北京大学
PEKING UNIVERSITY

From Entropy to Epiplexity: Rethinking Information for Computationally Bounded Intelligence

Marc Finzi Shikai Qiu Yiding Jiang Pavel Izmailov J. Zico Kolter
Andrew Gordon Wilson

PRESENTER: LILANG LIN

2026/04/26

Outline

1 Background

2 Method

3 Experiments

4 Discussion

Outline

1 Background

2 Method

3 Experiments

4 Discussion

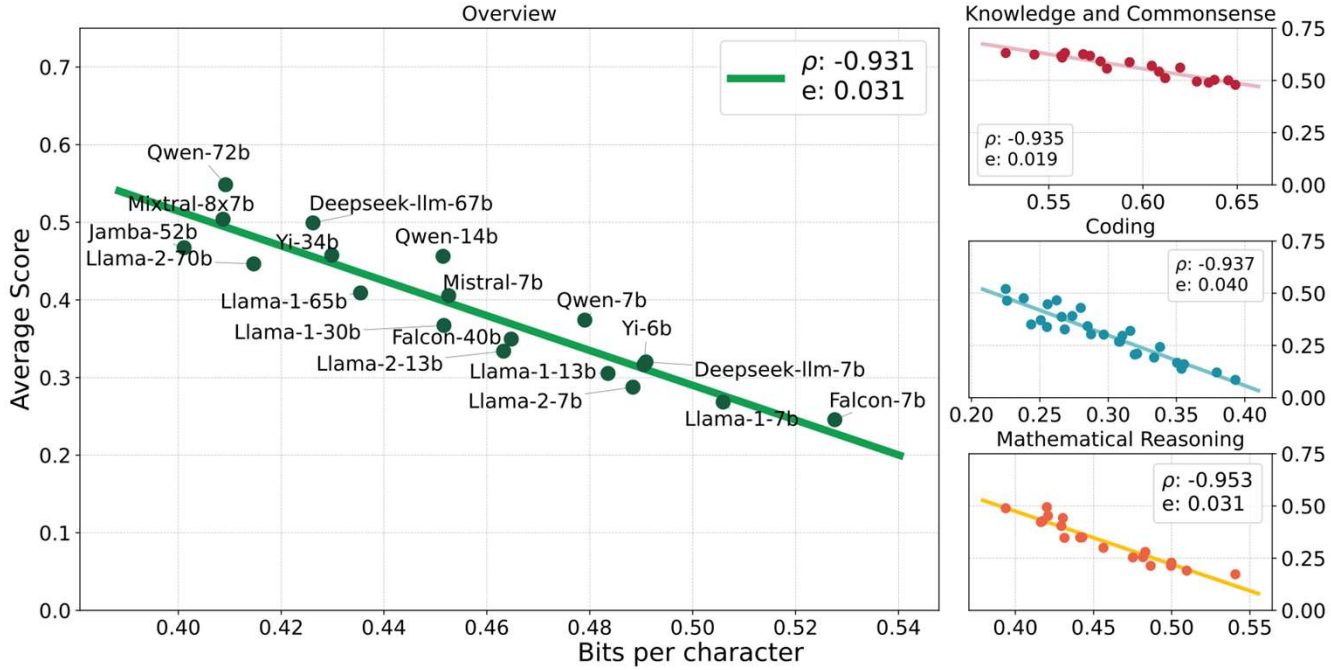
Background

Compression Represents Intelligence Linearly

Yuzhen Huang*[1] Jinghan Zhang*[1] Zifei Shan[2] Junxian He[1]
[1]The Hong Kong University of Science and Technology [2]Tencent
{yhuanghj,jzhangjv,junxianh}@cse.ust.hk

- **Compression Represents Intelligence Linearly (COLM 2024)**

$$\text{Optimal \# Bits on Average} = \mathbb{E}_{x\sim p_{data}}\left[\sum_{i=1}^{n} -\log_2 p_{model}(x_i|x_{1:i-1})\right],$$



Background

Kolmogorov Complexity

$$A = 4444444444$$
$$B = 2718281828$$
$$C = 1756475382$$

$$C_f(x) := \min\{|p| : f(p) = x\} .$$

$$K(x) := C_U(x) = \min\{|\langle M, w\rangle| : \text{TM } M \text{ halts on input } w \text{ and outputs } x\} .$$

A string x is *incompressible* or *Kolmogorov random* if $K(x) \geq |x|$.

"Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin." (Von Neumann, 1951)

Background

Kolmogorov Complexity

Theorem 6 (Euclid c. 300 BC)**.** *There are infinitely many prime numbers.*

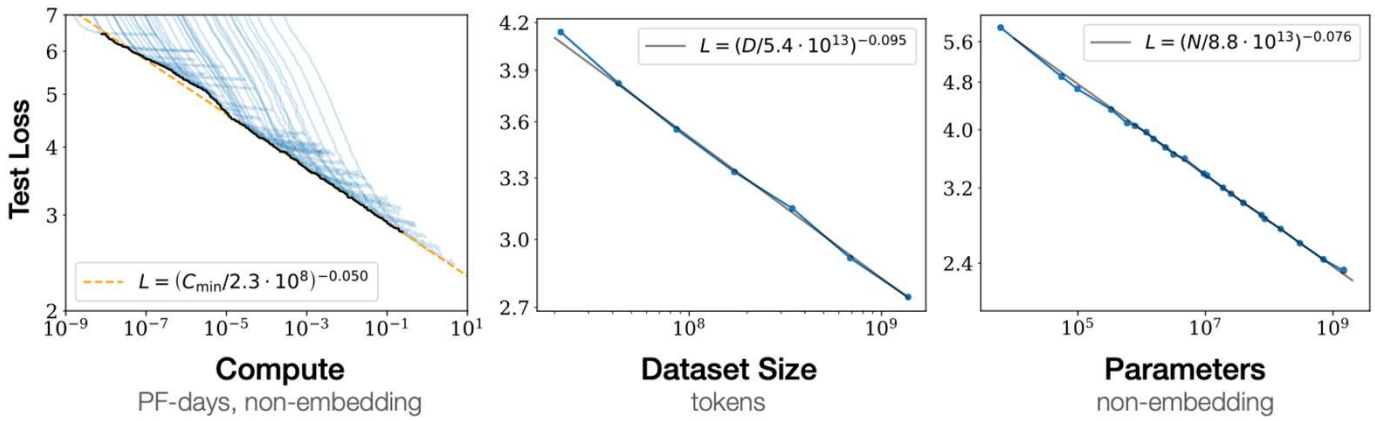
Proof. Suppose not. Let $p_1, p_2, \ldots, p_k$ be a list of all the primes for some integer k. Let $m \in \mathbb{N}$ be a natural number whose encoding in binary is incompressible,[2] and let $n = \lfloor \log_2 m \rfloor + 1$. Write m as a product of primes: for some non-negative integers $e_1, \ldots, e_k$,

$$m = p_1^{e_1} p_2^{e_2} \cdots p_k^{e_k} \ .$$

Note that we can describe m by providing $\langle e_1, e_2, \ldots, e_k \rangle$. We claim that this is a short description of m, contradicting the incompressibility of m. Note that each $e_i \leq \log_2 m \leq n$, so it takes only $\log_2 n$ bits to describe each e_i. Hence, we can descibe $\langle e_1, \ldots, e_k \rangle$ with at most $(2k-1)\log_2 n + 2(k-1) + O(1) = O(\log_2 n)$ bits. (Note that k is constant, independent of n.) But, we assumed that $K(m) \geq n$. Assuming n is large enough (which we may, as there are incompressible strings of every length), this is a contradiction. □

Background

- **Scaling Laws**



Outline

1 Background

2 **Method**

3 Experiments

4 Discussion

From Entropy to Epiplexity

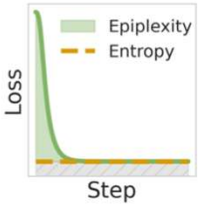
- **Paradox**
  - Information cannot be increased by deterministic processes.
  - Information is independent of factorization order.
  - Likelihood modeling is merely distribution matching.

