Associative memories

Daniel Hsu

COMS 6998-7 Spring 2025

Auto-associative memory

- Purpose of auto-associative memory M is to remember "patterns"
- Say pattern x is remembered by M if, upon prompting M with
$$x + \delta$$
for "small" corruption δ, the memory M returns x:
$$M(x + \delta) = x$$
- Question: How many patterns can be remembered?

Hopfield network

- Regard pattern as a particular setting of d neurons
$$x \in \{-1,1\}^d$$
- Hopfield network: "biologically plausible" associative memory where $M(x)$ is limiting state of discrete-time dynamics
$$x = x(1) \xrightarrow{T} x(2) \xrightarrow{T} \cdots \to x(\infty)$$
with update rule $T: \{-1,1\}^d \to \{-1,1\}^d$ defined by a neural net
$$T(x) = \operatorname{sign}[Wx]$$
- Theorem: Can remember $\sim d/\log d$ random patterns with d neurons
- Idea: Dynamics = "thresholded" gradient iteration on $x \mapsto -\frac{1}{2}x^\top Wx$

Related problem (but more relevant)

- <u>Associative memory</u> M: mechanism for remembering <u>associations</u>

$$(x, y) \in \{-1,1\}^d \times \{-1,1\}^d$$

  (For simplicity, assume input and output dimensions are the same)

- Say association (x, y) is <u>remembered</u> if, upon prompting M with

$$x + \delta$$

  for "small" corruption δ, the memory M returns y

- In fact, let's just consider $\delta = 0$
- Are there any "biologically plausible" solutions?

Intuition from elementary linear algebra

- Neural network starts with a linear map $x \mapsto Wx$ on $\mathbb{R}^d$
- If keys are <u>linearly independent</u>, and W has <u>full-rank</u>, then keys map to linearly independent "values"
- If keys are <u>right singular vectors</u> of W, and values are <u>left singular vectors</u> of W (scaled by corresponding singular value), then W correctly maps each key to desired value!

One-step Hopfield network

- Assume "keys" $x^{(1)}, \ldots, x^{(n)}$ are drawn independently and u.a.r. from $\{-1,1\}^d$, but allow "values" $y^{(1)}, \ldots, y^{(n)}$ to be arbitrary
- <u>One-step Hopfield network</u>:

$$M(x) = \operatorname{sign}[Wx], \qquad W := \sum_{i=1}^{n} y^{(i)} x^{(i)\top}$$

- Question: How large can n be?
- Answer: $n \sim d / \log d$

Analysis of one-step Hopfield network

1. One-step Hopfield network:

$$M(x) = \operatorname{sign}\left[\sum_{i=1}^{n} \langle x, x^{(i)} \rangle y^{(i)}\right]$$

2. Inside sign for $M\left(x^{(1)}\right)_j$:

$$d y_j^{(1)} + \sum_{i=2}^{n} \left\langle x^{(1)}, x^{(i)} \right\rangle y_j^{(i)}$$

3. We have $M\left(x^{(1)}\right)_j = y_j^{(1)}$ iff

$$-\sum_{i=2}^{n} \left\langle x^{(1)}, x^{(i)} \right\rangle y_j^{(1)} y_j^{(i)} < d$$

4. LHS is sum of $(n-1)d$ independent Rademacher r.v.'s

5. Probability that LHS is $< d$ is

$$\geq 1 - \exp\left(-\frac{d^2}{2(n-1)d}\right)$$

6. Apply union bound

One-step modern Hopfield network

- Krotov and Hopfield (2021) suggest that the following one-step mechanism (proposed and analyzed by Demircigil et al, 2017) is also "biologically plausible":

$$M(x) = \mathrm{sign}\left[\sum_{i=1}^{n} \exp\left(\langle x, x^{(i)} \rangle\right) y^{(i)}\right]$$

- Again, let's assume "keys" $x^{(1)}, \ldots, x^{(n)}$ are drawn independently and u.a.r. from $\{-1,1\}^d$, but allow "values" $y^{(1)}, \ldots, y^{(n)}$ to be arbitrary
- Question: How large can n be?
- Answer: $n \sim \exp(\Omega(d))$

Analysis of one-step modern Hopfield network

1. With probability at least
$$1 - \binom{n}{2} e^{-\epsilon^2 d},$$
for all $i \neq k$,
$$\langle x^{(i)}, x^{(k)} \rangle < \epsilon d$$

2. Inside sign for $M\left(x^{(1)}\right)_j$:
$$\sum_{i=1}^{n} e^{\langle x^{(1)}, x^{(i)} \rangle} y_j^{(i)}$$

3. Let $\alpha_i := e^{\langle x^{(1)}, x^{(i)} \rangle} / \sum_{k=1}^{n} e^{\langle x^{(1)}, x^{(k)} \rangle}$

4. We have $M\left(x^{(1)}\right)_j = y_j^{(1)}$ iff
$$\alpha_1 > -\sum_{i=2}^{n} \alpha_i y_j^{(1)} y_j^{(i)}$$

5. RHS is at most $1 - \alpha_1$

6. So suffices to have $\alpha_1 > 1/2$, i.e.,
$$\frac{1}{1 + \sum_{k=2}^{n} e^{\langle x^{(1)}, x^{(k)} \rangle - d}} > 1/2$$

Connection to transformers

- One-step modern Hopfield network, again:

$$M(x) = \operatorname{sign}\left[\sum_{i=1}^{n} \frac{e^{\langle x, x^{(i)} \rangle}}{\sum_{j=1}^{n} e^{\langle x, x^{(j)} \rangle}} y^{(i)}\right]$$

- Inside the sign is the attention mechanism!
  - Query: x
  - Keys: $x^{(1)}, \ldots, x^{(n)}$
  - Values: $y^{(1)}, \ldots, y^{(n)}$
- Correct operation for n key/value pairs with dimension $d = \Theta(\log n)$