Reversing Hebbian Learning: From Connectivity to Stored Patterns

BAMS project / course insight - Mathematical Modelling for Machine Learning

Abstract.

In MMML, you have seen how a Hopfield network can store patterns through a symmetric Hebbian learning rule and retrieve them from incomplete or corrupted inputs. This project explores the reverse problem: given only the network's weight matrix, how much can we infer about the patterns used to construct it?

An idea for a starting point could be to focus on a single-pattern model and use spectral methods to estimate the stored pattern from the leading eigenvector of the weight matrix. Introducing noise leads naturally to the spiked Wigner model: a rank-one signal hidden in a random matrix. From this starting point, the project explores what information the weights preserve, when reconstruction is possible, and how knowledge of the patterns' structure can improve recovery. It also connects to statistical physics through energy-based descriptions, inference in disordered systems, and phase transitions in reconstruction performance.

Prerequisites / Difficulty.

- Familiarity with Hebbian learning and Hopfield networks from MMML.
- Linear algebra: eigenvalues, eigenvectors, matrix rank and matrix factorisation.
- Basic probability and Bayes' theorem.
- Basic Python and NumPy for computational work.

Intermediate, with room for more advanced exploration. The spectral starting point builds on familiar linear algebra. Bayesian reconstruction and approximate message passing can be explored progressively, without assuming prior knowledge of these methods.

Starting materials.

Main reference. Goldt, Krzakala, Zdeborová and Brunel (2023). Bayesian reconstruction of memories stored in neural networks from their connectivity. Start with the introduction, the problem formulation and the subsection on reconstructing a single binary pattern.

A refresher on Hebbian learning. lecture notes on Energy-Based Models and Recurrent Neural Networks, for a refresh on the Hebbian learning concepts used in the project.

Spectral methods. Antoine Maillard (2026). Lecture notes on statistical and computational phase transitions in high-dimensional statistics. Read §2.5.1 (pp. 19–21) for the spiked Wigner model; §§3.1–3.3.1 (pp. 24–29) for the noise spectrum, the outlier and the recovery transition. For the eigenvector-correlation proof, see §3.3.3 (pp. 31–32).

Possible directions.

- **What can be recovered in principle?** Move from one pattern to several and examine why recovering the subspace they span does not necessarily identify the individual patterns. Investigate which ambiguities remain when the patterns are known to be binary or sparse.

- **Reconstruction using eigenvectors.** Implement a spectral estimator and study how its alignment with the original pattern changes with noise and network size. Explore the transition between informative and uninformative estimates, and the limitations of eigenvector-based reconstruction when several patterns are stored.

- **Using prior knowledge.** Formulate reconstruction as a Bayesian inference problem and investigate how assumptions about the patterns help constrain the answer. For small networks, enumerate candidate binary patterns to find the most probable reconstruction, and compare it with spectral estimates or local refinement.

- **Approximate inference and the reference paper.** Explore approximate message passing as an iterative approach to Bayesian reconstruction. Study how the reference paper applies it to memory recovery, and investigate selected results through numerical experiments or the authors' implementation.

- **Beyond the standard Hebbian model.** Investigate how the inverse problem changes under different learning rules, correlated patterns or other network architectures. Consider whether the methods can be adapted to asymmetric connectivity, for example by using singular vectors, and identify which assumptions and recovery guarantees need to be reconsidered.

Expected outcome.

The project offers a way to connect associative memory with inverse problems, random matrix theory, Bayesian inference and statistical physics. It can develop an understanding of what a network's connectivity reveals about its stored patterns, how noise and prior information affect reconstruction, and why an effective method in one setting may fail in another.

A computational investigation could establish when different estimators succeed and how they compare, while a reading-based exploration could clarify the ideas behind the reference paper. An advanced version could compare two methods or reproduce one result from the paper. A more theoretical version could focus on when reconstruction is uniquely possible.

Contacts.

For questions, ideas or further discussion, contact Arturo Zecchina:
arturo.zecchina@studbocconi.it
Phone / WhatsApp: +39 333 494 9202.