UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

A Formal Approach to Modeling the Cost of Cognitive Control

Permalink

https://escholarship.org/uc/item/6sb336sz

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 39(0)

Authors

Ozcimder, Kayhan
Dey, Biswadip
Musslick, Sebastian
et al.

Publication Date

2017

Peer reviewed

A Formal Approach to Modeling the Cost of Cognitive Control

Kayhan Özcimder [1,2,*], Biswadip Dey[2,*], Sebastian Musslick[1,*],
Giovanni Petri[3], Nesreen K. Ahmed[4], Theodore L. Willke[4], Jonathan D. Cohen[1]

[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08540, USA.
[2]Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544, USA.
[3] ISI Foundation, Via Alassio 11/c, 10126 Torino, Italy.
[4]Intel Labs, Santa Clara, CA 95054, USA.
*Equal Contribution, Corresponding Author: *ozcimder@princeton.edu*

Abstract

This paper introduces a formal method to model the level of demand on control when executing cognitive processes. The cost of cognitive control is parsed into an *intensity cost* which encapsulates how much additional input information is required so as to get the specified response, and an *interaction cost* which encapsulates the level of interference between individual processes in a network. We develop a formal relationship between the probability of successful execution of desired processes and the control signals (additive control biases). This relationship is also used to specify optimal control policies to achieve a desired probability of activation for processes. We observe that there are boundary cases when finding such control policies which leads us to introduce the interaction cost. We show that the interaction cost is influenced by the relative strengths of individual processes, as well as the directionality of the underlying competition between processes.

Keywords: cognitive control; multi-tasking; intensity; identity

Introduction

A long standing focus in cognitive research has been towards understanding the ability to execute tasks/processes[1] that demand *cognitive control*. In this context, cognitive control is defined as the set of mechanisms required to pursue a goal, especially when distraction or strong competing responses (interferences) must be overcome (Posner & Snyder, 1975; Shiffrin & Schneider, 1977; Cohen, Dunbar, & McClelland, 1990). Earlier work (Posner & Snyder, 1975; Shiffrin & Schneider, 1977; Cohen et al., 1990; Botvinick & Cohen, 2014) has argued that the processes demanding control can be distinguished from *automatic processes* in terms of the strength of the associations in the pathways underlying processing: automatic processes are characterized by pathways with associations strong enough to resist interference from competing processes, whereas controlled processes are weaker, and therefore rely on input from control mechanisms to support their execution against interference.

Another longstanding observation is that the allocation of cognitive control is costly (often discussed in terms of "mental effort" (Posner & Snyder, 1975; Botvinick & Braver, 2015; Shenhav et al., 2017)). This cost has been interpreted in physical terms (such as metabolic demands (Muraven, Tice,

& Baumeister, 1998)) or in terms of an opportunity cost reflecting the allocation of a limited resource (Kurzban, Duckworth, Kable, & Myers, 2013). Elsewhere (Feng, Schwemmer, Gershman, & Cohen, 2014; Musslick et al., 2016), we have proposed that limitations in the capacity for control-dependent processing reflect the *purpose* of control to diminish interference rather than any intrinsic limitation in the mechanism responsible for control. This view suggests that the architecture of the processing system as a whole constrains the opportunities for control-dependent processing, resulting in opportunity costs associated with allocating control to any particular task(s).

Here, we build on a closely related proposal, by Koechlin and Summerfield (2007), to define the cost of control in terms of internal representational requirements to insure that a given stimulus (or a set of stimuli) produces the desired response (or a set of responses), given the intrinsic architecture of the system. Their work focused on a single task. Here, we extend this to consider an arbitrary number of tasks and thus accommodate their possibility for, and costs of, multitasking (i.e.parallel processing of task pathways). To do so, we follow the framework proposed by Shenhav, Botvinick, and Cohen (2013) that distinguishes two components of control signals: *intensity* and *identity*. Specifically, Shenhav et al. (2013) defined the intensity of a control signal as the strength of the signal needed to insure performance of a particular task, and the identity as which control signal should be selected to achieve a desired objective given environmental conditions. Here we build on that distinction to define two corresponding components of the cognitive control costs - a cost associated with intensity, and a cost associated with interaction. Furthermore, we define the interaction cost to capture the level of interference between the processes in a network.

