a(\beta), \tag{2}$$

因此， $M(\beta) \coloneqq e^{S(\beta)}$ 是每个符号的有效规则数（图2）。 S_a 从 $S_a(0) = 2\ln V$（所有 V^2 规则等概率）到 $S_a(\infty) = 0$（每个符号只有一条规则），呈单调递减。根据 [21], 中确定的一种映射，每个 $\mathcal{Q}_a$ 都是一个具有 V^2个能级 [22]的随机能量模型的吉布斯测度。因此，当 $V \to \infty$时，

$$S_a(\beta)/(2\ln V) \to \begin{cases} 1-\beta^2/4 & \beta \le 2, \\ 0 & \beta > 2. \end{cases} \Rightarrow M(\beta) = \begin{cases} O\left(V^{2-\beta^2/2}\right) & \beta \le 2, \\ O(1) & \beta > 2. \end{cases} \tag{3}$$

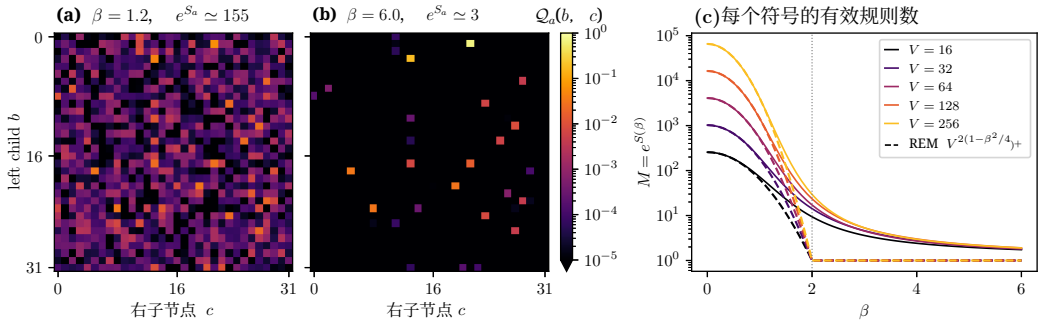

图 2：**一个参数控制文法的稀疏度。(a, b)** 在共享对数颜色刻度下，父符号在 $V = 32$ 和 $\beta = 1.2$,6.0 处的规则矩阵 $\mathcal{Q}_a(b, c)$。在 $\beta = 1.2$ 处，许多符号对都具有显著权重，因此文法是稠密的；在 $\beta = 6$ 处，权重集中于少数几个符号对，文法则是稀疏的。**(c)** 式 (2) 中每个符号的有效规则数 $M(\beta)$，对父符号及耦合的实现取平均。虚线表示式 (3) 的$V \to \infty$ 预测，$\beta = 2$ [21] 处的冻结转变以竖线标出。

3 学习调度与时间尺度，$L = 2$

任务。 我们固定 $L = 2$，使得观测序列（X_1，…，X_4）从根节点 R出发，经由两个潜在符号 Z_1 和 Z_2生成，它们分别是（X_1，X_2）和（X_3，X_4）的父节点（图1左）。任务是根据上下文（X_1，X_2，X_3）预测最后一个词元 X_4。我们使用 Adam 在交叉熵上训练一个两层仅解码器 Transformer $\hat{}p_\theta(x_4 \mid x_1,\ x_2,\ x_3)$，每个训练步都使用一批新序列，因此训练步 t 也能按批量大小计数已见样本数；当 V和 β变化时，所有超参数均保持不变（附录 A）。我们监测测试损失

$$\mathcal{L}(\theta) := -\mathbb{E}\big[\ln \hat{p}_\theta(X_4 \mid X_1, X_2, X_3)\big] \ \geq\ H_2 := H(X_4 \mid X_1, X_2, X_3), \tag{4}$$

当且仅当 $\hat{}p_\theta$ 几乎必然与真实条件分布一致时，等号成立。还有两个参考值可以帮助理解学习曲线：$\ln V$，即均匀预测器的损失，也是网络开始时的损失；以及 $H_1 := H(X_4 \mid X_3) \geq H_2$，即仅凭兄弟项 X_3 所能达到的最佳损失。这两个条件分布在性质上有所不同：X_3 和 X_4 共享父节点 Z_2，因此它们之间的依赖关系相当于一条规则；而表亲项（X_1, X_2）只能通过 Z_1 为 X_4 提供信息，差值 $H_1 - H_2$ 衡量潜变量所携带的信息。两种熵均根据每种实现的规则精确计算（附录 A）。

两种学习计划。图1（右）显示了稠密（$\beta = 1.2$）和稀疏（$\beta = 6$）文法在 $V = 48$ 处的测试损失，其归一化形式为 $\mathcal{L} := (\mathcal{L} - H_1)/(\ln V - H_1)$。稠密文法分两个阶段习得，与 [16]中观察到的现象相呼应：首先，损失下降至H_1，此时网络尽可能根据 X_3 预测 X_4，同时仍忽略（X_1，X_2）；随后，在这一平台期训练约三个数量级的时间后，损失下降至 H_2，此时终于利用了表亲项。其顺序遵循树结构：将 X_3 与 X_4 相连的第1层规则先于连接序列两半部分的根节点规则习得。稀疏文法没有平台期：损失从 $\ln V$一步降至 H_2，也就是说，兄弟项和表亲项同时得到利用。这并非因为稀疏文法中的第二阶段信息量较少：相较于第一阶段，差距 $H_1 - H_2$ 在 $\beta = 6$ 处大于在 $\beta = 1.2$ 处（比较图1中的下限）。

时间复杂度。 为了量化学习速度，我们将 t_1^* 定义为训练步数，即 $\mathcal{L}$ 首次越过 1/2时所处的训练步；1/2位于 $\ln V$与 H_1 之间的中点（图1中的圆圈）。我们将 t_2^* 定义为 $\mathcal{L}$首次越过 H_1 与 H_2 之间中点时所处的训练步（方块）；无论是否存在平台期，这两个定义都适用。内嵌图展示了 t_1^*（上）和 t_2^*（下）随 V的变化。在稠密文法中，$t_1^* \propto V$和 $t_2^* \propto V^3$：每个阶段的代价都与词汇表大小的某个幂成正比；由于两个幂次不同，随着 V增长，平台期会变宽。在稀疏文法中，两种时间都按 $\sqrt{V}$缩放，且只相差一个 $O(1)$ 因子：整个文法一次性习得，其指数甚至小于稠密文法第一阶段的指数。附录B展示了所有 V的学习曲线相对于 t/V^s的塌缩；在 $\beta = 1.2$ 时拟合指数为 $s_1 = 1.00 \pm 0.02$，在 $\beta = 6$时为 $s_2 = 3.04 \pm 0.34$。此外还展示了中间情形 $\beta = 2$，此时两个阶段仍然彼此区分，但二者之间只相差 V 的较小幂次。

