Representational Strengths and Limitations of Transformers

Clayton Sanford

July 18th, 2023

Joint work with Daniel Hsu and Matus Telgarsky

Transformer architecture

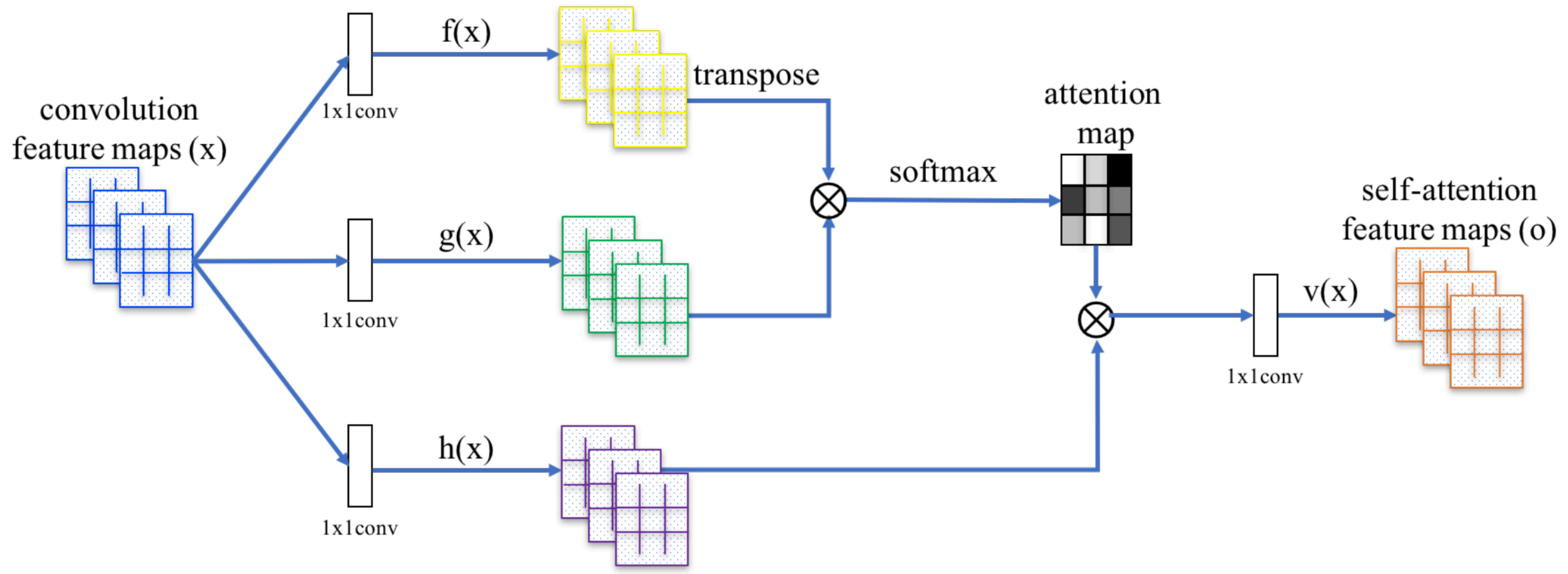
What is it?

- **Self-attention unit:**
  $f(X) = \mathrm{softmax}(XQK^T X^T)XV$ for input $X \in \mathbb{R}^{N \times d}$, model parameters $Q, K, V \in \mathbb{R}^{d \times m}$.

- **Multi-headed attention:**
  $$L(X) = X + \sum_{h=1}^{H} f_h(X)$$

- **Element-wise multi-layer perceptron (MLP):**
  $\phi(X) = (\phi(x_1), \ldots, \phi(x_N))$

- **Full transformer:**
  $T(X) = (\phi_D \circ L_D \circ \ldots \circ L_1 \circ \phi_0)(X)$



Source: https://lilianweng.github.io/posts/2018-06-24-attention/

Transformer architecture

What is it?

- **Self-attention unit:** $f(X) = \text{softmax}(XQK^TX^T)XV$ for input $X \in \mathbb{R}^{N\times d}$, model parameters $Q, K, V \in \mathbb{R}^{d\times m}$.
- **Multi-headed attention:** $$L(X) = X + \sum_{h=1}^{H} f_h(X)$$
- **Element-wise multi-layer perceptron (MLP):** $\phi(X) = (\phi(x_1), \ldots, \phi(x_N))$
- **Full transformer:** $T(X) = (\phi_D \circ L_D \circ \ldots \circ L_1 \circ \phi_0)(X)$

Our questions

Can the strengths and limitations of transformers be understood via function approximation?

1. Power of transformers over fully-connected & recurrent NNs?
2. Representational impact of model parameters m, H, D?
3. Tasks that transformers struggle with?

Transformer architecture

Our questions

Can the strengths and limitations of transformers be understood via function approximation?

1. Power of transformers over fully-connected & recurrent NNs for sequential tasks?
2. Representational impact of model parameters m, H, D?
3. Tasks that transformers struggle with?

Our contributions

Provide two “natural” tasks that exhibit key separations between transformers and other models:

- **Sparse averaging** is efficient for transformers, inefficient for RNNs, FNNs.
- **Pair finding** is easy for transformers, **triple finding** is not.

What is already known theoretically?

- **Universality:** Turing completeness of sufficiently large transformers [PMB19, YBR+20, WCM22]
- **Formal language recognition:**
  - Recognize counter languages [BAG20], bounded-depth Dyck languages [YPPN21], bounded-size automata [LAG+22]
  - **Fixed-size** transformer cannot represent infinite-depth Dyck languages [HAF22]
- **Learnability:** Generalization bounds via covering numbers [EGKZ22, BPKP22]
- **Optimization:** Convergence to OLS in-context learning (linear self-attention) [ZFB23]
- **Graph neural networks:**
  - Message-passing analogue to attention, equivalence to CONGEST distributed communication model [Lou19]
  - Different order GNNs related to graph isomorphism testing [XHLG18, CVCB19, MRF+19]

Transformer architecture

Our questions

Can the strengths and limitations of transformers be understood via function approximation?

1. Power of transformers over fully-connected & recurrent NNs for sequential tasks?
2. Representational impact of model parameters m, H, D?
3. Tasks that transformers struggle with?

Modeling decisions

Model	Context length (N)	#layers (D)	#heads (H)	*#param self-attn* (m)	*#param MLP* (k)
GPT-3	2048	96	96	128	12288
GPT-4	32k	🙃	🙃	🙃	🙃

- Context length $N \gg$ #params in self-attention unit (depth D, heads H, and embedding dim m)

  $\Longrightarrow$ **restricted pairwise computation between elements, model size independent of N**