In this paper, we begin by introducing formal constructs for intensity and interaction costs by using the graph theoretic representation of a neural network and terms/notions adopted from probability theory. We describe an intensity cost that represents the control signals (as biases infused into a neural network), above and beyond the specified strength of the signal (stimulus) itself. This is achieved by developing a formal relationship between the probability of successful execution of desired processes and the control signals. In turn, this defines an optimization problem, which can then be solved to find optimal control signals that achieve a specified activation for desired processes. However, we observe that there

---

[1]A task/process/input-output mapping is defined as a unique mapping from all possible vectors in the input subspace to corresponding vectors in the output subspace, that is independent of the mappings for all other combinations of input and output components in the network.
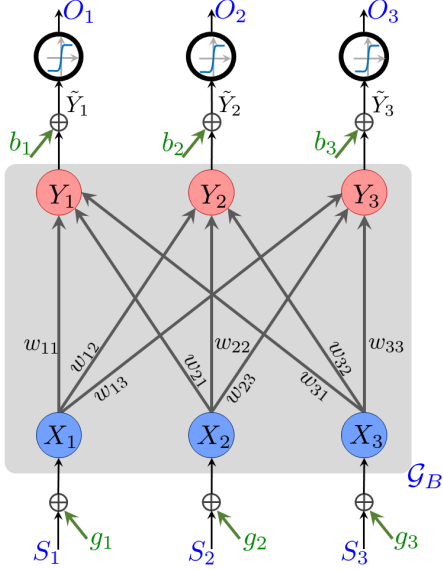
Figure 1: Illustration of a single-layered, feed-forward network with 3-input and 3-output layer components, wherein the individual features are scalar in nature.

are boundary cases in which this optimization problem can not be solved. These cases reflect situations in which the simultaneous execution of the processes is not feasible due to interference. Hence, interaction cost analysis motivates an additional investigation towards finding a proper metric that continuously measures the level of interference between processes. To achieve this we introduce the definition of interaction cost associated with process mappings in a network configuration. Specifically, it measures the level of interference introduced by competing processes that interfere with the tasks of interest. In their study, Koechlin and Summerfield (2007) have already used information theoretic terms to measure cognitive control. However, in order to apply these metrics to neural networks while considering parallelism, we augment some of these measures, and through simulations we will demonstrate how interaction cost can be used to predict interference in neural network architectures. Finally, we will discuss general research directions revealed by the analysis presented here.

Intensity: the cost of control

Cognitive control is defined as the underlying mechanism that biases the processing of a task in order to maximize the reward (Botvinick, Braver, Barch, Carter, & Cohen, 2001; Botvinick, Cohen, & Carter, 2004; Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Botvinick, 2007). Here, we adapt the notion of intensity cost from Shenhav et al. (2013) as a function of the amount of control bias that cognitive control applies to the system. However, Shenhav et al. (2013) described this function in qualitative rather than quantitative terms. In this work we provide an explicit characterization of the cost of cognitive effort in terms of a set of physically meaningful parameters, which allow the manipulation of the response of a cognitive architecture.

