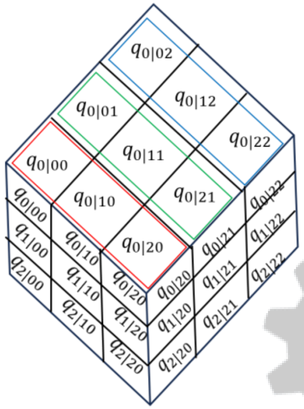
言語の統語と意味をつなぐ n-gramの幾何代数

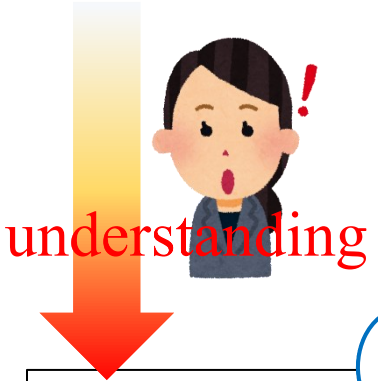
Shohei Hidaka

Japan Advanced Institute of Science and Technology (JAIST)

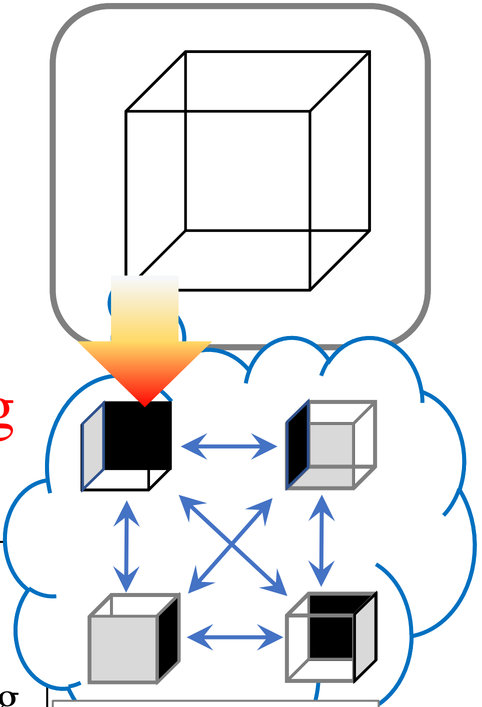


Motivation: understanding our understanding as structure finding from an ambiguous data

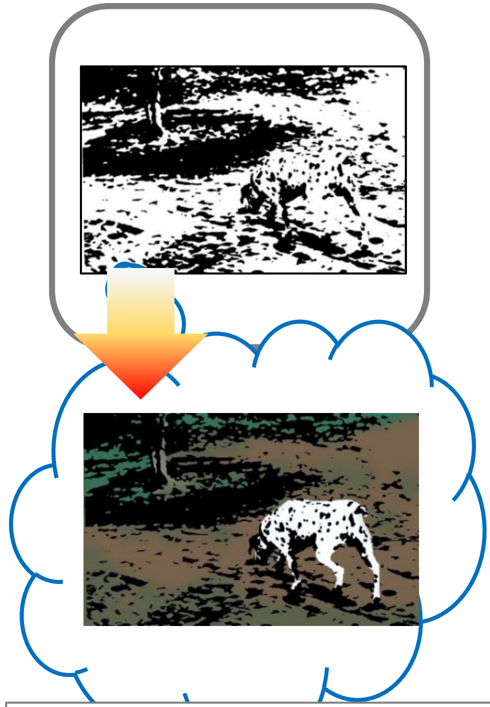
Data with ambiguity



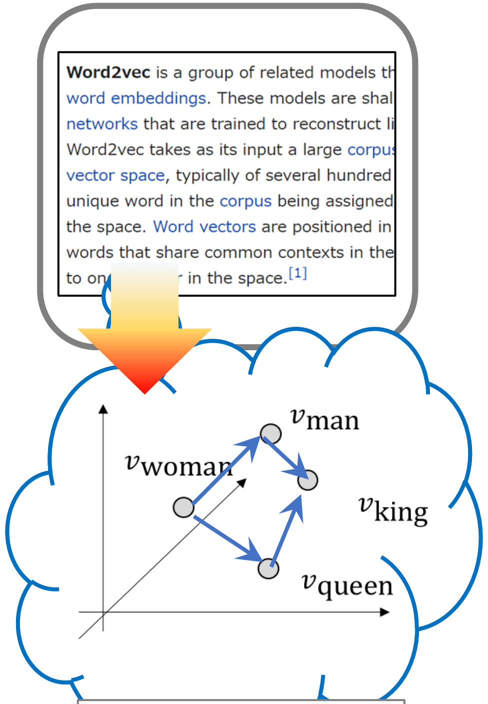
Interpretation
Sense making
Structure finding



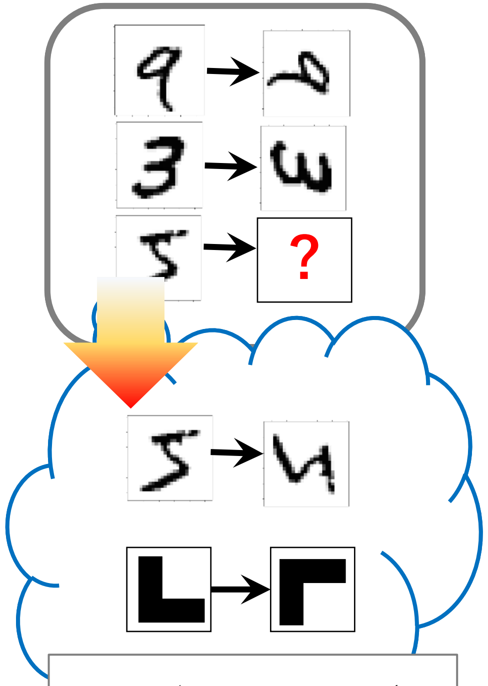
Perception of missing structure



Object perception



Grammar
Metaphor/ analogy



Finding a rule
Hypothesis building

Sprugnoli+(2017);

Outline & Summary

- I propose *n-gram algebra* to characterize math. theory of n-grams.
  - Motivation
    - Considering grammar learning (AGL), not clear how n-gram statistics should be handled.
    - Word vectors represent “semantic structure” as its geometric shapes e.g., “parallelogram”.
    - To understand LLMs, n-gram algebra is probability useful.
- n-gram algebra
  - n-tensor of non-negative real values with tensor convolution.
  - Hierarchy of n-gram formed by ideals.
  - It integrates “distributional hypothesis” and “universal grammar”.
- Analyzing Transformer as a n-gram estimator

Hidaka, S. (2025) Skip-bigrams reconstruct trigrams in 2-word languages. NLP2025

Motivation 1: Artificial Grammar Learning

- Sabbatical (2023) @ City, Univ. of London with Prof. E. Pothos

Psychological Bulletin
2007, Vol. 133, No. 2, 227–244

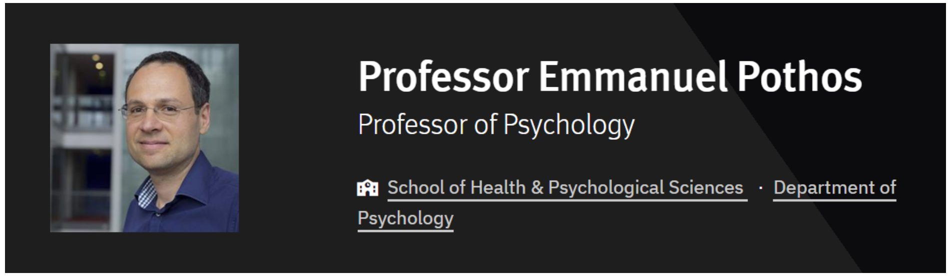
0033-2909/07/$12.00 DOI: 10.1037/0033-2909.133.2.227

Theories of Artificial Grammar Learning

Emmanuel M. Pothos
Swansea University

Artificial grammar learning (AGL) is one of the most commonly used paradigms for the study of implicit learning and the contrast between rules, similarity, and associative learning. Despite five decades of extensive research, however, a satisfactory theoretical consensus has not been forthcoming. Theoretical accounts of AGL are reviewed, together with relevant human experimental and neuroscience data. The author concludes that satisfactory understanding of AGL requires (a) an understanding of implicit knowledge as knowledge that is not consciously activated at the time of a cognitive operation; this could be because the corresponding representations are impoverished or they cannot be concurrently supported in working memory with other representations or operations, and (b) adopting a frequency-independent view of rule knowledge and contrasting rule knowledge with specific similarity and associative learning (co-occurrence) knowledge.

Keywords: artificial grammar learning, implicit learning, rules, similarity, associative learning

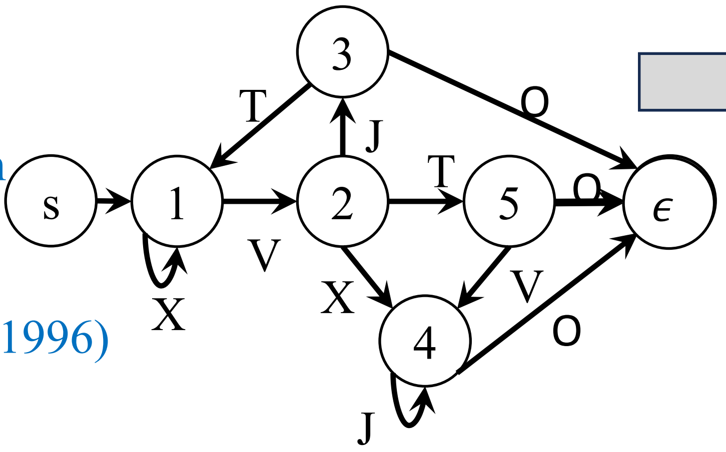


CITY
UNIVERSITY OF LONDON
EST 1894

AGL task (in a typical form)

- *Learning session*: Participants go through a set of stimuli (with some properties following some set of rules).
- *Test session*: Participants answer whether each of new set of stimuli follows underlying rules common in the learned stimuli.
- *Analysis*: They analyze what properties of stimuli or underlying rules would predict their participants' grammaticality judgements.

Finite state automaton as an underlying artificial grammar (Knowlton & Squire, 1996)



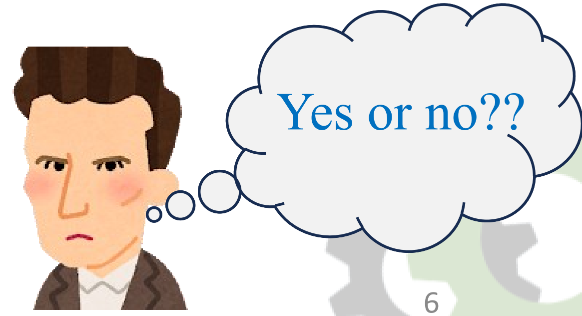
Learning session

Stimuli to be learned

"VJ0"
"VJTVJ0"
"VJTVJTVJ0"
"VJTVJTVJTVJTXVJTVTVJ0"
"VJTVJTVJTVT0"
"VJTVJTVJTVX0“
"VJTVJTVJTVXJJ0"
"VJTVJTVJTXVJTVT0"

Test session

Does "JX" follow the rule?

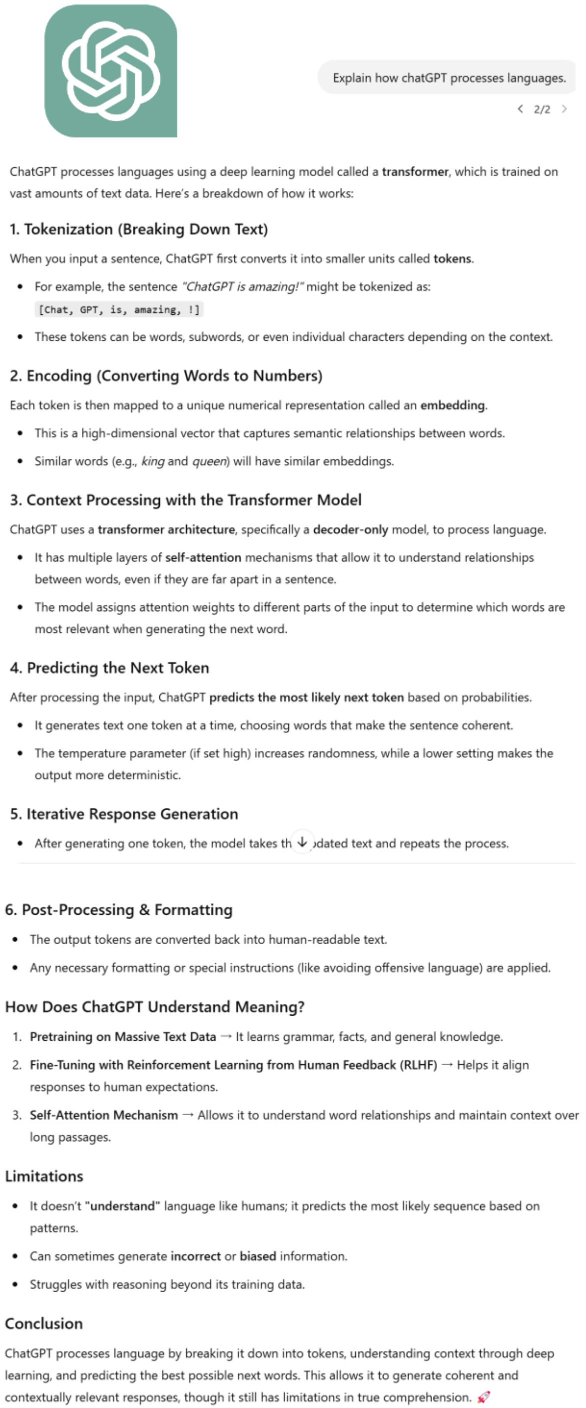


Is “chunk strength” reasonable? (Knowlton & Squire, 1996)

- Chunk strength is defined by likelihood of test sequence given the training sample.
- Psychological studies of AGLs often employ the sum of chunk strength of *bigram and trigram*.
- I felt so uncomfortable to have “sum of bigram and trigram”, as they are qualitatively different in general.