- #params in MLP $k \gg$ #params in self-attention

  $\Longrightarrow$ **unlimited element-wise computational power**

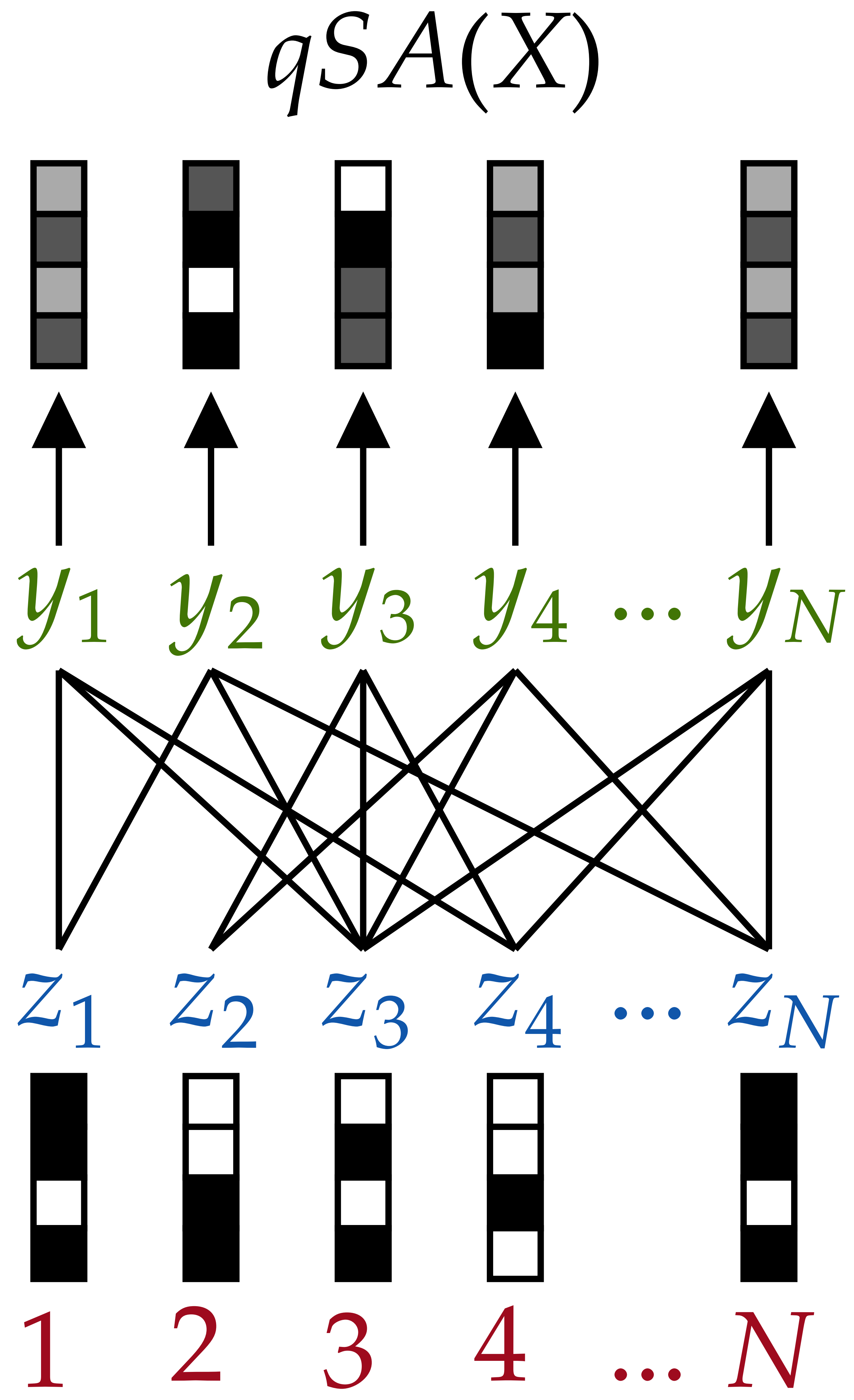
Part 1: Sparse averaging

The task

Input: $X = ((y_1, z_1), \ldots, (y_N, z_N))$

- $y_i \in \binom{[N]}{q}$
- $z_i \in \mathbb{R}^d$

Output: $qSA(X)_i = \frac{1}{q} \sum_{j \in y_i} z_i$



Part 1: Sparse averaging

The task

Input: $X = ((y_1, z_1), \ldots, (y_N, z_N))$

- $y_i \in \binom{[N]]}{q}$
- $z_i \in \mathbb{R}^d$.

Output: $qSA(X)_i = \frac{1}{q} \sum_{j \in y_i} z_i$

Results

1. Inefficient representation with FNNs or RNNs.
   - Any FNN requires width $\Omega(Nd)$.
   - Any RNN requires $\Omega(N)$-bit hidden state.
2. **Exists self-attention unit approximating $qSA(X)$ iff embedding dim $m \gtrsim q$.**

Part 2: Pair and triple finding

The tasks

Input: $X = (x_1, \ldots, x_N) \in [M]^N$.

$\mathrm{Match2}(X)_i = 1\{\exists j : x_i + x_j \equiv_M 0\}$

$\mathrm{Match3}(X)_i = 1\{\exists j_1, j_2 : x_i + x_{j_1} + x_{j_2} \equiv_M 0\}$

Results

1. Efficient representation of $\mathrm{Match2}$ with self-attention unit.
2. No efficient representation of $\mathrm{Match3}$ with multi-headed self-attention.
3. Efficient representation of $\mathrm{Match3}$ under 3-order attention.

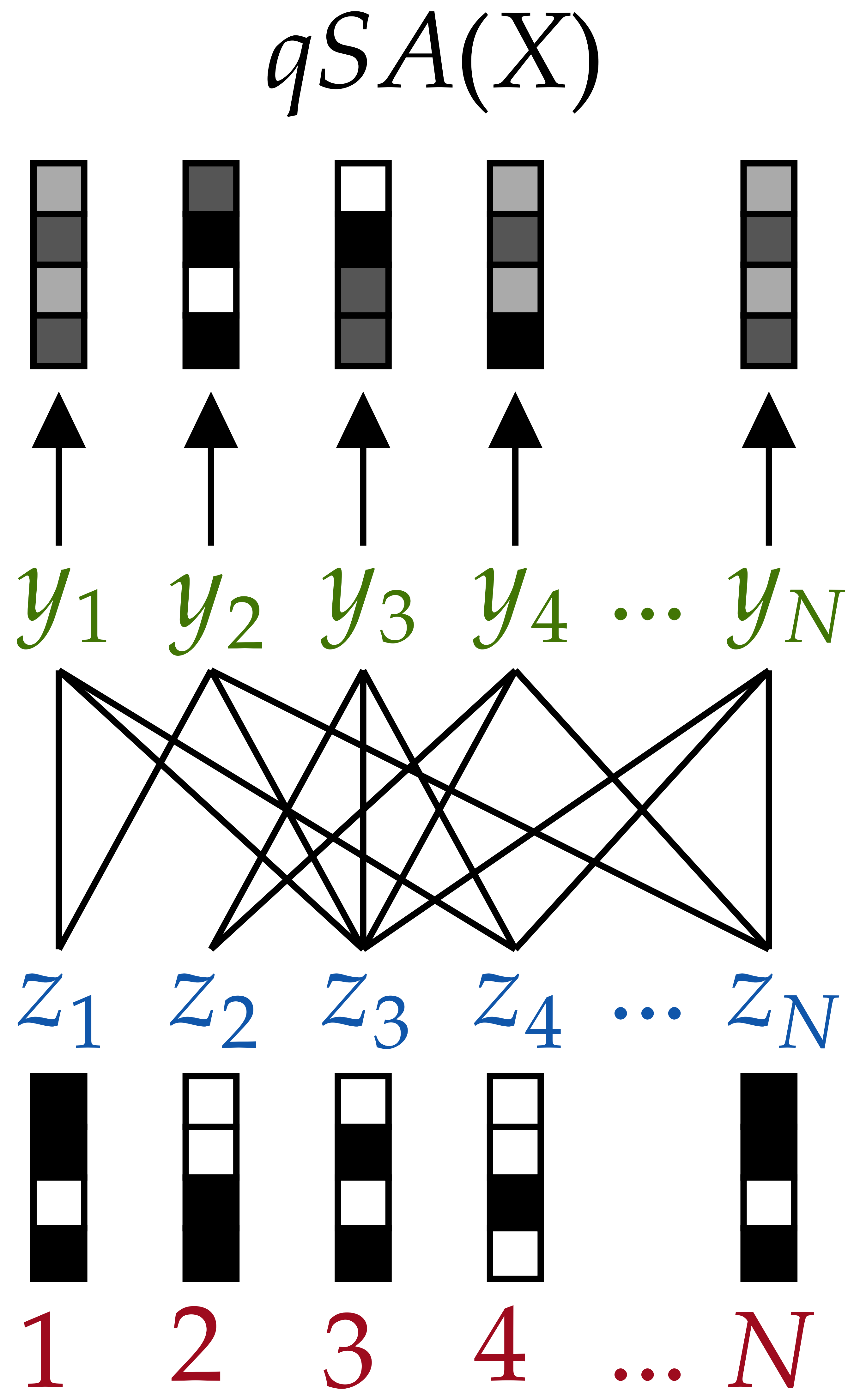
Part 1: Sparse averaging

The positive result

Theorem: For all q, there exists a self-attention unit f with embedding dimension $m = O(d + q \log N)$ that approximates qSA at all X with $\log(N)$-bit precision* arithmetic.

