Learning the symmetric group

Jan de Gier (University of Melbourne)

Guttmann-fest, 30 June 2025, Melbourne

Alexandr Garbali
Max Petschack

MATRIX, a word of thanks

- 10th Anniversary of MATRIX
- Key role in the early stages of MATRIX
- Chair of the MATRIX Advisory Board for 9 years
- Listening ear and sage advice
- Advocacy for Australian mathematical sciences research institute for more than 30 years
- Knowledge, expertise, enthusiasm and community spirit in advancing MATRIX

MATRIX, a word of thanks

https://www.youtube.com/watch?v=PDWcANR-nLQ

Personal memories

Half around the bay.

Wishing you many healthy years to come!

Knot = Unknot?

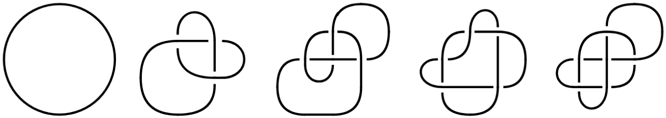
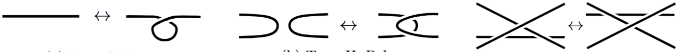


Figure 1: Examples of knots. From left to right: unknot (0_1), trefoil (3_1), figure-eight (4_1), 5_1, and 5_2.



(a) Type I: Twist (b) Type II: Poke (c) Type III: Slide

Figure 2: Reidemeister moves.

How to determine whether a knot is equivalent to the unknot?

Learning to Unknot, Gukov et al., arXiv:2010.16263

Braids

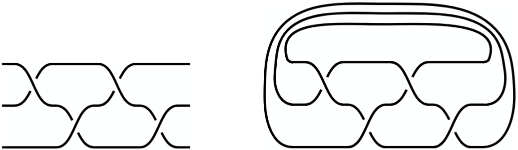
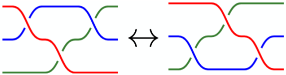
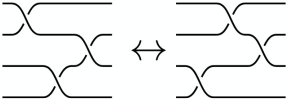


Figure 4: A braid $\sigma_1\sigma_2^{-1}\sigma_1\sigma_2^{-1}$ (left) and its closure (right).



(a) Relation 1: $\sigma_1\sigma_2\sigma_1 = \sigma_2\sigma_1\sigma_2$



(b) Relation 2: $\sigma_1\sigma_3\sigma_2 = \sigma_3\sigma_1\sigma_2$

Figure 5: Braid relations

A knot is the closure of a braid

A braid is a permutation if we forget about under/over crossings

Word problem

Let $p \in S_n$ be a permutation and σ_i an elementary transposition

$$(p_1, p_2, \ldots, p_i, p_{i+1}, \ldots, p_n) \cdot \sigma_i = (p_1, p_2, \ldots, p_{i+1}, p_i, \ldots, p_n)$$

What is the the permutation corresponding to

$$\sigma_5\sigma_8\sigma_7\sigma_4\sigma_8\sigma_3\sigma_2\sigma_7\sigma_3\sigma_9\sigma_5\sigma_7\sigma_3\sigma_4\sigma_5\sigma_8\sigma_1\sigma_9\sigma_3\sigma_5?$$

Word problem

Let $p \in S_n$ be a permutation and σ_i an elementary transposition

$$(p_1, p_2, \dots, p_i, p_{i+1}, \dots, p_n) \cdot \sigma_i = (p_1, p_2, \dots, p_{i+1}, p_i, \dots, p_n)$$

What is the the permutation corresponding to

$$\sigma_5\sigma_8\sigma_7\sigma_4\sigma_8\sigma_3\sigma_2\sigma_7\sigma_3\sigma_9\sigma_5\sigma_7\sigma_3\sigma_4\sigma_5\sigma_8\sigma_1\sigma_9\sigma_3\sigma_5?$$

Answer:

$$(6, 1, 5, 2, 3, 4, 9, 10, 7, 8)$$

Word problem

Let $p \in S_n$ be a permutation and σ_i an elementary transposition

$$(p_1, p_2, \ldots, p_i, p_{i+1}, \ldots, p_n) \cdot \sigma_i = (p_1, p_2, \ldots, p_{i+1}, p_i, \ldots, p_n)$$

What is the the permutation corresponding to

$$\sigma_5\sigma_8\sigma_7\sigma_4\sigma_8\sigma_3\sigma_2\sigma_7\sigma_3\sigma_9\sigma_5\sigma_7\sigma_3\sigma_4\sigma_5\sigma_8\sigma_1\sigma_9\sigma_3\sigma_5?$$

Answer:

$$(6, 1, 5, 2, 3, 4, 9, 10, 7, 8)$$

Another example

$$\sigma_4\sigma_2\sigma_6\sigma_7\sigma_7\sigma_9\sigma_6\sigma_2\sigma_7\sigma_9\sigma_4\sigma_3\sigma_4\sigma_5\sigma_3\sigma_5\sigma_4\sigma_7\sigma_3\sigma_4$$

