From Correlation to Grammar (Cagnetta)

How NN acquire hierarchical structure From data!

Q. Why does DL works?

- Curse of dimensionality (Poverty of the stimulus)

↓

difficult to sample enough

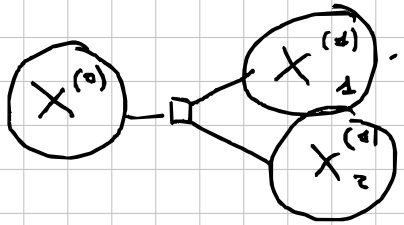
Intelligence becomes experimentally accessible.

- Introduce Random Hierarchies

Hierarchical Compositionality: complex meaning From simple building blocks

→ Tree-like Generative Model

Goal: a controlled generator of sentences/sequences



- Draw a root symbol
- Draw a children given the parent

CONTEXT FREE GRAMMARS

$$Q\left((a,b) \mid X^{(0)} = r\right) = P_r\left(r \to (a,b)\right)$$

Finite, non recursive : all valid sentences have finite length

↓

LAYER PROBABILISTIC CONTEXT FREE GRAMMAR

RHH :

$X^{(0)} = a, b, c \ldots$

$X^{(1)}_1 = (a, a)$

$X^{(1)}_2 = (a, b)$

$Q\left((a,b) \mid X^{(0)} = c\right) = ?$

↓

PRODUCTION RULE

PARENTS	SIBLING PAIRS	RULE BOOK
(a)	$(a,a)\ (a,b)\ (a,c)$	$a \to (a,a)$
(b)	$(b,a)\ (b,b)\ (b,c)$	$a \to (b,c)$
(c)	$(c,a)\ (c,b)\ (c,c)$	⋮

Assumption : Tree of depth L, independent & random rules at each layer

Data properties are available analytically and depend on L, vocab size v, m rules

sequence length $d = 2^L$

sequences $V^{(d-1)}$

LATENT variables lower the dimensionality

$$V^d \to V^{d/2} \to V^{d/4} \to \dots \to V$$

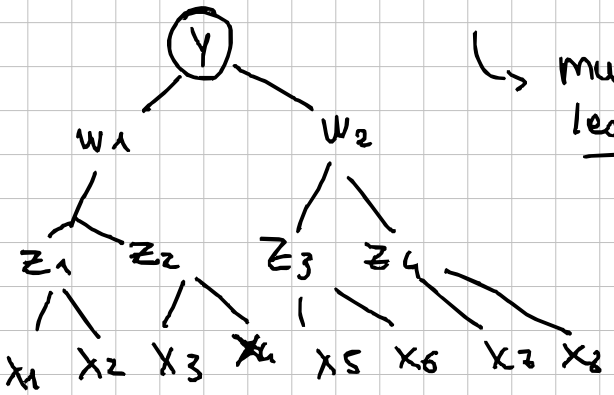
Can the learner infer them from labelled examples?

$Y = r$ $\quad X_1$ $\quad X_2$

$$P(Y = r \mid X_{1:8}) = 1/V$$

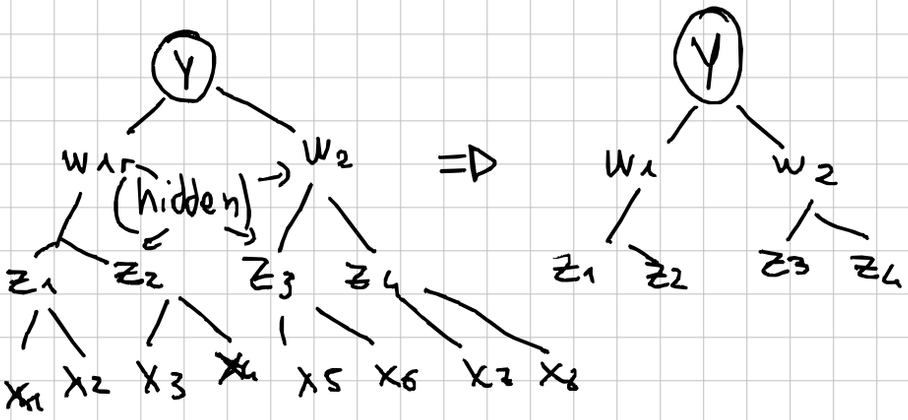
$$P(Y = r \mid X_{1:2}) = \frac{1}{V} + \varepsilon$$

↳ mutual info root leaves



↳ i can cluster groups of symbols to the extent variables

We can reduce



Training examples → Low-order models

↳ hidden symbol models

sample complexity #{data} s.t moments

>> Samling data

Method of moments

IF $(X_i, Y_{i+\ell})$ are generated by H

$$C_{ab}(y): \Pr((Y_i, X_{i+1}); Y=y)$$

$$= \sum \underbrace{Q((a,b|H=h)}_{v^2\text{-by-}v} \underbrace{\Pr(H=h; Y=y)}_{v\text{-by-}v}$$