Motivation 2: Why and how LLMs work so well?

- Obviously, it requires a whole lot of techniques, but one of keys to build an internal architecture is the next token prediction task.



Explain how chatGPT processes languages.

2/2

ChatGPT processes languages using a deep learning model called a **transformer**, which is trained on vast amounts of text data. Here's a breakdown of how it works:

1. Tokenization (Breaking Down Text)

When you input a sentence, ChatGPT first converts it into smaller units called **tokens**.

- For example, the sentence *"ChatGPT is amazing!"* might be tokenized as: `[Chat, GPT, is, amazing, !]`
- These tokens can be words, subwords, or even individual characters depending on the context.

2. Encoding (Converting Words to Numbers)

Each token is then mapped to a unique numerical representation called an **embedding**.

- This is a high-dimensional vector that captures semantic relationships between words.
- Similar words (e.g., *king* and *queen*) will have similar embeddings.

3. Context Processing with the Transformer Model

ChatGPT uses a **transformer architecture**, specifically a **decoder-only** model, to process language.

- It has multiple layers of **self-attention** mechanisms that allow it to understand relationships between words, even if they are far apart in a sentence.
- The model assigns attention weights to different parts of the input to determine which words are most relevant when generating the next word.

4. Predicting the Next Token

After processing the input, ChatGPT **predicts the most likely next token** based on probabilities.

- It generates text one token at a time, choosing words that make the sentence coherent.
- The temperature parameter (if set high) increases randomness, while a lower setting makes the output more deterministic.

5. Iterative Response Generation

- After generating one token, the model takes th dated text and repeats the process.

6. Post-Processing & Formatting

- The output tokens are converted back into human-readable text.
- Any necessary formatting or special instructions (like avoiding offensive language) are applied.

How Does ChatGPT Understand Meaning?

1. **Pretraining on Massive Text Data** → It learns grammar, facts, and general knowledge.
2. **Fine-Tuning with Reinforcement Learning from Human Feedback (RLHF)** → Helps it align responses to human expectations.
3. **Self-Attention Mechanism** → Allows it to understand word relationships and maintain context over long passages.

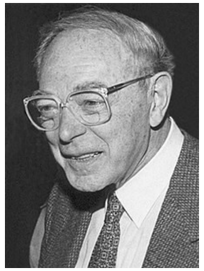
Limitations

- It doesn't "**understand**" language like humans; it predicts the most likely sequence based on patterns.
- Can sometimes generate **incorrect** or **biased** information.
- Struggles with reasoning beyond its training data.

Conclusion

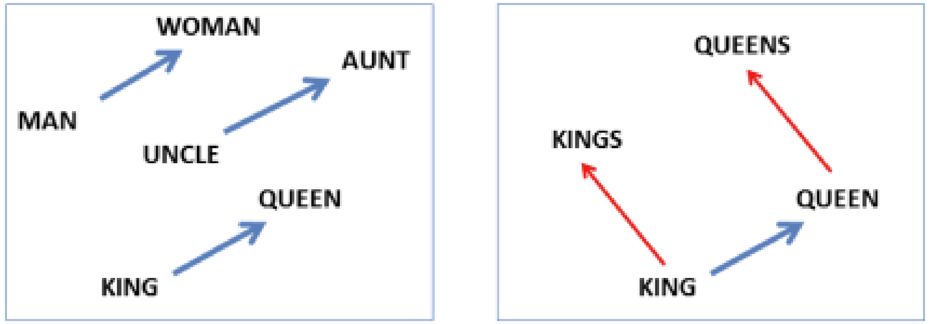
ChatGPT processes language by breaking it down into tokens, understanding context through deep learning, and predicting the best possible next words. This allows it to generate coherent and contextually relevant responses, though it still has limitations in true comprehension. 🚀

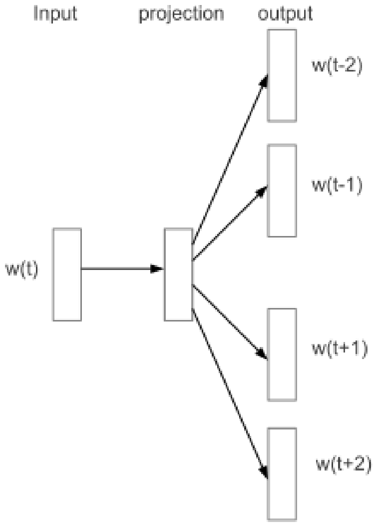
Distributional hypothesis (Harris, 1954)
"You shall know a word by the company it keeps."



- What the word means is understood by its context words
  - “she eats ____ every morning.” “____ is a fruit.” accepts both “apple” & “banana”.

- Word2vec (Mikolov+, 2013) learns word vectors by “filling-in-blank” problem, and construct “parallelograms”.

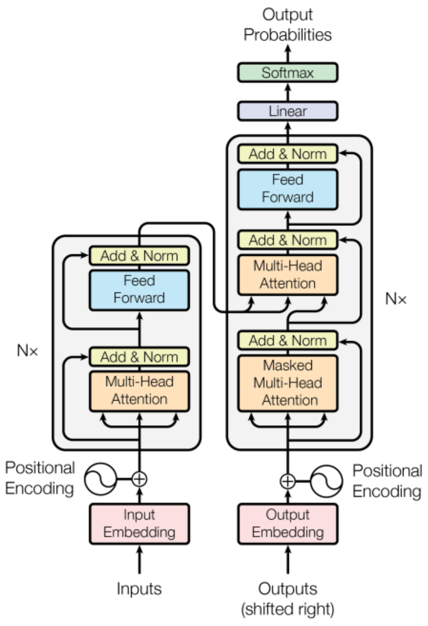




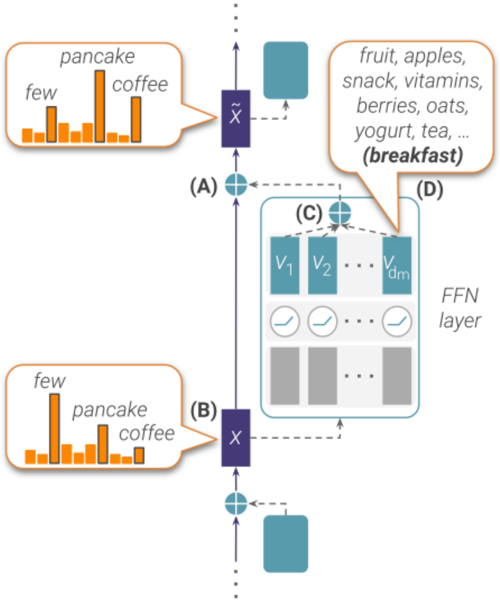
$$P(y|x) = \frac{\exp(v_y \cdot v_x)}{\sum_{w \in W} \exp(v_w \cdot v_x)}$$

Transformer is likely to build contexturalized word vectors.

- A series of updating of context through layers.

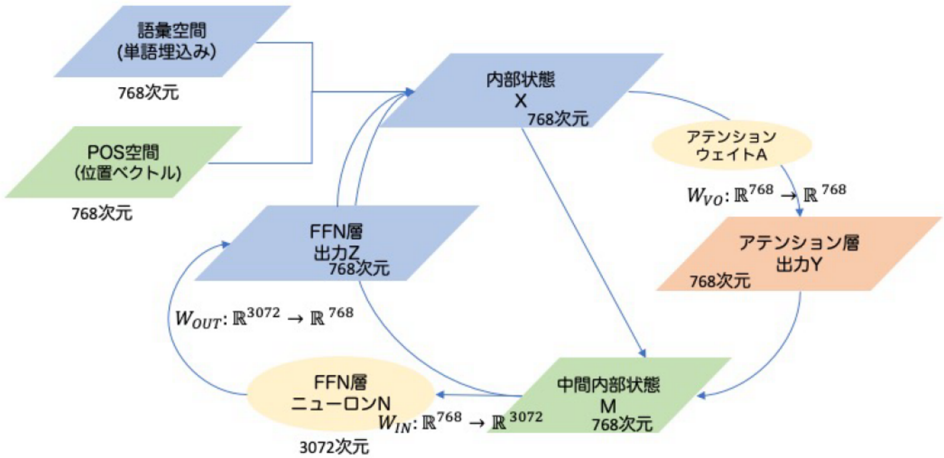


(Vaswani+ 2017)



(Geva+ 2022)

② [仮説]Transformer各層は擬似直交している



- 予想１：アテンション層とFFN層は直交している
- 予想２：FFN層は単語埋め込みと同じ部分空間にある
- 予想３：FFN層の出力は解釈可能であり，文脈化を担っている

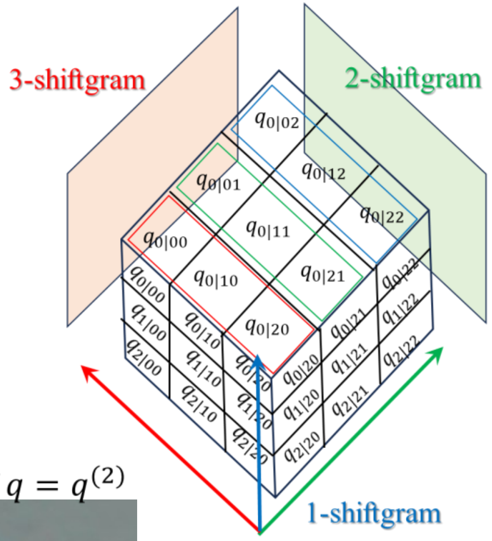
(前田+, P2-12@NLP2025)
部分空間の擬似直交性によるTransformer言語モデルの内部表現の解釈

LLMs process higher order statistics as whole, while it process second order one locally.

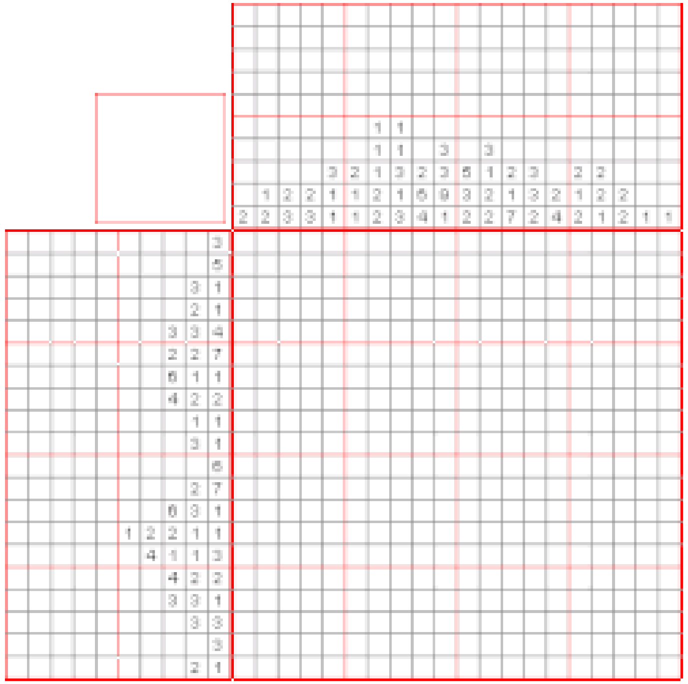
- General observation
  - LLMs seem to process **more than pairwise (2-gram) statistics as a whole**.
  - However, LLMs processes **only pairwise statistics** (i.e., inner products) in each step and layer.
- Theoretical background
  - Bigrams are **NOT sufficient** in general to infer all the statistical structure in 3-grams.
  - A high order statistics such as probability over trigrams $P(X_t, X_{t-1}, X_{t-2})$ *cannot be reduced to any set of skip-grams* $P(X_t, X_{t-1}), P(X_t, X_{t-2}), P(X_{t-1}, X_{t-2})$ *in general.*
- Two scenarios on natural languages and their coding
  - H0: Languages are simple enough to be coded by bigrams.
  - H1: Skip-bigrams are complex enough to code languages.
    - There is some unidentified theoretical pathway to connect a set of skip-bigrams to statistical structure over 3- or higher n-grams.

