Fundamental Law of Memory Recall

Code at `https://github.com/Einlar/memory_reco`

Francesco Manzali

Università degli Studi di Padova

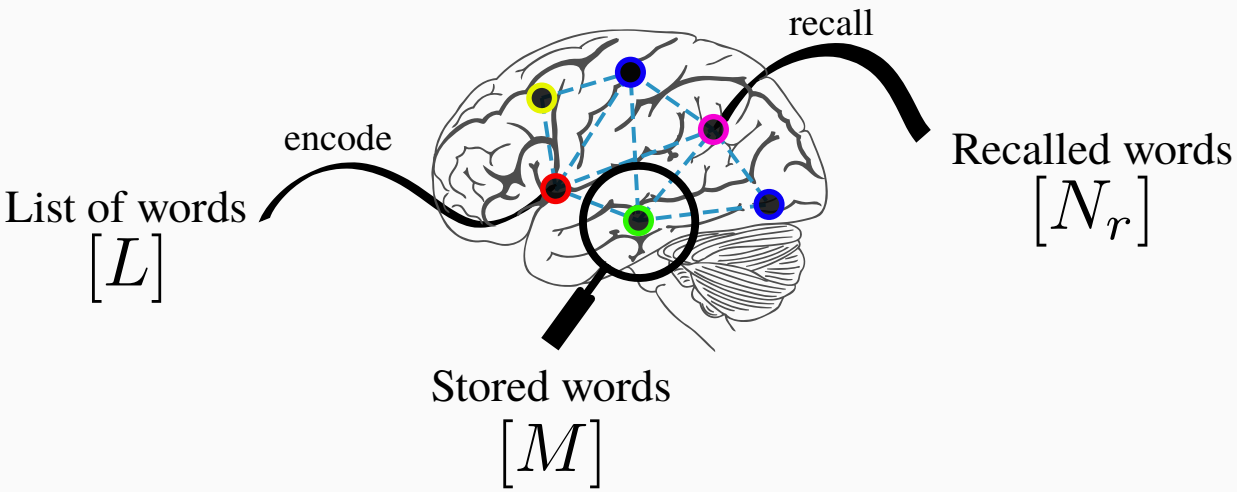
April 15, 2021

Table of contents

How does the brain **store** knowledge?

- **Free recall paradigm**

  Show L words one at a time, then ask to recall all of them in **any** order, and count how many are correctly remembered (N_r).



Recalled items N_r scale as a **power-law** [1]:

$$N_r \sim L^{\alpha} \qquad \alpha \approx .5$$

- Same scaling even if no list is presented, and subjects recall words starting from a given letter.

  $\rightarrow$ This suggests some **universal** underlying principle for recall.

- However, N_r depends also on presentation time Δt and other experimental details.

  $\rightarrow$ Need to separate the **encoding** step (dependent on procedure) from the **recall** step.

Two main approaches for modelling free recall

Bottom-up

Model the neuron dynamics.

- Network model of the hippocampus (Hasselmo and Wyble, 1997).

! **Fails** to reproduce $N_r \sim L^{\alpha}$ even qualitatively.
The **details** of recall are not clear, so it is not possible to model them directly (yet).

Top-down

Model cognitive processes

- *Search of associative memory* (SAM, by Raaijmakers and Shiffrin, 1980)
- *Temporal context model* (TCM, by Howard and Kahana, 2002)

! Very good fit of data, but **many** parameters to be **tuned** to the specific experimental setting (e.g. they vary on Δt, L...).

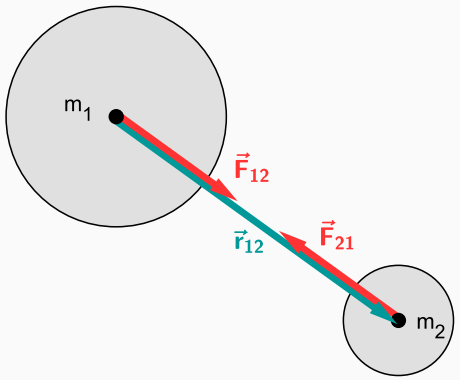
The **physicist**'s approach:

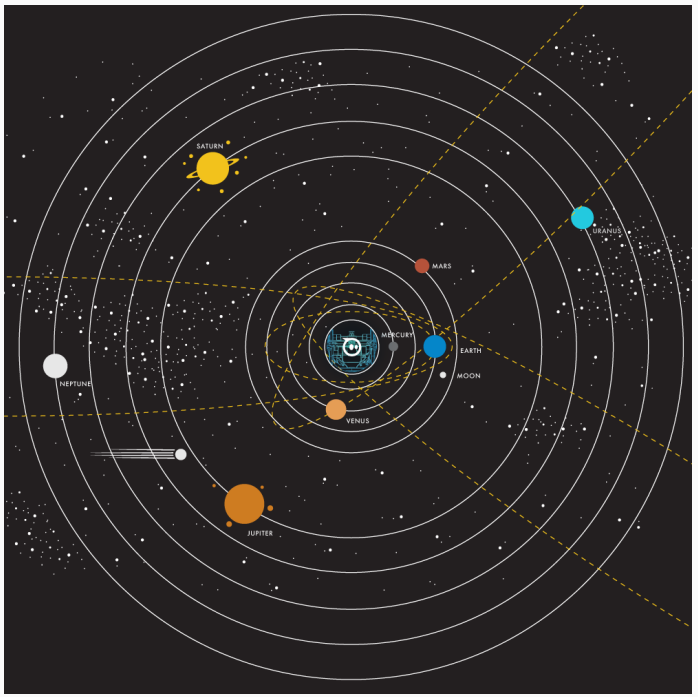
Formulate a few good **first principles**, develop a mathematical **framework** and test their implications with **experiments**.