Word problem

Let $p \in S_n$ be a permutation and σ_i an elementary transposition

$$(p_1, p_2, \ldots, p_i, p_{i+1}, \ldots, p_n) \cdot \sigma_i = (p_1, p_2, \ldots, p_{i+1}, p_i, \ldots, p_n)$$

What is the the permutation corresponding to

$$\sigma_5\sigma_8\sigma_7\sigma_4\sigma_8\sigma_3\sigma_2\sigma_7\sigma_3\sigma_9\sigma_5\sigma_7\sigma_3\sigma_4\sigma_5\sigma_8\sigma_1\sigma_9\sigma_3\sigma_5?$$

Answer:

$$(6, 1, 5, 2, 3, 4, 9, 10, 7, 8)$$

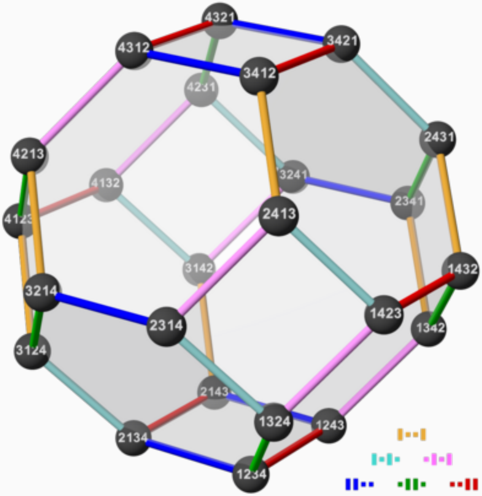
Another example

$$\sigma_4\sigma_2\sigma_6\sigma_7\sigma_7\sigma_9\sigma_6\sigma_2\sigma_7\sigma_9\sigma_4\sigma_3\sigma_4\sigma_5\sigma_3\sigma_5\sigma_4\sigma_7\sigma_3\sigma_4$$

Answer:

$$(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$$

Combinatorics: walks on Permutohedron



- Can a machine predict permutations from words in the symmetric group through statistical learning?
- How to train a machine when n is large?
- What does a machine learn? Does it memorise or learn mathematics?

Input

Let $\boldsymbol{x} = (x_1, \ldots, x_N) \in \mathcal{X}^N$ with $\mathcal{X} = \{1, \ldots, n(n-1)/2\}$ represent a word in S_n, i.e.

$$w = \sigma_{x_1} \cdots \sigma_{x_N}.$$

Output

Probability distribution on sequences of integers $\in [1, n]$,

$$\Phi : \mathcal{X}^N \to ([0,1]^n)^n, \qquad \Phi : \boldsymbol{x} \to \mathbb{P}(\boldsymbol{p})$$

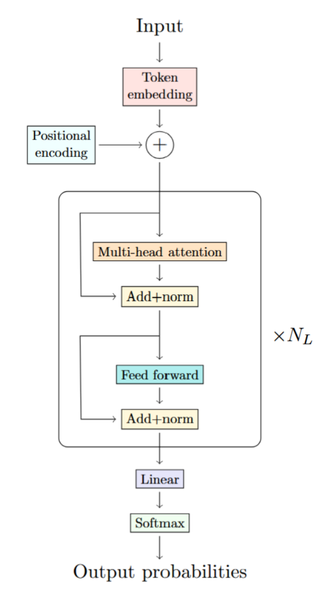
with $\boldsymbol{p} = (p_1, \ldots, p_n)$?

Set up

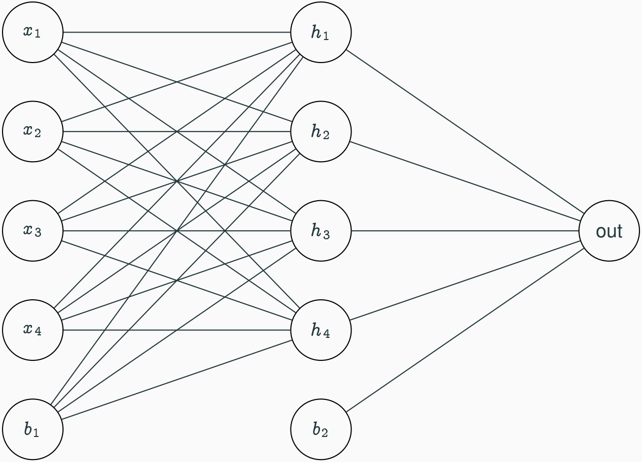
Train a neural network

- Randomly generate a (small) subset of points $\boldsymbol{x}$ and corresponding permutations $\boldsymbol{p}$
- Train a Transformer network architecture to learn $\Phi : \boldsymbol{x} \to \mathbb{P}(\boldsymbol{p})$ on this data
- For large n train only using words that do not permute more than m elements with $m < n$
- In our experiment $n = 16$ ($N = 60$) and $m = 10$: $16^{60} \approx 1.8 * 10^{72}$ posible words, $16! = 2.0 * 10^{13}$ permutations; training data was $1.6 * 10^{7}$
- Test Φ on new words $\boldsymbol{x}$ that were not used in training
- What is the performance on words that permute more than m elements (out of distribution learning)?