From Entropy to Epiplexity

Random vs structural information

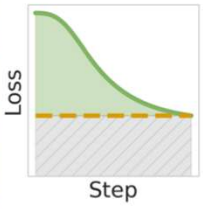
Low random info, low structural info

```
def is_even(n):
  if n == 0: return True
  elif n == 1: return False
  elif n == 2: return True
  elif n == 3: return False
  elif n == 4: return True
  elif n == 5: return False
  ...
```



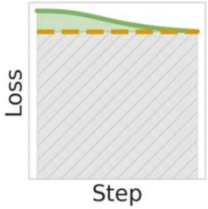
Moderate random info, high structural info

```
def dijkstra(g, s):
  D = defaultdict(lambda:float('inf'))
  D[s] = 0; q = [(0, s)]
  while q:
    d, u = pop(q)
    if d == D[u]:
      for v, w in g.get(u, []):
        if (nd := d + w) < D[v]:
          D[v] = nd; push(q, (nd, v))
  return D
```

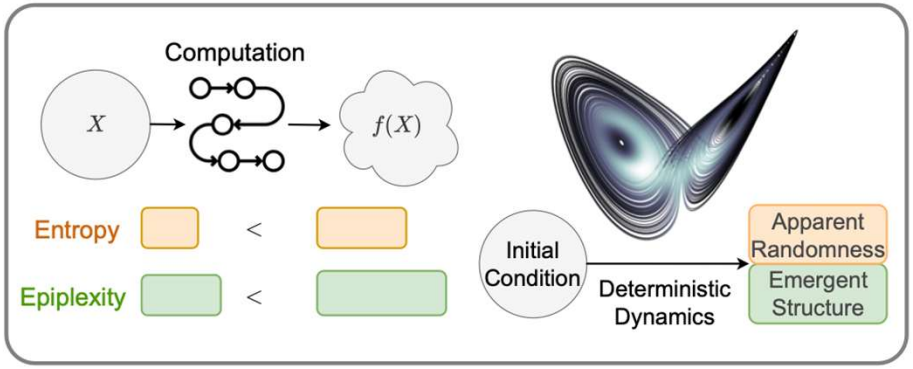


High random info, low structural info

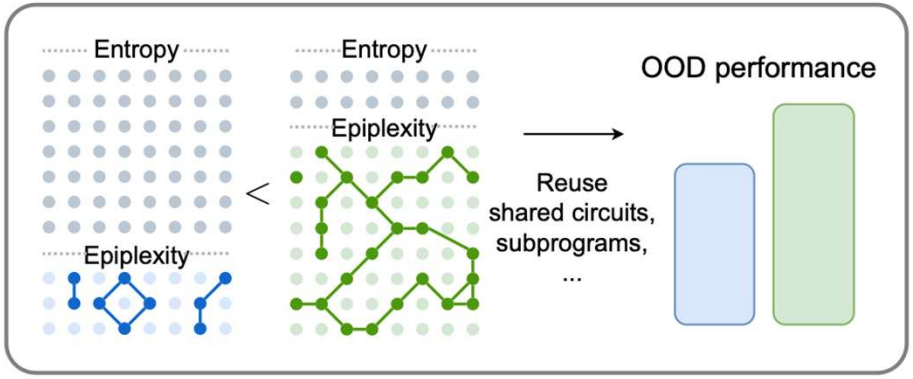
```
API_KEY = "sk_7aF2jK1ycP9LmvYzz34"
USER_ID = "usr_4f8a2c1e9b7d3065"
BUCKET = "s3://data-8a3f1b-west-prod"
SAVE_DIR = "/mnt/marc/exp_7f2a/ckpts"
SAVE_CKPT = True
DEBUG = False
SEED = 9284715
...
```