Think: $\log(N), d \ll q \ll N$

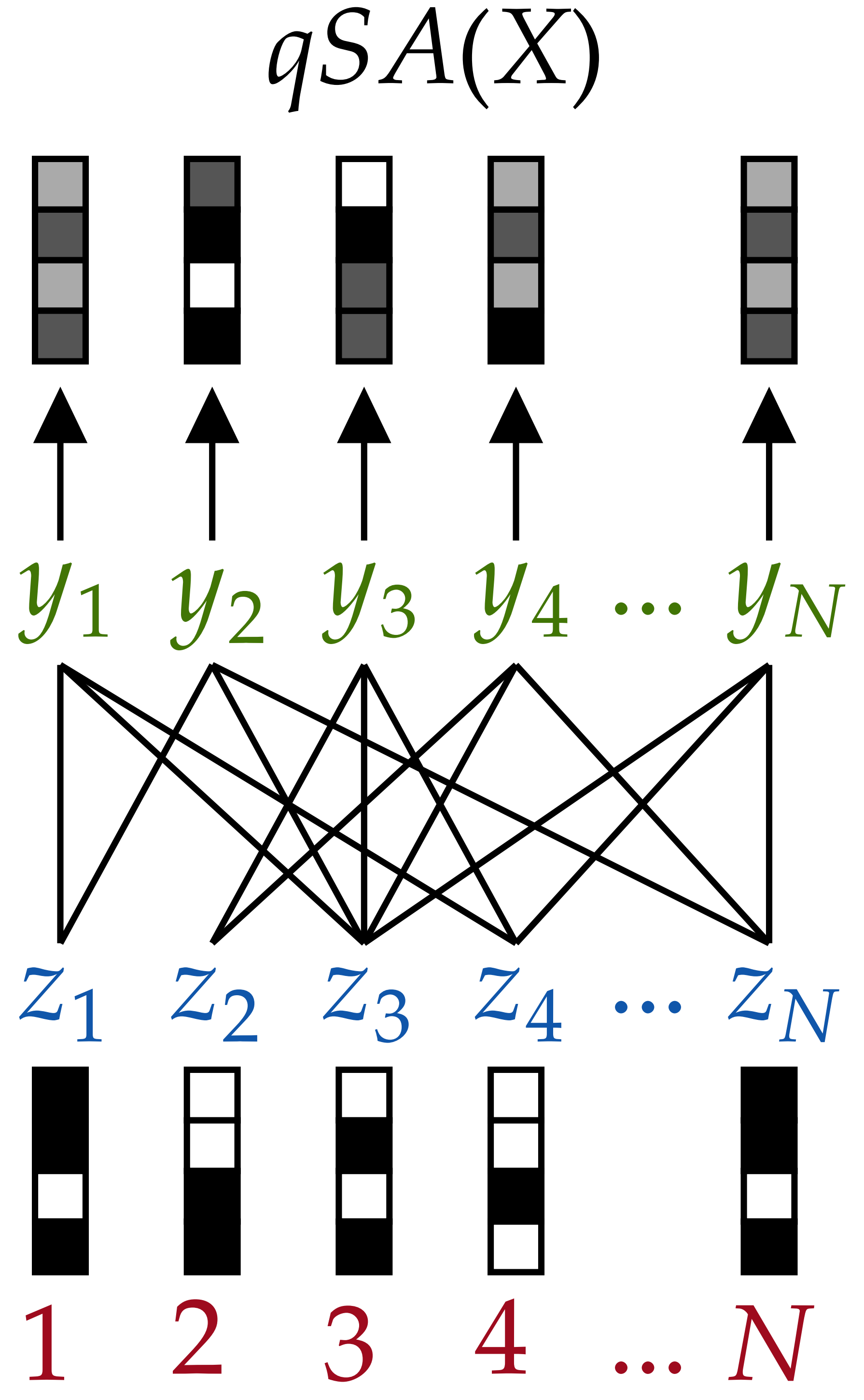
*The $\log N$ factor can be eliminated by using infinite-bit precision.

