# *Generalized Capacity Estimation of Hebbian Auto-associative Networks*

*CH*

*October 2, 2025*

Dense Associative Memory models dramatically expand the capacity of traditional Hopfield networks. This is demonstrated by initializing the network with random patterns, and estimating the signal to noise ratio of recalling a desired pattern. Using similar methods, we provide a general theory of auto-associative networks trained using Hebbian learning rules.

## *Introduction*

Dense Associative Memory (DAM) networks are content-addressable memory models which generalize Hopfield networks (Hopfield, 1982; Krotov & Hopfield, 2016). Let $K$ be the number of patterns that the network will store, and let $D$ be the dimensionality of the patterns. Further, $F_n(\cdot) = \frac{1}{n}(\cdot)^n$ be the *polynomial* function. The energy of the network, with current state buffer $\sigma$, and memories $\xi^\mu$, $\mu = 1, 2, \ldots, K$ is given by the following general formula:

$$E(\sigma) = -\sum_{\mu}^{K} F_n(\xi_i^\mu \sigma_i). \tag{1}$$

It is shown in Demircigil et al. (2017) and Krotov and Hopfield (2016) that capacity of the memory model, $K^{\max}$, scales exponentially as a function of $D$, in particular:

$$K^{\max} \propto D^{n-1}. \tag{2}$$

This substantially improves the capacity of traditional Hopfield networks, where $K^{\max} \approx 0.14D$ (Hopfield, 1982).

However, one thing lost with the general DAM model is the intuitive weight update rule of Hopfield networks. Recall that a Hopfield network contains $D$ units with full, lateral connections between each unit. Weighting these connections is a $D \times D$ matrix, $W$, with a vanishing diagonal, as the network has no recursive connections. Single-pass recall for the traditional network is:

$$\begin{aligned} \sigma^{(t+1)} &= W\sigma^{(t)} = \Xi^\top(\Xi x) \\ &= \xi^1(\xi^1 \cdot \sigma) + \xi^2(\xi^2 \cdot \sigma) + \cdots + \xi^K(\xi^K \cdot x), \end{aligned} \tag{3}$$

where $\Xi$ is the $K \times D$ pattern, or, design matrix. Naively reproducing this single-pass update rule can be achieved with the *activity product rule* (Haykin, 2009):

$$\Delta W = \xi^\mu \otimes \xi^\mu, \tag{4}$$

where $(\cdot \otimes \cdot)$ denotes the outer product. Informally, for each time step $t = 1, 2, \ldots, K$ we present a new pattern $\xi^{(\mu=t)}$ to the network, and update the weights by its self-correlation (hence, the name of *correlation matrix* memories (Kohonen, 1988)).

This simple weight update rule however opens a new door for potentially many different kinds of weight update rules. In particular, we will be investigating Hebbian weight update rules. A Hebbian learning rule is entirely local (Gerstner & Kistler, 2002), which means that they are biologically plausible.

## *Related Work*

(Hoover et al., 2024; Hu et al., 2024; Kozachkov et al., 2024; Salvatori et al., 2024).

## *Hebbian Learning Rules*

A *Hebbian* learning rule is a general characterization of local update rules:

**Definition 1** (Local layers; Hebbian Learning Rule)**.** For a neural network of layers $x_i$, $i = 1, 2, \ldots, K$ layers of dimensions $d_i$, layer $x_i$ is *local* to layer $x_j$ if $i = 1 \pm j$. A *Hebbian learning rule* is an update rule for the $d_i \times d_j$ weight matrix connection layers $x_i$ and $x_j$:

$$W^{(t+1)} = f(W^{(t)};\ x_i, x_j).$$

Gerstner and Kistler (2002) generalizes the Hebbian learning rule using a Taylor expansion of $f$ about $(0, 0)$:

**Definition 2** (Expanded Hebbian Learning Rule)**.** Expanding $f$ from Definition 1, we get:

$$\begin{aligned} f(W; x_i, x_j) = f(W; 0, 0) &+ \frac{\partial f}{\partial x_i}\Big|_{(0,0)} x_i + \frac{\partial f}{\partial x_j}\Big|_{(0,0)} x_j \\ &+ \frac{1}{2}\frac{\partial^2 f}{\partial v_i^2}\Big|_{(0,0)} x_i^2 + \frac{1}{2}\frac{\partial^2 f}{\partial v_j^2}\Big|_{(0,0)} x_j^2 \\ &+ \frac{\partial^2 f}{\partial x_i \partial x_j}\Big|_{(0,0)} x_i x_j + \mathcal{O}(v). \end{aligned}$$

More simply denoted, we can refer to the coefficients using $c_i$, getting us:

$$\begin{aligned} W^{(t+1)} = f(W^{(t)}; x_i, x_j) &= c_0(W^{(t)}) + c_1^{\text{pre}}(W^{(t)}) x_j + c_1^{\text{post}} x_i \\ &+ c_2^{\text{pre}}(W^{(t)}) x_j^2 + c_2^{\text{post}}(W^{(t)}) x_i^2 + c_2^{\text{corr}}(W^{(t)}) x_i x_j \\ &+ \mathcal{O}(x). \end{aligned}$$

**Example 1** (Activity Product Rule)**.** *The Activity Product Rule (Haykin, 2009),*

$$W^{(t+1)} = \eta x_i x_j^\top,$$

*where $\eta$ is the learning rate, is identical to Definition 2 with $c_2^{corr}$ set to some value, and all other coefficients set to 0:*

$$W^{(t+1)} = c_2^{corr}(W^{(t)}) x_i x_j^T.$$

## *References*


Demircigil, M., Heusel, J., Löwe, M., Upgang, S., & Vermet, F. (2017). On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, *168*(2), 288–299. https://doi.org/10.1007/s10955-017-1806-y

Gerstner, W., & Kistler, W. M. (2002). Mathematical formulations of hebbian learning. *Biological Cybernetics*, *87*(5), 404–415. https://doi.org/10.1007/s00422-002-0353-y

Haykin, S. S. (2009). *Neural networks and learning machines* (3. ed). Prentice-Hall.

Hoover, B., Chau, D. H., Strobelt, H., Ram, P., & Krotov, D. (2024). Dense associative memory through the lens of random features [Version Number: 1]. https://doi.org/10.48550/ARXIV.2410.24153

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*, 2554–2558.

Hu, J. Y.-C., Wu, D., & Liu, H. (2024, October 31). Provably optimal memory capacity for modern hopfield models: Transformer-compatible dense associative memories as spherical codes. https://doi.org/10.48550/arXiv.2410.23126

Kohonen, T. (1988). Correlation matrix memories. In J. A. Anderson & E. Rosenfeld (Eds.), *Neurocomputing, volume 1* (pp. 174–180). The MIT Press. https://doi.org/10.7551/mitpress/4943.003.0075

Kozachkov, L., Slotine, J.-J., & Krotov, D. (2024, July 23). Neuron-astrocyte associative memory. https://doi.org/10.48550/arXiv.2311.08135

Krotov, D., & Hopfield, J. J. (2016, September 27). Dense associative memory for pattern recognition. https://doi.org/10.48550/arXiv.1606.01164

Salvatori, T., Millidge, B., Song, Y., Bogacz, R., & Lukasiewicz, T. (2024, February 16). Associative memories in the feature space. https://doi.org/10.48550/arXiv.2402.10814