A Formal Approach to Modeling the Cost of Cognitive Control

Biswadip Dey
Princeton University
Princeton, New Jersey, USA

joint work with:

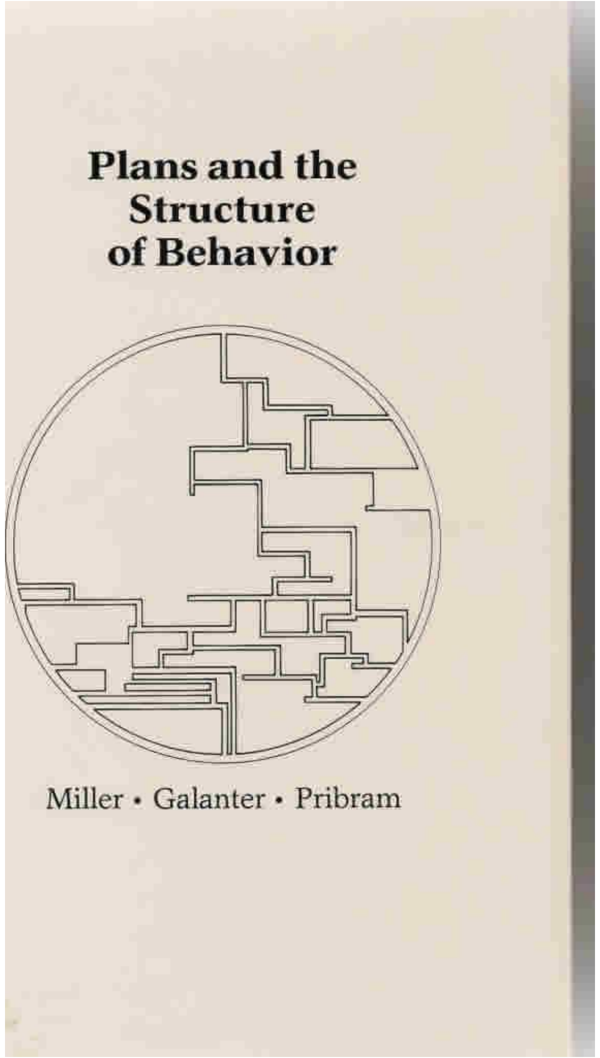
Kayhan Ozcimder (Princeton University, USA)
Sebastian Musslick (Princeton University, USA)
Giovanni Petri (ISI Foundation, Italy)
Nesreen K. Ahmed (Intel Corporation, USA)
Theodore L. Willke (Intel Corporation, USA)
Jonathan D. Cohen (Princeton University, USA)

PRINCETON Neuroscience INSTITUTE

ISI
ISI Foundation
& ISI Global Science Foundation

39th Annual Meeting of the Cognitive Science Society (CogSci) – July 29, 2017, London, UK

Motivation and Background



[New York : Holt, Rinehart and Winston, 1960]

- Cognitive control is broadly defined as the set of mechanisms required to pursue a goal.

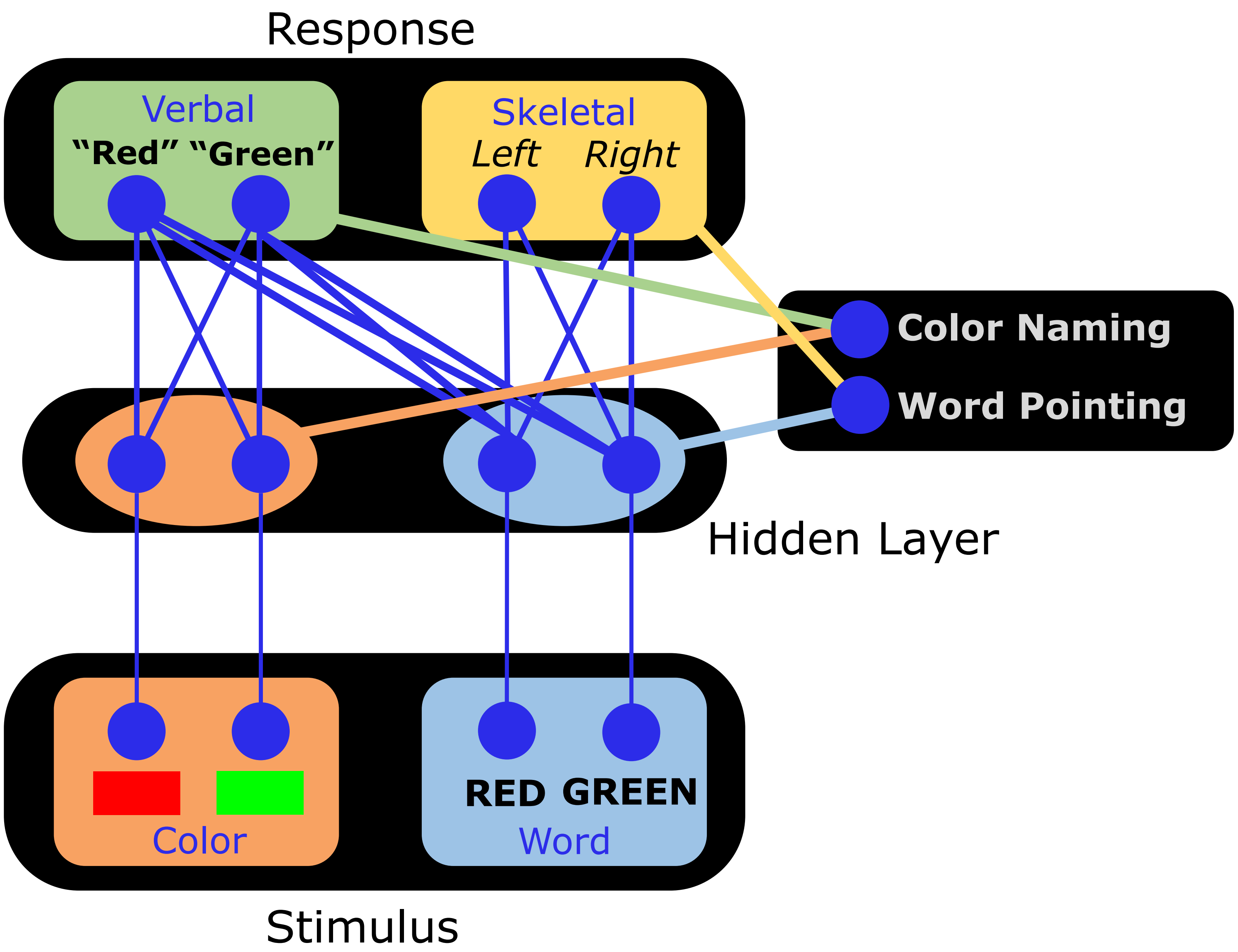
- Control and information theoretic approaches towards cognitive control can potentially lead to an AI that can mimic human cognition.