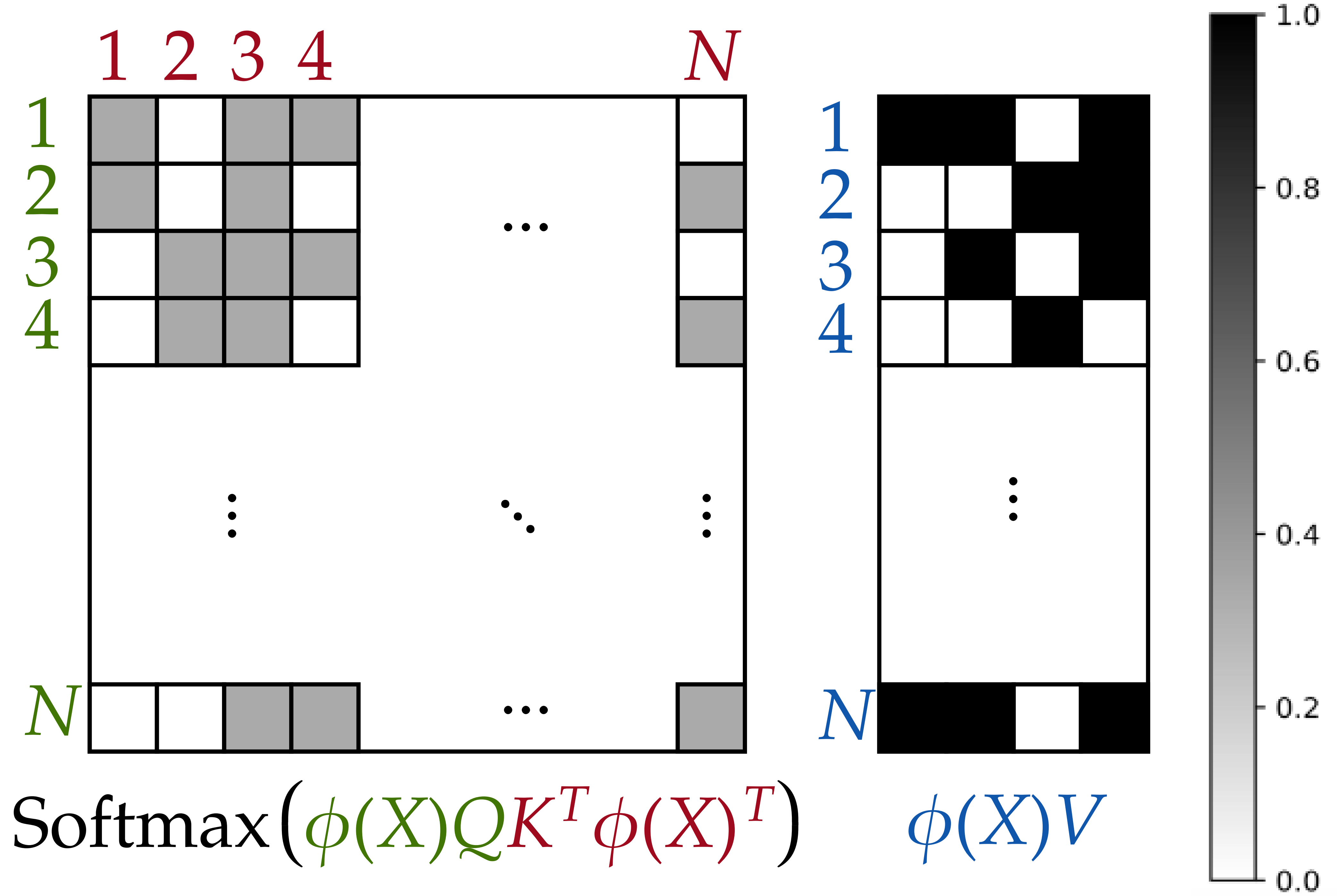


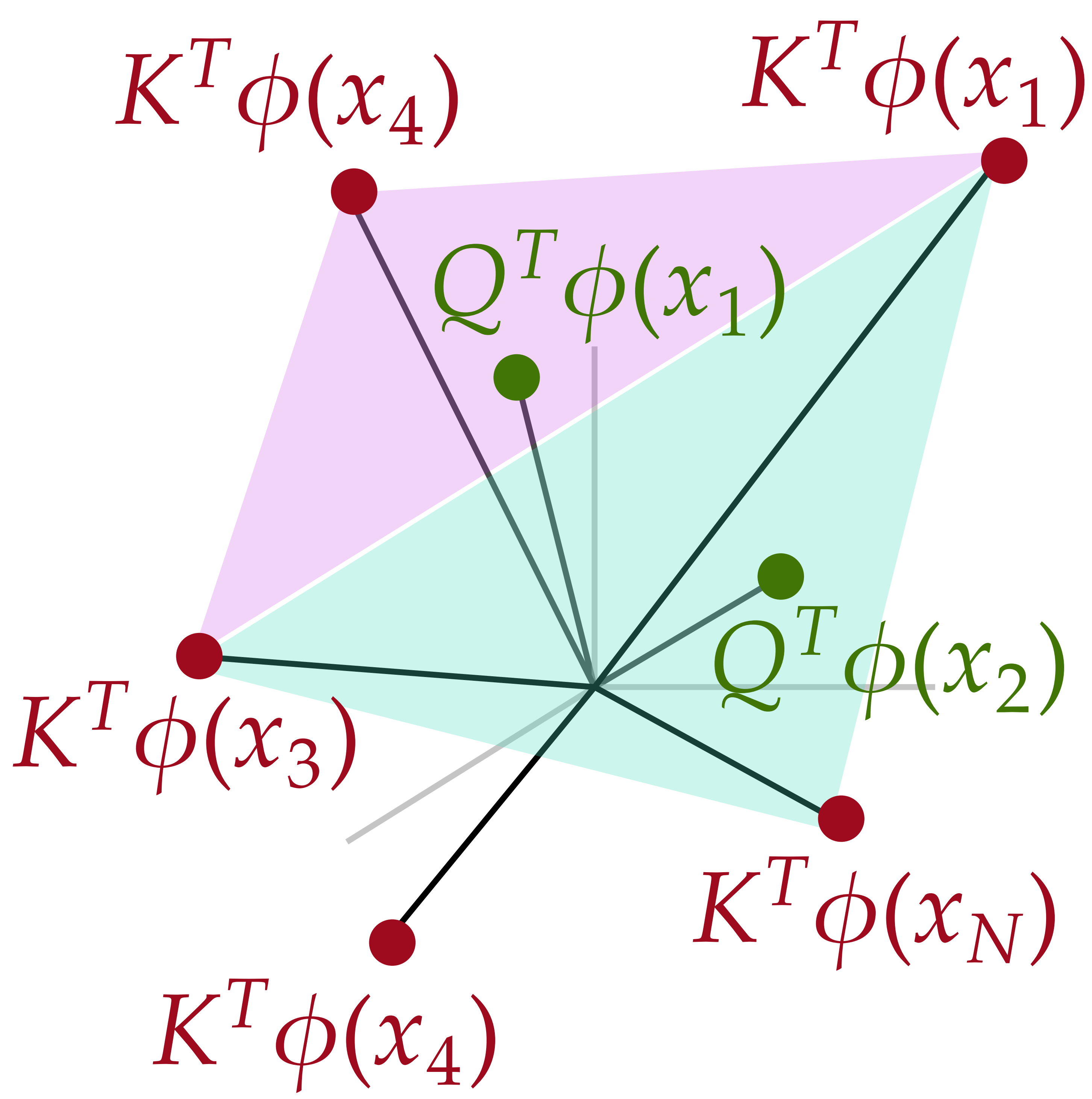
Part 1: Sparse averaging

The positive result: proof by picture

Theorem: For all q, there exists a self-attention unit f with embedding dimension $m = O(d + q \log N)$ that approximates qSA at all X with $\log(N)$-bit precision arithmetic.



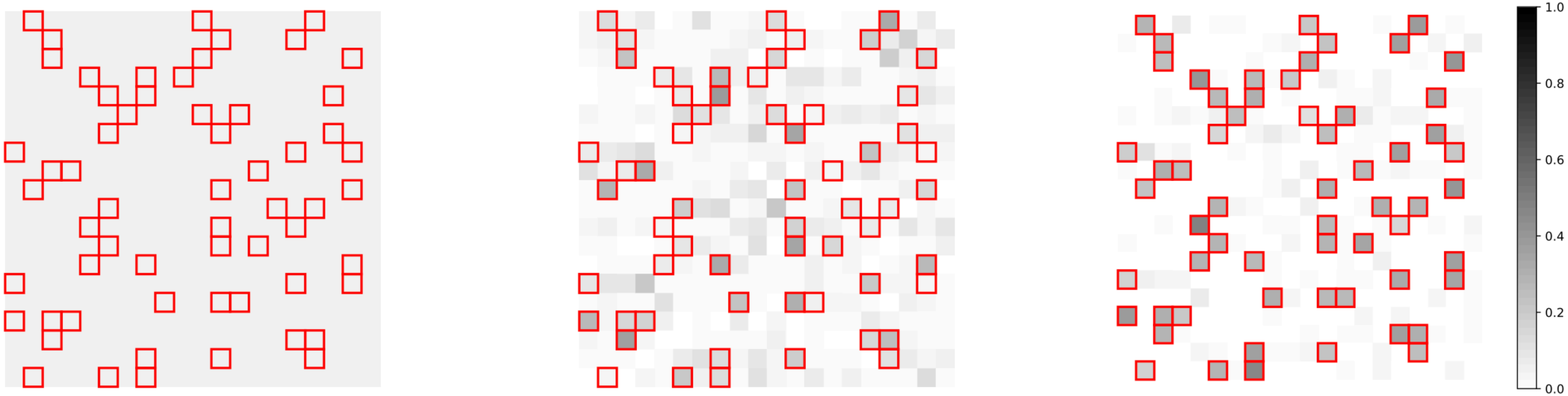




Part 1: Sparse averaging

The positive result: proof by picture

Theorem: For all q, there exists a self-attention unit f with embedding dimension $m = O(d + q \log N)$ that approximates qSA at all X with $\log(N)$-bit precision arithmetic.

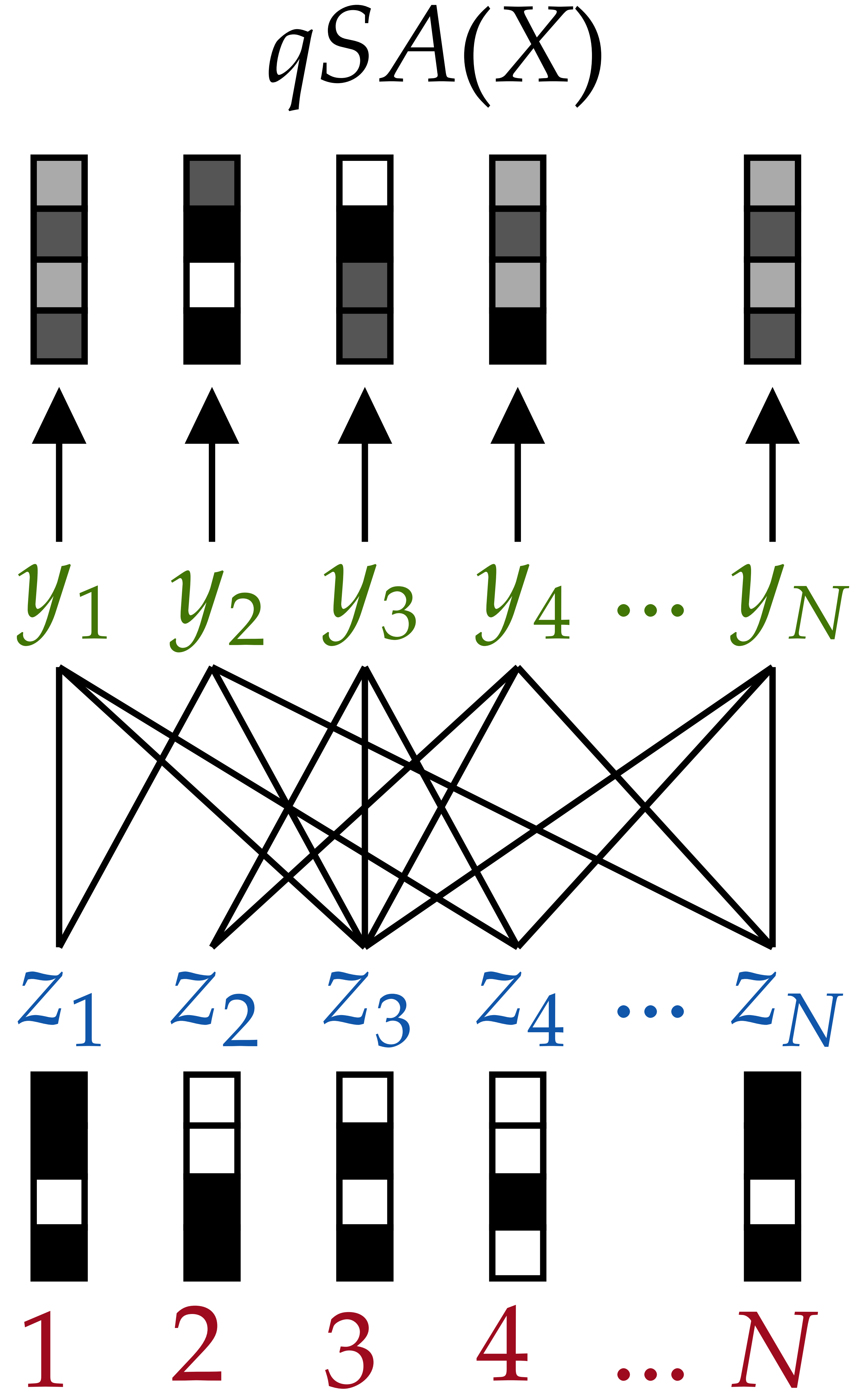


(a) $T = 0$. (b) $T = 1000$. (c) $T = 40000$.