周辺化情報から構造を推定

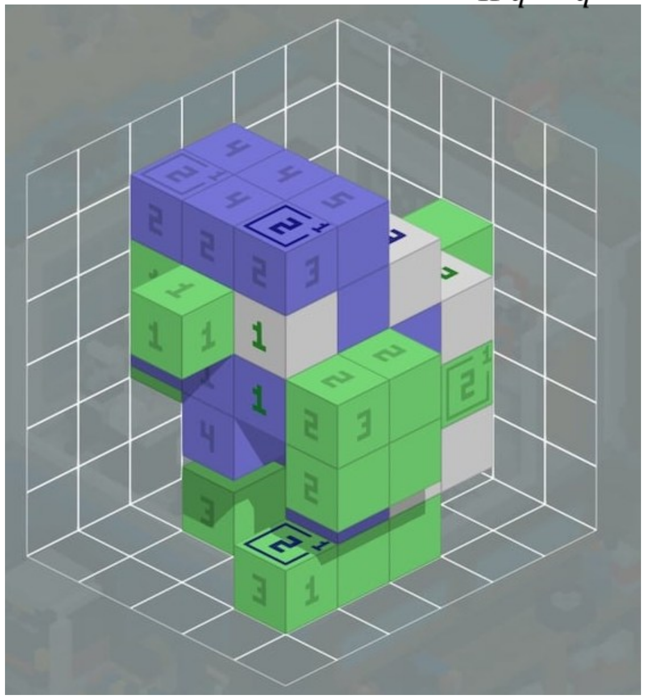
必ずしも一意に決まらない→どんな条件で構造が特定可能か？



お絵描きロジック/ nonogram



https://ja.wikipedia.org/wiki/%E3%81%8A%E7%B5%B5%E3%81%8B%E3%81%8D%E3%83%AD%E3%82%B8%E3%83%83%E3%82%AF



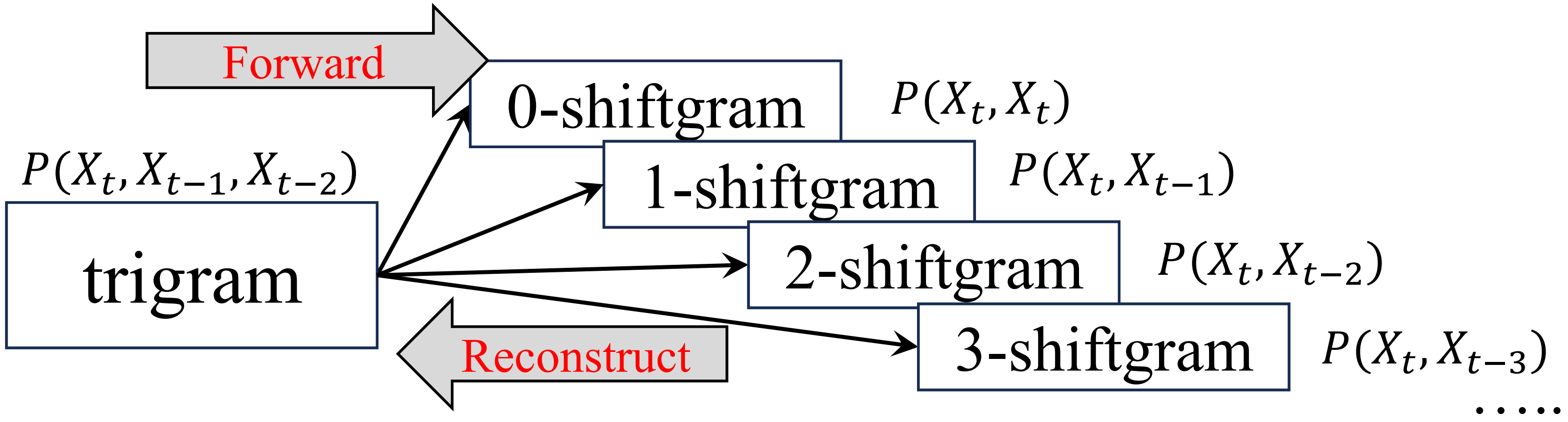
https://automaton-media.com/articles/newsjp/voxelgram-2-20250311-331077/

Hypothesis: n-grams can be decomposed into skip-bigrams.

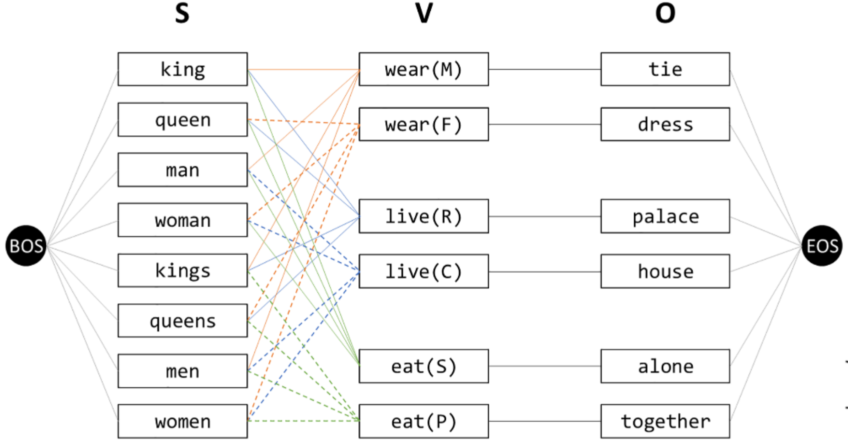
- Questions
  - Can we estimate trigram or higher n-grams from a series of “skip-grams”?
  - Skip-grams (Mikolov+ 2013)
    - pairs of words $(X_t, X_{t-1}), (X_t, X_{t-2})$, ... are said *skip-grams* for a series of words X_t, X_{t-1}, X_{t-2}, ..., and they have been used as a form of inputs to word vector models and LLMs.

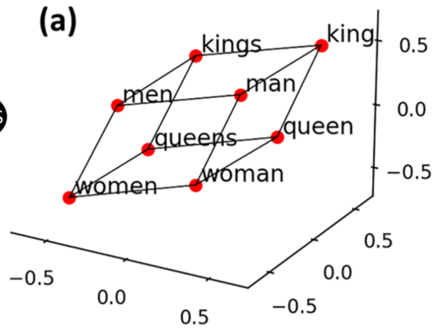
What is the relationship between skip-grams and trigram probabilities?

- Hypothesis: an essential isomorphism between a series of m-shiftgrams and trigram.
  - *m-shiftgrams* $P(X_t, X_{t-m})$ $(m = 0,1, \ldots)$: a mathematical notion formalizes "skip-grams".
    - 0-shiftgram = unigram, 1-shiftgram = bigram, 2-shiftgram = 1-skip-gram, and so forth.
- Is it possible to reconstruct trigram from the series of bigrams?
  - Forward: Given Ps over trigrams, one can compute m-shiftgrams for any m.
  - Inverse: It is unclear whether a set of shiftgrams is enough to identify 3-gram.



A mini corpus with 3-gram statistics

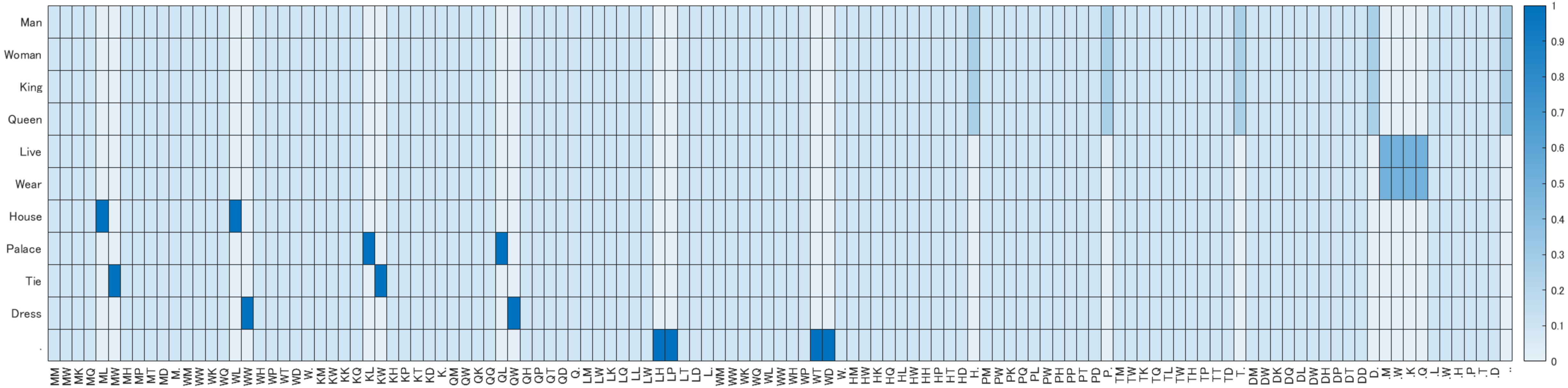




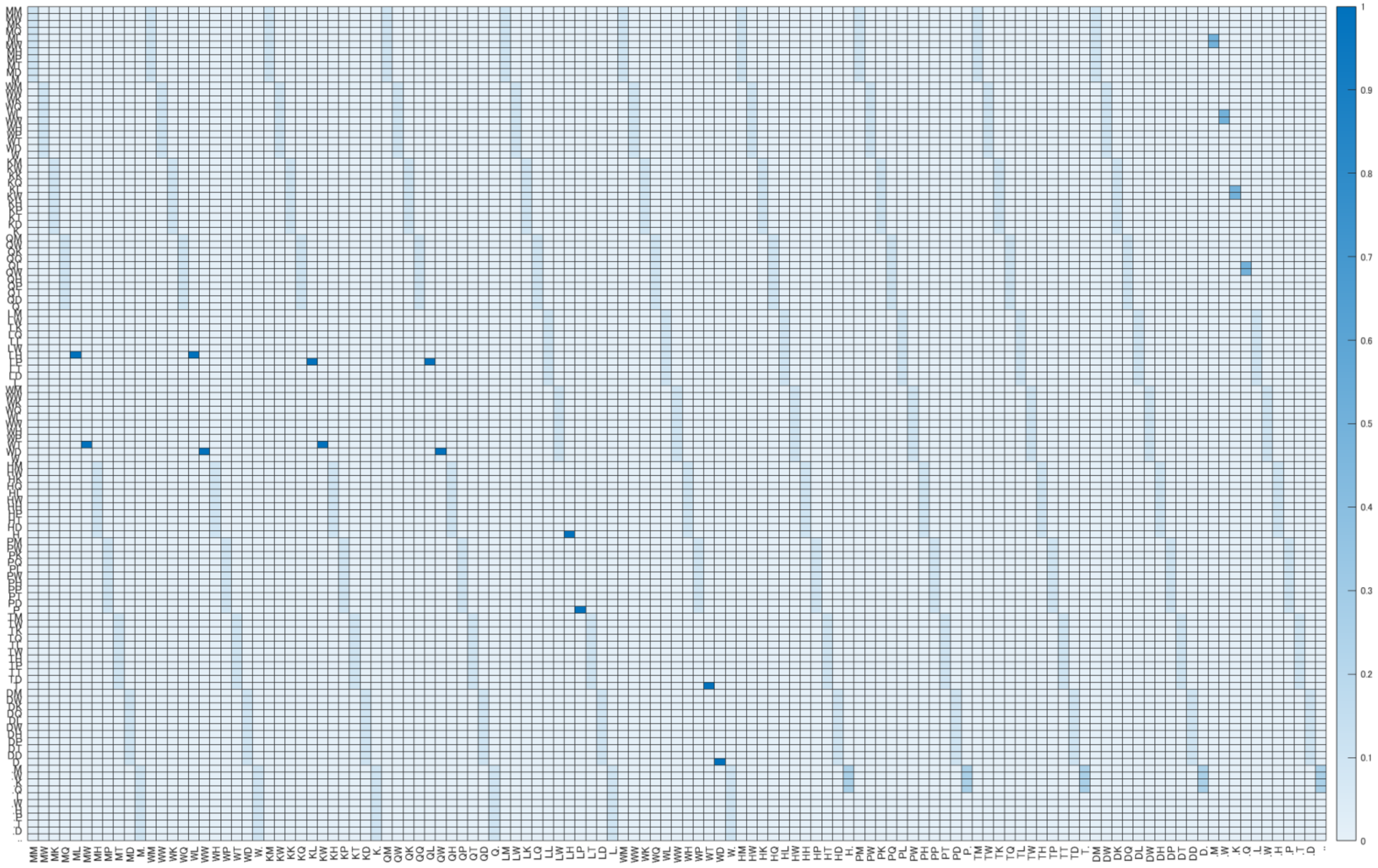
- Corpus with 11 words *"man"*, *"woman"*, *"king"*, *"queen"*, *"live"*, *"wear"*, *"house"*, *"palace"*, *"tie"*, *"dress"* and period *"."*.
- Grammar
  - Any of the 4 subject nouns "man", "woman", "king", and "queen" follows ".".
  - Only one of the 2 verbs "live" or "wear" follows each noun.
  - Each of the 4 object nouns "house", "palace", "tie" and "dress" follows after a particular SV combination: "king live palace".
  - The period "." follows any VO pair, and some subjective noun follows "O.".
- Outcomes:
  - Only 8 sentences appear in this corpus with some probability: 'KLP.', 'KWT.', 'MLH.', 'MWT.', 'QLP.', 'QWD.', 'WLH.', and 'WWD.'.