- Related Literature:
  - Posner and Snyder (1975). *Attention and cognitive control*.
  - Shiffrin and Schneider (1977). *Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory*.
  - Shenhav, Botvinick and Cohen (2013). *The expected value of control: an integrative theory of anterior cingulate cortex function*.
  - Botvinick and Cohen (2014). *The computational and neural basis of cognitive control: Charted territory and new frontiers*.

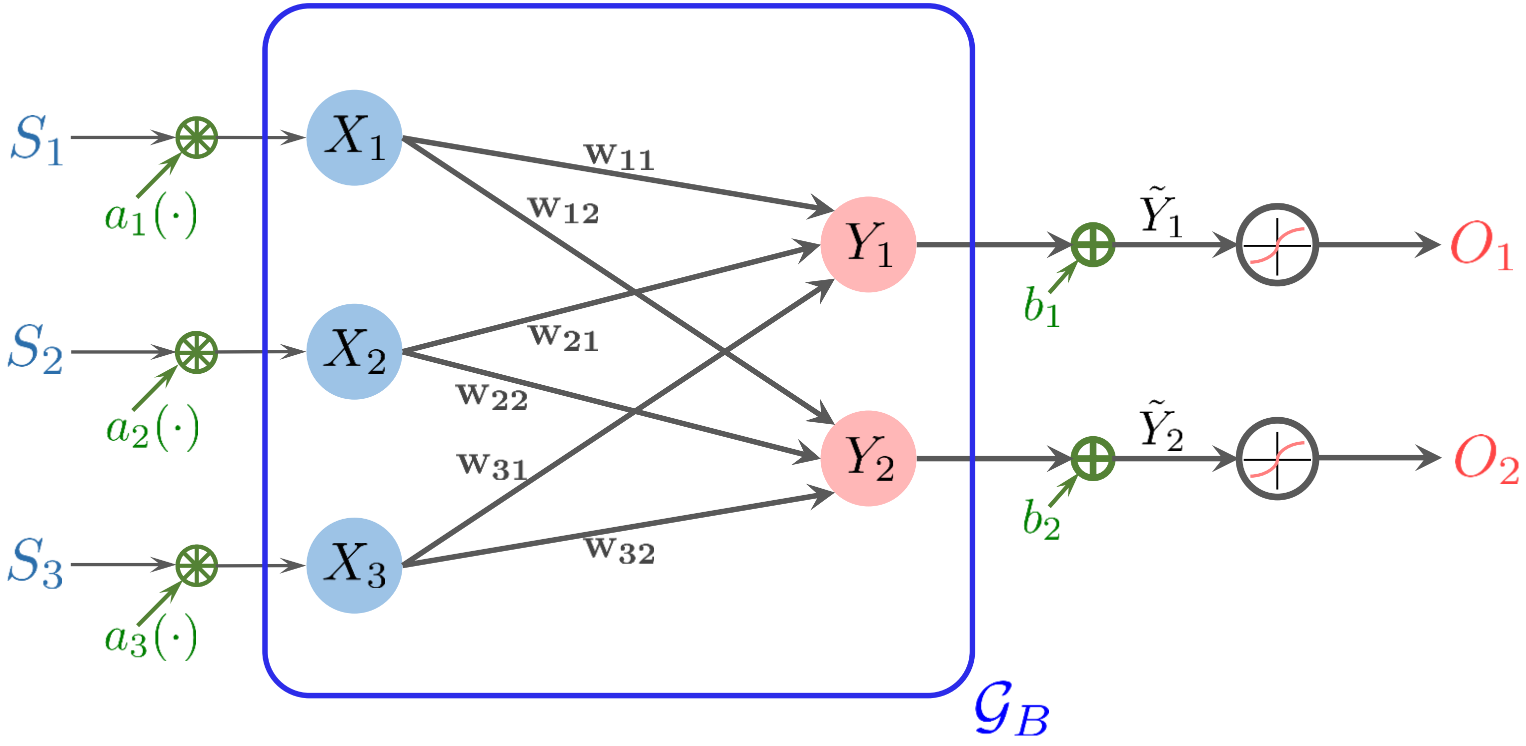
Outline

- Exact Model and Intensity Cost
  - Additional control to get a desired response
- An Abstraction and Interaction Cost
  - Captures the level of interference between the tasks/processes
- Neural Network Simulation
  - Interaction cost captures essential aspects of task performance