Information can be created by computation



Structural information → OOD generalization



From Entropy to Epiplexity

Definition 2 (Martin–Löf random sequence (Martin-Löf, 1966)) *An infinite sequence* $x_{1:\infty} \in \{0,1\}^{\mathbb{N}}$ *is Martin–Löf random iff there exists a constant* c *such that for all* n, $K(x_{1:n}) \geq n - c$. *Using this criterion, all computable randomness tests are condensed into a single incomputable randomness test concerning Kolmogorov complexity.*

Definition 5 (Naive Sophistication (Mota et al., 2013)) *Sophistication, like Kolmogorov complexity, is defined on individual bitstrings, and it uses the compressibility criterion from Martin-Löf randomness to carve out the random content of the bitstring. Sophistication is defined as the smallest Kolmogorov complexity of a set* S *such that* x *is a random element from that set (at randomness discrepancy of* c*).*

$$\text{nsoph}_c(x) = \min_S : \{K(S) : K(x \mid S) > \log|S| - c\} \tag{2}$$

$$K(x) \approx K(S) + \log|S|$$

$$K(x) \approx \text{nsoph}(x) + \text{randomness}(x)$$

From Entropy to Epiplexity

Definition 6 (Two-part MDL (Rissanen, 2004; Grünwald, 2007)) *Let $x \in \{0,1\}^{n\times d}$ be the data and $\mathcal{H}$ be a set of candidate models. The two-part MDL is:*

$$L(x) = \min_{H\in\mathcal{H}} L(H) - \log P(x \mid H),$$

where $L(H)$ specifies the number of bits required to encode the model H, and $-\log P(x \mid H)$ is the number of bits required to encode the data given the model.

Definition 8 (Epiplexity and Time-Bounded Entropy) *Consider a random variable X on $\{0,1\}^n$. Let*

$$\mathrm{P}^\star = \arg\min_{\mathrm{P}\in\mathcal{P}_T} \{|\mathrm{P}| + \mathbb{E}[\log 1/P(X)]\} \tag{3}$$

be the program that minimizes the time bounded MDL with ties broken by the smallest program, and expectations taken over X. $|\mathrm{P}|$ denotes the length of the program P in bits, and logarithms are in base 2. We define the T-bounded epiplexity S_T *and* entropy H_T *of the random variable X as*

$$\mathrm{S}_T(X) := |\mathrm{P}^\star|, \quad \textit{and} \quad \mathrm{H}_T(X) := \mathbb{E}[\log 1/P^\star(X)]. \tag{4}$$

From Entropy to Epiplexity

Definition 11 (Conditional epiplexity and time-bounded entropy) *For a pair of random variables X and Y, define $\mathcal{P}^X_{T(n)}$ as the set of probabilistic models P such that for each fixed x, the conditional model $\mathrm{P}_{Y|x}$ is in $\mathcal{P}_{T(n)}$. The optimal conditional model with access to X is:*

$$\mathrm{P}^\star_{Y|X} = \arg\min_{\mathrm{P}\in\mathcal{P}^X_T} \left\{ |\mathrm{P}| + \mathbb{E}_{(X,Y)} \left[-\log P(Y \mid X) \right] \right\}. \tag{5}$$

The conditional epiplexity *and* time-bounded entropy *are defined as:*

$$\mathrm{S}_T(Y \mid X) := \left| \mathrm{P}^\star_{Y|X} \right|, \quad \mathrm{H}_T(Y \mid X) := \mathbb{E}_{(X,Y)} \left[-\log P^\star_{Y|X}(y \mid x) \right]. \tag{6}$$