4 在 $L = 1$ 处的时间复杂度缩放

图1中的两个指数并非层级结构所致。图 3 展示了在深度为1的RLM上训练的 Transformer，它们根据第一个词元预测第二个词元：对于每个 β，损失都在一步内从 $\ln V$ 降至 $H := H(X_2 \mid X_1)$（面板a）；然而，归一化损失降至一半时所需的时间 t^* 仍将两个阶段区分开来（面板b），在稠密相中按 V 增长，在稀疏相中按 $\sqrt{V}$ 增长——这分别对应于第一阶段和深度为二时单步过程的指数。

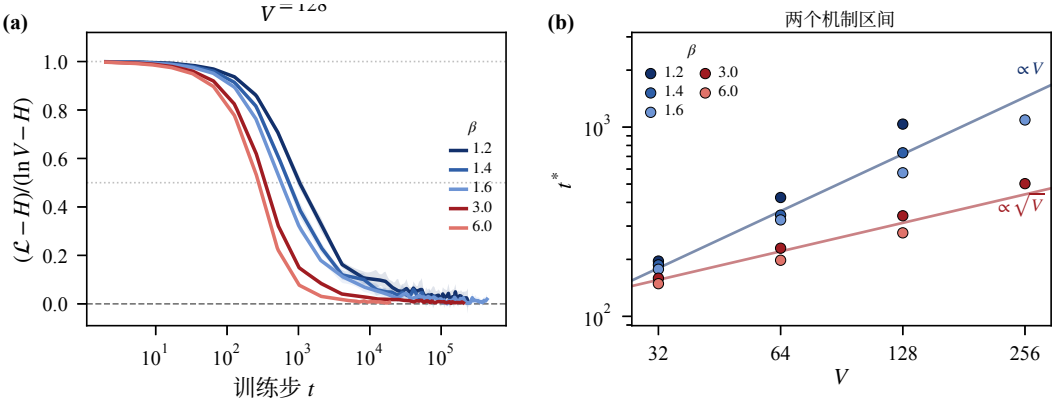

图3：**在 $L = 1$ 处，两个机制区间均存在**。使用深度为1的RLM训练的 Transformer，从 X_1 预测 X_2。**(a)** 在 $V = 128$ 处的归一化测试损失 $(\mathcal{L} - H)\ /\ (\ln V - H)$，其中 $H = H(X_2 \mid X_1)$ 根据每个实现的规则精确计算；稠密相中取 β 的三个值（1.2、1.4、1.6；蓝色），稀疏相中取两个值（3、6；红色）。点线标示数值 1 和 $1/2$。**(b)**对于 $V = 32$ 至 256，交叉时间 t^* 随 V 的变化。实线表示附录 C 中为线性模型使用 Adam 训练所得的两个缩放关系 $t^* \propto V$ 和 $t^* \propto \sqrt{V}$；在同一机制区间内，不同的 β 会使 t^* 发生 $O(1)$ 倍的平移，但不会改变指数。结果为多次重复实验的平均值，带状区域表示一个标准差；实验方案见附录 A。

两个指数都源自一个线性模型：一个用于预测 a 之后的 b 的 logits 表 θ_{ab}。该模型使用 Adam（批量大小为 B，学习率为 η）在交叉熵上进行训练（详情见附录 C）。将 Adam 的运行矩替换为其总体值后，只要该信噪比较小，梯度均值为 μ_{ab}、标准差为 σ_{ab} 的坐标每步移动 $\eta\,\mu_{ab}\sqrt{B}/\sigma_{ab}$。因此，覆盖到目标的 logit 距离 D_{ab} 需要 $t^* \asymp D_{ab}\,\sigma_{ab}/(\eta\,|\mu_{ab}|\sqrt{B})$ 步。在稀疏相中，一个 logit 仅从首个词元正确的样本中占比为 $1/V$ 的那部分样本获得梯度，这使信号缩小了 $1/V$ 倍，而噪声仅缩小 $1/\sqrt{V}$ 倍，因此除对数因子外为 $t^* \propto \sqrt{V}$。在稠密相中，信号和需要 traversed 的距离都与两个词元之间的微弱相关性成正比，这两者相互抵消；而由联合概率决定的噪声使该阶段每个 β 都保持为 $t^* \propto V$。相比之下，不会除以噪声的普通随机梯度下降将分别得到 V 和 V^2（附录 C）。

5 结论

我们发现，对于在稀疏度受控的层级文法上训练的 Transformer，稀疏度决定了数据的潜在结构习得的顺序和速度。稠密语法是逐层习得的，每一层都以 V 的更高次幂习得，因此阶段之间的平台期会随着 V 增宽；稀疏语法则以单一步骤习得，且指数较小。仍有两个方向有待探索。第一，深度在习得过程中的作用：附录 D 的初步证据表明，在稀疏相深处，单个注意力层就足够；而在接近转变时，则需要第二个注意力层才能达到完整上下文下限。第二，潜变量的习得机制决定了第二阶段的指数：我们的论证仅涵盖第一阶段，第二阶段是否遵循 [16], 的基于相关性的机制仍有待研究；按照这种机制，层级结构的每一层都会在其引起的词元间相关性从采样噪声中被分辨出来后习得。

致谢与资金披露

本项目始于 EAHM 在 SISSA 的研究实习，实习是 EAHM 在蒙特雷科技高等教育学院（蒙特雷科技大学）攻读工程物理学学士学位期间的一部分；大部分工作也在此期间完成。作者感谢 Sebastian Goldt 提供有益的讨论。在本项目的部分工作中，FC 得到了欧盟“地平线欧洲”计划的资助，资助依据为玛丽·斯克沃多夫斯卡-居里资助协议 No. 101154584。