Role of Control in Extended Stroop Setting



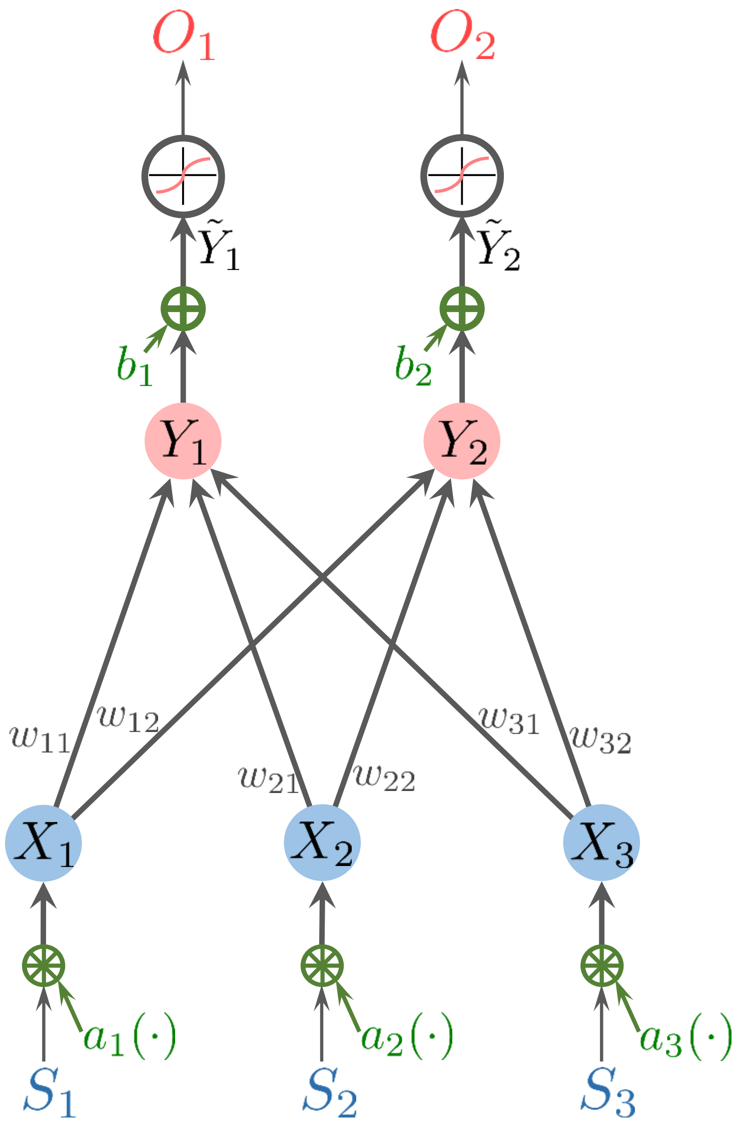
Abstraction as a 2-layer Neural Network



- Pre-interaction (or Hidden Layer) Bias:
  $X_i = a_i(S_i) = a_i^m S_i + a_i^a \mathbf{1}_{n_i}$

- Post-interaction (or Output) Bias:
  $\tilde{Y}_j = Y_j + b_j \mathbf{1}_{l_j}$

Likelihood of a Desired Response



- Logistic nonlinearity via $\tilde{Y}_i \to O_i$
  - O_i has a logit-normal distribution

- The response O_i should overcome a specified threshold in order to execute the corresponding task (process) [Shenhav et. al. (2013)]
  - Activation Threshold: $\alpha_i \in (0,1)$

- This allows us to compute the probability of a response being active in terms of network parameters and prior distribution.