$$\mathbf{p} = m\mathbf{v} \quad (1)$$

$$\mathbf{F} = \dot{\mathbf{p}} \quad (2)$$

$$\mathbf{F}_{12} = -\mathbf{F}_{21} \quad (3)$$





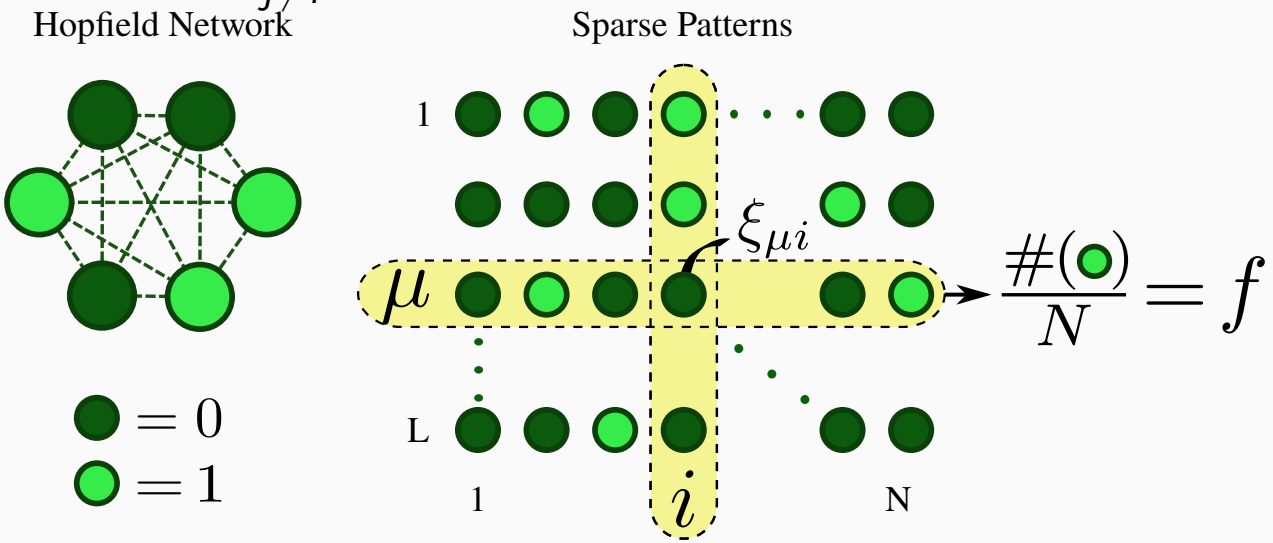
Network Model Hopfield network [3]

- N **visible** binary neurons $V_i \in \{0,1\}$, $i = 1, \dots, N$.
- L **sparse** [2] binary **patterns** $\xi_\mu \in \{0,1\}^N$, $\mu = 1, \dots, L$ and $\mathbb{P}[\xi_{\mu i} = 1] = f$.
- Fully connected with **weights** J_{ij} given by:

$$J_{ij} = \frac{1}{Nf(1-f)} \sum_{\mu=1}^{L} (\xi_{\mu i} - f)(\xi_{\mu j} - f) \tag{1}$$

- **Dynamics**:

$$V_i(t+1) = \Theta\Big(\sum_{\substack{j=1 \\ j \neq i}}^{N} J_{ij} V_j(t) - \mathrm{th}_i \Big) \qquad \mathrm{th}_i \sim \mathcal{U}[-T, T]: \text{ local threshold}$$



This choice of J_{ij} is such that $\xi_{\mu i}$ are **stable attractors** for the dynamics [3].

Proof

The input induced by the μ-th pattern on the i-th neuron is, on average:

$$h_{\mu i} = \frac{1}{Nf(1-f)} \sum_{j \neq i}^{N} (\xi_{\mu i} - f)\xi_{\mu j}(\xi_{\mu j} - f) + \qquad (\nu = \mu)$$

$$+ \frac{1}{Nf(1-f)} \sum_{j \neq i}^{N} \sum_{\nu \neq \mu}^{L} \xi_{\mu j}(\xi_{\nu i} - f)(\xi_{\nu j} - f) \qquad (\nu \neq \mu)$$

The second sum averages to 0, with variance $\approx (L/N)f$. So for $L \ll N$:

$$\langle h_{\mu i} \rangle = \frac{\xi_{\mu i} - f}{Nf(1-f)} N \langle \xi_{\mu j}(\xi_{\mu j} - f) \rangle = \xi_{\mu i} - f$$

Since $\xi_{\mu j} \in \{0, 1\} \Rightarrow \xi_{\mu j}^2 = \xi_j^\mu$, $\langle \xi_{\mu j}(\xi_{\mu j} - f) \rangle = (1-f)\langle \xi_{\mu j} \rangle = f(1-f)$.

$V_i = \xi_{i\mu}$ is **stable** if $\xi_{\mu i} = 1 \Leftrightarrow h_{\mu i} > \mathrm{th}_i$:

$$\max_i(\mathrm{th}_i) < f < \min_i(1 - \mathrm{th}_i) \Leftrightarrow T < f < 1 - T$$

To induce **transitions** between different attractors, we need to **control** their stability.

- Add a **global inhibition** parameter J_0:

$$V_i(t+1) = \Theta\Big(\sum_{\substack{j=1 \\ j \neq i}}^{N} \mathrm{J}_{ij} V_j(t) - \frac{J_0}{Nf} \sum_{j=1}^{N} V_j(t) - \mathrm{th}_i(t) \Big)$$

- The average activation from pattern μ is:

$$\langle h_{\mu i} \rangle = \xi_{\mu i} - f - J_0$$

- The **stability** condition for patterns becomes:

$$T - f < \mathrm{J}_0 < 1 - T - f$$

Setting $J_0 > 1 - T - f$ makes the patterns **unstable**.

- But **intersections** of patterns may still be stable!
  If $V_i = \xi_{\mu i}\xi_{\nu i}$, the average activation (to first order) is:

  $$\langle h_{i;\mu\nu} \rangle = f(\xi_{\mu i} + \xi_{\nu i} - 2f)$$

  Which can be stable at higher J_0:

  $$1 - 2f - \frac{T}{f} < J_0 < 2 - 2f - \frac{T}{f}$$

  with $2 - 2f - T/f > 1 - T - f$ if $f \in [T, 1]$.

  (Because intersections have less active neurons $\sim Nf^2$ than patterns $\sim Nf$, so they experience less inhibition)

By **cycling** J_0 between values below and above the pattern stability threshold, we can induce transitions:

$$\texttt{Pattern} \xrightarrow{\text{Rise } J_0} \texttt{Intersection of patterns} \xrightarrow{\text{Lower } J_0} \texttt{Pattern}$$

How to transition to **new** patterns?