参考文献

[1] 诺姆·乔姆斯基．《句法结构》 ．Mouton ，海牙 1957 ，.[2] Martin B. H. Everaert、Marinus A. C. Huybregts、诺姆·乔姆斯基、Robert C. Berwick 和 Johan J. Bolhuis。《结构，而非字符串：语言学作为认知科学的一部分》。《认知科学趋势》，19(12):729–743，2015 年。[3] Nai Ding、Lucia Melloni、Hang Zhang、Xing Tian 和 David Poeppel。《连续语流中层级语言结构的皮层追踪》。《自然-神经科学》，19:158–164，2016 年。[4] 诺姆·乔姆斯基。《规则与表征》。《行为与脑科学》，3(1):1–15，1980 年。[5] Jenny R. Saffran、Richard N. Aslin 和 Elissa L. Newport。《8 个月大婴儿的统计学习》。《科学》，274(5294):1926–1928，1996 年。[6] Jenny R. Saffran 和 Natasha Z. Kirkham。《婴儿的统计学习》。《心理学年度评论》，69:181–203，2018 年。[7] Richard Futrell 和 Kyle Mahowald。《语言学如何学会不再忧虑并爱上语言模型》。《行为与脑科学》，2025 年。[8] Richard Futrell 和 Kyle Mahowald。《你们不能在这里打架！这里可是 BBS！》*arXiv* 预印本，2026 年。对《行为与脑科学》目标文章评论文章的回应。[9] Ian Tenney、Dipanjan Das 和 Ellie Pavlick。《BERT 重新发现了经典自然语言处理流程》。载于计算语言学协会第 *57* 届年会论文集（*ACL*），第 4593–4601 页，2019 年。[10] Christopher D. Manning、Kevin Clark、John Hewitt、Urvashi Khandelwal 和 Omer Levy。《自监督训练的人工神经网络中涌现的语言结构》。《美国国家科学院院刊》，117(48):30046–30054，2020 年。[11] Haoyu Zhao、Abhishek Panigrahi、Rong Ge 和 Sanjeev Arora。《Transformer 在预测掩码词时会进行句法分析吗？》。载于*2023* 年自然语言处理实证方法会议论文集（*EMNLP*），第 16513–16542 页，2023 年。[12] Zeyuan Allen-Zhu 和 Yuanzhi Li。《语言模型的物理学：第 1 部分，学习层级语言结构》。*arXiv* 预印本，2023 年。[13] Elchanan Mossel。《深度学习与层级生成模型》。*arXiv* 预印本，2016 年。[14] Eran Malach 和 Shai Shalev-Shwartz。《一种可证明正确且实际有效的深度学习算法》。*arXiv* 预印本，2018 年。[15] Francesco Cagnetta、Leonardo Petrini、Umberto M. Tomasini、Alessandro Favero 和 Matthieu Wyart。《深度神经网络如何学习组合数据：随机层级模型》。《物理评论*X*》，14:031001，2024 年。[16] Francesco Cagnetta 和 Matthieu Wyart。《探究深度神经网络如何习得语言结构的理论》。载于神经信息处理系统进展（*NeurIPS*），2024 年。[17] Jack T. Parley、Francesco Cagnetta 和 Matthieu Wyart。《深度网络从局部统计中学会解析均匀深度的上下文无关语言》。*arXiv* 预印本，2026 年。

[18] Daniel J. Korchinski、Alessandro Favero 和 Matthieu Wyart。《从自身潜变量而非词元中学习：样本复杂度理论》。*arXiv* 预印本，2026 年。[19] Eric DeGiuli。《随机语言模型》。《物理评论快报》，122:128301，2019 年。[20] Eric De Giuli。《随机语言中的有序性涌现》。《物理学杂志 *A*：数学与理论》，52:504001，2019 年。[21] Giorlandino、Eric De Giuli 和 Sebastian Goldt。《随机语言模型的相结构》。*arXiv* 预印本，2026 年。[22] Bernard Derrida。《随机能量模型：一种可精确求解的无序系统模型》。《物理评论 *B*》，24:2613–2626，1981 年。

A 模拟的详细信息

本文所有图表的复现代码和配置文件将在论文接收后公开。

数据。 序列从第 2 节的深度为 L 的 RLM 中在线采样：每个训练步都会从文法中抽取一个新的批次，因此不存在有限的训练集，也没有训练轮次。根节点在词汇表上均匀分布，耦合的一次实现（式 (1)）定义了一个生成规则张量，该张量由树的所有层级共享。每次重复实验都使用独立的文法、独立的权重初始化和独立的数据流。我们报告 8 次重复实验的平均值，阴影带表示重复实验间的一个标准差。

架构。 我们使用仅解码器因果 Transformer。词元和绝对位置通过学习得到的嵌入映射到维度为 d 的表示，嵌入按 $\mathcal{N}(0, 0.02_2)$ 初始化后相加。主干网络由一系列预层归一化残差块组成，每个块结合了具有 4 个注意力头的因果多头自注意力（键、查询、值和输出投影均无偏置）以及一个宽度为 $4d$ 的两层感知机，该感知机使用单个 ReLU 且不含偏置，并采用神经切线核参数化。最后接一个层归一化层和一个映射到 V 个 logits 的线性预测头；该预测头含偏置，且不与嵌入权重绑定，因此初始化时的损失接近 $\ln V$。随机失活已禁用。只有最后一个输入位置的 logits 会计入损失：对于 $L = 2$，输入为上下文（X_1, X_2, X_3），目标为 X_4；对于 $L = 1$，输入为 X_1，目标为 X_2。

训练。 我们使用 Adam（默认 $\beta_1 = 0.9$, β_2 =0.999, $\epsilon = 10^{-8}$，不使用权重衰减），以恒定学习率 10^{-4} 最小化最后一个词元的交叉熵，不进行预热或衰减。测试损失是在同一文法的新采样序列上计算的平均交叉熵；训练初期按对数间隔步骤评估，之后则按固定间隔评估。步骤数的设定确保每次运行都能达到其下限。改变 V 和 β 时，所有超参数均保持不变。

使用三种设置，不同设置的曲线绝不会绘制在同一组坐标轴上 ：

	块	d	批次	测试样本	用于
正文（$\beta = 1.2$，6.0）	2	256	128	4096	图 1、4、5、8、11
较小模型（$\beta = 2.0$）	2	128	64	512	图 6、7、10
深度为一的任务（$L = 1$）	1	128	128	4096	图 3

第一个的参数为 1.60×10^6，取值为 $V = 48$；第二个的参数为 4.03×10^5，取值为 $V = 32$。只有在第二种设置中才有取值为 $\beta = 2$ 的运行。在附录 D 中，块的数量在其他设置完全相同的情况下分别设为一块和两块：取值为 $\beta = 2$ 的成对运行使用较小模型，取值为 $\beta = 6$ 的成对运行使用主模型。

熵下限与归一化损失。 参考熵是通过对潜变量求和，根据每次重复实验的规则张量精确计算得出的，并非根据样本估算。对于$L = 2$，其值为 $H_1 = H(X_4|X_3)$ 和 $H_2 = H(X_4|X_1,\ X_2,\ X_3)$；对于 $L = 1$，则为 $H = H(X_2|X_1)$。第一阶段的归一化损失为 $(\mathcal{L} - H_1)/(\ln\ V - H_1)$，第二阶段为 $(\mathcal{L} - H_2)/(H_1 - H_2)$，而稀疏区间内及在 $L = 1$ 处的单阶段归一化损失分别为 $(\mathcal{L} - H_2)/(\ln\ V - H_2)$ 和 $(\mathcal{L} - H)/(\ln V - H)$。每个阶段开始时损失为 1，结束时为 0。