Part 1: Sparse averaging

The negative result

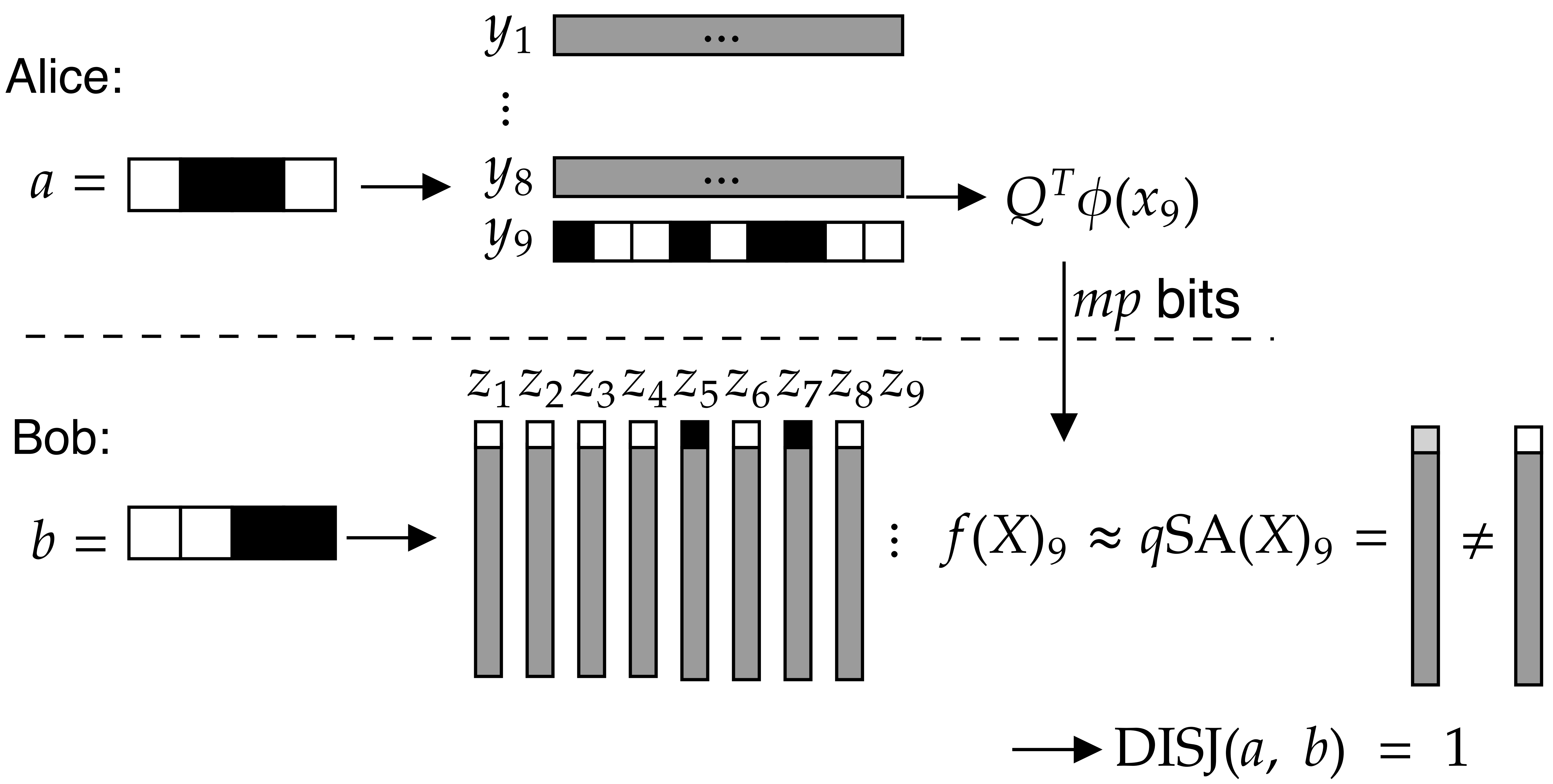
Theorem: Any self-attention unit f that approximates qSA with $\log(N)$-bit precision arithmetic requires embedding dimension $m \geq q/\log N$.

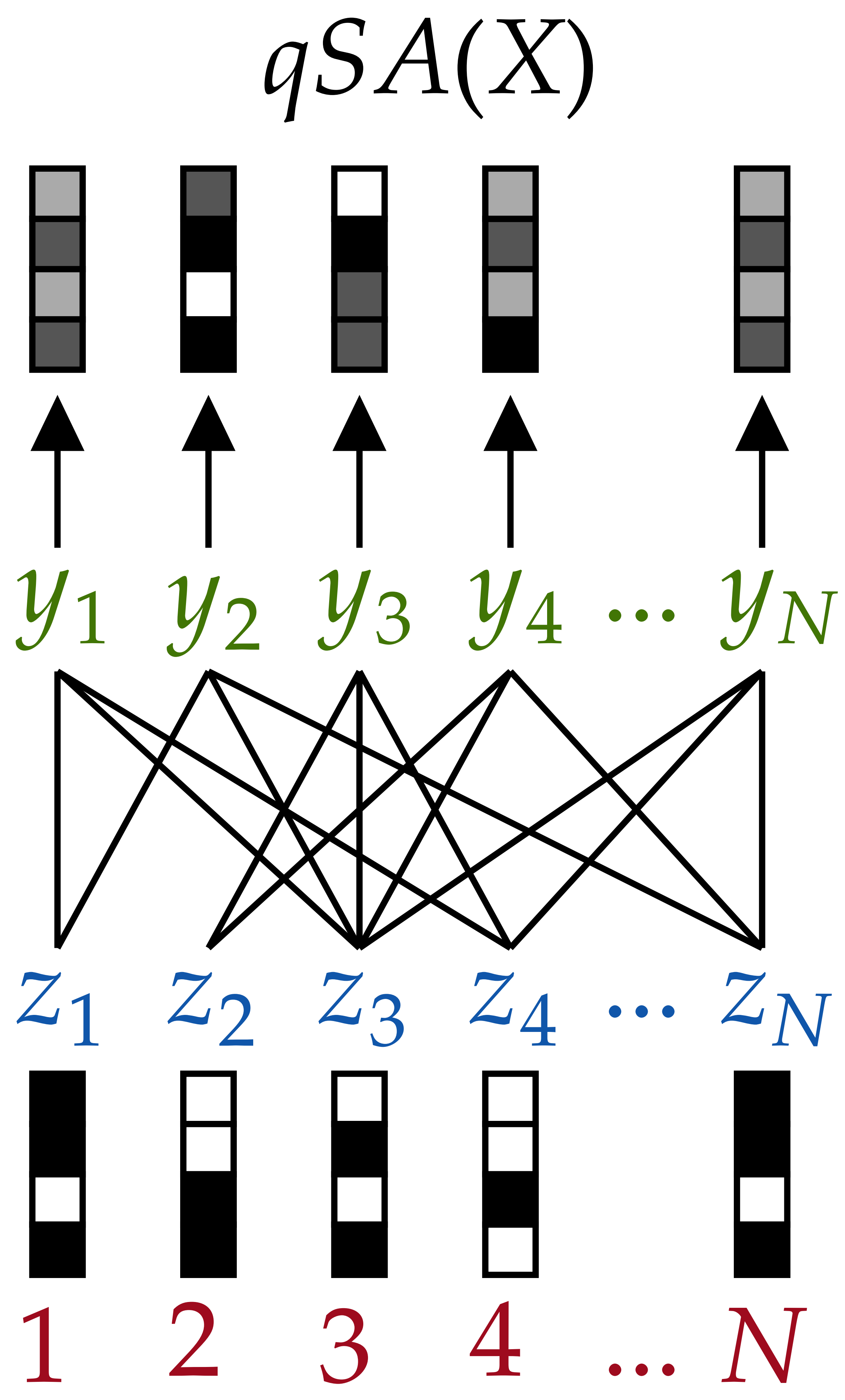


Part 1: Sparse averaging

The negative result: proof by picture

Theorem: Any self-attention unit f that approximates qSA with $\log(N)$-bit precision arithmetic requires embedding dimension $m \geq q/\log N$.





Part 1: Sparse averaging

The task

Input: $X = ((y_1, z_1), \ldots, (y_N, z_N))$

- $y_i \in \binom{[N]}{q}$
- $z_i \in \mathbb{R}^d$.

Output: $qSA(X)_i = \frac{1}{q} \sum_{j \in y_i} z_i$

Results

1. Inefficient representation with FNNs or RNNs.
   - Any FNN requires width $\Omega(Nd)$.
   - Any RNN requires $\Omega(N)$-bit hidden state.
2. Exists self-attention unit approximating $qSA(X)$ iff embedding dim $m \gtrsim q$.

Part 2: Pair and triple finding

The tasks

Input: $X = (x_1, \ldots, x_N) \in [M]^N$.

$\mathrm{Match2}(X)_i = 1\{\, \exists j : x_i + x_j \equiv_M 0\}$

$\mathrm{Match3}(X)_i = 1\{\, \exists j_1, j_2 : x_i + x_{j_1} + x_{j_2} \equiv_M 0\}$

Results

1. Efficient representation of $\mathrm{Match2}$ with self-attention unit.
2. No efficient representation of $\mathrm{Match3}$ with multi-headed self-attention.
3. Efficient representation of $\mathrm{Match3}$ under 3-order attention.