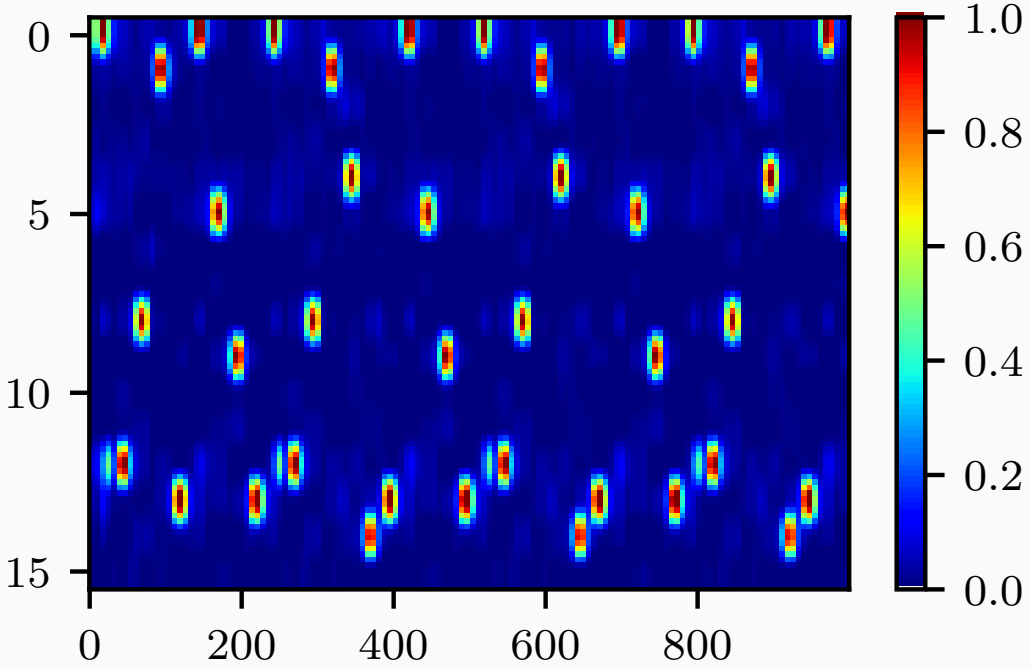
- **Adapt** the thresholds!

$$\mathrm{th}_i(t+1) = \mathrm{th}_i(t) + \underbrace{\frac{D_{\mathrm{th}} V_i(t)}{T_{\mathrm{th}}}}_{\text{Makes current state less stable, incentivizing jumps to new states}} - \underbrace{\frac{\mathrm{th}_i(t) - \mathrm{th}_i(0)}{T_{\mathrm{th}}}}_{\text{Stay close to initial values (limit adaptation)}}$$

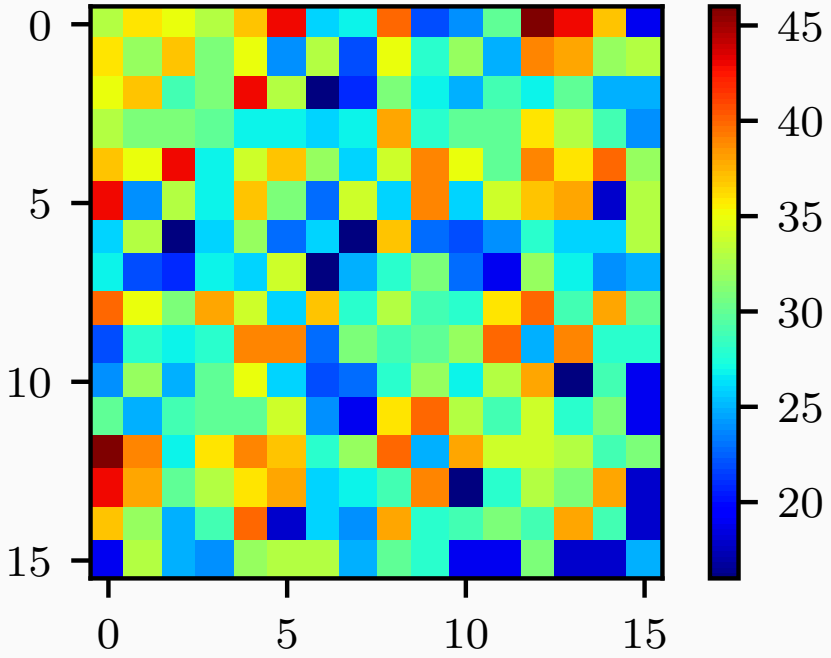
where T_{th} is the timescale of adaptation.

Order parameter: **overlap** with the μ-th pattern

$$m_\mu(t) = \frac{1}{Nf(1-f)} \sum_{i=1}^{N} (\xi_{\mu i} - f)\, V_i(t) \qquad \mu = 1, \ldots, L$$



(a) Plot of $m_\mu(t)$. The dynamics form a loop, and not all patterns are recalled (even if they are indeed stored in J_{ij}).



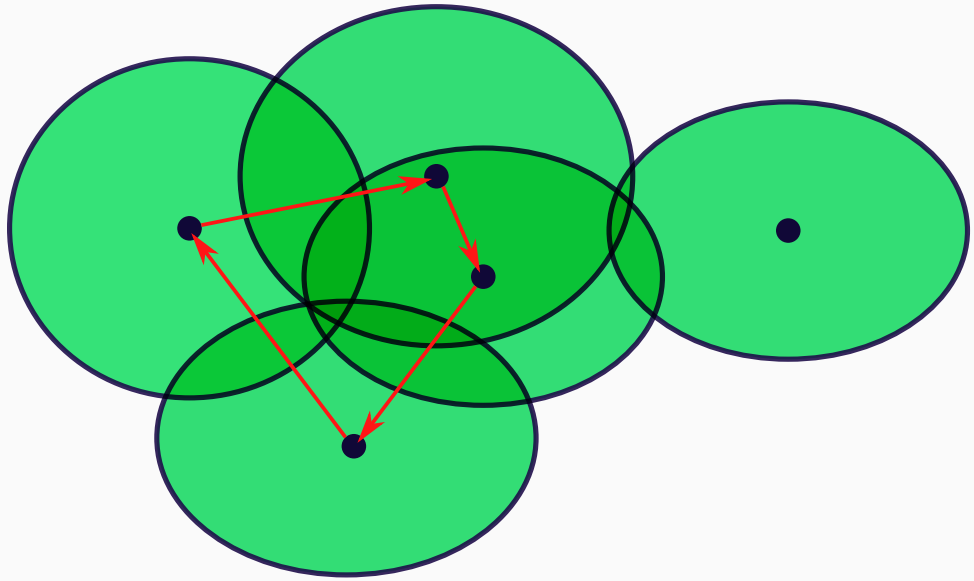
(b) Overlap between each pair of the original L patterns.