时间尺度。 将重复实验插值到 t 中的共同对数网格上并取平均； t^*_ℓ 是平均归一化曲线首次向下穿过 $1/2$ 的位置，通过测量点之间的对数线性插值获得。 $t^*_\ell \propto V^{s_\ell}$ 中的指数 s_ℓ 是对 $\ln t^*_\ell$ 与 $\ln V$ 进行最小二乘拟合得到的；所报告的不确定性是对重复实验进行 1000 次自助重采样所得的标准差，每次重采样都重新执行平均和拟合。指数在一定程度上取决于用于定义 t^*的交叉水平：在 $\beta = 1.2$ 处，第一阶段指数从中点处的 1.05 到 10% 水平处的 1.45 不等，这表明曲线并未完全重合。全文均使用中点。

B 数值结果详情

本附录报告了图1中时间尺度背后的学习曲线，按词汇表大小进行区分，并展示中间语法 $\beta = 2$；每种文法的每个阶段各有一张图。每张图都以两种方式展示相同的数据：左侧为归一化损失随训练步的变化，右侧为其随重缩放时间 t/V^s 的变化，其中 s 为拟合指数。恰好在 $t^* \propto V^s$ 时，右侧面板会收敛到一条主曲线上。

纵轴的归一化方式采用附录 A 中的定义。对于一个从损失 $\mathcal{L}_\mathrm{i}$ 开始、以损失 $\mathcal{L}_\mathrm{f}$ 结束的阶段，我们绘制 $(\mathcal{L} - \mathcal{L}_\mathrm{f})/(\mathcal{L}_\mathrm{i} - \mathcal{L}_\mathrm{f})$；该值从阶段开始时的 1 变化到阶段结束时的 0，并消除了熵对 V 和 β 的依赖。具体而言，第一阶段从均匀预测器开始，以兄弟熵结束，因此 $\mathcal{L}_\mathrm{i} = \ln V$ 和 $\mathcal{L}_\mathrm{f} = H_1$；第二阶段从兄弟熵开始，以全上下文熵结束，因此 $\mathcal{L}_\mathrm{i} = H_1$ 和 $\mathcal{L}_\mathrm{f} = H_2$；稀疏区间的单阶段则从均匀预测器直接延伸到全上下文熵，$\mathcal{L}_\mathrm{i} = \ln V$ 和 $\mathcal{L}_\mathrm{f} = H_2$。两种熵均根据每个实现的规则精确计算。曲线是在 t 的公共对数网格上对各实现取平均得到的，阴影带表示各实现之间的一个标准差。

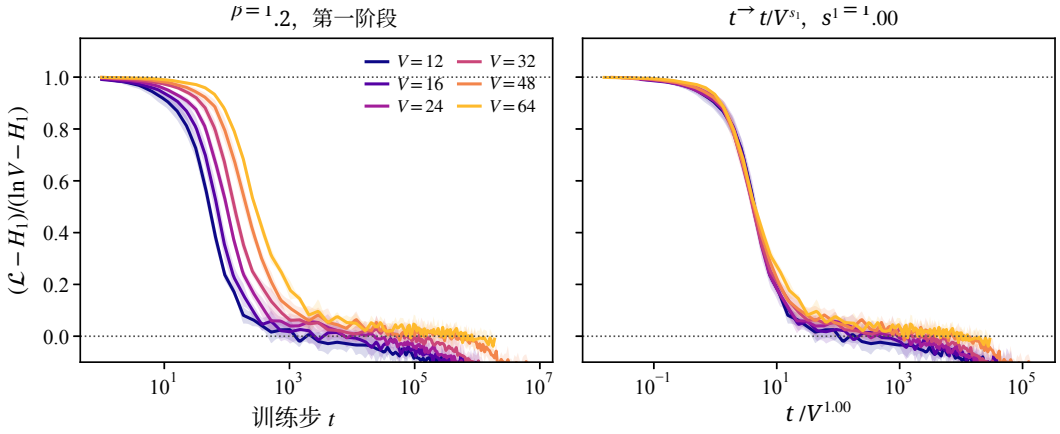

图 4：**稠密语法的第一阶段，** $\beta = 1.2$。 </style}对于 $V = 12$ 到 64（每个取 8 种实现），在 $\ln V$ 与 $H_1 = H(X_4|X_3)$ 之间的归一化损失，随训练步数的变化（左图）以及随t/V^{s_1} 的变化（右图， $s_1 = 1.00 \pm 0.02$ ）。这种塌缩表明，第一个时间尺度与词汇表大小呈线性关系。

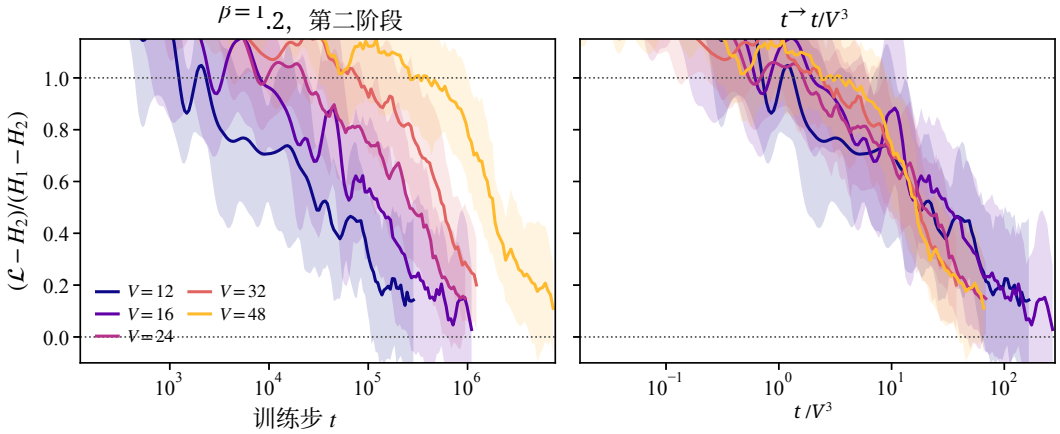

图5：**稠密语法的第二阶段，** $\beta = 1.2$。 H_1 与$H_2 = H$之间的归一化损失（$X_4|X_1,\ X_2,\ X_3$），针对 $V = 12$ 至 48（每个各运行 8 次），分别相对于训练步（左）以及相对于带有 $s_2 = 3.04 \pm 0.34$ 的 t/V^{s_2}（右）绘制。曲线在 $t \ll V^3$ 处处于平台期，只有在 $t \gg V^3$ 时才达到下限，比第一阶段晚两个 V 的幂次。省略 $V = 64$：在可用预算内无法完成此阶段。曲线采用 $\ln t$ 中的 5 点移动平均进行平滑；两个阶段中，第二阶段的噪声更大。