Part 2: Pair and triple finding

Positive result for $\mathrm{Match2}$

$$\mathrm{Match2}(X)_i = 1\{\, \exists j : x_i + x_j \equiv_M 0\}$$

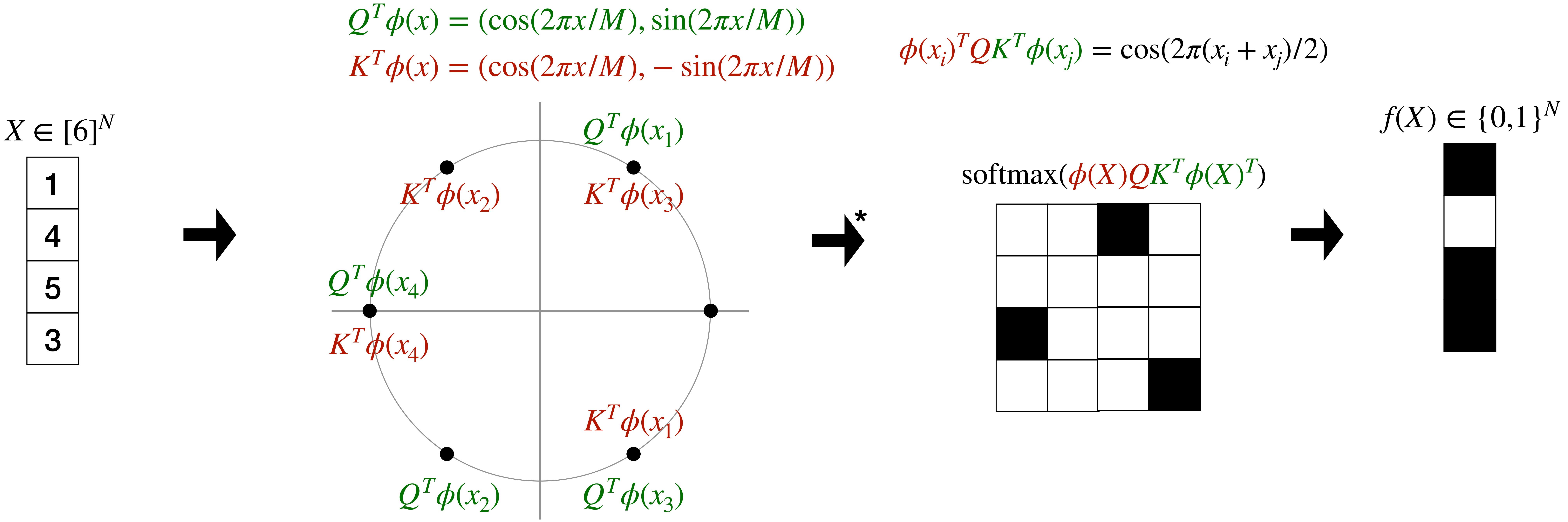
Theorem: There exists self-attention unit f with input MLPs and embedding dimension $m = O(1)$ such that $f(X) = \mathrm{Match2}(X)$.

Part 2: Pair and triple finding

Positive result for $\mathrm{Match2}$: proof by picture

$$\mathrm{Match2}(X)_i = 1\{\exists j : x_i + x_j \equiv_M 0\}$$

Theorem: There exists self-attention unit f with input MLPs and embedding dimension $m = O(1)$ such that $f(X) = \mathrm{Match2}(X)$.



Part 2: Pair and triple finding

Negative result for $\mathrm{Match3}$

$$\mathrm{Match3}(X)_i = 1\{\, \exists j_1, j_2 : x_i + x_{j_1} + x_{j_2} \equiv_M 0\}$$

Theorem: Any H-headed self-attention with input and output MLPs and embedding dimension m and $O(\log N)$-bit precision arithmetic approximating $\mathrm{Match3}$ has $mH = \Omega(N / \log N)$.

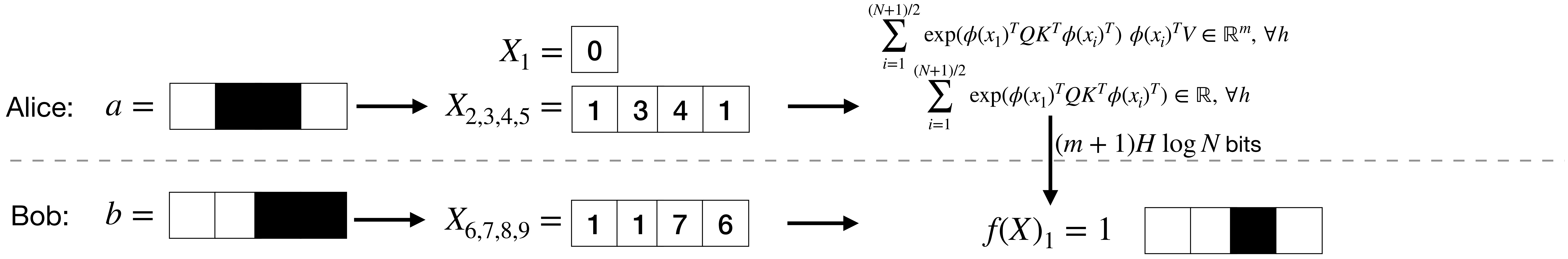
Part 2: Pair and triple finding

Negative result for $\mathrm{Match3}$: proof by picture

$$\mathrm{Match3}(X)_i = 1\{\, \exists j_1, j_2 : x_i + x_{j_1} + x_{j_2} \equiv_M 0\}$$

Theorem: Any H-headed self-attention with input and output MLPs and embedding dimension m and $O(\log N)$-bit precision arithmetic approximating $\mathrm{Match3}$ has $mH = \Omega(N/\log N)$.

- Consider $\mathrm{Match3}(X)_1 = 1\{\, \exists j_1, j_2 \; : x_{j_1} + x_{j_2} \equiv_M 0\}$ ($x_1 = 0$) for $M = N + 2$.
- Suppose exists H-head self-attention layer $f(X)_1 = \psi(\sum_h f_h(\phi(X))_1 = \mathrm{Match3}(X)_1$ having attention units f_h with Q_h, K_h, V_h.
- Reduce (again) from set disjointness with $a, b \in \{0,1\}^n$, $n = (N-1)/2$.



Part 2: Pair and triple finding

Positive result for $\mathrm{Match3}$ (with 3-order attention)

$$\mathrm{Match3}(X)_i = 1\{\, \exists j_1, j_2 : x_i + x_{j_1} + x_{j_2} \equiv_M 0\}$$

3-order attention:

$$f_{Q,K^1,K^2,V^1,V^2}(X) = \mathrm{softmax}(\underbrace{XQ}_{\mathbb{R}^{N\times m}} \underbrace{((XK^1)\otimes(XK^2))^T}_{\mathbb{R}^{m\times N^2}})\underbrace{((XV^1)\otimes(XV^2))}_{\mathbb{R}^{N^2}}$$

$$X \in \mathbb{R}^{N\times d},\ Q, K^1, K^2 \in \mathbb{R}^{d\times m},\ V_1, V_2 \in \mathbb{R}^d$$

Theorem: There exists 3-order self-attention unit f with input MLPs and embedding dimension $m = O(1)$ such that $f(X) = \mathrm{Match3}(X)$.

Part 2: Pair and triple finding

Positive result for $\mathrm{Match3}$ (with 3-order attention): proof sketch

$\mathrm{Match3}(X)_i = 1\{\exists j_1, j_2 : x_i + x_{j_1} + x_{j_2} \equiv_M 0\}$

3-order attention: $f_{Q,K^1,K^2,V^1,V^2}(X) = \mathrm{softmax}(\underbrace{XQ}_{\mathbb{R}^{N\times m}} \underbrace{((XK^1)\otimes(XK^2))^T}_{\mathbb{R}^{m\times N^2}})\underbrace{((XV^1)\otimes(XV^2))}_{\mathbb{R}^{N^2}}$

Theorem: There exists 3-order self-attention unit f with input MLPs and embedding dimension $m = O(1)$ such that $f(X) = \mathrm{Match3}(X)$.

$$Q^T\phi(x) = (\cos(2\pi x/M), -\cos(2\pi x/M), \sin(2\pi x/M), \sin(2\pi x/M))$$
$$K^{1T}\phi(x) = (\cos(2\pi x/M), \sin(2\pi x/M), -\cos(2\pi x/M), \sin(2\pi x/M))$$
$$K^{2T}\phi(x) = (\cos(2\pi x/M), \sin(2\pi x/M), \sin(2\pi x/M), -\cos(2\pi x/M))$$

⬇

$$(\phi(X)Q((\phi(X)K^1)\otimes(\phi(X)K^2))^T)_{i,j_1,j_2} = \cos(2\pi(x_i + x_{j_1} + x_{j_2})/M)$$

Part 2: Pair and triple finding

The tasks

Input: $X = (x_1, \ldots, x_N) \in [M]^N$.

$\mathrm{Match2}(X)_i = 1\{\exists j : x_i + x_j \equiv_M 0\}$

$\mathrm{Match3}(X)_i = 1\{\exists j_1, j_2 : x_i + x_{j_1} + x_{j_2} \equiv_M 0\}$

Results

1. Efficient representation of $\mathrm{Match2}$ with self-attention unit.
2. No efficient representation of $\mathrm{Match3}$ with multi-headed self-attention.
3. Efficient representation of $\mathrm{Match3}$ under 3-order attention.
4. Efficient representation of “assisted” $\mathrm{Match3}$ with standard transformer.