Figure 1 – Parameters are: $N = 3000$, $L = 16$, $f = 0.1$, $T = 0.015 \ll f$, $D_{\text{th}} = 4.2T > T$, $T_{J0} = 25$, $T_{\text{th}} = 30 \gtrsim T_{J0}$, $\min J_0(t) = 0.6 < 1 - 2f + T/f$, $\max J_0(t) = 1.2 > 1 - T - f$.

The previous model suggests the following two basic principles [4] [5]:

1. **Encoding**. Memory items are stored by *overlapping* random *sparse* neural ensembles.
2. **Associativity**. During free recall of memory items, the next item to be recalled is the one with the **highest overlap** with the previously recalled item.

   $\rightarrow$ If the highest overlap is with the item that was recalled at the previous step, the *second highest overlap* is chosen instead. This avoids two-item loops.



Analytical Model

Simplified model [5, Supp. material][3]:

- Construct a matrix of overlaps/similarities S_{ij}.
- **First** recalled item is $\mathrm{rec}(1) = i$, chosen at random from the L patterns.
- The **next** recalled item is:

$$\mathrm{rec}(n+1) = \underset{\substack{j \neq \mathrm{rec}(n) \\ j \neq \mathrm{rec}(n-1)}}{\arg\max} \; S_{\mathrm{rec}(n),j}$$

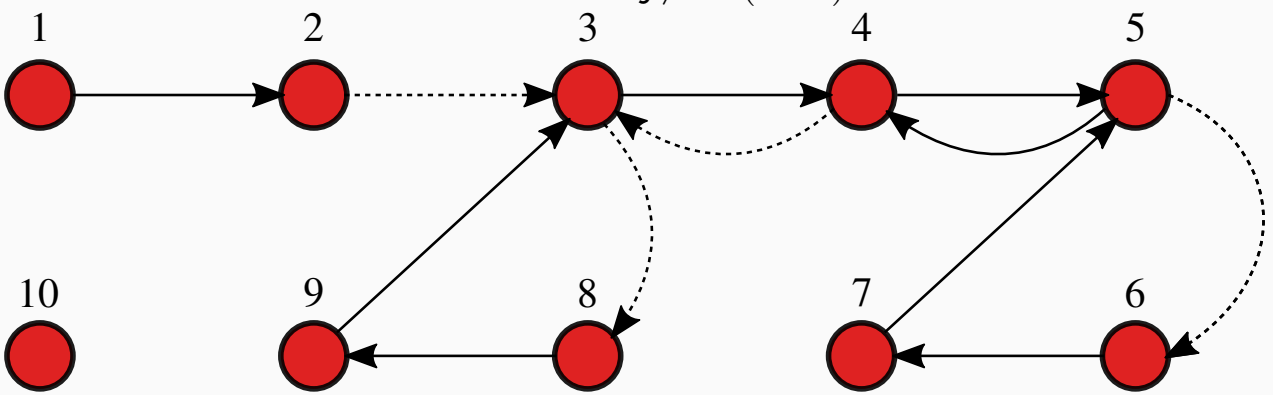


Figure 2 – Example of dynamics. Transitions to items with the maximum overlap are shown as **solid** arrows ("strong" links). If the item with the highest overlap has been visited in the previous step, a transition to the item with the second highest overlap occurs instead (**dashed** arrows, representing "weak" links). Only items inside the **loop** are recalled.

The number of recalled items N_r is the length of the **loop**.

Simplest case: S_{ij} are all i.i.d. random variables (**asymmetric** matrix).

All transitions are **equally** probable:

$$\mathbb{P}[\mathrm{rec}(t+1) = j] = \frac{1}{L-2} \underset{L \gg 1}{\approx} \frac{1}{L} \equiv p_0$$

(because $\mathrm{rec}(t) \to \mathrm{rec}(t)$ and $\mathrm{rec}(t) \to \mathrm{rec}(t-1)$ are not possible).

If k nodes have been visited, the probability of going back to one of them is $(k-2)p_0$.

Then, the probability of a k-loop given L elements in the list is:

$$P(k; L) = \underbrace{(1-p_0)(1-2p_0)(1-3p_0)\cdots(1-(k-3)p_0)}_{\text{Do not enter loops of } \leq k \text{ elements}} \underbrace{(k-2)p_0}_{\substack{\text{Enter a loop of} \\ k \text{ elements}}}$$

$$= \prod_{n=1}^{k-3}(1-np_0)(k-2)p_0 \underset{\substack{L \gg 1 \\ p_0 \ll 1}}{\approx} \exp(-ip_0)kp_0 \approx \frac{k}{L}\exp\left(-\frac{k^2}{2L}\right) \qquad (2)$$

And so $\langle k \rangle \approx \sqrt{\pi L/2}$.

But if S_{ij} represent overlaps, then S_{ij} is a **symmetric** matrix. This introduces **correlations** between elements in the same row/column.