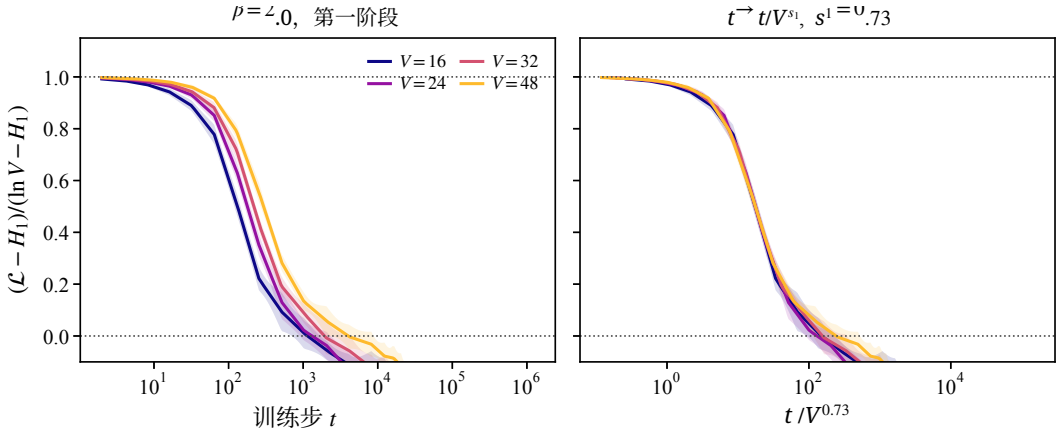

图6：**第一阶段，** $\beta = 2$。数量与图4相同，范围为 $V = 16$ 至 48，使用 $s_1 = 0.73 \pm 0.04$。这些运行使用附录 A 中较小的模型，不能与 $\beta = 1.2$ 和 $\beta = 6$ 曲线放在相同坐标轴上比较。

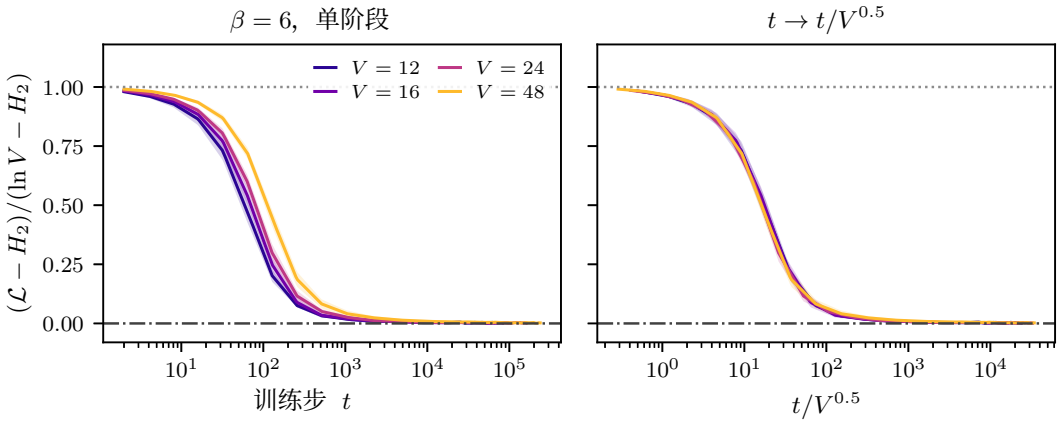

图8：**稀疏语法的单阶段，** $\beta = 6$。对于$V = 12$ 至48（各有8种实现），在$\ln V$与 H_2之间的归一化损失随训练步的变化（左），以及随 t/V^s 和 $s = 0.47$的变化（右）；直接拟合单阶段曲线得到 $s = 0.44 \pm 0.03$。在 H_1 处看不到平台期：兄弟项和表亲项同时习得，整个下降过程可用一个指数描述。

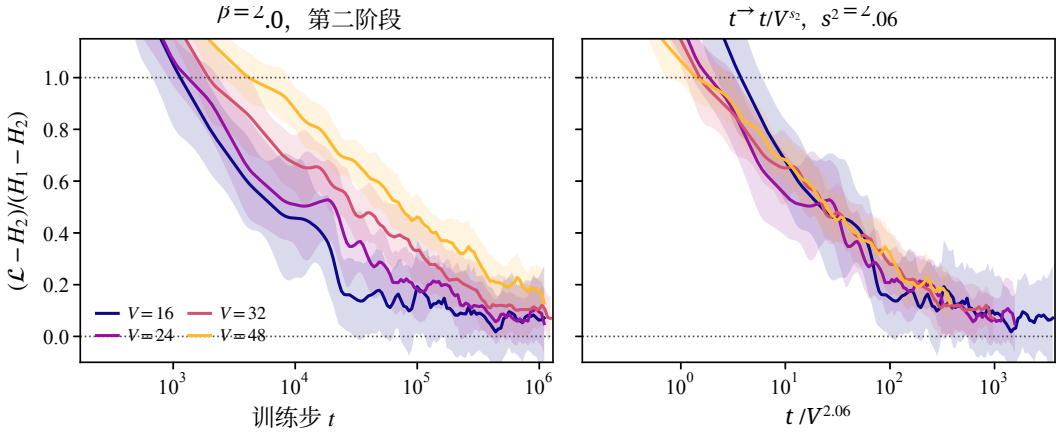

图 7：**第二阶段，** $\beta = 2$。数量与图 5 相同，使用了 $s_2 = 2.06 \pm 0.37$。在 $V = 48$，该阶段在可用预算内仅能部分完成，因此拟合依赖于中途交叉点，而非趋近下限的过程，不确定性在 s_2 上也相应更大。在 $\beta = 2$，两个阶段相隔约 1.3 个 V 数量级，而在 $\beta = 1.2$ 时相隔 2 个数量级：调度已开始收缩。

C Adam 在 $L = 1$ 处线性模型的时间尺度

第一时间尺度的指数 $1/2$ 和 1（图 3）可以通过分析使用 Adam 训练的线性模型如何摆脱初始化状态而得到。我们考虑从 X_1 预测 X_2 的 $L = 1$ 任务；在该任务中，线性模型实质上就是 Transformer 的嵌入层；在 $L = 2$ 处的第一阶段进行相同的计算，只需用（X_3, X_4）替换（X_1, X_2）。这里的分析关注的是下降初期，而非趋近熵下限的过程。