Outline

1 Background

2 Method

3 **Experiments**

4 Discussion

Experiments

Measuring Epiplexity

给定训练序列

$$Z_{0:M-1} = \{Z_0, Z_1, \ldots, Z_{M-1}\},$$

定义一个模型序列：

$$P_0 \to P_1 \to \cdots \to P_M$$

其中：

- P_0：随机初始化模型
- P_{i+1}：在样本 Z_i 上训练 P_i 后得到

Experiments

Measuring Epiplexity

对每个样本 Z_i：

1. 用当前模型 P_i 预测
2. 用编码长度：

$$\log \frac{1}{P_i(Z_i)}$$

   对 Z_i 进行熵编码
3. 再用 Z_i 更新模型

Experiments

Measuring Epiplexity

于是得到：

$$L(Z_{0:M-1}, P_M) = \sum_{i=0}^{M-1} \log \frac{1}{P_i(Z_i)}$$

这个量同时编码了：

- 数据 Z
- 最终模型 P_M

因为：

> 解码端同步训练，可以重建整个训练过程 → 自动得到 P_M

Experiments

Measuring Epiplexity

利用 Kolmogorov complexity 中的性质：

$$K(P_M) = K(Z, P_M) - K(Z \mid P_M) + O(1)$$

对应到编码长度：

$$|P_M| \approx L(Z, P_M) - L(Z \mid P_M)$$

定义：

$$L(Z \mid P_M) = \sum_{i=0}^{M-1} \log \frac{1}{P_M(Z_i)}$$

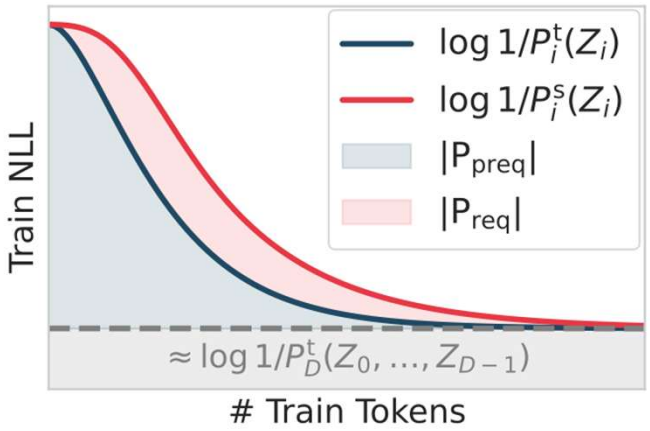
即：

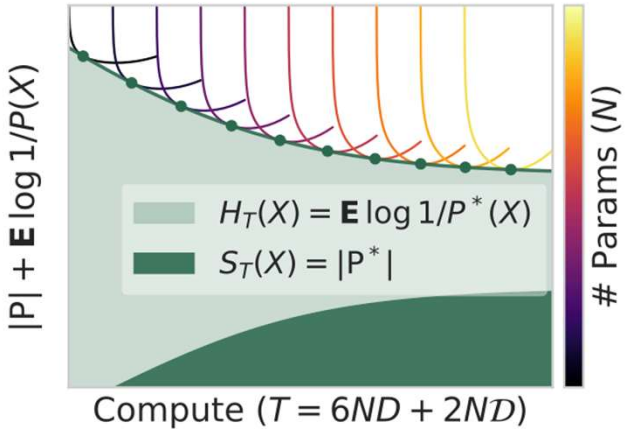
> 用最终模型 P_M 来重新编码整个训练数据

Experiments

Measuring Epiplexity

$$|P_{\text{preq}}| \approx \sum_{i=0}^{M-1} \left(\log \frac{1}{P_i(Z_i)} - \log \frac{1}{P_M(Z_i)} \right)$$