- Consider L sparse patterns, with $\mathbb{P}[\xi_{\mu i} = 1] = f$, and define:

$$S_{\mu\nu} = \sum_{i=1}^{N} \xi_{\mu i}\xi_{\nu i} \Rightarrow S_{\mu\nu} = S_{\nu\mu}$$

Elements in the same row/column are **correlated**. In fact:

$$\langle S_{\mu\nu} \rangle = Nf^2$$

$$\langle S_{\mu\nu}^2 \rangle = \sum_{i,j=1}^{N} \xi_{\mu i}\xi_{\nu i}\xi_{\mu j}\xi_{\nu j} = \sum_{i=1}^{N} \xi_{\mu i}^2\xi_{\nu i}^2 + \sum_{i=1}^{N}\sum_{j\neq i}^{N} \xi_{\mu i}\xi_{\nu i}\xi_{\mu j}\xi_{\nu j} = Nf^2 + N(N-1)f^4$$

$$\mathrm{Var}(S_{\mu\nu}) = \langle S_{\mu\nu}^2 \rangle - \langle S_{\mu\nu} \rangle^2 = Nf^2(1-f^2)$$

$$\langle S_{\mu\nu}S_{\mu\delta} \rangle = \sum_{i,j=1}^{N} \xi_{\mu i}\xi_{\nu i}\xi_{\mu j}\xi_{\delta j} = \sum_{i=1}^{N} \xi_{\mu i}^2\xi_{\nu i}\xi_{\delta i} + \sum_{i=1}^{N}\sum_{j\neq i}^{N} \xi_{\mu i}\xi_{\nu i}\xi_{\mu j}\xi_{\delta j} = Nf^3 + N(N-1)f^4$$

$$\rho(S_{\mu\nu}, S_{\mu\delta}) = \frac{\langle S_{\mu\nu}S_{\mu\delta} \rangle - \langle S_{\mu\nu} \rangle \langle S_{\mu\delta} \rangle}{\sqrt{\mathrm{Var}(S_{\mu\nu})\,\mathrm{Var}(S_{\mu\delta})}} = \frac{f}{1+f} > 0$$

In the symmetric case, **not** all transitions are equally likely!

- If S_{nk} is the highest in row n, then $S_{kn} = S_{nk}$ is likely the highest in row k too.

In the two rows there are $2L$ elements. Of these, $2L-1$ are assumed to be i.i.d. random variables $\{x_n\}_{n=1,\dots,2L-1}$, and the last is fixed from $S_{kn} = S_{nk} \equiv x_1$.

$$\begin{aligned}
&\mathbb{P}[S_{nk} = \max \text{row } k | S_{nk} = \max \text{row } n] = \\
&= \mathbb{P}[x_1 > \max\{x_n\}_{n=L+1,\dots,2L-1} | x_1 = \max\{x_n\}_{n=1,\dots,L}] = \\
&= \frac{\mathbb{P}[x_1 = \max\{x_n\}_{n=1,\dots,2L-1}]}{\mathbb{P}[x_1 = \max\{x_n\}_{n=1,\dots,L}]} = \frac{1/(2L-1)}{1/L} = \frac{L}{2L-1} \underset{L\gg 1}{\approx} \frac{1}{2}
\end{aligned}$$

- This also means that weak/strong transitions are equally likely.
- All other entries are equally likely to be the highest:

$$\mathbb{P}[S_{jk} = \max \text{row } k, j \neq k | S_{nk} = \max \text{row } n] \approx \frac{1}{2(L-2)}$$

Consider a sequence that visits n ("old" node) and reaches k ("new" node) at the end:

$$\underbrace{(\alpha_1 \to \cdots \to \alpha_s)}_{\alpha_i \neq n,k} \to n \to \underbrace{(\beta_1 \to \cdots \to \beta_t)}_{\beta_j \neq n,k} \to k$$

What is the probability p_0 that the next transition "loops back" to n, i.e. that $k \to n$ happens?

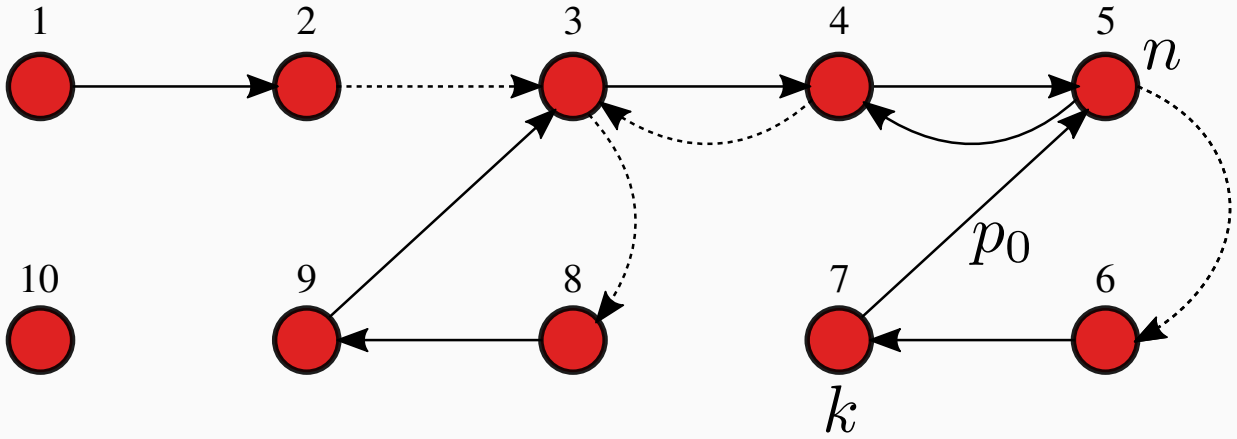


Figure 3 – The path goes from $1 \to 2 \to 3 \to 4 \to 5 \to 6 \to 7 \equiv k$ and then back to $n = 5$.

For $k \to n$ to be allowed, $\mathrm{prec}(k) \neq n$, i.e. $n \to k$ must have **not** happened, meaning that S_{nk} is likely small, and since $S_{kn} = S_{nk}$, and so $k \to n$ is *less* likely.

In fact, it is *half* as likely as the same transition in the case of S_{ij} asymmetric:

$$p_0 \approx \frac{1}{2L}$$

Proof for p_0 scaling (expanded from supp. material of [5])

Since $n \to k$ has **not** happened, then $\mathrm{S}_{nk} < \max_j \mathrm{S}_{nj}$.
(If $n \to \mathrm{succ}(n)$ is a "weak" link, we can say even that $\mathrm{S}_{nk} < \text{second}\max_j \mathrm{S}_{nj}$. But for $L \gg 1$, $\text{second}\max_j \mathrm{S}_{nj} \approx \max_j \mathrm{S}_{nj}$, so the two events are mostly the same).

Then, $k \to n$ can happen either as a "weak" or "strong" transition, with equal chances $(1/2)$.