$$\mathbb{P}[O_i \geq \alpha_i] = \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{\log\left(\frac{\alpha_i}{1-\alpha_i}\right) - b_i - \sum_{j=1}^{N} w_{ji}(a_j^m \mu_j + a_j^a)}{\sqrt{2\sum_{j=1}^{N}\sum_{k=1}^{N} a_j^m a_k^m w_{ji} w_{ki} \sigma_{kj}}}\right)$$

Performance Optimization

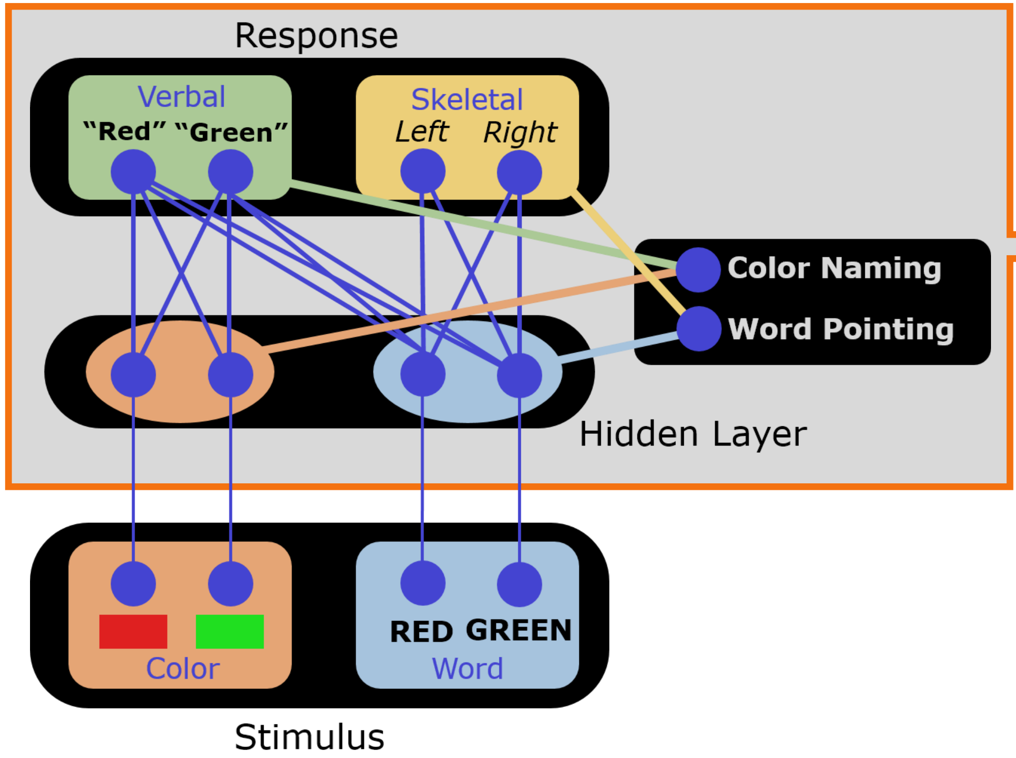
$$\mathbb{P}[O_i \geq \alpha_i] = \frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{\log\left(\frac{\alpha_i}{1-\alpha_i}\right) - b_i - \sum_{j=1}^{N} w_{ji}(a_j^m \mu_j + a_j^a)}{\sqrt{2\sum_{j=1}^{N}\sum_{k=1}^{N} a_j^m a_k^m w_{ji} w_{ki} \sigma_{kj}}}\right)$$