Following earlier works (Feng et al., 2014; Musslick et al.,

2016), we consider a single-layered, feed-forward network with N input and M output layer components to formalize the notion of intensity cost in a cognitive control context (Fig. 1 shows a simple example of such a network). In this framework, each component represents an input/stimulus or output/response dimension (vector subspace), and the connection from an input to an output component constitutes the processing pathway for a given task. This allows us to define an abstraction of the network as a directed bipartite graph $\mathcal{G}_B = (\mathcal{V}, \mathcal{E})$, wherein the set of vertices $\mathcal{V}$ can be partitioned into two disjoint sets $\mathcal{V}_{in}$ and $\mathcal{V}_{out}$, representing the input and output layer components respectively. Moreover, a directed edge $(i, j) \in \mathcal{E} \subseteq \mathcal{V}_{in} \times \mathcal{V}_{out}$ represents a connection from the vertex i in the input layer to vertex j in the output layer (i.e., a task). In this setting, we represent the processing pathway by introducing a weight matrix W with elements w_{ij}. As we will see later, this abstraction plays an important role in formalizing the interaction cost of cognitive control.

In this setting, we assume that control signals bias the processing of a stimulus towards a specified response at two different levels, i.e. g_i and b_j, which we refer to as pre-interaction and post-interaction control biases, respectively. This complies with early computational models of cognitive control in which control signals act as an increase in gain of non-linear processing units (Cohen et al., 1990; Botvinick et al., 2001) and allows us to treat such control biases as key contributing factors towards the intensity cost for cognitive control. It is worth noting here that for simplicity the only sources of nonlinearity in this setting are the logistic[2] activation functions which act upon the linearized output vector $\tilde{\mathbf{y}} = [\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_M]$. Without loss of generality, in what follows we consider the individual features to be scalar, and carry out a formal investigation on how these control biases $\mathbf{g} = [g_1, \ldots, g_N]$ and $\mathbf{b} = [b_1, \ldots, b_M]$ influence the response from this cognitive architecture. In our formulation, the corresponding magnitude, i.e. $\|\mathbf{g}\|^2 + \|\mathbf{b}\|^2$, can be treated as a measure of control intensity applied to the system, and therefore the cost for cognitive control.

We begin our analysis by assuming the vector of features $\mathbf{s} = [S_1, S_2, \ldots, S_N]$ to be an N-dimensional multivariate Gaussian random variable with mean μ^S and covariance Σ^S. (The assumption of Gaussianity is motivated by the technical tractability). With this assumption, the vector $[X_1, X_2, \ldots, X_N]$ becomes an N-dimensional multivariate Gaussian random variable with a shifted mean and same covariance. Furthermore, as all the transformations (before the nonlinear logistic activation function) are linear in nature, $[Y_1, Y_2, \ldots, Y_N]$ also remains a multivariate Gaussian random variable whose mean and covariance are given by $\mu^Y = W(\mu^S + \mathbf{g})$ and $\Sigma^Y = W\Sigma^S W^T$, respectively. Similarly, the vector of linearized outputs $[\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_N]$ is also a multivariate Gaussian, with a shifted mean and the same covariance.

---

[2]Although we are restricting ourselves to logistic functions with unit steepness, in a more general setting one can use the steepness as another design parameter.

Hence, each individual linearized output $\tilde{Y}_i$ is itself a Gaussian random variable with

$$\text{mean:} \quad \mu_i^{\tilde{Y}} = \sum_{j=1}^{N} w_{ji}(\mu_j + g_j) + b_i$$

$$\text{and variance:} \quad \sigma_{ii}^{\tilde{Y}} = \sum_{j=1}^{N} \sum_{k=1}^{N} w_{ji}w_{ki}\sigma_{kj},$$

where μ_j is the mean of stimulus S_k and σ_{kj} is the covariance between stimuli S_k and S_j. As a consequence, the corresponding output (response) will have a logit-normal distribution, and this leads us to our key result in this section.

As outlined by Shenhav et al. (2013), the response O_i should overcome a specified threshold in order to execute the corresponding process (task). Then, by letting $\alpha_i \in (0,1)$ represent this activation threshold associated with output O_i, the corresponding probability of task execution (probability of the output O_i surpassing the threshold α_i) is expressed as

$$P[O_i \geq \alpha_i] = \frac{1}{2} - \frac{1}{2}\,\mathrm{erf}\underbrace{\left(\frac{\log\left(\frac{\alpha_i}{1-\alpha_i}\right) - b_i - \sum_{j=1}^{N} w_{ji}(\mu_j + g_j)}{\sqrt{2\sum_{j=1}^{N}\sum_{k=1}^{N} w_{ji}w_{ki}\sigma_{kj}}} \right)}_{f(\alpha_i, b_i, w, \mu, \mathbf{g}, \Sigma^S)}.$$

Here we have exploited the monotonicity of the logistic function to compute its inverse. Then the result follows from the cumulative distribution function of $\tilde{Y}_i$.