Consider the "strong" case, where we need $\mathrm{S}_{kn} = \max \boldsymbol{S_k}$, and denote with $\boldsymbol{S_i}$ the i-th row of S:

$$p_{0,\text{strong}} = \mathbb{P}(\mathrm{S}_{kn} = \max \boldsymbol{S_k} | \mathrm{S}_{kn} < \max \boldsymbol{S_n}) = \frac{\mathbb{P}[\mathrm{S}_{kn} = \max \boldsymbol{S_k}, \mathrm{S}_{kn} < \max \boldsymbol{S_n}]}{\mathbb{P}[\mathrm{S}_{kn} < \max \boldsymbol{S_n}]} =$$

$$= \frac{\mathbb{P}[\mathrm{S}_{kn} = \max \boldsymbol{S_k}]\mathbb{P}[\mathrm{S}_{kn} < \max \boldsymbol{S_n} | \mathrm{S}_{kn} = \max \boldsymbol{S_k}]}{\mathbb{P}[\mathrm{S}_{kn} \neq \max \boldsymbol{S_n}]} =$$

$$= \frac{\mathbb{P}[\mathrm{S}_{kn} = \max \boldsymbol{S_k}]\mathbb{P}[\max \boldsymbol{S_k} < \max \boldsymbol{S_n}]}{\mathbb{P}[\mathrm{S}_{kn} \neq \max \boldsymbol{S_n}]} \approx \frac{(1/L)(1/2)}{1 - 1/L} \approx \frac{1}{2L}$$

The same scaling holds for $p_{0,\text{weak}} = \mathbb{P}(\mathrm{S}_{kn} = \text{second}\max \boldsymbol{S_k} | \mathrm{S}_{kn} < \max \boldsymbol{S_n})$, since max and second max are similar for $L \gg 1$.

Thus, $p_0 = p_{0,\text{strong}}/2 + p_{0,\text{weak}}/2 \approx 1/(2L)$.

However p_0 is **not** the probability to enter a cycle!

In fact, after a transition $k \to n$ to an old element, there is still a chance (p_1) to **escape**:

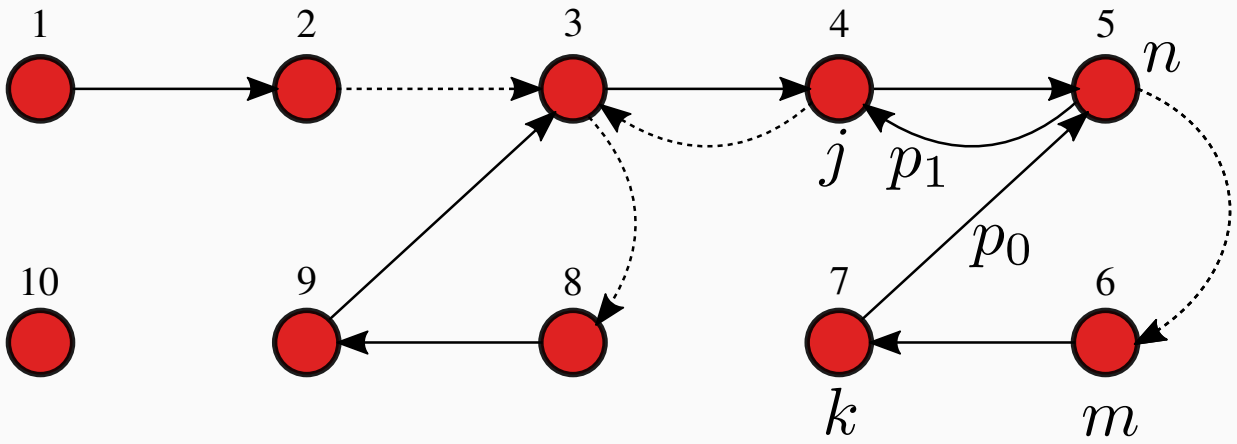


Figure 4 – Since $5 \to 6$ was a "weak" transition, after re-entering 5 from 7, the "backward transition" $5 \to 4$ is taken, which avoids entering a cycle!

- $n \to j$ happens for sure if $n \to m$ was a "weak" link.
- By itself, any transition has a probability $1/2$ of being "weak", i.e. $S_{nj} = \max \boldsymbol{S_n}$, and $S_{nj} = \text{second max } \boldsymbol{S_n}$. However we are still building on $k \to n$, which is possible only if $n \to k$ has not happened ($S_{nk} \neq \max \boldsymbol{S_n}$).
- In the limit $f \to 0$ ("very sparse code"):

$$p_1 = \frac{1}{3}$$

Proof for p_1 (expanded from supp. material of [5])

We know that $j \to n$, $n \to m$ and $k \to n$ have happened. This means that:

$$S_{jn} = \max_{\alpha \neq \text{prec}(j)} S_{j\alpha}$$

$$S_{nm} = \max_{\alpha \neq j} S_{n\alpha}$$

$$S_{kn} = \max_{\alpha \neq m} S_{k\alpha}$$

When computing the maxima we remove the previous item to avoid "backward" links. We also set $S_{ii} < 0$ to avoid self-connections.

Now, $S_{jn} = S_{nj}$ and $S_{kn} = S_{nk}$ by symmetry.

Note that the "backward" transition $n \to j$ can happen only if $S_{nj} > S_{nm}$, otherwise we would get $n \to m$ again instead. We know that $n \to m$ has happened, but not $n \to k$, and so $S_{nm} > S_{nk}$. Thus, we need the probability:

$$p_1 = \mathbb{P}[S_{nj} < S_{nm} | S_{nm} > S_{nk}]$$

Since all of these terms are maxima of equal length vectors (rows), any ordering of S_{nj}, S_{nm} and S_{nk} is equally likely (if we can neglect correlations, i.e. $f \to 0$). There are $3! = 6$ orderings possible, and only 2 satisfy the requirements ($S_{nj} < S_{nk} < S_{nm}$ and $S_{nk} < S_{nj} < S_{nm}$), and so $p_1 = 2/6 = 1/3$.

To first order, the probability of entering a loop after visiting m items is $(m-2)p_0(1-p_1)$.

By just substituting $p_0 \to p_0(1-p_1) = 1/(3L)$ in (2) we get:

$$\langle k \rangle = \sqrt{\frac{3\pi}{2}L}$$

- This model of free recall, in the limit $f \to 0$, has **zero parameters**!

