Systematic Explorations of Measures

Juan S. Rojas

17 de mayo de 2025

This document focuses on organize the trials of measures behing the simplest dynamical layer until now and HFM model. Each trial will be given by an equation of the measure, assumptions, result, Limitations and why we failed.

Global Objective: Estimate the self-awareness in a machine that learns, it must be a layer that returns a value about *how much the machine knows about itself.*

Framework

1. Hierarchical Feature Model (HFM)

The HFM defines a probability distribution over binary feature vectors $s \in \{0,1\}^n$ based on their hierarchical resolution m_s, interpreted as the index of the most complex feature present in s. The energy function $\mathcal{H}(s)$ penalizes high-resolution configurations, with a cost scaled by the coupling parameter g.

$$p(s) = \frac{1}{Z(g)}\, e^{-g\,\mathcal{H}(s)}, \qquad \mathcal{H}(s) = \text{máx}(m_s - 1, 0), \quad m_s = \text{máx}\{i \mid s_i = 1\}$$

This model satisfies the *Principle of Maximal Relevance*: Lower-resolution states are favored under high coupling ($g \gg g_c$), while lower g allows exploration of richer, high-resolution feature sets. The critical coupling $g_c = \log 2$ separates phases.

2. Layer Dynamics Model

To interpret how hierarchical states s_t are processed, we define a layered system where each layer j integrates its input over time. Each w^j is a fixed weights w^j are drawn from $\mathcal{N}(0,1)$ independently, and $\sigma(\cdot)$ is a nonlinear activation function (e.g., sigmoid, perceptron).

$$a_t^j = (1-\epsilon)a_{t-1}^j + \epsilon \cdot \sigma\left(\sum_i w_i^j \cdot s_i(t)\right)$$

The update rule for each layer is a leaky integrator (exponential moving average), where the parameter $\epsilon \in [0,1]$ controls memory:

- $\epsilon \approx 1$: fast adaptation, little memory.
- $\epsilon \ll 1$: slow dynamics, strong temporal smoothing.

The combination of the HFM (as the data source) and the Layer Model (as the observer) provides a setup to investigate the relationship between internal feature organization and emergent learning behavior.

3. Measure 1: Projection onto Principal Component

This measure evaluates how strongly the system's state at time a_t aligns with its *dominant mode of variation across layers.*

$$f(t) = |v_1 \cdot a_t|$$

- $a_t = (a_t^1, ..., a_t^J)$ is the vector of activations across all J layers at time t
- v_1 is the first principal component (leading eigenvector of the covariance matrix of $\{a_t\}$)
- $f(t)$ quantifies the projection of the system onto its collective direction of maximal variance

Interpretation: A high value of $f(t)$ indicates that the system's activation state is well-aligned with the dominant internal structure. Low values suggest misalignment or dispersion. This measure captures the degree of organization along the principal axis of collective behavior.

4. **Measure 2: Static dispersion metric (variance among activations)**
5. **Measure 3: Bin collisions**
6. **Measure 4: Heterogeneity in activation patterns**
7. **Measure 5: Product of (3) and (4) as a hybrid score**