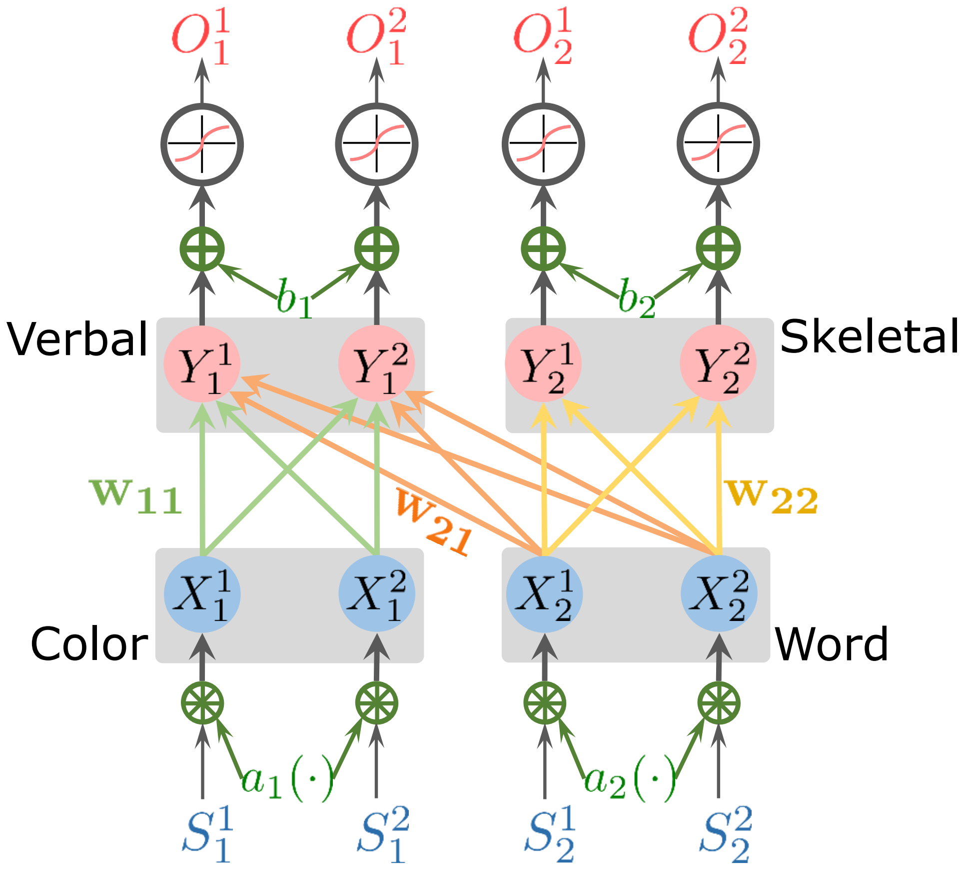
$$\underset{\substack{a^a, a^m \in \mathbb{R}^N \\ b \in \mathbb{R}^L}}{\text{Minimize}} \quad \sum_{i=1}^{N} (a_i^a)^2 + \sum_{i=1}^{N} (a_i^m)^2 + \sum_{j=1}^{L} b_l^2$$

$$\text{subject to:} \quad \mathbb{P}[O_k \geq \alpha_k] \geq \tau_k$$

◇ This optimization minimizes the intensity cost of cognitive control for a desired probability of activation of the response.

