Systematic Explorations of Measures

Juan S. Rojas

18 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

To deep info see this tutorial. From a set of layers $a_t \in \mathbb{R}^J$ (where J is the number of layers) for each instant $t \in \{1, \ldots, T\}$.

Our total data matrix applying the transpose (T) is:

$$A = \begin{bmatrix} - & a_1^\top & - \\ - & a_2^\top & - \\ & \vdots & \\ - & a_T^\top & - \end{bmatrix} \in \mathbb{R}^{T \times J}$$

Let's analyze **how the system activity varies over time** $\Rightarrow$ Then we do PCA on the rows of A. The layers can help us to increases the dimensionality of the data and fix the time.

1. **Center the data:** $\tilde{A} = A - \bar{A}, \quad \bar{A} = \frac{1}{T}\sum_{t=1}^{T} a_t$
2. **Covariance matrix:** $C = \frac{1}{T}\tilde{A}^\top \tilde{A} \in \mathbb{R}^{J \times J}$

   This matrix tells us how *co-varies* the layers over time.

4. Measure 2: Layer–Sampling Alignment

This measure captures how well the observed differences between layers' activations align with the differences induced by the instantaneous stimulus at a fixed time t.

Specifically, we define the Layer–Sampling Alignment score $Y(t)$ as:

$$Y(t) = \sum_{j<j'} \left(a_t^{(j)} - a_t^{(j')}\right) \cdot \left(\sigma\left(\mathbf{w}^{(j)\top}\mathbf{s}_t\right) - \sigma\left(\mathbf{w}^{(j')\top}\mathbf{s}_t\right)\right)$$

Note that when $Y(t)$ is large and positive, the relative differences between layer outputs are aligned with differences in their respective inputs.

On the simulation, the measure grows with g, indicating a progressive alignment between the stimulus and the layer 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