Experiments

Measuring Epiplexity

(1) Teacher（教师模型）

$$P_0^t, P_1^t, \ldots, P_{M-1}^t$$

- 通常来自真实训练过程的 checkpoint
- 近似真实分布 P_X

(2) Student（学生模型）

$$P_0^s \to P_1^s \to \cdots \to P_M^s$$

训练方式：

$$Z_i \sim P_i^t$$

Experiments

Measuring Epiplexity

每一步编码成本：

$$\mathrm{KL}(P_i^t \| P_i^s) + \log(1 + \mathrm{KL}) + O(1)$$

因此总编码长度：

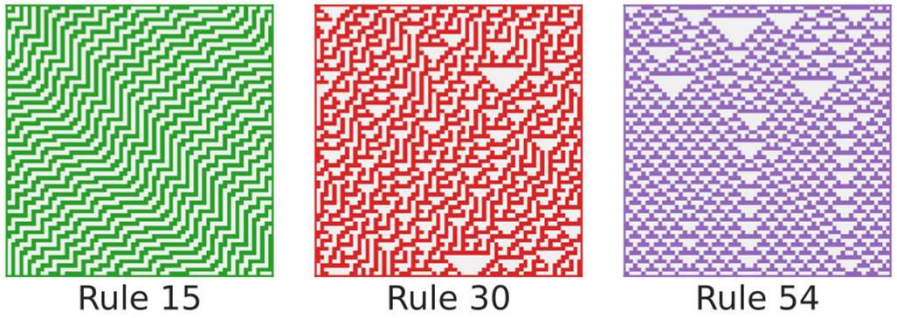
$$|P_{\text{req}}| \approx \sum_{i=0}^{M-1} \mathrm{KL}(P_i^t \| P_i^s)$$

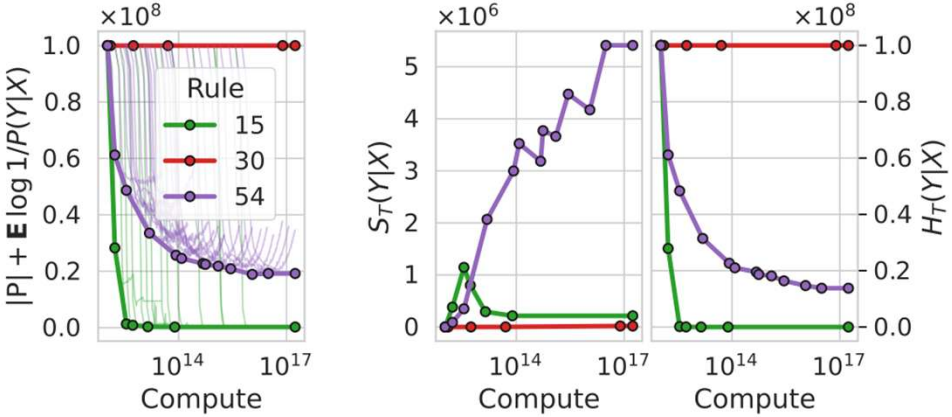
$$\mathrm{KL}(P_i^t \| P_i^s) \approx \log \frac{1}{P_i^s(Z_i)} - \log \frac{1}{P_i^t(Z_i)}$$