Torii, Maeda, & Hidaka (2024) Distributional hypothesis as isomorphism between word-word co-occurrence and analogical parallelogram. PLoS ONE 19(10): e0312151.

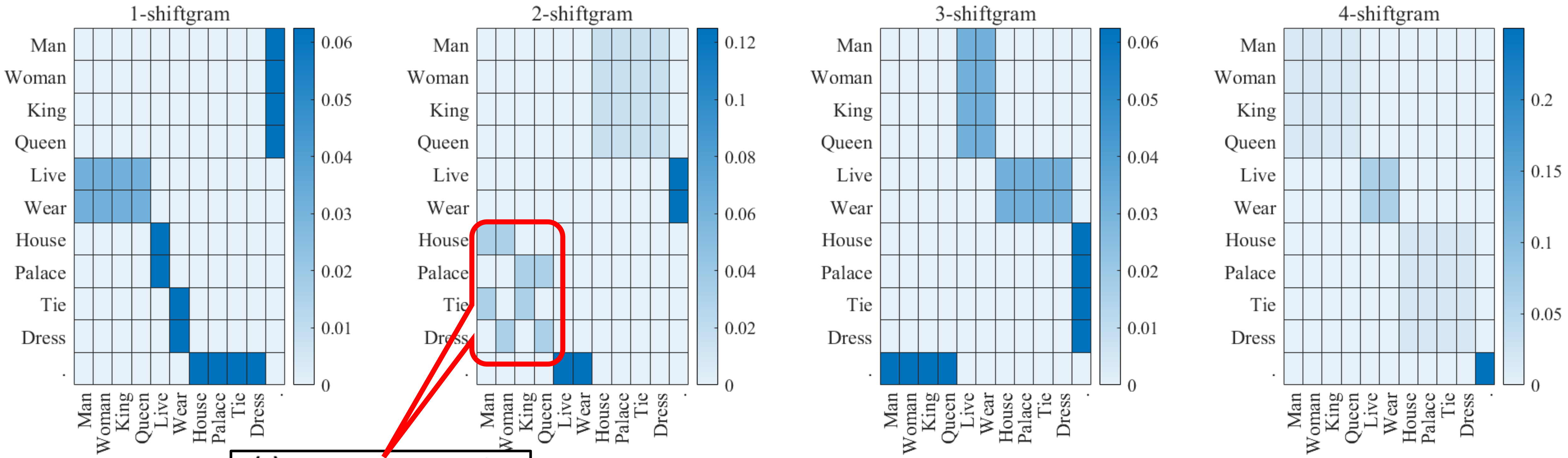
11×11^2 conditional trigram matrix $P(X_t \mid X_{t-1}, X_{t-2})$

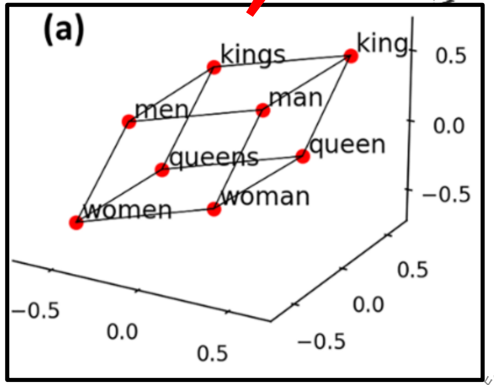


$11^2 \times 11^2$ trigram transition matrix $P(X_t, X_{t-1} \mid X_{t-1}, X_{t-2})$



11×11 conditional m-shiftgrams $P(X_t \mid X_{t-m})$



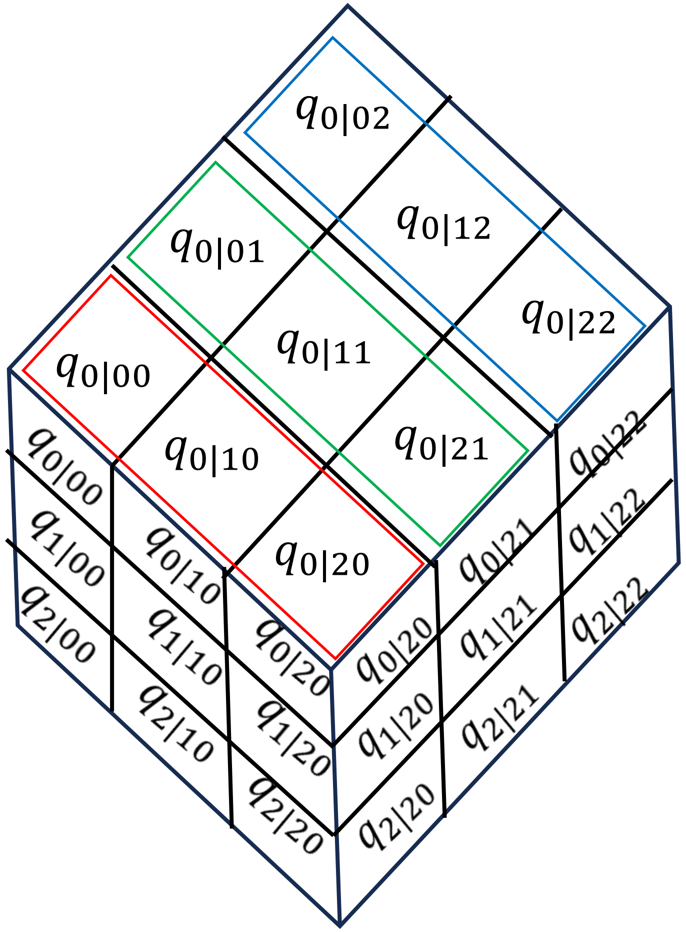


Conditional: each column-sum is normalized to be 1.
Joint (omitted): non-negative vector, matrix or tensor.
Most of (joint) tensors is defined over non-negative cone, while conditional tensors is a projective space of it.

Quick mid summary

- Question: is it possible to represent n-gram by some set of m-shiftgrams?
- Probably, yes: Demonstration of m-shiftgrams, with which it found geo. shape by a small toy corpus.
  - A set of m-shiftgrams seems to sufficiently represent n-grams, as its shows a characteristic parallelopiped of “man-king-woman-queen”.
- Is this finding generalizable? → math analysis of m-shiftgrams.

Trigram probabilities as a third order tensor

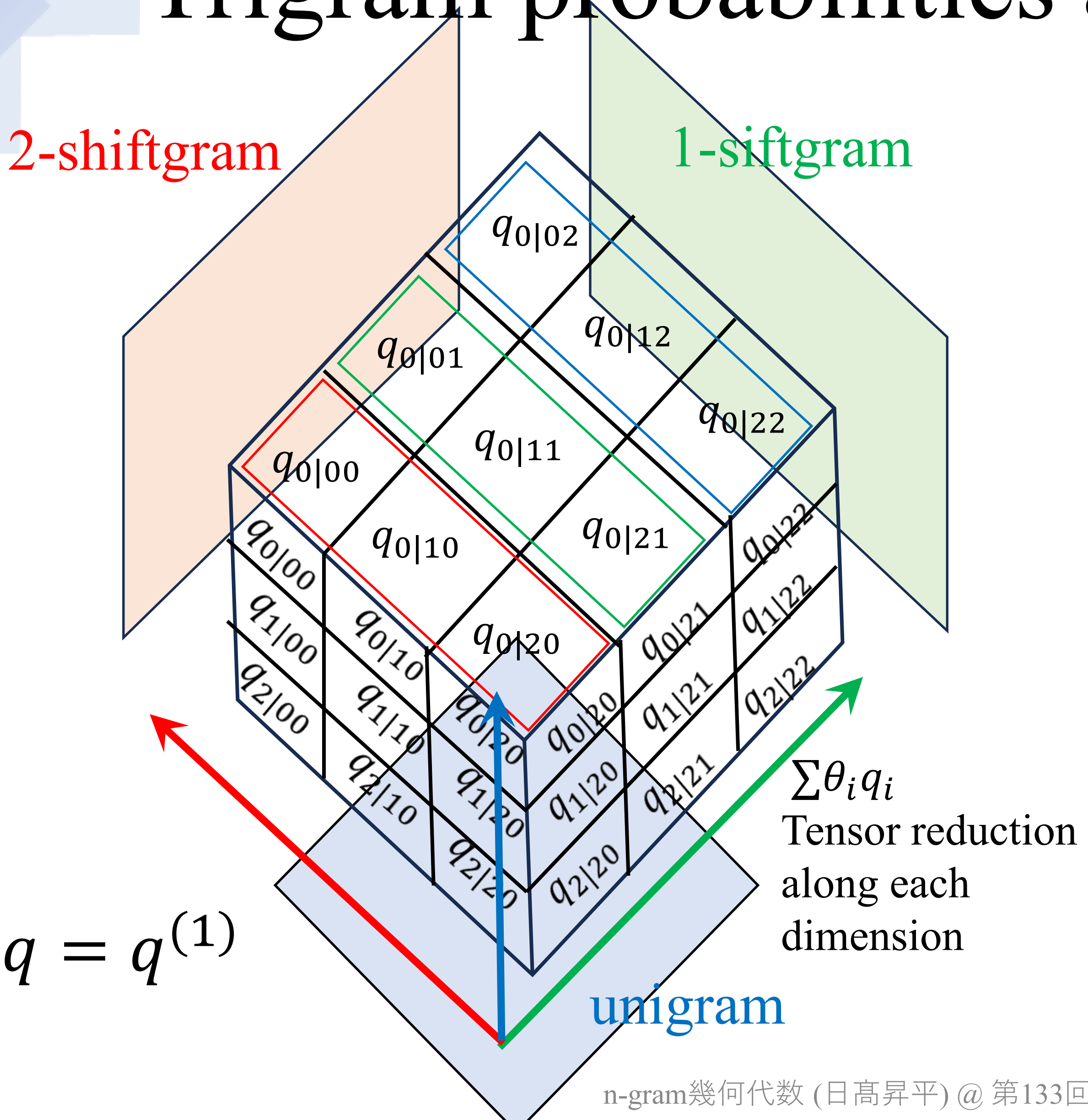


$q_{0\|00}$	$q_{0\|10}$	$q_{0\|20}$
$q_{1\|00}$	$q_{1\|10}$	$q_{1\|20}$
$q_{2\|00}$	$q_{2\|10}$	$q_{2\|20}$

$q_{0\|01}$	$q_{0\|11}$	$q_{0\|21}$
$q_{1\|01}$	$q_{1\|11}$	$q_{1\|21}$
$q_{2\|01}$	$q_{2\|11}$	$q_{2\|21}$

$q_{0\|02}$	$q_{0\|12}$	$q_{0\|22}$
$q_{1\|02}$	$q_{1\|12}$	$q_{1\|22}$
$q_{2\|02}$	$q_{2\|12}$	$q_{2\|22}$

Trigram probabilities as a third order tensor



If $q = q^{(1)}$

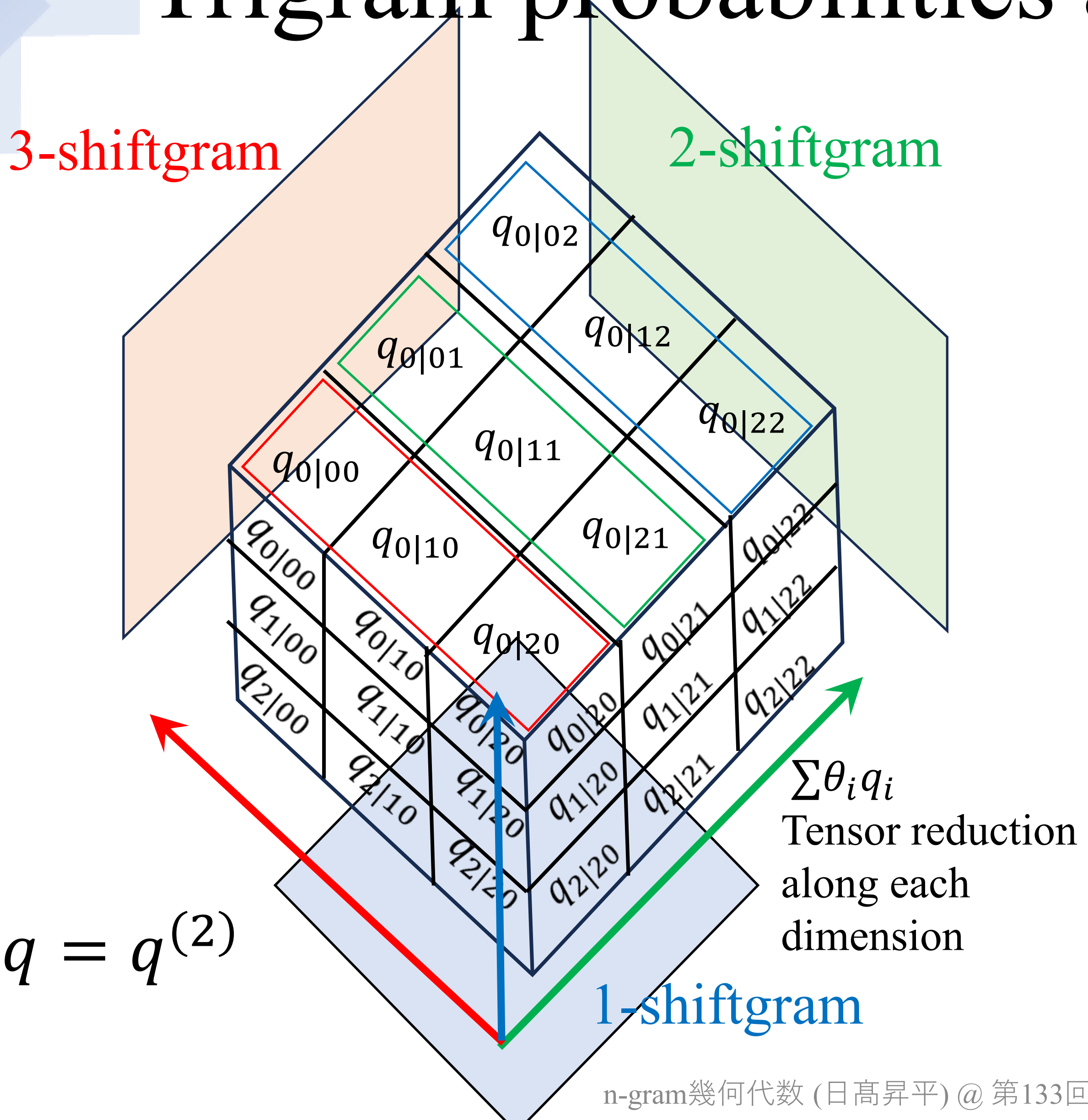
$q_{0\|00}$	$q_{0\|10}$	$q_{0\|20}$
$q_{1\|00}$	$q_{1\|10}$	$q_{1\|20}$
$q_{2\|00}$	$q_{2\|10}$	$q_{2\|20}$