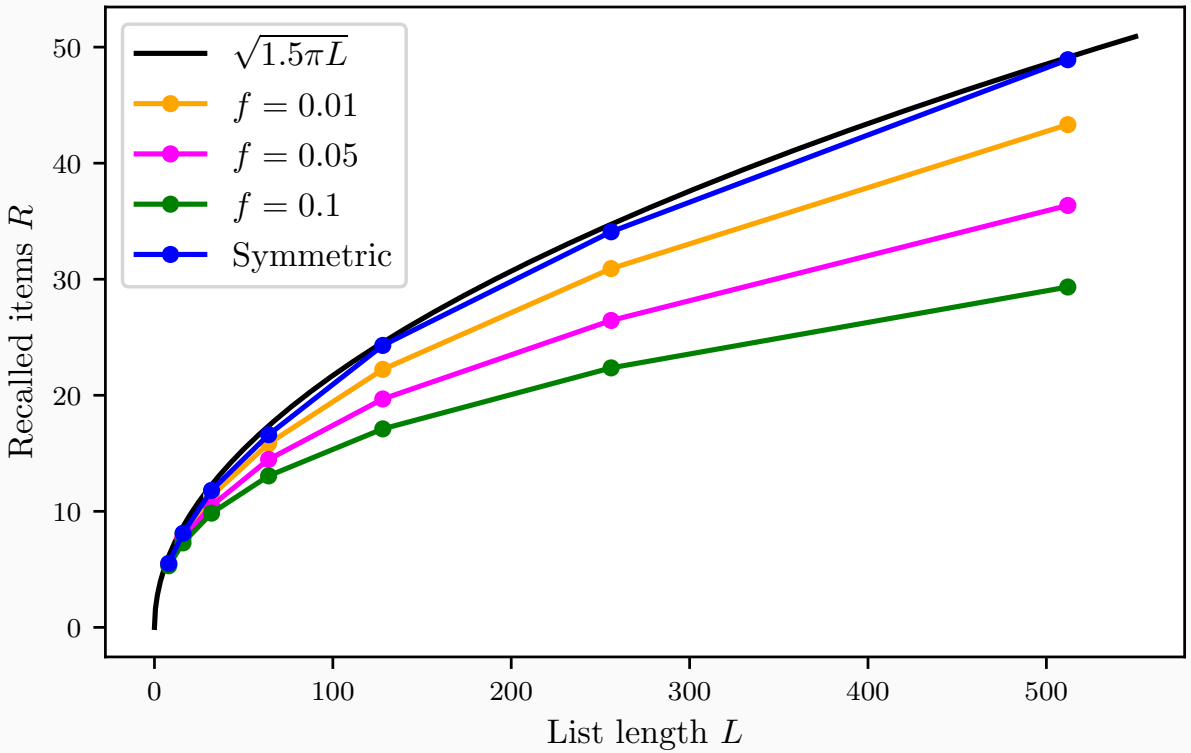
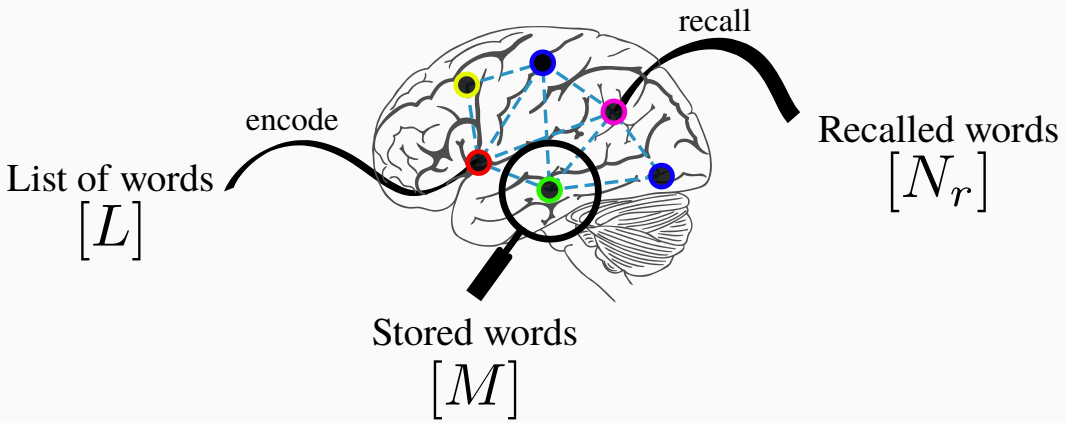


Figure 5 – Average number of recalled items (i.e. length of loops) obtained from a numerical simulation. For the 3 curves at finite f, the similarity matrix S_{ij} was computed from the overlaps of random sparse binary patterns of $N = 100000$ neurons, while for the "Symmetric" curve, S_{ij} was built by first sampling $L(L-1)/2$ i.i.d. variables from $\mathcal{U}[0, 10]$ for the $i \geq j$ entries, and then using $S_{ji} = S_{ij}$ to fill the entries with $i < j$. For each value of L, the shown points are average of 1000 trials. In each trial, the loop length is computed for all possible L starting positions, and the results are averaged.

- In the analytical model, *L* is the number of patterns **stored** in the network.
- But in actual experiments, *L* is the number of words shown. The words that are actually “stored”, i.e. that “they can be recalled”, are instead $M \neq L$.



How to estimate *M*?

Dataset: L words chosen at random from a pool of 751 "common" (frequency > 10 per million) English words.

1. Pick a **list length** $L \in \{8, 16, 32, 64, 128, 256, 512\}$ and a **presentation time** (either $1\,\mathrm{s/word}$ or $1.5\,\mathrm{s/word}$, including a $0.5\,\mathrm{s}$ blank frame as separation).
2. **Estimate** N_r.
   - 2.1 Show a list of length L.
   - 2.2 Each participant **writes recalled words** in any order.
     Already typed words are removed from the screen.
     Repetitions/intrusions are ignored, misspellings are corrected.
3. **Estimate** M.

   "If a word is in memory, it can be recognized between unrelated words."