Experiments

- **Information Cannot be Created by Deterministic Transformations**

$$K(f(x)) \leq K(x) + K(f) + c$$

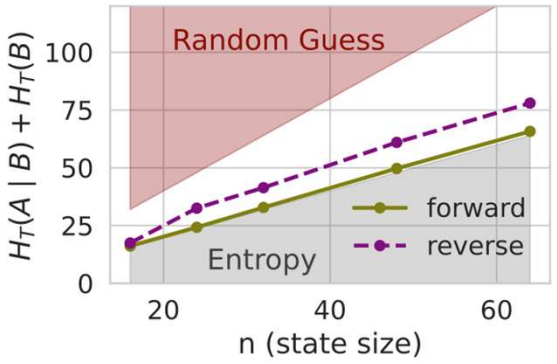




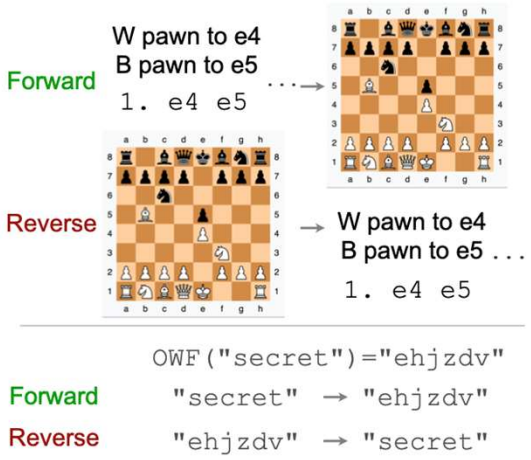
Experiments

- **Information Content is Independent of Factorization**

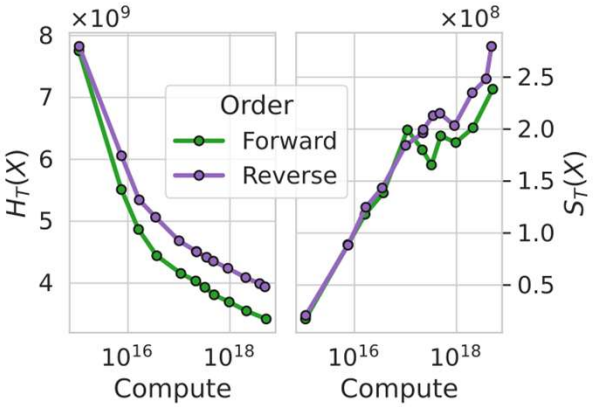
$$K(y \mid x)+K(x) = K(x \mid y)+K(y)+O(1)$$



(a) One way functions



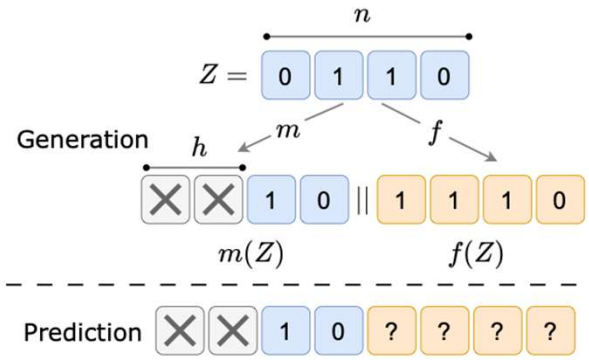
(b) Factorization order



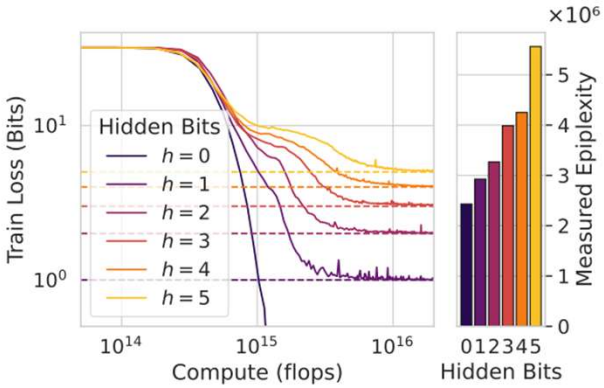
(c) Chess orderings

Experiments

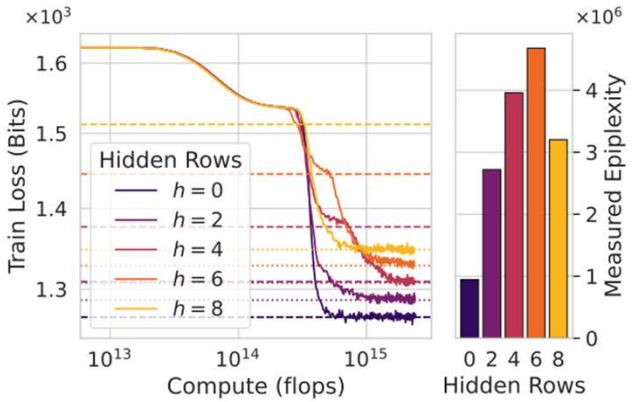
- **Likelihood Modeling is Merely Distribution Matching**



(a) Data generating process



(b) Induction (hard)



(c) Induction (easy)

Outline

1 Background

2 Method

3 Experiments

4 **Discussion**

Discussion

$$X = \text{structure} + \text{noise}$$

Thanks!