Transformer

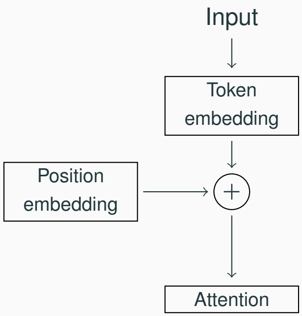


Multi Layer Perceptron (MLP)



$$L_i(\boldsymbol{x}) = \theta_i(W_i\boldsymbol{x} + \boldsymbol{b}_i)$$

Attention



Query and Key matrices,

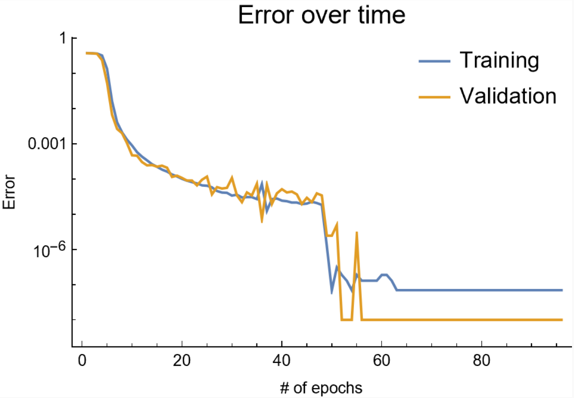
$$Q = XQ_0^T \qquad K = XK_0^T$$

$$A = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{C}}\right) \qquad (\mathrm{softmax} = \mathrm{Boltzmann})$$

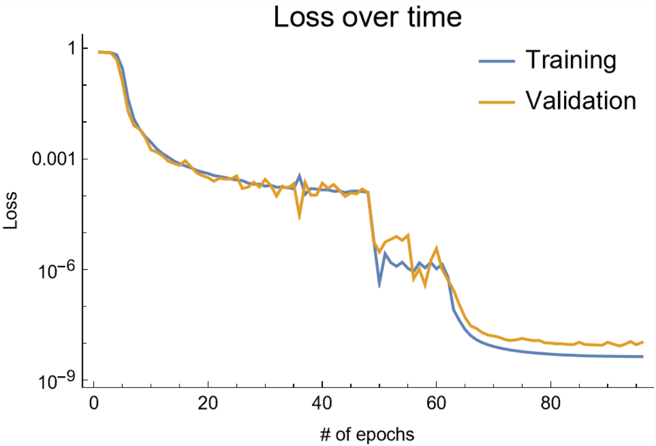
$A_{i,j}$ is how much attention is paid by token i to token j

Error

$$\text{error} = \frac{\#\text{incorrect predictions}}{\#\text{test datapoints}}.$$



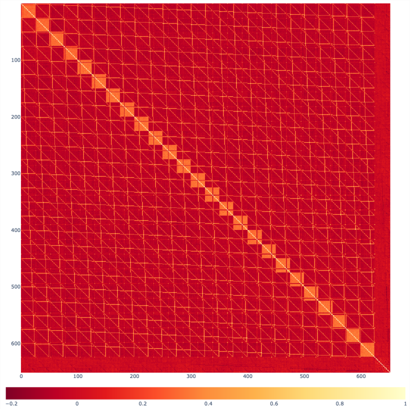
Training (blue) and validation (orange) error (log scale).



Training (blue) and validation (orange) loss (log scale).

Interpretability

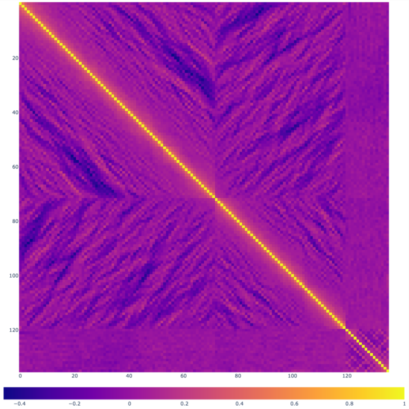
Token embedding



Token embedding for general transpositions

Interpretability

Position embedding



Can embedding structure be interpreted?

Conclusion

- Machine learning of mathematics is fun!
- No noise in data $\Rightarrow$ overfitting can be good (grokking)
- Can learn S_n from training on embeddings of S_m with $m < n$
- Applications to pattern avoidance?
- Extend to braid group and knots

Conclusion

- Machine learning of mathematics is fun!
- No noise in data $\Rightarrow$ overfitting can be good (grokking)
- Can learn S_n from training on embeddings of S_m with $m < n$
- Applications to pattern avoidance?
- Extend to braid group and knots

Happy birthday Tony!