$q_{0\|01}$	$q_{0\|11}$	$q_{0\|21}$
$q_{1\|01}$	$q_{1\|11}$	$q_{1\|21}$
$q_{2\|01}$	$q_{2\|11}$	$q_{2\|21}$

$q_{0\|02}$	$q_{0\|12}$	$q_{0\|22}$
$q_{1\|02}$	$q_{1\|12}$	$q_{1\|22}$
$q_{2\|02}$	$q_{2\|12}$	$q_{2\|22}$

The number of remaining free variables is $(K-1)^n$, after fixing unigram, 1- and 2-shift-gram

Trigram probabilities as a third order tensor



If $q = q^{(2)}$

$q_{0\|00}$	$q_{0\|10}$	$q_{0\|20}$
$q_{1\|00}$	$q_{1\|10}$	$q_{1\|20}$
$q_{2\|00}$	$q_{2\|10}$	$q_{2\|20}$

$q_{0\|01}$	$q_{0\|11}$	$q_{0\|21}$
$q_{1\|01}$	$q_{1\|11}$	$q_{1\|21}$
$q_{2\|01}$	$q_{2\|11}$	$q_{2\|21}$

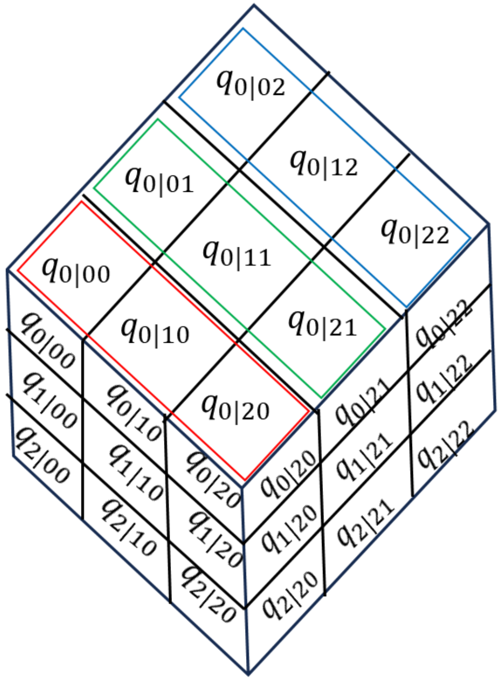
$q_{0\|02}$	$q_{0\|12}$	$q_{0\|22}$
$q_{1\|02}$	$q_{1\|12}$	$q_{1\|22}$
$q_{2\|02}$	$q_{2\|12}$	$q_{2\|22}$

The number of remaining free variables is $(K-1)^n$, after fixing unigram, 0- and 1- skip Bigram,

Mathematical notations

- $1_k \coloneqq \{1\}^k$ is k-dimensional *one vector*.
- $0_k \coloneqq \{0\}^k$ is k-dimensional *zero vector*.
- $\mathrm{e}_{k,i} \coloneqq \left(0_{i-1}^T, 1, 0_{k-i}^T\right) \in \{0,1\}^k$ is k-dimensional i-th *unit vector*.
- $I_k = \sum_{i=1}^{k} \mathrm{e}_{k,i}\mathrm{e}_{k,i}^T \in \{0,1\}^{k\times k}$ is k-dimensional identity matrix.
- $\mathbb{S}_{k-1} \coloneqq \left\{ x \in \mathbb{R}^k \mid 1_k^T x = 1, x \geq 0 \right\}$ is (k-1)-simplex in k-dimensional vector space.
- $\mathbb{S}_{k-1}^m \coloneqq \{ (x_1, \dots, x_m) \mid x_i \in \mathbb{S}_{k-1} \}$ is a set of matrix with m tuple of (k-1)-simplex vectors.
- $\otimes : \mathbb{R}^{m_0\times n_0} \times \mathbb{R}^{m_1\times n_1} \to \mathbb{R}^{m_0 m_1 \times n_0 n_1}$ is the Kronecker product of two matrices.

Math Formulation 1: n-gram probabilities over k words.



- *Conditional 3-gram probability tensor* over 2 words $\{0,1\}$ $(n = 3, k = 2)$:
  - $Q_3 := (q_{00}\ q_{10} | q_{01}\ q_{11}) = \left(\begin{array}{cc|cc} q_{0|00} & q_{0|10} & q_{0|01} & q_{0|11} \\ q_{1|00} & q_{1|10} & q_{1|01} & q_{1|11} \end{array}\right) \in \mathbb{S}_{k-1}^{k^{n-1}}$.
  - $q_{ij} = \begin{pmatrix} q_{0|ij} \\ q_{1|ij} \end{pmatrix} \in \mathbb{S}_{k-1}$. With identity $Q_3 = (1_k \otimes I_k)\overline{Q}_3$.
- *3-gram transition probability matrix*:
  - $\overline{Q}_3 := \begin{pmatrix} q_{00} & 0_k & q_{01} & 0_k \\ 0_k & q_{10} & 0_k & q_{11} \end{pmatrix} = \begin{array}{c} {\scriptstyle (t,t-1):} \\ (0,0) \\ (1,0) \\ (0,1) \\ (1,1) \end{array} \!\!\begin{array}{c} \begin{array}{cccc} (0,0) & (1,0) & (0,1) & (1,1) \end{array} : (t-1, t-2) \\ \begin{pmatrix} q_{0|00} & 0 & q_{0|01} & 0 \\ q_{1|00} & 0 & q_{1|01} & 0 \\ 0 & q_{0|10} & 0 & q_{0|01} \\ 0 & q_{1|10} & 0 & q_{1|01} \end{pmatrix} \end{array} \in \mathbb{S}_{k^{n-1}-1}^{k^{n-1}}$.

Math Formulation 2: $(n$-1)-gram prob. tensor is isomorphic to stationary vector of n-gram prob. tensor

- $\theta_2 = \overline{Q}_3\theta_2 \in \mathbb{S}_{k^{n-1}-1}$ denotes *stationary vector* of $\overline{Q}_3$:
  - $\overline{Q}_3\theta_2 = \begin{pmatrix} q_{0|00} & 0 & q_{0|01} & 0 \\ q_{1|00} & 0 & q_{1|01} & 0 \\ 0 & q_{0|10} & 0 & q_{0|01} \\ 0 & q_{1|10} & 0 & q_{1|01} \end{pmatrix} \begin{pmatrix} \theta_{00} \\ \theta_{10} \\ \theta_{01} \\ \theta_{11} \end{pmatrix} = \begin{pmatrix} \theta_{00} \\ \theta_{10} \\ \theta_{01} \\ \theta_{11} \end{pmatrix} = \theta_2 \in \mathbb{S}_{k^{n-1}-1}.$

- Lemma 1: Any joint *2-gram matrix* $\Theta_2 = \begin{pmatrix} \theta_{00} & \theta_{01} \\ \theta_{10} & \theta_{11} \end{pmatrix} \in \mathbb{R}^{k^{n-2} \times k^{n-2}}$ satisfies $\overline{Q}_3 \mathrm{vec}(\Theta_2) = \mathrm{vec}(\Theta_2) = \theta_2$.

Math Formulation 3: Recursive construction of lower *n*-grams

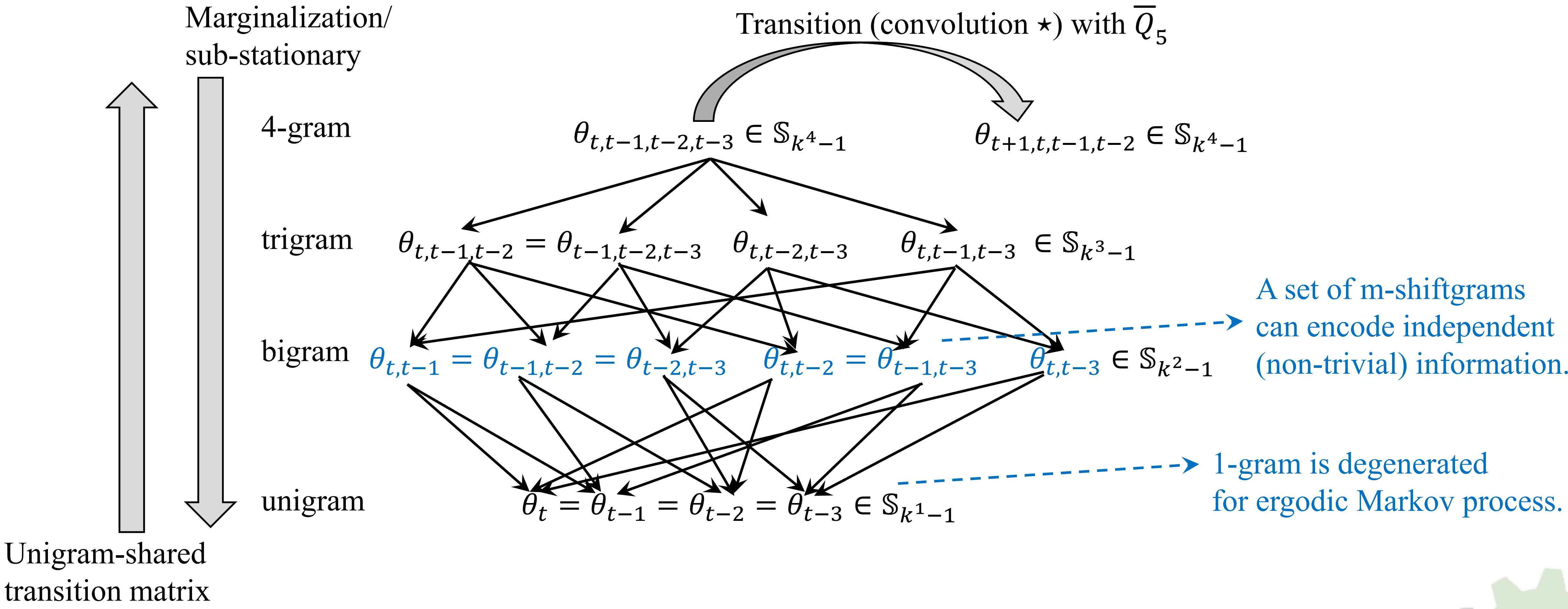
- *2-gram joint probability tensor* $\Theta_2 = \begin{pmatrix} \theta_{00} & \theta_{01} \\ \theta_{10} & \theta_{11} \end{pmatrix} \in \mathbb{R}^{k^{n-2} \times k^{n-2}}$ is stationary vector of $\overline{Q}_3$: $\overline{Q}_3 \mathrm{vec}(\Theta_2) = \mathrm{vec}(\Theta_2)$.
- *1-gram joint probability tensor* $\theta_1 = (\theta_0, \theta_1)^T \in \mathbb{R}^{k^{n-2}}$ is stationary vector of $\overline{Q}_2 \coloneqq \Theta_2 \mathrm{diag}(\Theta_2^T 1_k)^{-1}$: $\overline{Q}_2 \theta_1 = \theta_1$.

Time-shift invariance of stationary vectors

- Def: n-gram is *time-shift invariance* if $P(X_t, X_{t-1}, \dots, X_{t-n+1}) = P(X_{t-1}, X_{t-2}, \dots, X_{t-n})$ for any $(X_t, \dots, X_{t-n}) \in \{0, \dots, k-1\}^{n+1}$.
- Def: n-gram vector $\theta \in \mathbb{S}_{k^n}$ is *(n, k)-stationary*, if there is a transition matrix $Q \in \mathbb{S}_{k^n-1}^{k^n}$ s.t. $Q\theta = \theta$.