Revisiting the Stroop task

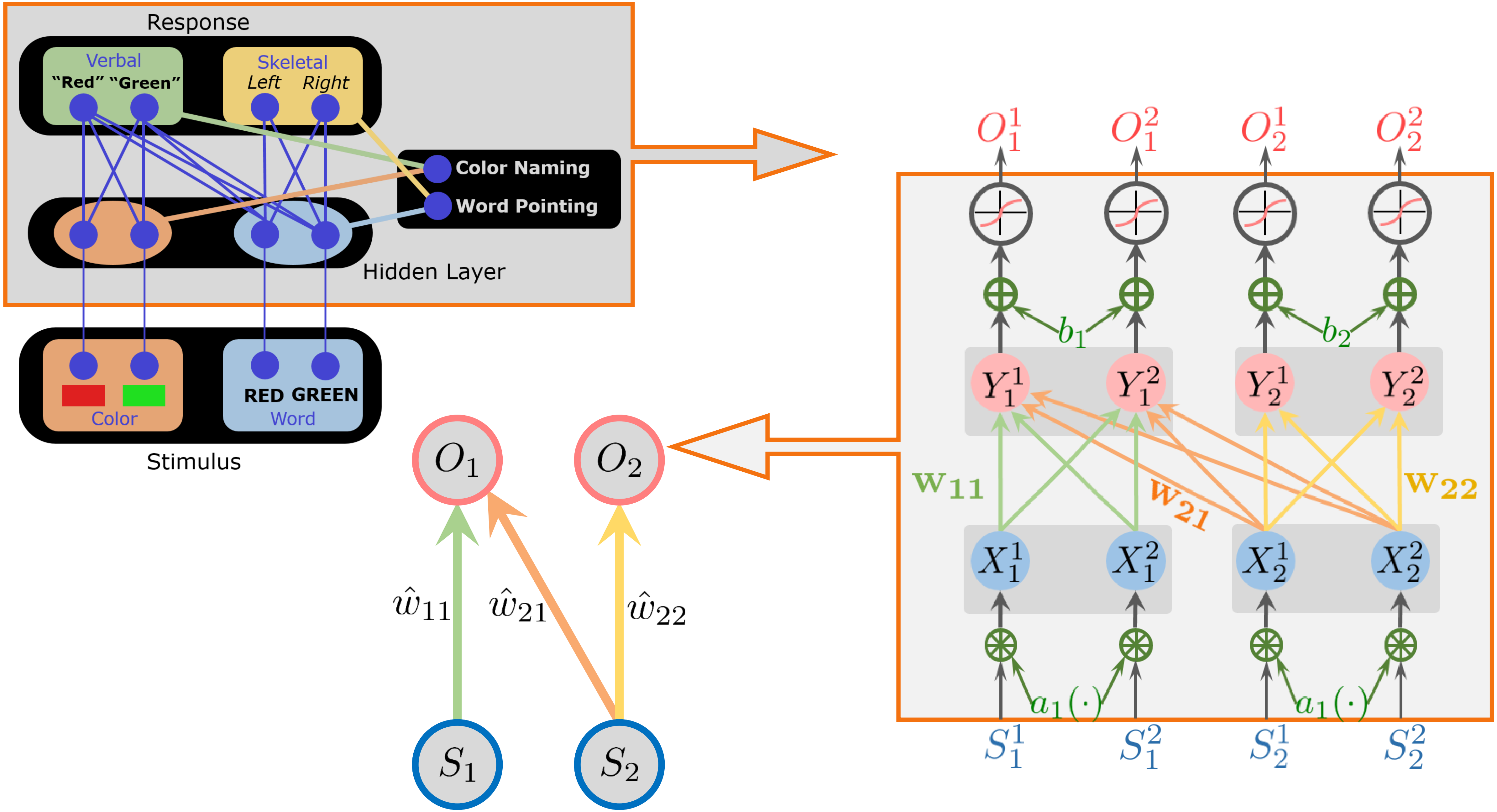




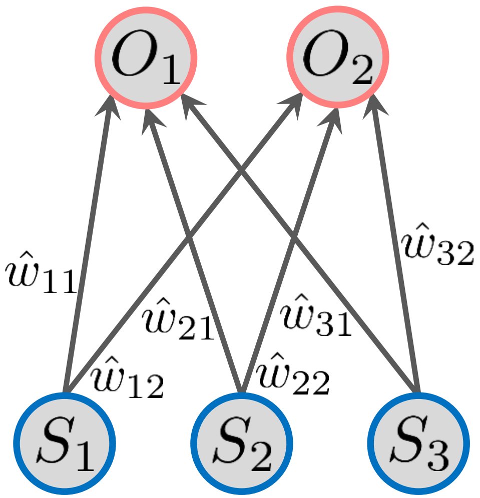
$$\mathbb{P}[O_1^1 \geq \alpha] = \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{log(\frac{\alpha}{1-\alpha}) - b_1 - w_{11}(a_1^m\mu_1 + a_1^a) - w_{21}(a_1^m\mu_2 + a_1^a) - w_{31}(a_2^m\mu_3 + a_2^a) + w_{41}(a_2^m\mu_4 + a_2^a)}{\sqrt{2\sum_{j=1}^{N}\sum_{k=1}^{N} a_j^m a_k^m w_{j1} w_{k1}\sigma_{kj}}}\right)$$

$$\mathbb{P}[O_2^1 < \gamma] = 1 + \frac{1}{2}\text{erf}\left(\frac{log(\frac{\gamma}{1-\gamma}) - b_1 - w_{12}(a_1^m\mu_1 + a_1^a) - w_{22}(a_1^m\mu_2 + a_1^a) - w_{32}(a_2^m\mu_3 + a_2^a) + w_{42}(a_2^m\mu_4 + a_2^a)}{\sqrt{2\sum_{j=1}^{N}\sum_{k=1}^{N} a_j^m a_k^m w_{j2} w_{k2}\sigma_{kj}}}\right)$$

Further Abstraction Leads to Interaction Cost



Interaction Cost



$$T_j = \begin{cases} 1 & \text{Output } O_j \text{ responds to stimulus } S_1 \\ 2 & \text{Output } O_j \text{ responds to stimulus } S_2 \\ \vdots & \\ N & \text{Output } O_j \text{ responds to stimulus } S_N \\ 0 & \text{Output } O_j \text{ does not respond at all} \end{cases}$$

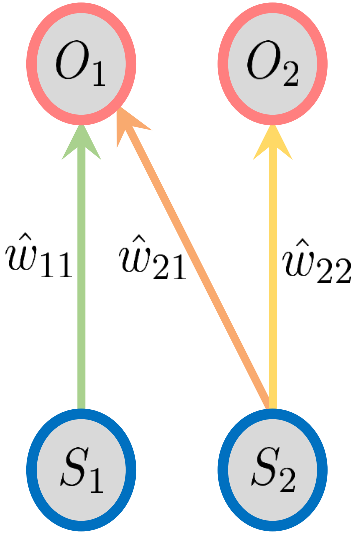
$$\mathbb{P}[T_j = i] = \frac{e^{\hat{w}_{kj}} \mathbb{1}(S_i)}{M + \sum_{k=1}^{N} e^{\hat{w}_{kj}} \mathbb{1}(S_k)}$$

◇ This is an indicator function which represents whether a particular stimulus is active or not.