线性模型。模型是一张 logits 表 $\theta \in \mathbb{R}^{V\times V}$，其中 $\hat{}p_\theta(b \mid a) = e^{\theta_{ab}} / \sum_{b'} e^{\theta_{ab'}}$，初始化为 $\theta = 0$，即采用均匀预测器，此时损失为 $\ln V$。对于一个样本（X_1, X_2），交叉熵 $\ell(\theta) = -\ln \hat{}p_\theta(X_2 \mid X_1)$ 的梯度为

$$g_{ab} \coloneqq \partial_{\theta_{ab}} \ell(\theta) = \mathbf{1}\{X_1 = a\}\big(\hat{p}_\theta(b \mid X_1) - \mathbf{1}\{X_2 = b\}\big). \tag{5}$$

一批包含 B 个独立样本的数据给出小批量梯度 $\bar{g}_{ab} = \frac{1}{B}\sum_{i=1}^{B} g_{ab}^{(i)}$，其均值和方差为

$$\mathbb{E}\,\bar{g}_{ab} = \mu_{ab} \coloneqq \mathbb{E}\,g_{ab}, \qquad \operatorname{Var}\bar{g}_{ab} = \frac{\sigma_{ab}^2}{B}, \qquad \sigma_{ab}^2 \coloneqq \operatorname{Var} g_{ab} = \mathbb{E}\,g_{ab}^2 - \mu_{ab}^2. \tag{6}$$

优化模型。 Adam 通过 $-\eta\, m_{ab}/\sqrt{v_{ab}}$ 更新每个坐标，其中 m 和 v 是 $\bar{g}$ 和 $\bar{g}^2$ 的指数移动平均。只要参数在平均窗口内变化不大，且批次是新鲜的，就可以用总体矩替代运行平均值；我们的实验方案（附录 A）便符合这一条件。于是，更新公式为

$$\Delta\theta_{ab} = -\eta\,\frac{\mu_{ab}}{\sqrt{\mu_{ab}^2 + \sigma_{ab}^2/B}} = -\eta\,\frac{\mathrm{SNR}_{ab}}{\sqrt{1 + \mathrm{SNR}_{ab}^2}}, \qquad \mathrm{SNR}_{ab} \coloneqq \frac{\mu_{ab}\sqrt{B}}{\sigma_{ab}}. \tag{7}$$

在噪声主导区间 $|\mathrm{SNR}_{ab}| \ll 1$ 内（当批次足够小时，该条件成立），坐标每步以 $\eta\,|\mathrm{SNR}_{ab}|$ 的速度漂移。若要达到目标，该坐标必须移动 logit 距离 D_{ab}，所需步数为

$$t^* \asymp \frac{D_{ab}}{\eta}\,\frac{1}{|\mathrm{SNR}_{ab}|} = \frac{D_{ab}}{\eta}\,\frac{\sigma_{ab}}{|\mu_{ab}|\sqrt{B}}. \tag{8}$$

普通随机梯度下降会以 $\eta|\mu_{ab}|$ 的速度漂移，并需要 $D_{ab}/(\eta|\mu_{ab}|)$ 步：唯一的区别在于 Adam 会按梯度均方根进行归一化，这使得速度取决于比值 μ/σ，而非 μ。接下来将计算文法的两个阶段中，在 $\theta = 0$ 处的 μ_{ab}、σ_{ab} 和 D_{ab}。

稀疏相。对于 $\beta \to \infty$，每个父节点只有一条规则。令符号 a 为k 的左子节点，因此 $\mathbb{P}(X_1 = a) = k/V$；以高概率，$k$ 个后续项各不相同，因此，对于 k 活跃对中的 b，$\mathbb{P}(X_2 = b \mid X_1 = a) = 1/k$，其他情况下则为 0。在 $\theta = 0$ 处，$\hat{}p_\theta = 1/V$，由式（5）可得，对于一个活跃对，

$$\mu_{ab} = \frac{k}{V}\Big(\frac{1}{V} - \frac{1}{k}\Big) \simeq -\frac{1}{V}, \qquad \mathbb{E}\, g_{ab}^2 = \frac{k}{V}\Big[\frac{1}{k}\Big(1 - \frac{1}{V}\Big)^2 + \Big(1 - \frac{1}{k}\Big)\frac{1}{V^2}\Big] \simeq \frac{1}{V}, \tag{9}$$

使得 $\sigma_{ab}^2 \simeq 1/V$和

$$|\mathrm{SNR}_{ab}| \simeq \sqrt{\frac{B}{V}}. \tag{10}$$

梯度仅在满足 $X_1 = a$ 的比例为 $1/V$ 的样本上非零，其量级为 1：信号被 $1/V$ 稀释，而噪声仅被 $1/\sqrt{V}$ 稀释。同一行的非活跃坐标具有 $\mu_{ab'} = k/V^2$ 和 $\sigma_{ab'}^2 \simeq k/V^3$，因此其信噪比为 $|\mathrm{SNR}| \simeq \sqrt{kB/V}$，量级相同，并以相同速度向下漂移。目标逻辑值在活跃后续项上的概率为 $1/k$，在其他项上的概率为 0：当活跃与非活跃 logits 之间的差距为 $\ln V$ 量级时，第 a 行的损失从 $\ln V$ 降至 $\ln k$；当差距达到 $\frac{1}{2}\ln V$ 量级时，损失达到用于定义 t^* 的中点。因此$D \asymp \ln V$，并且由公式（8）可得，

$$t^* \asymp \frac{\ln V}{\eta}\sqrt{\frac{V}{B}} \;\propto\; V^{1/2} \tag{11}$$

忽略对数因子后，这就是稀疏区间的指数。

稠密相。对于有限的 β，两个词元的联合分布为 $\mathcal{P}_{ab}^{(1)} = \frac{1}{V}\sum_z \mathcal{Q}_z(a,\ b)$。令 C_{ab}：$= \mathcal{P}_{ab}^{(1)} - \mathbb{P}(X_1 = a)\ \mathbb{P}(X_2 = b)$ 表示其连通部分，并令 $\bar{C} := \|C\|_F/V$ 表示其均方根元素。当条件分布与均匀分布的偏离很小时，我们称该相为稠密相，即典型的相对偏离

$$\epsilon := V\left|\mathbb{P}(X_2 = b \mid X_1 = a) - \tfrac{1}{V}\right|_{\mathrm{typ}} \simeq V^2\bar{C} \tag{12}$$

至多为一阶；根据精确相关性，$\epsilon \approx 0.8$在 $\beta = 1$ 和 ϵ 处介于 0。95 与 1。55 之间，且 $\beta = 1.2$ 对于介于 16 和 128 之间的 V 成立，因此下面的估计适用于稠密运行中其有效性的边界处。在 $\theta = 0$ 处，式 (5) 给出

$$\mu_{ab} = \frac{\mathbb{P}(X_1 = a)}{V} - \mathcal{P}_{ab}^{(1)} =: -\Delta_{ab}, \tag{13}$$