- Lemma 2: Under ergodic condition, (n, k)-stationary n-gram prob. tensor implies $(n-1, k)$-stationary (n-1)-gram prob tensor, which is t.s. invariant.

Hierarchy of stationary vectors:
m-shiftgrams may give a minimal representation.



3-gram probabilities over k words to m-shiftgrams

- m-shiftgram probability matrix for n-gram conditional tensor $Q \in \mathbb{R}^{k^n}$:
  - $S_m \coloneqq M_{k,n}^T \overline{Q}_n^m \mathrm{diag}(\mathrm{vec}\Theta_{n-1}) M_{k,n}$.
    - Marginalization matrix: $M_{k,n} \coloneqq 1_{k^{n-2}} \otimes I_k$.
    - $S_0 = \mathrm{diag}(\mathrm{vec}\theta_1), S_1 = \Theta_2 = \begin{pmatrix} \theta_{00} & \theta_{01} \\ \theta_{10} & \theta_{11} \end{pmatrix}, S_2 = \begin{pmatrix} \theta_{0\blacksquare 0} & \theta_{0\blacksquare 1} \\ \theta_{1\blacksquare 0} & \theta_{1\blacksquare 1} \end{pmatrix}, \in \mathbb{R}^{k\times k}$.

3-gram tensor
$Q_3 := (q_{00}\ q_{10} | q_{01}\ q_{11})$

3-gram transition matrix
$\overline{Q}_3$ $(Q_3 = M_{k,n}^T \overline{Q}_3)$

→

2-gram transition matrix
$\mathrm{vec}\Theta_2 = \overline{Q}_3 \mathrm{vec}\Theta_2$
$\overline{Q}_2 = \Theta_2 \mathrm{diag}(\Theta_2^T 1_k)^{-1}$
$S_1 = \Theta_2$ (1-shiftgram)

→

1-gram transition matrix
$\theta_1 = \overline{Q}_2 \theta_1$
$S_1 = \mathrm{diag}(\theta_1)$ (0-shiftgram)

↘

m-shiftgram for $m \geq 2$:
$S_m \coloneqq M_{k,n}^T \overline{Q}_n^m \mathrm{diag}(\mathrm{vec}\Theta_{n-1}) M_{k,n}$

Properties of m-shiftgrams 1: 1-gram sharing

- Any m-shiftgram S_m for a given Q_n and any $m \geq 0$ holds:
  - $\overline{S}_m \theta_1 = \theta_1$, where $\overline{S}_m \coloneqq S_m \mathrm{diag}(S_m^T 1_k)^{-1}$.
  - It implies any m-shiftgram for the same n-gram tensor Q_n share the same unigram vector as its stationary.

- Corollary 1: any m-shiftgrams has 1-gram as its component.
  - $S_m(B) = \theta_1 \theta_1^T + \mathrm{diag}(\theta_1) A B A^T \mathrm{diag}(\theta_1)$,
    where $A_k = \left(a_{k,1}, a_{k,2}, \dots, a_{k,k-1}\right) \in \{-1,0,1\}^{k \times (k-1)}$ denotes *anti-one matrix*,
    $a_i = e_{k,i-1} - e_{k,i} \in \{-1,0,1\}^k$ denotes *anti-one vectors*
    with some matrix $B \in \mathbb{R}^{(k-1)\times(k-1)}$.

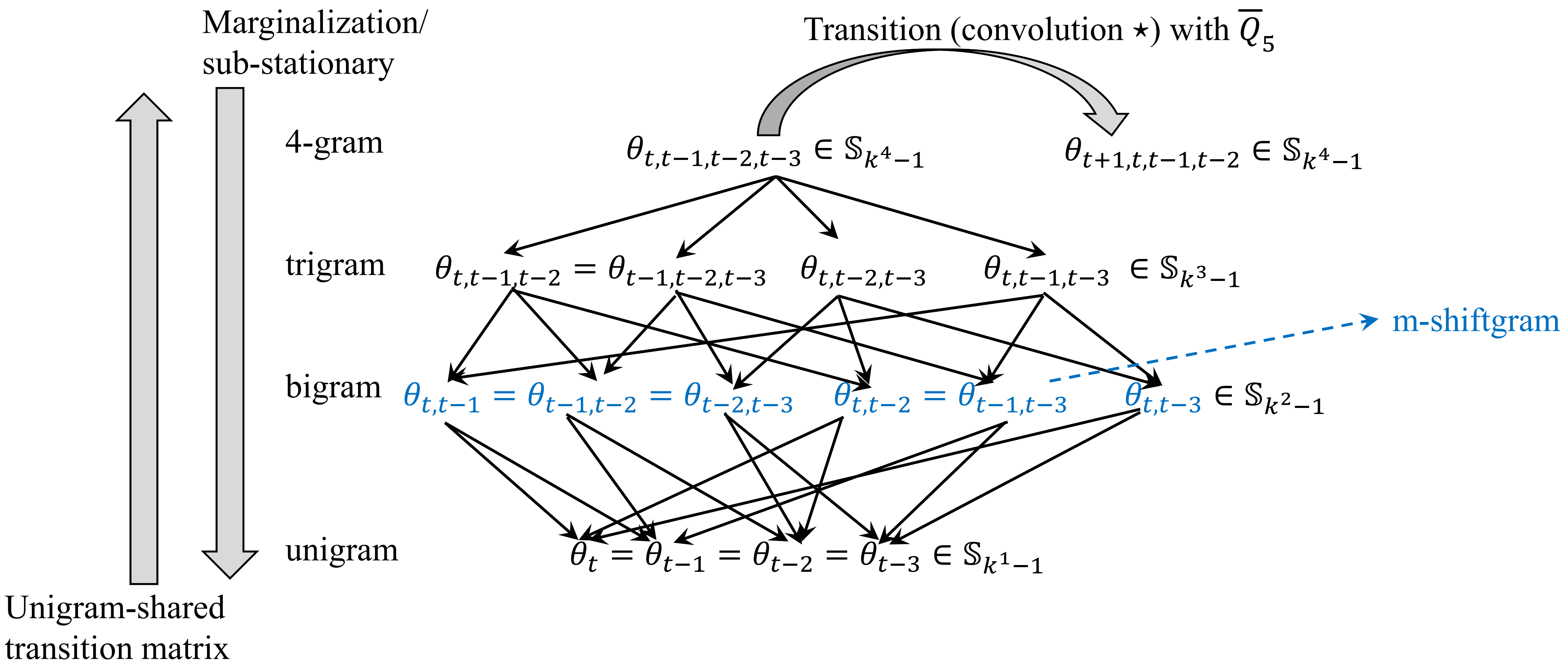
- It explains why Pointwise Mutual Information (PMI; Church & Hanks, 1990) is useful. PMI$=S_m - \theta_1 \theta_1^T$
  - Many studies in NLP implicitly employ (log-)PMI as its preprocess of bigrams.

Properties of m-shiftgrams 2: time-shift invariance $q_{t,t-m} = q_{t+s,t+s-m}$

- Any m-shiftgram S_m for a given Q_n and any $m \geq 0$ has a set of variations:
  - $S_{m,d,i} \coloneqq M_{k,n,i}^T \overline{Q}_n^d \mathrm{diag}(\mathrm{vec}\Theta_{n-1}) M_{k,n,i+d-m}$.
  - $M_{k,n,i} \coloneqq 1_{k^i} \otimes I_k \otimes 1_{k^{n-2-i}}$.
  - $S_{m,d,i} = S_m$ for any $i, m, d \geq 0$ such that $\max(0, m-d) \leq i \leq \min(n-2, n-2-d+m)$.

  - This properties follow the nature of the stationary vector.

Hierarchy of stationary vectors



A set of equations of m-shiftgrams 1: linear equations

- Linear equations up to $m \leq 2$ for $k = 2, n = 3$
  - $$\begin{pmatrix} S_1 \\ S_2 \\ S_1 \end{pmatrix} = \begin{pmatrix} 1_k^T \otimes I_k \otimes I_k \\ I_k \otimes 1_k^T \otimes I_k \\ I_k \otimes I_k \otimes 1_k^T \end{pmatrix} \mathrm{vec}(Q_3).$$

marginalization $\theta_{t,t-1,t-2} \in \mathbb{S}_{k^3-1}$

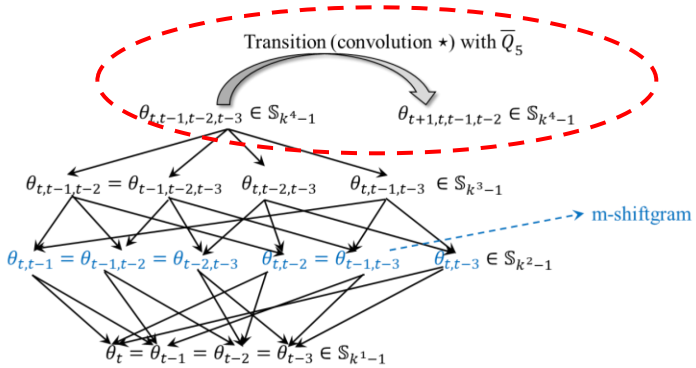
$\boxed{1_k^T \otimes I_k \otimes I_k}$ $\boxed{I_k \otimes I_k \otimes 1_k^T}$ $\boxed{I_k \otimes 1_k^T \otimes I_k}$

$\theta_{t,t-1} = \theta_{t-1,t-2}$, $\theta_{t,t-1} \in \mathbb{S}_{k^3-1}$

S_1 S_1 S_2

  - The root of this equation is with any arbitrary $X \in \mathbb{R}^{(k-1)^n}$
    $\mathrm{vec}(Q_3) = (A_k \otimes A_k \otimes A_k)X + Y$, with *anti-one matrix* $A_k \in \{-1,0,1\}^{k \times (k-1)}$,
  - $Y = \left(\frac{1_k}{k} + \hat{A}_k\theta_1\right) \otimes \left(\frac{1_k}{k} + \hat{A}_k\theta_1\right) \otimes \left(\frac{1_k}{k} + \hat{A}_k\theta_1\right) - \hat{A}_k\theta_1 \otimes \hat{A}_k\theta_1 \otimes \hat{A}_k\theta_1$
    $+ \frac{1}{k}\left((1_k \otimes A_k \otimes A_k + A_k \otimes A_k \otimes 1_k)b_1 + (A_k \otimes 1_k \otimes A)b_2\right)$,
    with $\hat{A}_k := A_k\left(A_k^T A_k\right)^{-1} A_k^T$, $\mathrm{vec}(S_i) = \theta_1 \otimes \theta_1 + (A_k \otimes A_k)b_i$ for $i = 1, 2$
    (Cor. 1).

For $m > 2$: the vector space spanned by convolution



- For $Q_n \in \mathbb{S}_{k-1}^{k^{n-1}}$, $\theta_n := \mathrm{vec}(Q_n \Theta_{n-1})$ with $\Theta_{n-1} = \mathrm{diag}(\theta_{n-1})$, $\theta_{n-1} = \overline{Q}_n \theta_{n-1}$, m-shiftgram is

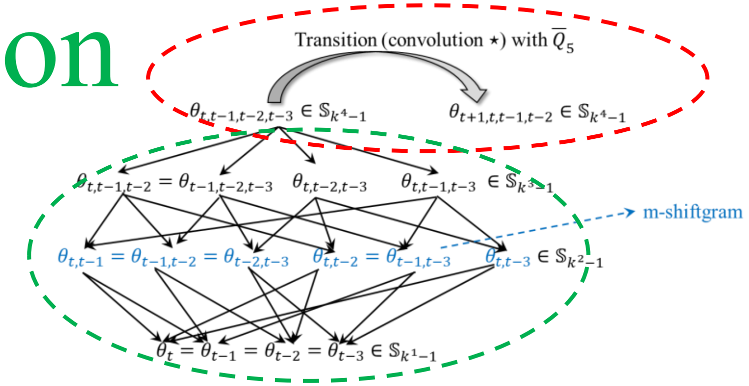
$$S_{n+m-2} = \boxed{\left(I_k \otimes 1_{k^{n-2}}^T\right)} \otimes I_k\Big)\Big(\boxed{\left(\Theta_{n-1}^{-1} Q_n^{m-1} \Theta_{n-1}\right)^T} \otimes I_k\Big)\boxed{\theta_n.}$$

- It naturally induces a tensor convolution from right hand side

$$\left(I_k \otimes 1_{k^{n-2}}^T\right) X_n^m = \Big(\big((Q_n \star Q_n) \star Q_n\big) \star \cdots\Big) \star Q_n$$
$$X_n = \left(\Theta_{n-1}^{-1} Q_n^{m-1} \Theta_{n-1}\right)^T$$

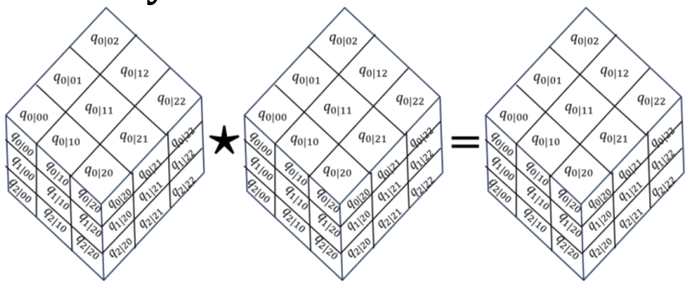
Tensor convolution $\star: \mathbb{R}^{k^n} \times \mathbb{R}^{k^n} \to \mathbb{R}^{k^n}$ is defined for $Q = (q_{i_1 i_2 \dots i_{n-1}}), R = (r_{i_1 i_2 \dots i_{n-1}})$ by $(Q \star R)_{i_1 i_2 \dots i_{n-1}} = (q_{1,i_2,\dots,i_{n-1}}, \dots, q_{k,i_2,\dots,i_{n-1}}) r_{i_1 i_2 \dots i_{n-1}}$.