Interaction Cost:

$$\Psi(T_j = i) = -\log\left(\mathbb{P}[T_j = i]\right)$$

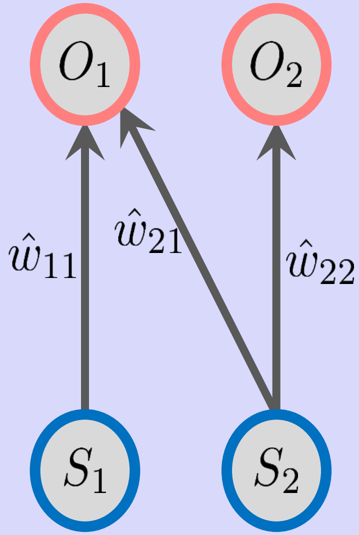
Interaction Cost – Some Case Studies



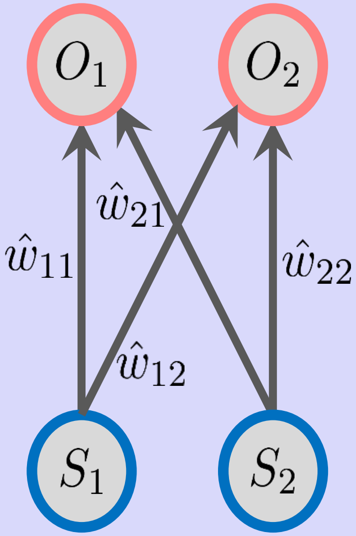
$$\mathbb{P}[T_1 = 1] = \frac{e^{\hat{w}_{11}}}{M + e^{\hat{w}_{11}} + e^{\hat{w}_{21}}}$$

$$\mathbb{P}[T_1 = 2] = \frac{e^{\hat{w}_{21}}}{M + e^{\hat{w}_{11}} + e^{\hat{w}_{21}}}$$

$$\mathbb{P}[T_1 = 0] = \frac{M}{M + e^{\hat{w}_{11}} + e^{\hat{w}_{21}}}$$

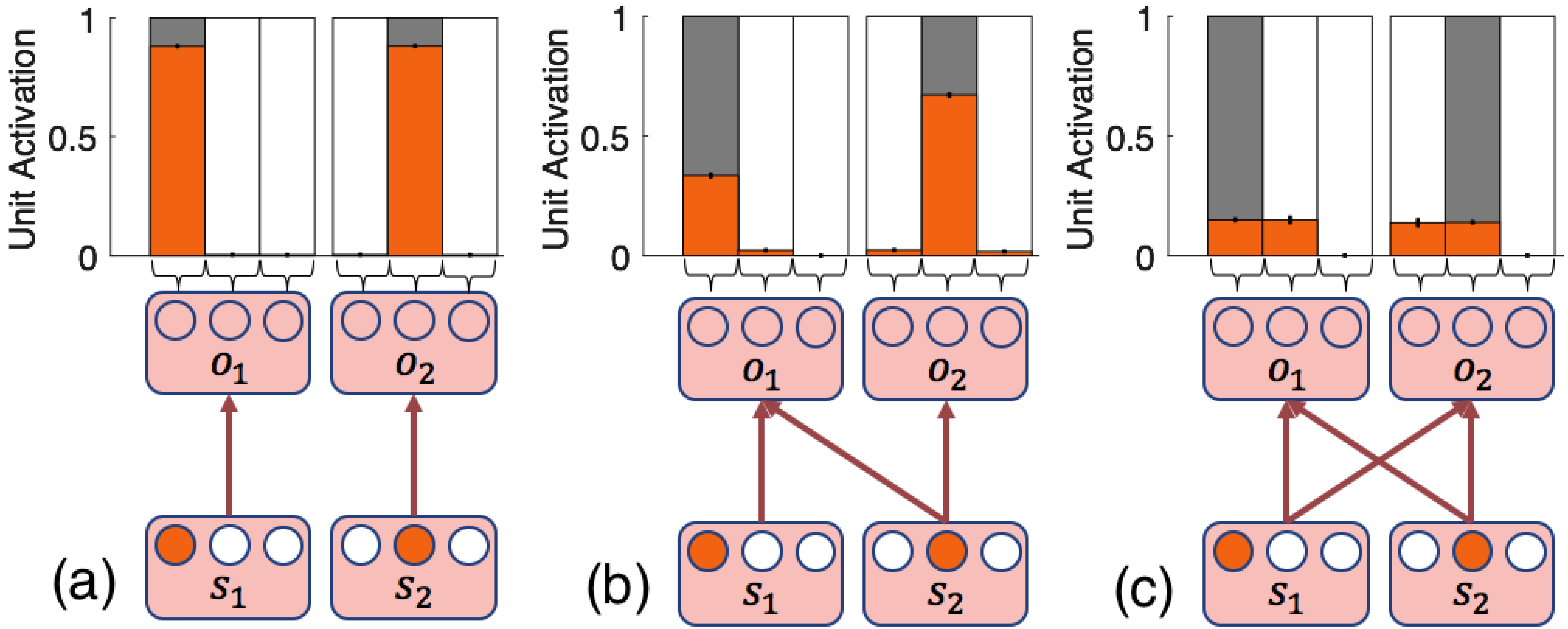


$$\mathbb{P}[T_1 = 1, T_2 = 2] = \frac{e^{\hat{w}_{11}}}{M + e^{\hat{w}_{11}} + e^{\hat{w}_{21}}} \cdot \frac{e^{\hat{w}_{22}}}{M + e^{\hat{w}_{22}}}$$