Part 2: Pair and triple finding

Negative conjecture for $\mathrm{Match3}$

$$\mathrm{Match3}(X)_i = 1\{\, \exists j_1, j_2 : x_i + x_{j_1} + x_{j_2} \equiv_M 0\}$$

Conjecture: Any D**-depth** H-headed transformer with embedding dimension m and $O(\log N)$-bit precision arithmetic approximating $\mathrm{Match3}$ has $mHD = \Omega(N/\log N)$.

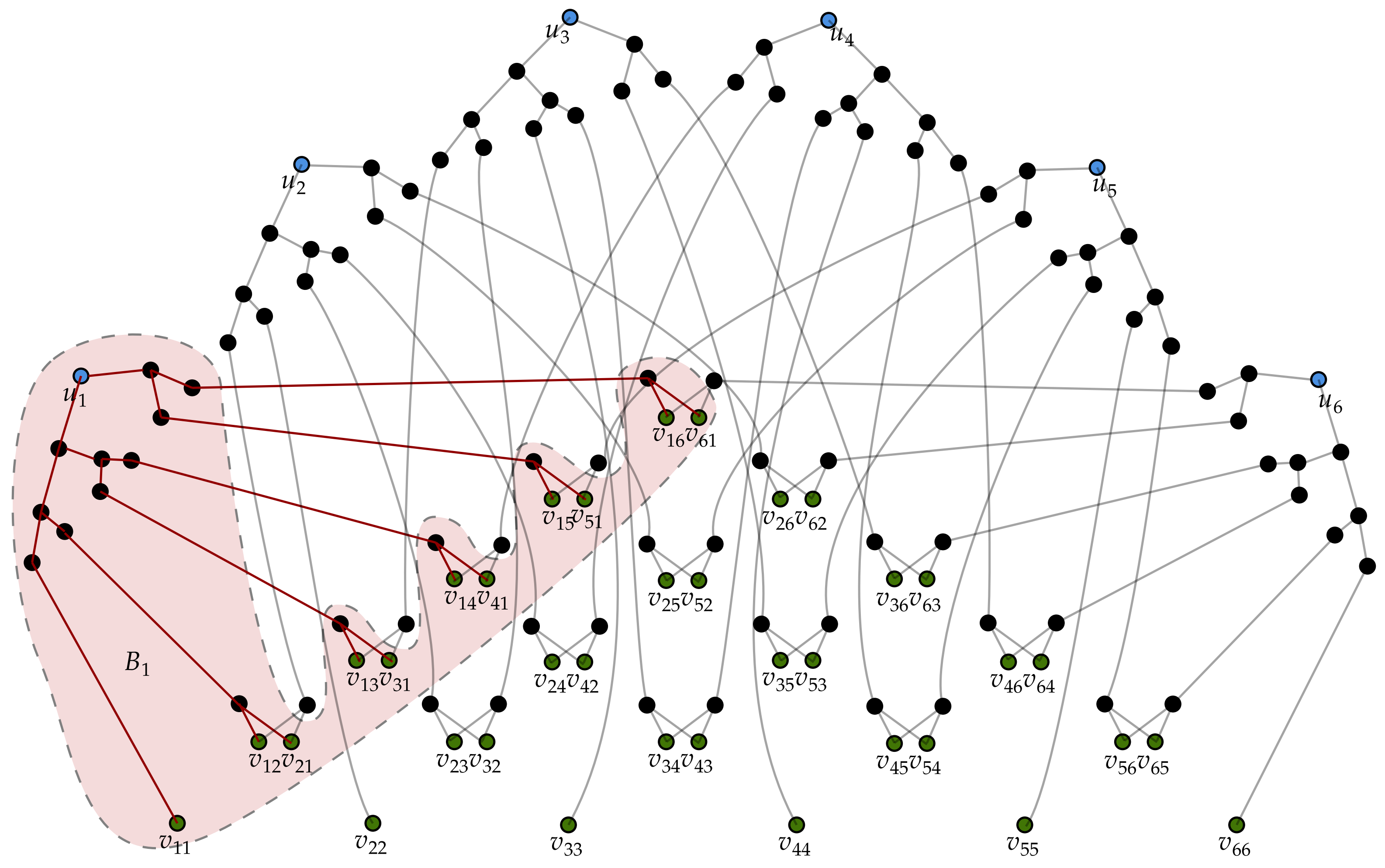
Part 2: Pair and triple finding

Negative conjecture for $\mathrm{Match3}$: hazy intuition

$\mathrm{Match3}(X)_i = 1\{\exists j_1, j_2 : x_i + x_{j_1} + x_{j_2} \equiv_M 0\}$

Conjecture: Any D**-depth** H-headed transformer with embedding dimension m and $O(\log N)$-bit precision arithmetic approximating $\mathrm{Match3}$ has $mHD = \Omega(N/\log N)$.

- Any transformer can be simulated with $O(mHD \log N)$ rounds of communication on a degree-3 CONGEST network with $O(N^2)$ nodes.
- Distribution over inputs with $M = N^4$:
  (1) With probability $1/2$, draw $x_i \sim [M]$ iid. (WHP $\mathrm{Match3(X)} = \overrightarrow{0}$.)
  (2) With probability $1/2$, $x_i \equiv_M -x_{j_1} - x_{j_2}$ for $i, j_1, j_2 \sim [N]$. ($\mathrm{Match3}(X) \neq \overrightarrow{0}$.)
- Indistinguishable unless some node "knows" all of x_i, x_{j_1}, x_{j_2} (?), WP $\approx 1/N^3$
- With $O(N^2)$ total nodes, need $\approx N$ rounds for distinction to occur.



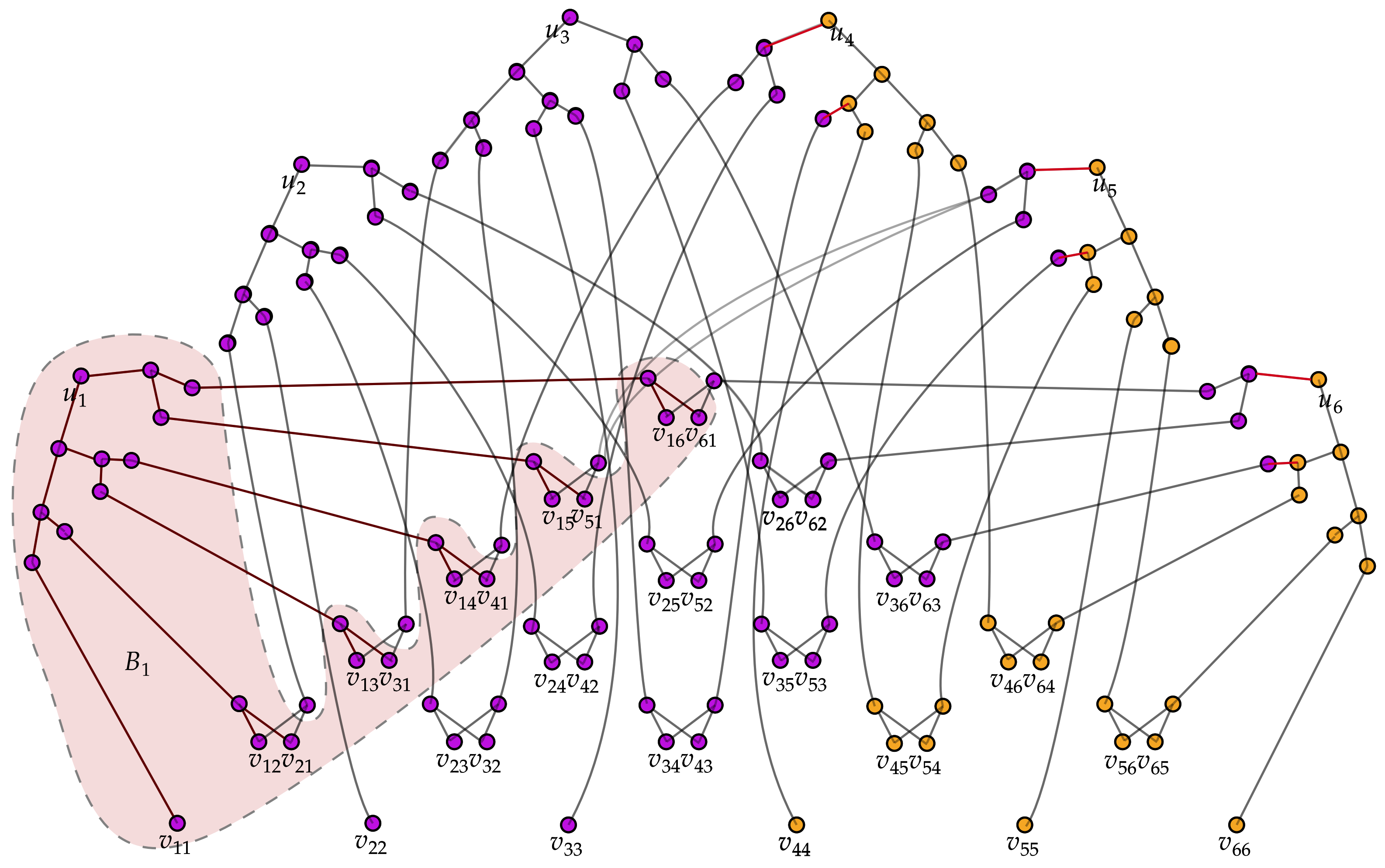
Part 2: Pair and triple finding

Negative conjecture for $\mathrm{Match3}$: a comparable proof

For adjacency matrix $X \in \{0,1\}^{N \times N}$, $\mathrm{Cycle3}(X)_i = 1\{\, \exists j_1, j_2 : (i, j_1, j_1)$ is a cycle$\}$.