其中，Δ_{ab} 是联合分布相对于其行均匀值的偏离，除去边缘分布的非均匀性后等于 C_{ab}，其典型大小为 $\bar{C}$。$\mathcal{P}_{ab}^{(1)}$ 的均匀部分会抵消，因为 $\theta = 0$ 已经是均匀预测器：唯一的信号就是相关性。二阶矩为

$$\mathbb{E}\, g_{ab}^2 = \mathcal{P}_{ab}^{(1)}\Big(1 - \frac{1}{V}\Big)^2 + \frac{\mathbb{P}(X_1 = a) - \mathcal{P}_{ab}^{(1)}}{V^2} \simeq \mathcal{P}_{ab}^{(1)} \simeq \frac{1}{V^2}, \tag{14}$$

因此 $\sigma_{ab} \simeq 1/V$：噪声由联合概率本身决定，而不是由其偏差决定。因此

$$|\mathrm{SNR}_{ab}| \simeq V\bar{C}\sqrt{B}. \tag{15}$$

目标逻辑值为 $\theta_{ab}^* = \ln \mathbb{P}(X_2 = b \mid X_1 = a) + \mathrm{const}$，因此坐标必须按照条件分布相对于均匀分布的偏差移动，

$$D_{ab} = \Big|\ln\big(V\,\mathbb{P}(X_2 = b \mid X_1 = a)\big)\Big| \simeq V\Big|\mathbb{P}(X_2 = b \mid X_1 = a) - \frac{1}{V}\Big| = \frac{V|\Delta_{ab}|}{\mathbb{P}(X_1 = a)} \simeq V^2\bar{C}. \tag{16}$$

信号和距离都与相关性成正比，而相关性在公式（8）中相互抵消：

$$t^* \asymp \frac{V^2\bar{C}}{\eta V\bar{C}\sqrt{B}} = \frac{V}{\eta\sqrt{B}} \;\propto\; V. \tag{17}$$

¯在整个稠密相中，指数始终为 1，与 C 的取值无关，因此也与 β 的取值无关。在速度为 $\eta|\mu_{ab}| \simeq \eta\bar{C}$ 的随机梯度下降下，达到相同距离需要 V^2/η 步。

说明。(i) 两个结果是同一公式在两种极限情况下的表现。当 V 固定而 β 增大时，相对偏差 $\epsilon = V^2\bar{C}$ 超过 1，logit 距离会饱和于 $D = O(\ln V)$，而信噪比 $V\bar{C}\sqrt{B}$ 仍持续增长；若 $\bar{C} \propto V^{-a_1}$，公式（8）给出 $t^* \propto V^{a_1-1}$；对于 $\beta \geq 2$，当 $a_1 = 3/2$ 时，这就是公式（11）。因此，随着稀疏度增加，第一阶段指数会从 1 插值变化至 $1/2$。(ii) 公式（8）所假设的噪声主导区间要求 $B \lesssim \sigma_{ab}^2/\mu_{ab}^2$，即在稀疏相中为 $B \lesssim V$，在稠密相中为 $B \lesssim 1/(V\bar{C})^2$。超过这些值后，Adam 会退化为符号下降，$\Delta\theta_{ab} = -\eta \text{sign}(\mu_{ab})$，而 $t^* \asymp D/\eta$ 不再随 V 按幂律增长。为检验这些标度关系在论证适用的情形下是否成立，图 9 展示了批量大小为$B = 2$时线性模型本身的表现；在整个 V范围内，备注（ii）的两个条件均成立：稀疏相中 t^* 随 $\sqrt{V\ln V}$增长，而在稠密相中则随 V增长，正如方程 (11) 和 (17) 所预测；按相应比例缩放时间后，损失的下降曲线也重合。前几节中的 Transformer 使用更大的批量大小进行训练，此时稀疏相中 $B \lesssim V$不成立；对于这些模型，我们仅将该论证用作解释这两个指数的启发式方法。

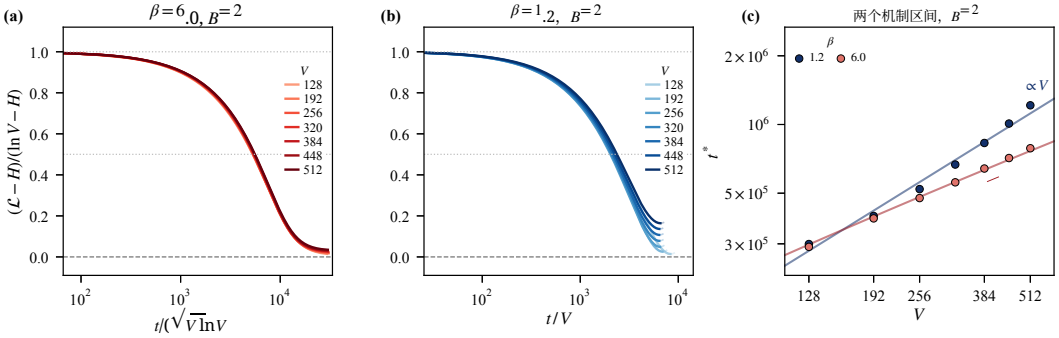

图 9：**小批量大小下的线性模型，** $B = 2$。线性模型的归一化损失（$\mathcal{L} - H$）/($\ln V - H$)，对 8 种文法取平均，$V = 128$ 至 512：（a）$\beta = 6$ 对比 $t/(\sqrt{V\ln V})$，（b）$\beta = 1.2$ 对比 t/V。（c）平均损失首次达到 $1/2$时的时间 t^*，采用方程 (11) 和 (17) 中的 $t^* \propto \sqrt{V\ln V}$和 $t^* \propto V$，并拟合前因子。Adam 使用默认参数（$\beta_1 = 0.9$、$\beta_2 = 0.999$、$\epsilon = 10^{-8}$，无权重衰减；附录 A），学习率为 $\eta = 3 \times 10^{-5}$，每一步都使用新样本。在 $\beta = 1.2$ 时，损失在较晚时刻趋于平稳，停留在下限之上，随后上升（浅色曲线）；较大的 V时这一现象更明显，但不影响 t^*。

D 深度的作用

层次文法分阶段习得这一现象表明，网络需要具备内部深度，才能表示树的中间层级。我们通过训练具有一个和两个注意力块的 Transformer 来检验这一点，并使用完全相同的文法：对于每种词汇表大小和每次重复实验，深度为 1 和深度为 2 的网络看到的规则张量都相同，因此这是配对比较，差异不能归因于文法的抽样。下方两个面板均显示归一化损失（$\mathcal{L} - H_2$）/（$\ln V - H_2$），其中 $H_2 = H$（$X_4|X_1,\ X_2,\ X_3$）为该次实现的全上下文熵，按重复实验取平均，带状区域表示一个标准差。