   - 3.1 Show a list of length L (different from the previous one!).
   - 3.2 Show pairs of words on screen. One word is from the list, the other is unrelated, and their order is randomized. The participant tries to **recognize** the word from the list.
   - 3.3 If the word is remembered, then the participant will pick the correct one with $p = 1$. Otherwise, they will choose at random, giving the correct answer with $p = .5$. Thus, if the probability of remembering a word is $m = M/L$:

$$\mathbb{P}(\text{Answer is correct}) \equiv c = \frac{M}{L} \cdot 1 + \left(1 - \frac{M}{L}\right) \cdot \frac{1}{2} \Rightarrow M = L(2c - 1)$$

Experiments Results

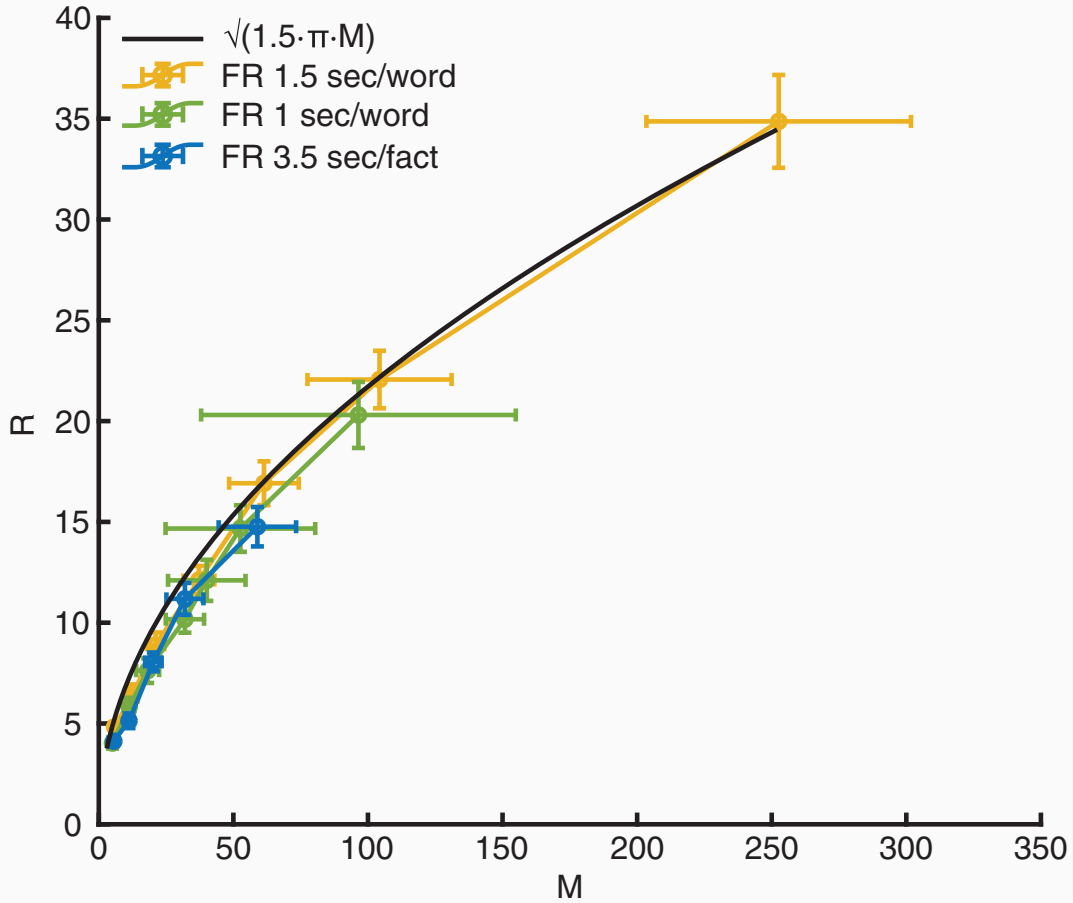


Figure 6 – Results from the experiment: average number of recalled items R ($\equiv N_r$) versus estimate M of items stored in memory, reconstructed from the probability c of correct recognition. Error in R is the standard error of the mean, while error in M is computed with a bootstrap procedure. Three experiments are done with different presentation times. **Yellow**: $1.5\,\mathrm{s/word}$, 348 participants. **Green**: $1\,\mathrm{s/word}$, 375 participants. **Blue**: free recall of brief sentences, $3.5\,\mathrm{s/fact}$, 331 participants. A fact is considered correctly recalled if the mean word embedding S_1 is "close to" the true average word embedding S_2, in the sense that $\cos(S_1, S_2) > 0.9$. Taken from fig. 1 of [5].

Conclusions and Perspectives

The average number of recalled words follows the theoretical predictions, without needing **any** tuning!

- Both M and N_r highly depend on experimental details, but $N_r(M) \approx \sqrt{3\pi L/2}$ in any case!

  $\rightarrow$ Different curves *collapse* together when $N_r(M)$ is plotted ($\sim$ systems at criticality).
- Memory recall may follow some **universal** rules.

  $\rightarrow$ Indeed, the analytical model does not depend on the *details* of the similarity matrix (as long as it is symmetric, and $f \rightarrow 0$).
- This suggests a new **framework** to explore the link between biophysics and cognition.

  $\rightarrow$ Are all the hypotheses necessary?
  $\rightarrow$ Do similar relations apply also to different systems (e.g. ANN)?
  $\rightarrow$ Can trauma impact the recall scaling?

References I

[1] D. J. Murray, Carol Pye, and W. E. Hockley.
Standing's power function in long-term memory: Further research.
Psychological Research, 38(4):319–331, 1976.

[2] Rodrigo Quian Quiroga and Gabriel Kreiman.
Measuring sparseness in the brain: Comment on Bowers (2009).
Psychological Review, 117(1):291–297, 2010.

[3] Sandro Romani, Itai Pinkoviezky, Alon Rubin, and Misha Tsodyks.
Scaling laws of associative memory retrieval.
Neural Comput., 25(10):2523–2544, October 2013.

[4] M. Katkov, S. Romani, and M. Tsodyks.
Memory retrieval from first principles.
Neuron, 94(5):1027–1032, 2017.

[5] Michelangelo Naim, Mikhail Katkov, Sandro Romani, and Misha Tsodyks.
Fundamental law of memory recall.
bioRxiv, 2019.