Shiftgrams = convolution⋆ + reduction



For any two 3rd order tensors, $P = (P_{ijk}), Q = (Q_{ijk})$,
The star product (convolution) is defined by

$$(P \star Q)_{ijk} \coloneqq \sum_l P_{ilj} Q_{ljk} .$$



If $P = Q = (q_{i|jk})$, the star product updates the transition probability

$$(q_{x_{t+1}|x_t,x_{t-1}}, q_{x_t|x_{t-1},x_{t-2}}) \mapsto q_{x_{t+1}|x_{t-1},x_{t-2}}$$

$n = 3$ $\quad P(X_{t-1}, X_{t-2})$ $\quad P(X_t, X_{t-1}, X_{t-2})$ $\quad P(X_{t+1}, X_{t-1}, X_{t-2})$ $\quad P(X_{t+2}, X_{t-1}, X_{t-2})$ …

$$\bar{I}_k \xrightarrow{\text{convolution}} Q \xrightarrow{\text{convolution}} Q \star Q \xrightarrow{\text{convolution}} (Q \star Q) \star Q \ \ldots$$

reduction Σ_1 Σ_2 Σ_1 Σ_2 Σ_1 Σ_2 Σ_1

m-shiftgrams $\quad S_0 \quad S_1 \quad S_2 \quad S_3 \ \ldots$

$P(X_t)$ $\quad P(X_t|X_{t-1})$ $\quad P(X_t|X_{t-2})$ $\quad P(X_t|X_{t-3})$ …

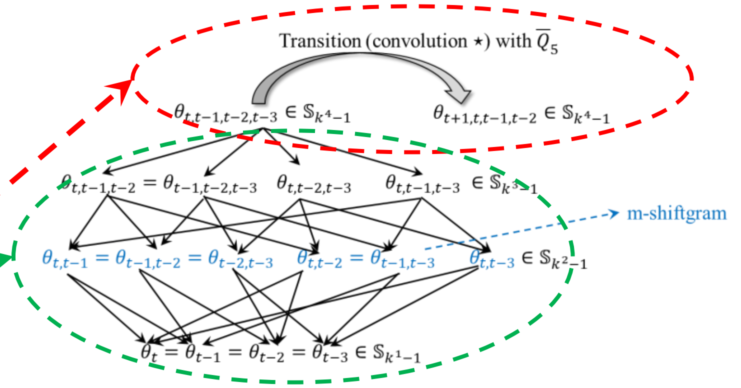
Shiftgram equations ($n = 3$)

$\frac{n(n-1)}{2}$

Linear equations

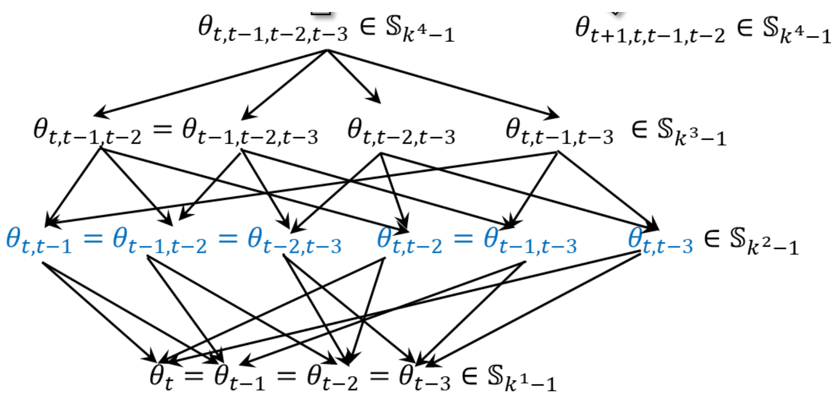
$$
\begin{pmatrix} S_1 \\ S_2 \\ S_1 \\ S_3 \\ S_4 \\ \vdots \end{pmatrix} = \begin{pmatrix} 1_k^T \otimes I_k \otimes I_k \\ I_k \otimes 1_k^T \otimes I_k \\ I_k \otimes I_k \otimes 1_k^T \\ (I_k \otimes 1_{k^{n-2}}^T) X_n^1 \otimes I_k \\ (I_k \otimes 1_{k^{n-2}}^T) X_n^2 \otimes I_k \\ \vdots \end{pmatrix} \mathrm{vec}(Q_n),
$$

$$
X_n := \left(\Theta_{n-1}^{-1} Q_n \, \Theta_{n-1}\right)^T
$$



- The next question is whether the tensor convolution X_n^m kept producing a "new" component constraining Q_n.
- In other words, what is the (right) *ideals* of this convolution?
  - Right ideal I of a ring R is a sub ring closed under addition, and satisfies $I \star r = I$ for any $r \in R$.

n-gram tensor as a non-associative algebra

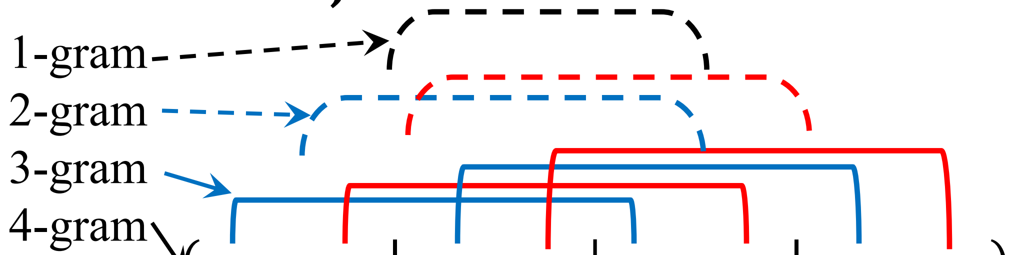


- $\mathbb{N}_{n,k} := \left(\mathbb{S}_{k-1}^{k^{n-1}}, +, \star\right)$ is an algebra without associativity of its product.
  - $\mathbb{N}_{n,k}$ is a commutative group under $+$.
  - $\mathbb{N}_{n,k}$ is a monoid-like without associativity under $\star$.
    - But only convolution we consider is right-most product $\left((Q \star Q) \star Q\right) \star Q$.
  - $\mathbb{N}_{n,k}$ is distributive: $(Q_0 + Q_1) \star R = Q_0 \star R + Q_1 \star R$.
  - $\mathbb{N}_{n,k}$ has a scalar (real value) product.

- Theorem 2 (conjecture): isomorphism theorem of n-gram algebra
  - Any ideal of n-gram ring $\mathbb{N}_{n,k}$ is isomorphic to some $\mathbb{N}_{m,k}$ for $m \leq n$.

- This theorem implies “irreduceable” n-gram algebra, which is not isomorphic to any $\mathbb{N}_{m,k}$ for $m < n$, is learnable through a set of m-shiftgrams.

Subalgebras of n-gram algebra $\mathbb{N}_{n,k}$.



$$\left(q_{000}\ q_{100} \middle| q_{010}\ q_{110} \middle| q_{001}\ q_{101} \middle| q_{011}\ q_{111}\right)$$

- $Q_4 :\in \mathbb{N}_{4,2}$

$$= \left(q_{000}\ q_{100} \middle| q_{010}\ q_{110} \middle| q_{001}\ q_{101} \middle| q_{011}\ q_{111}\right)$$

$$= \left(\begin{array}{cc|cc|cc|cc} q_{0|000} & q_{0|100} & q_{0|010} & q_{0|110} & q_{0|001} & q_{0|101} & q_{0|011} & q_{0|111} \\ q_{1|000} & q_{1|100} & q_{1|010} & q_{1|110} & q_{1|001} & q_{1|101} & q_{1|011} & q_{1|111} \end{array}\right)$$

$$\left(Q_{00\_} \middle| Q_{10\_} \middle| Q_{01\_} \middle| Q_{11\_}\right) = \left(q_{000}\ q_{001} \middle| q_{100}\ q_{101} \middle| q_{010}\ q_{011} \middle| q_{110}\ q_{111}\right)$$

$$\left(R_{\_00} \middle| R_{\_10} \middle| R_{\_01} \middle| R_{\_11}\right) = \left(r_{000}\ r_{100} \middle| r_{010}\ r_{110} \middle| r_{001}\ r_{101} \middle| r_{011}\ r_{111}\right)$$

- $Q_4 \star R_4 = \left(Q_{00\_} r_{000}\ Q_{00\_} r_{100} \middle| Q_{10} r_{010}\ Q_{10} r_{110} \middle| Q_{01} r_{001}\ Q_{01} r_{101} \middle| Q_{11} r_{011}\ Q_{11} r_{111}\right)$

$$= \left(Q_{00\_} R_{\_00} \middle| Q_{10\_} R_{\_10} \middle| Q_{01\_} R_{\_01} \middle| Q_{11\_} R_{\_11}\right)$$

- The “copy” subset $S \coloneqq \{Q_4 \mid q_{000} = q_{001}\}$ is subalgebra:
  - If $S_4, R_4 \in S \coloneqq \{Q_4 \mid q_{000} = q_{001}\}$, $S_4 + R_4 \in S$ and $S_4 \star R_4 \in S$.
  - $S \coloneqq \left\{Q_4 \mid q_{ij0} = q_{ij1}\right\}$ for $i, j \in \{0,1\}$ is a subalgebra isomorphic to $\mathbb{N}_{3,2}$. (3-gram)
  - $S \coloneqq \left\{Q_4 \mid q_{ij0} = q_{ij1}\ \forall i, j \in \{0,1\}, q_{000} = q_{010}\right\}$ is a subalgebra isomorphic to $\mathbb{N}_{2,2}$. (2-gram)
  - $S \coloneqq \left\{Q_4 \mid q_{ij0} = q_{ij1}, q_{olo} = q_{olo} \forall i, j, l \in \{0,1\}, q_{000} = q_{100}\right\}$ is a subalgebra isomorphic to $\mathbb{N}_{1,2}$. (1-gram)

Simultaneous identification of subalgebras and probability parameters.

- Unknown probability tensor Q_n can be identified (up to finite) by the following equation, but the rank of W_m (X_n^m) depends also on Q_n.
- $$\begin{pmatrix} S_1 \\ S_2 \\ S_1 \\ S_3 \\ S_4 \\ \vdots \\ S_m \end{pmatrix} = W_m \operatorname{vec}(Q_n), W_m = \begin{pmatrix} 1_k^T \otimes I_k \otimes I_k \\ I_k \otimes 1_k^T \otimes I_k \\ I_k \otimes I_k \otimes 1_k^T \\ \left(I_k \otimes 1_{k^{n-2}}^T\right) X_n^1 \otimes I_k \\ \left(I_k \otimes 1_{k^{n-2}}^T\right) X_n^2 \otimes I_k \\ \vdots \end{pmatrix}$$

Emergence of Clifford algebra:

every equation is written with exterior $\wedge$ and inner product $\cdot$,
which roughly means the the root of the equation is represented by geometric objects (points, lines, ..) or simplicial complexes.