这种对比与正文中的图景一致。当文法足够稠密、使各层能够逐一习得时，第二阶段需要组合两个潜在符号，而只有一个注意力块的网络无法充分表示这种组合，因此无法达到贝叶斯下限。当文法稀疏时，整个序列会在一个步骤中习得，无需中间表示，因此第二个块并无助益。因此，只有在学习过程具有时间层次性时，增加深度才会带来收益。

我们指出一项局限：正文配置中的 $\beta = 2$ 没有成对的深度 1/深度 2 运行，因此图 10 使用较小模型，而图 11 使用主模型。每幅图内的深度间比较均为配对且精确的；但两幅图之间的比较并非在固定架构下进行。

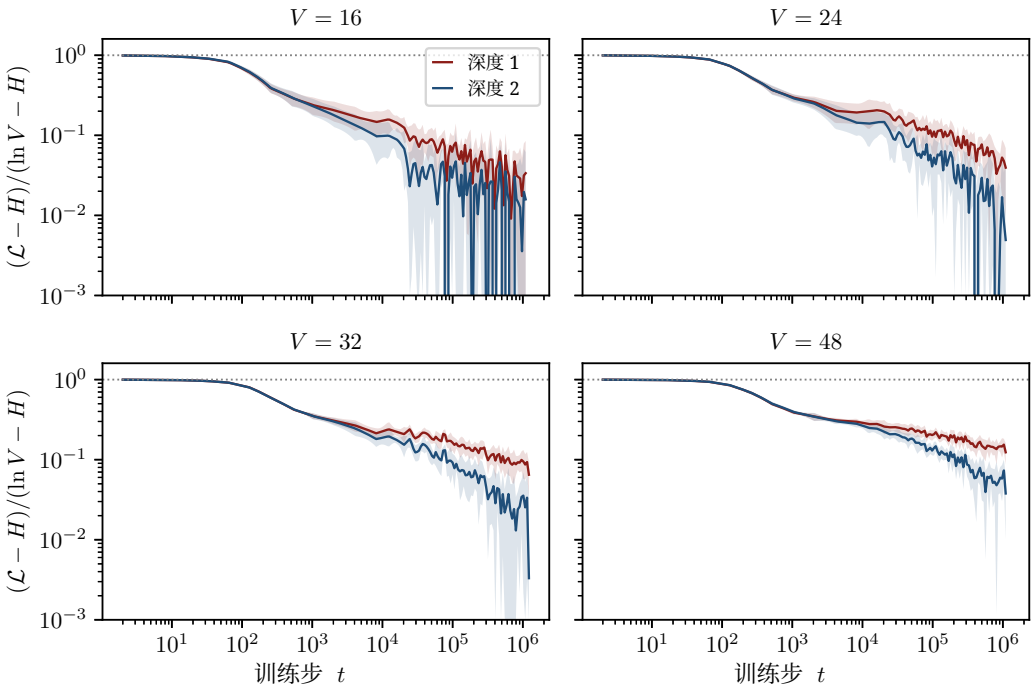

图 10：**在 beta 2.0 时，深度有所帮助**。 在 $\beta = 2$ 下训练的深度为 1 和深度为 2 的 Transformer，使用相同的文法，训练步数为 $V = 16$，24，32，48，每种设置重复实验 8 次。随着损失从均匀预测器水平下降，这两个网络难以区分；但在训练时间较长时，它们表现出差异：深度为 2 的网络达到完整上下文下限，而深度为 1 的网络则停滞在该下限之上。采用附录 A 中的较小模型（嵌入维度 128，批量大小 64）。

$L = 2$, $\beta = 6.0$ —深度 1 与深度 2 对比（6 次重复实验 $\pm$，每个 V 的标准差）

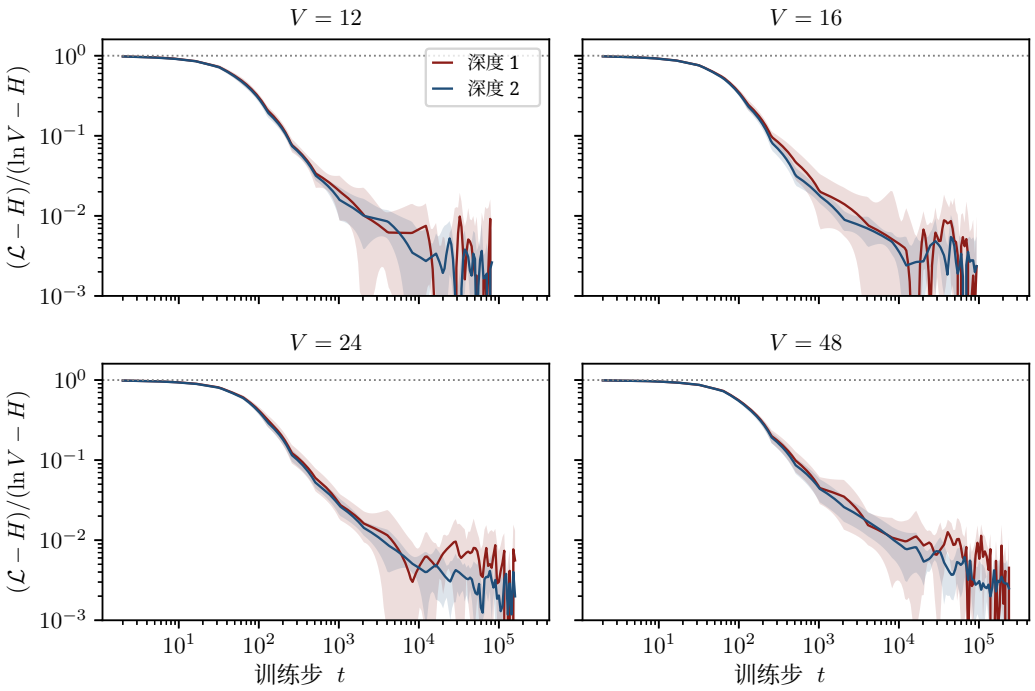

图 11：**在稀疏区间内，深度无关紧要**。在 $\beta = 6$、$V = 12$下进行相同的比较，16、24、48，每种设置各进行 6 次配对重复实验，采用对数纵轴，因此可以在三个数量级范围内观察到趋近下限的过程。两种深度在整个过程中都无法区分，深度为 1 的网络至多略慢一些。正文中的实验方案（嵌入维度 256，批量大小 128）；深度为 1 的运行使用从文件中读取的深度为 2 的运行的规则张量进行训练，因此文法完全一致。