SAMPLE COMPLEXITY REDUCTION

$$|C_{ab}(y)|^2 >> |C_{ab}(y) - \hat{C}_{ab}(y)|^2$$

$$= \Theta(1/\#\{data\})$$

Test on NN

Root classification

Given depth to store latent variables

↓

exp (d)

Signal $|C_{ab}(z)|^2$

$= \Theta(1/(vm)^\ell)$

Sample complexity

$\#\{data\} \sim vm^\ell$

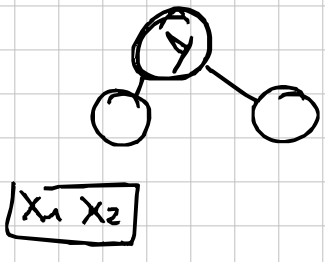
poly (d)

$(vm)^2$

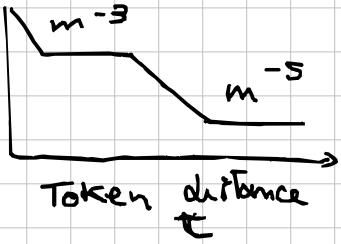
- VARYING TREE TOPOLOGY

Correlations reveal latent variables and local topology

Last-token prediction

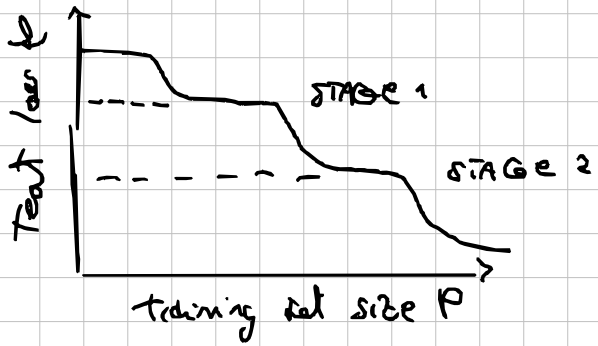


Signal decay with distance



→ Only bottleneck is DATASET size

Empirical learning curve



- All sequences V^d
- valid sequences $m d^v$
- sample size $V m^L \sim d^{\theta(\ln(m))}$

EXPONENTIAL CREATIVITY WITH POLYNOMIAL experience

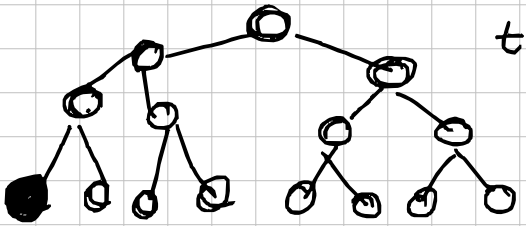
What has the network learned?

Linear Representation Hypothesis

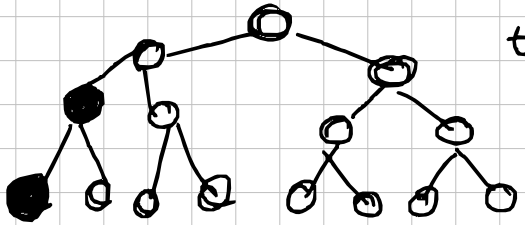
$$\phi_{\ell,t}(X) = A \cdot H + E$$

First-layer activations encode Z_1

$\delta_H(\phi)$ = Share of ϕ's variance due to H's code



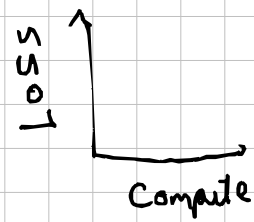
$t = 1$



$t > 2$

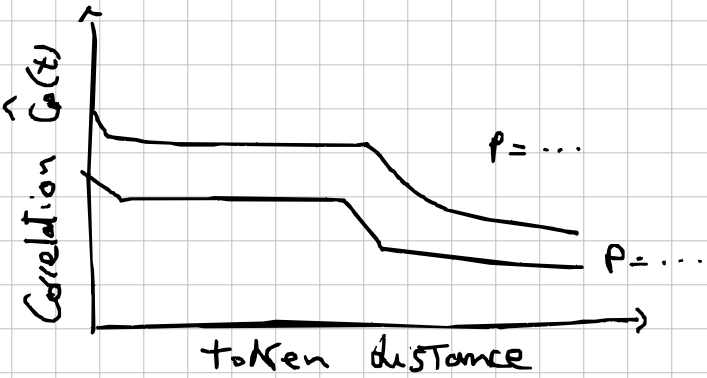
BEYOND SYNTHETIC DATA

→ Neural scaling law



Effective prediction time horizon

From the RHM: {Finite # of data} = P



turns out $n^*(P) \sim P^{1/2\beta}$

$\mathcal{L}_n^*(P) \sim P^{-\delta/2\beta}$

NO FITTING PARAMS, PURELY DATA STATISTICS

Persistence of latent variables

$$d_{e,t}(X) = A \cdot H + \varepsilon$$

- Deeper layer code for deeper latent variables
- Deeper latent variables influence longer portions of the context