Lemma 8. *There is some matrix* $A = (a_0, a_1, a_2) \in \mathbb{R}^{3\times 3}$ *such that*

$$\begin{cases} (A - (\theta_{00}, \theta_{01}, \theta_{02}))\,(\theta_{00}, \theta_{10}, \theta_{20}) = \mathbf{0}_{3,3} \\ (A - (\theta_{10}, \theta_{11}, \theta_{12}))\,(\theta_{01}, \theta_{11}, \theta_{21}) = \mathbf{0}_{3,3} \\ (A - (\theta_{20}, \theta_{21}, \theta_{22}))\,(\theta_{02}, \theta_{12}, \theta_{22}) = \mathbf{0}_{3,3} \end{cases}, \quad (93)$$

where $\mathbf{1}_k^\top \theta_{ij} = 1$, *if there are two triplets of the set of vectors* $(\{\alpha_{0,0}, \alpha_{0,1}\}, \{\alpha_{1,0}, \alpha_{1,1}\}, \{\alpha_{2,0}, \alpha_{2,1}\}), (\{d_{0,0}, d_{0,1}\}, \{d_{1,0}, d_{1,1}\}, \{d_{2,0}, d_{2,1}\}) \in \mathbb{R}^{3\times 3}$ *such that for* $i = 0, 1, 2$, *each of the set* $\{\alpha_{i,0}, \alpha_{i,1}\}$ *and* $\{d_{i,0}, d_{i,1}\}$ *is linearly independent, and*

$$A - (\theta_{i0}, \theta_{i1}, \theta_{i2}) = \sum_{j=0,1} \alpha_{i,j} d_{i,j}^\top. \quad (94)$$

A part of proof

361 There are two cases: In Case 0, the set of two vectors $d_{i,0}^\top(\theta_{0i}, \theta_{1i}, \theta_{2i}), d_{i,1}^\top(\theta_{0i}, \theta_{1i}, \theta_{2i})$ is linear
362 dependent, and in Case1, it is linearly independent. By Lemma 9, Case 0 implies

$$\boxed{(d_{i,0} \wedge d_{i,1}) \cdot (\theta_{0i} \wedge \theta_{1i}) = (d_{i,0} \wedge d_{i,1}) \cdot (\theta_{1i} \wedge \theta_{2i}) = (d_{i,0} \wedge d_{i,1}) \cdot (\theta_{0i} \wedge \theta_{2i}) = 0,} \quad (99)$$

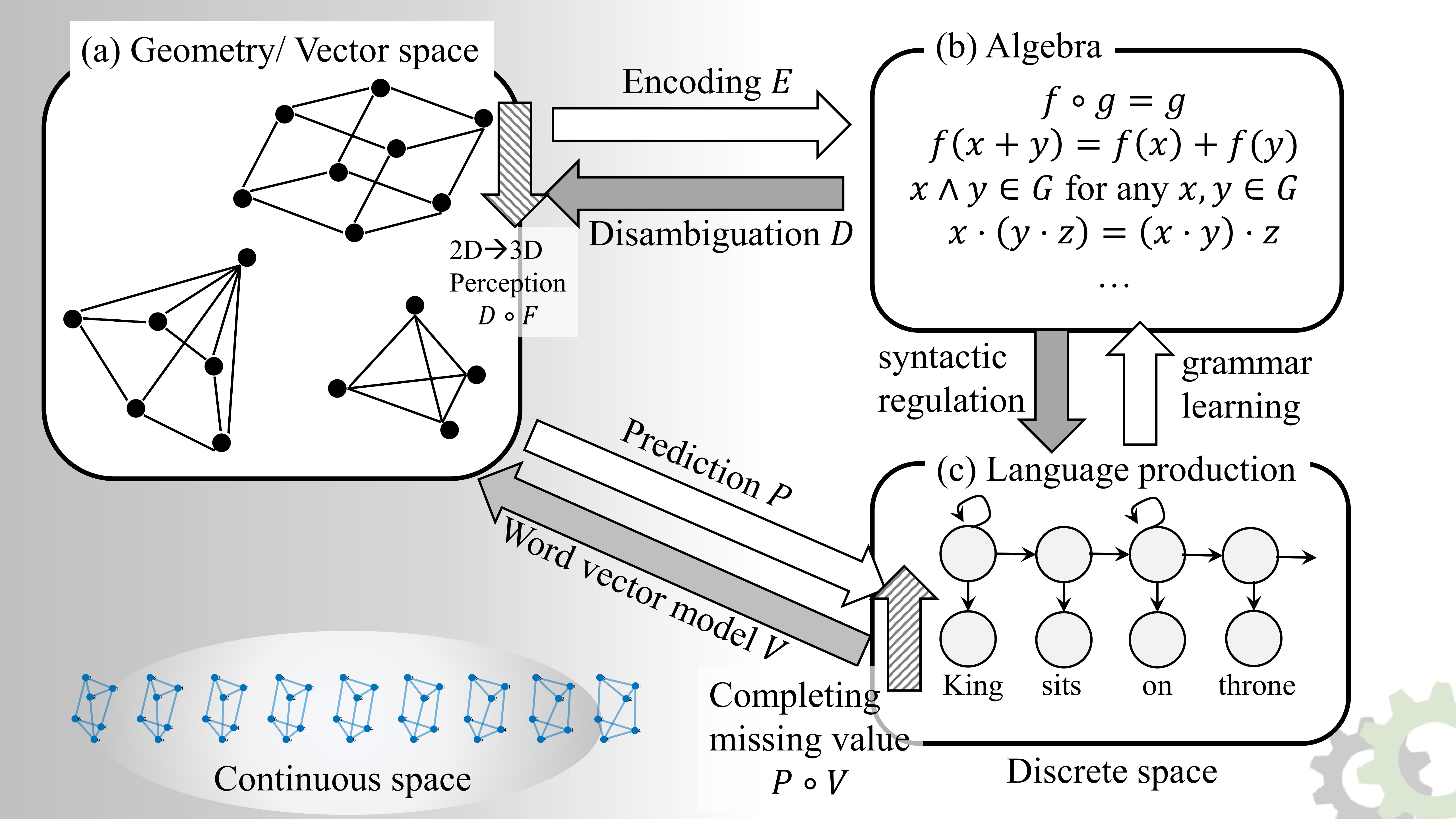
363 where $\wedge$ is the exterior product.

$$\theta_{0i} \wedge \theta_{1i} = \left(a_i - \alpha_{0,j} \mathrm{e}_{k,i}^\top d_{0,j}\right) \wedge \left(a_i - \alpha_{1,j} \mathrm{e}_{k,i}^\top d_{1,j}\right) = \left(\alpha_{0,j} \mathrm{e}_{k,i}^\top d_{0,j} - \alpha_{1,j} \mathrm{e}_{k,i}^\top d_{1,j}\right) \wedge a_i + \alpha_{0,j} \mathrm{e}_{k,i}^\top d_{0,j} \wedge \alpha_{1,j} \mathrm{e}_{k,i}^\top d_{1,j}. \quad (100)$$

364 Thus, eliminating a_i, we have

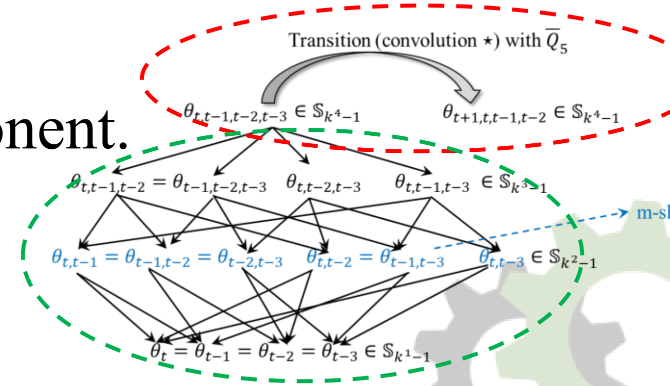
$$\boxed{(d_{i,0} \wedge d_{i,1}) \cdot \left(\alpha_{0,j} \mathrm{e}_{k,i}^\top d_{0,j} \wedge \alpha_{1,j} \mathrm{e}_{k,i}^\top d_{1,j} - \alpha_{0,j} \mathrm{e}_{k,i}^\top d_{0,j} \wedge \alpha_{2,j} \mathrm{e}_{k,i}^\top d_{1,j} + \alpha_{1,j} \mathrm{e}_{k,i}^\top d_{0,j} \wedge \alpha_{2,j} \mathrm{e}_{k,i}^\top d_{1,j}\right) = 0.} \quad (101)$$

This quadratic form indicates α_i is a "sphere" degerated into a line (co-linear).


(a) Geometry/ Vector space
Encoding E
(b) Algebra
$f \circ g = g$
$f(x+y) = f(x) + f(y)$
$x \wedge y \in G$ for any $x, y \in G$
$x \cdot (y \cdot z) = (x \cdot y) \cdot z$
$\ldots$
Disambiguation D
2D→3D
Perception
$D \circ F$
syntactic
regulation
grammar
learning
Prediction P
(c) Language production
Word vector model V
King
sits
on
throne
Completing
missing value
$P \circ V$
Continuous space
Discrete space

N-gram to geometric algebra, WHY?
Geometric algebra ~ multilinear algebra

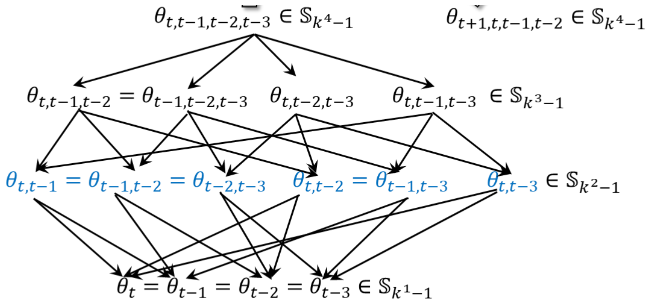
- Schur-Weyl duality
  - Space of multilinear maps $\cong$ Tensor space $\cong_{adjoint}$ Symmetry group
- Roughly speaking, …
  - Any multilinear map is representable by some linear map from tensor space.
  - Any tensor space can be well decomposed by "symmetric" and "anti-symmetric (alternating)" tensors, and such decomposition is closely related to conjugacy classes of some symmetry group.
- More specifically, …
  - Marginalization (sum/ averaging) ~ symmetric tensor component.
  - Anti-one ~ anti-symmetric (alternating) tensor component.

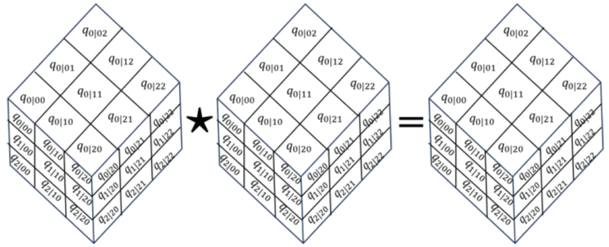


Summary & future work

- Is any n-gram finitely identifiable with m-shiftgrams?
  Yes, up to isomorphism (to be proven for sure).
- Conjecture
  - For any k-word & n-gram language, a set of 0- to $(n+k-2)$-shiftgrams is sufficient (and necessary in general) to reconstruct n-gram probabilities. (under category equivalence)
    - A new type of complexity integrating the spatial (k) and temporal (n) complexity.
- Future work
  - Complete the proofs.
  - Perhaps developing this math leads "n-gram algebra", which characterize basic natural linguistic operations.

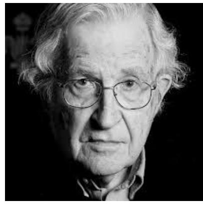
Findings and speculations





- "Pointwise Mutual Information (PMI)" is explained.
  - $S_m(B) = \theta_1\theta_1^T + \mathrm{diag}(\theta_1)ABA^T\mathrm{diag}(\theta_1)$ (Corollary 1).
- A non-associative algebra for general hierarchy of n-grams.
  - It is like a "onion skin pealing" to decompose n-grams into lower m-grams.
  - This requires "almost-linear but a bit nonlinear computation" (like order reasoning as Boolean module).

- An *integration* of "distributional hypothesis" and "universal grammar"



  - *n*-gram languages are representable by a "distributed basis" spanning its probabilistic tensor space.
    - Encoding: generating sentences by a n-gram.
    - Decoding: identifying the (minimal) basis of tensors by a given sample of sentences.
    - Algebra (coding principle): a "grammar" of the language.

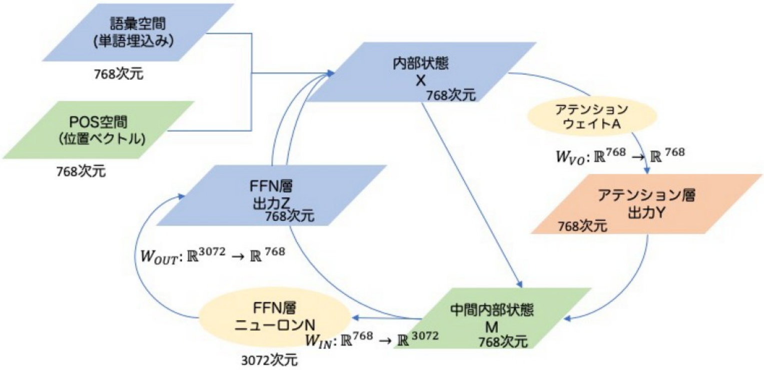
Putative computational implementation

- Subspace (attention)
- Evaluation (decode)
  - Lossy code recovery/ Correction (contextualization)
  - Eigenspace identification
- Generation (encode)
- Completion (prediction)

Do LLMs decode n-gram algebra?

Mr. A. Maeda (Ph.D. Student in JAIST)

- Solving a general set of m-shiftgrams for n-gram probs over k words needs to solve a series of "nonlinear eigenvalue problems."
  - Power method to numerically compute eigenvectors of a matrix $A =: A_1$
    - $x_{i,t+1} \coloneqq \frac{A_i\, x_{i,t}}{\alpha_{i,t}}, \alpha_{i,t} = \|A_i x_{i,t}\|$ and $y_i = \lim_{t\to\infty} x_{i,t}\,, \lambda_i = \lim_{t\to\infty} \alpha_{i,t}\,, A_{i+1} \coloneqq A_i - \alpha_i y_i y_i^T$.
- "nonlinear power method" is like Transformer (Vaswani+, 2017).
  - In each layer of Transformer, word vectors in FFN layers may decode the difference between n- and (n-1)-grams.



(前田, 鳥居, 日高, 井之上, 大関, 優秀賞@NLP2025)

Thank you for your attention!

- Acknowledgements
  - This work is supported by JSPS KAKENHI JP23H0369, JP22K19790, JST PRESTO JPMJPR20C9.
- I appreciate my colleagues' encouraging discussion with me.

Prof .E. Pothos
(City, University of London)

Prof. T. Torii
(Tokyo Denki U.)

Mr. A. Maeda
(Ph.D. student in JAIST)