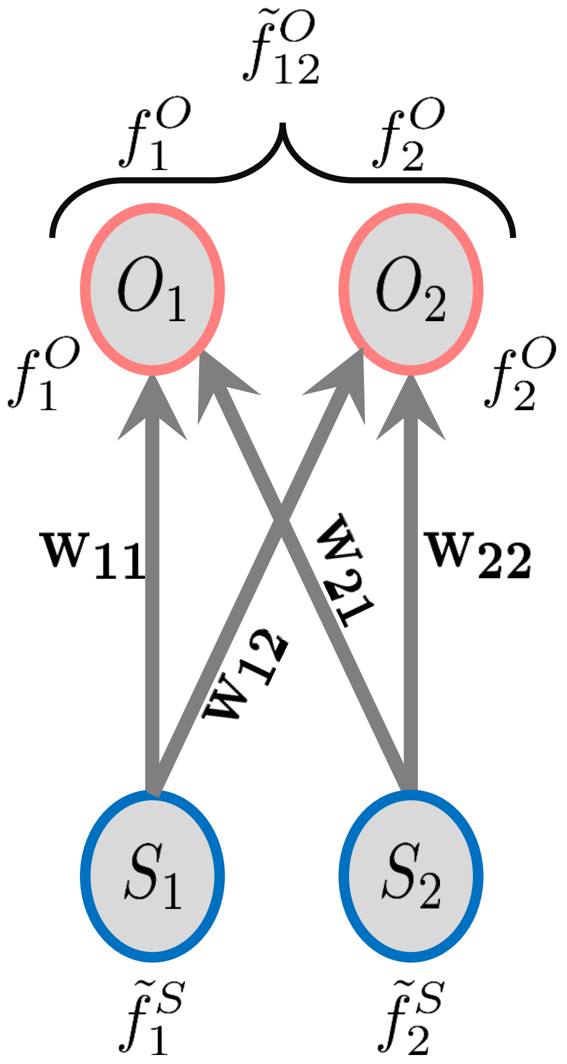
$$\mathbb{P}[T_1 = 1, T_2 = 2] = \frac{e^{\hat{w}_{11}}}{M + e^{\hat{w}_{11}} + e^{\hat{w}_{21}}} \cdot \frac{e^{\hat{w}_{22}}}{M + e^{\hat{w}_{12}} + e^{\hat{w}_{22}}}$$

Results from Network Simulation



$$\Psi(T_1 = 1, T_2 = 2)|_a \quad < \quad \Psi(T_1 = 1, T_2 = 2)|_b \quad < \quad \Psi(T_1 = 1, T_2 = 2)|_c$$

Future Directions – A Unified Approach



- Joint distribution of responses in absence of interference

  $f^O_1 f^O_2$

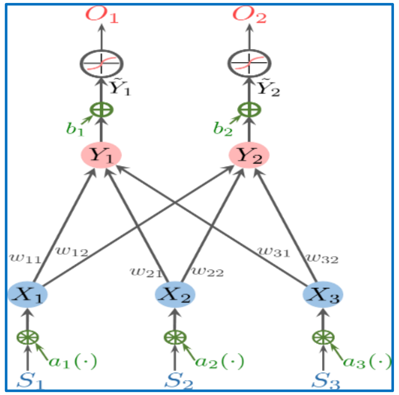
- Joint distribution of responses in presence of interference

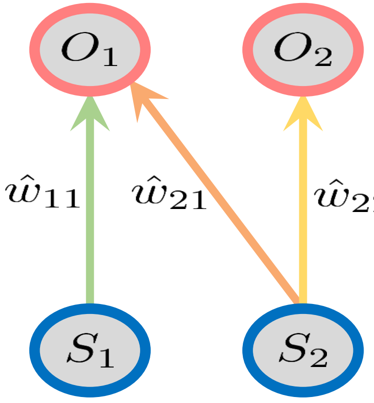
  $\tilde{f}^O_{12}$

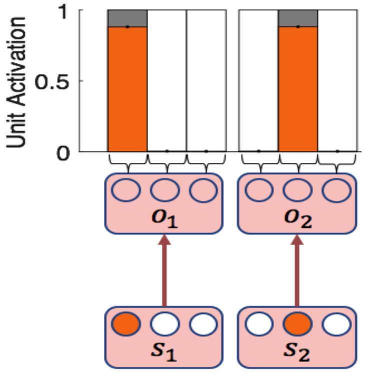
- An appropriate notion of distance between these two distributions (*joint* and the *product of the marginals*) can be used to measure the amount of dependency within a group of tasks

  $D_{KL}(\tilde{f}^O_{12} || f^O_1 f^O_2)$

Take-Home







- Intensity cost captures how much additional information is required so as to get the desired response.
- Interaction cost measures the level of interference between processes by means of their type of connections & weights.
- Simulations demonstrate the influence of directionality in interference between tasks.

Acknowledgements

- Supported by:
  - John Templeton Foundation
  - Intel Corporation

- Thanks to:
  - Z. Aminzare (Princeton University)
  - J. Pillow (Princeton University)

Thank You!