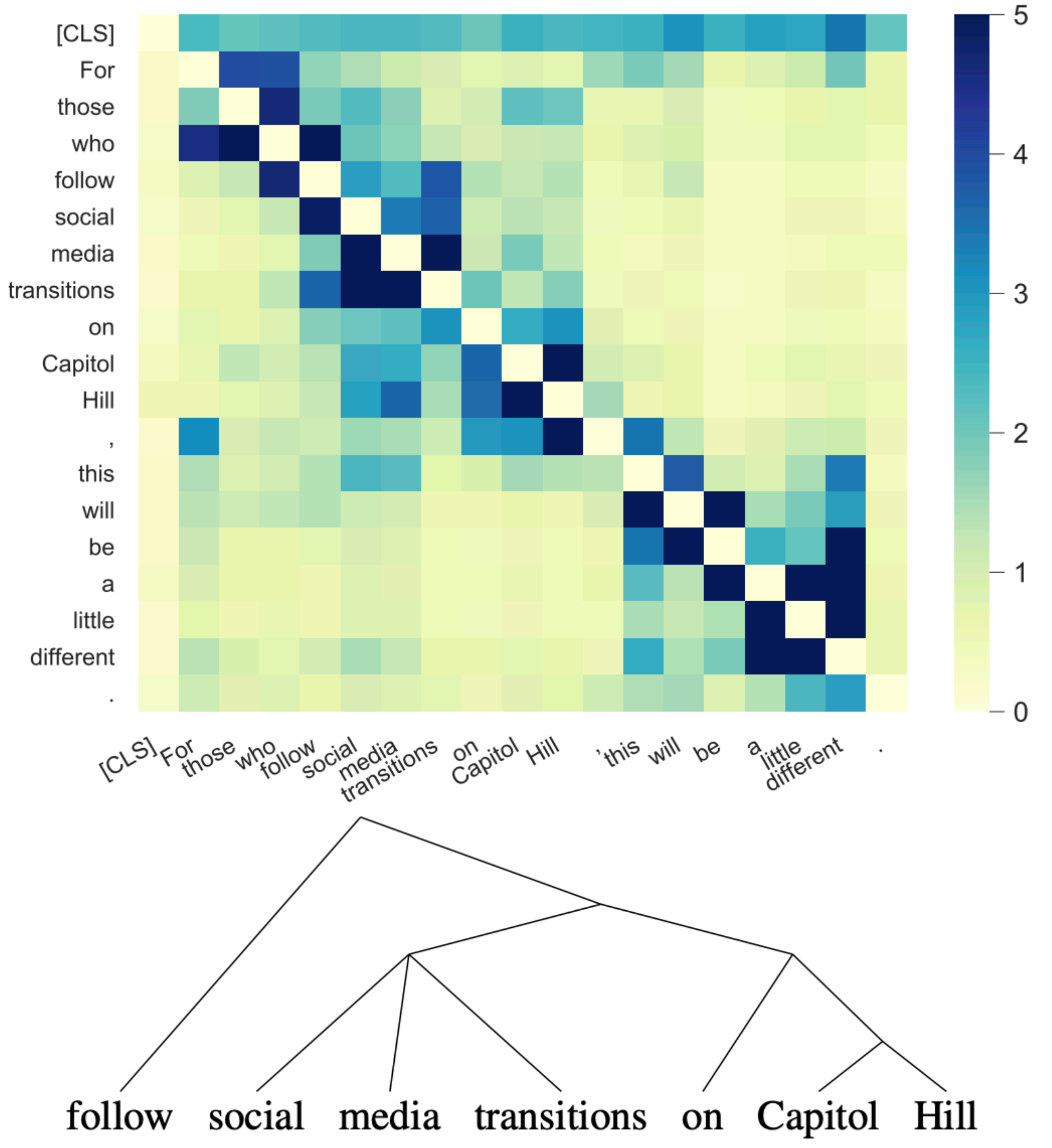
Theorem: Any D-depth H-headed transformer with embedding dimension m and $O(\log N)$-bit precision arithmetic approximating $\mathrm{Cycle3}$ has $mHD = \tilde{\Omega}(N)$.

- Any transformer can be simulated with $O(mHD \log N)$ rounds of communication on a degree-3 CONGEST network with $O(N^2)$ nodes.
- Once again, set-disjointness reduction.



Future work and open questions

- Can more advanced communication complexity and distributed computing techniques be used to resolve the conjecture?
- Can geometric approaches remove the dependence on bit-precision?
- How apt is the "sparse pairwise connectedness" framework for understanding language?
- Are there practical "intrinsically three-wise" learning tasks where modern transformers fail?



Thank you

References

[PMB19] Jorge Pérez, Javier Marinković, and Pablo Barceló. “On the Turing completeness of modern neural network architectures.” 2019.

[YBR+20] Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. “Are transformers universal approximators of sequence-to-sequence functions?” ICLR 2020.

[WCM22] Colin Wei, Yining Chen, and Tengyu Ma. “Statistically meaningful approximation: a case study on approximating turing machines with transformers.” NeurIPS 2022.

[BAG20] Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. “On the ability and limitations of transformers to recognize formal languages.” EMNLP 2020.

[YPPN21] Shunyu Yao, Binghui Peng, Christos H. Papadimitriou, and Karthik Narasimhan. “Self-attention networks can process bounded hierarchical languages.” ACL 2021.

[LAG+22] Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. “Transformers learn shortcuts to automata.” 2022.

[HAF22] Yiding Hao, Dana Angluin, and Robert Frank. “Formal language recognition by hard attention transformers: Perspectives from circuit complexity.” 2022.

[EGKZ22] Benjamin L. Edelman, Surbhi Goel, Sham M. Kakade, and Cyril Zhang. “Inductive biases and variable creation in self-attention mechanisms.” ICML 2022.

[BPKP22] Satwik Bhattamishra, Arkil Patel, Varun Kanade, and Phil Blunsom. “Simplicity bias in transformers and their ability to learn sparse boolean functions.” 2022.

[ZFB23] Ruiqi Zhang, Spencer Frei, Peter Bartlett. “Trained Transformers Learn Linear Models In-Context.” 2023.

[Lou19] Andreas Loukas. “What graph neural networks cannot learn: depth vs width.” 2019.

[XHLG18] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. “How powerful are graph neural networks?” 2018.

[CBCB19] Zhengdao Chen, Soledad Villar, Lei Chen, and Joan Bruna. “On the equivalence between graph isomorphism testing and function approximation with GNNs.” NeurIPS 2019.

[MRF+19] Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. “Weisfeiler and leman go neural: Higher-order graph neural networks.” AAAI 2019.

Appendix / Extra slides

Part 1: Sparse averaging

An aside on communication complexity

- Suppose Alice has $a \in \{0,1\}^n$ and Bob has $b \in \{0,1\}^n$ and they want to compute $\mathrm{DISJ}(a, b) = \max_i a_i b_i$.
- Unlimited computation, bounded communication:
  - Alice and Bob take turns sending single bits of information to one another.
- What is the minimum rounds of communication?
  - $\leq n$ (Alice sends all bits to Bob)
  - $\geq n$ (rank of characteristic matrix)

Part 1: Sparse averaging

The negative result: proof

Theorem: Any self-attention unit f that approximates qSA with $\log(N)$-bit precision arithmetic requires embedding dimension $m \geq q/\log N$.

- Create an $m \log N$-bit protocol for $\mathrm{DISJ}(a, b)$ with $n = q$, assuming the existence of f.
- Alice encodes her input in subset $y_{2q+1} = \{2i + a_i - 1 : i \in [q]\}$.
- Bob encodes his input as $z_{2i-1} = 2a_i - 1,\ z_{2i} = -1$. All other values set arbitrarily.
- Alice sends Bob her $m \log N$-bit query encoding $Q(x_{2q+1})$.
- Bob computes $f(X)$ and returns 1 iff $f(X)_{2q+1} \neq -1$.
- By CC bound, $m \log N \geq q$.