We characterized the activation probability of a given network in terms of the pre-interaction and post-interaction control biases. This is crucial because it provides new directions to incorporate the cost of control into the design of a cognitive network architecture. For example, the problem of allocating a limited amount of cognitive control into different components of the network to maximize the associated probability of activation can be formulated as an optimization problem in which the goal becomes minimizing $f(\alpha_i, b_i, w, \mu, \mathbf{g}, \Sigma^S)$ over $\mathbf{g}$ and $\mathbf{b}$ subject to the constraints $\sum_{i=1}^{N} g_i^2 \leq C_g$ and $\sum_{i=1}^{M} b_i^2 \leq C_b$, where C_g and C_b define the maximum amount of control that can be applied. Alternatively, in this setting, we can also approach the problem of minimizing the cost of control, while still maintaining a desired value of probability of activation.

One can consider the joint distribution of the processes of interest to incorporate the effects of interaction between tasks. To be consistent with (Feng et al., 2014; Musslick et al., 2016), it is reasonable to begin with a focus narrowed to the situation where the choice of interaction weights and the prior distribution of the stimuli render the interactions undesirable. Then the effect of multitasking can measured by introducing a suitable distance metric (for example, the Kullback-Leibler divergence (Ortega & Braun, 2013)) between the joint distributions of relevant processes and the product of corresponding marginals, and one can attempt to minimize this distance

at the expense of a limited amount of cognitive control. However, as one might expect, this optimization problem can have an empty solution set under certain values of the interaction weights and activation thresholds, meaning that certain network configurations strictly prohibit successful multitasking performance (Musslick et al., 2016). Before approaching this computation in detail, it would be beneficial to investigate how the interaction structure influences the solution space, and that leads us to our next section wherein we introduce the notion of interaction cost.

Interaction: the cost of mapping

In this section, we will introduce a detailed formalism of the interaction cost associated with process mappings in a network configuration to accommodate the possibility for multitasking. In our earlier work (Musslick et al., 2016), we have formalized three distinct types of interference (Fig. 2).

Convergent interference (Fig. 2a) occurs when two inputs/stimuli (e.g. S_1 and S_2) compete to determine a common output (e.g. O_1). We also consider *divergent interference* in our analysis (Fig. 2b). Although this does not pose an impediment to performance, i.e. it is possible to generate two distinct outputs (e.g. O_1 and O_2) to the same input (e.g. S_1), it represents a restriction on the number of independent stimuli (and therefore the number of tasks) that the system can process at once, and thus was treated formally as a type of interference due to this dependency in our analysis of parallel processing capability. Finally, we consider a third, *indirect interference* that supervenes on the first two (shown in Fig. 2c and Fig. 2d). In this case, the two tasks with strengths w_{11} and w_{22} in question do not directly interfere with one another. However, their simultaneous execution would necessarily engage a third task with strength w_{21} (also possibly a fourth task with strength w_{12}) that would produce interference in output O_1 (and O_2). While Musslick et al. (2016) treated these three types of interference identically in terms of their effect on the overall parallel processing capability of a network, the proposed interaction cost will also distinguish between these three types of interference.

In interaction cost analysis, we will assume that a stimulus is of value 1 when it is active, and 0 otherwise. Moreover, to increase tractability, we will consider linear activation at the output level, which also implies that without loss of generality the pre- and post-interaction biases can be assumed to be zero. A more detailed version of the interaction cost analysis, involving the strength of stimuli, as well as the nonlinear activation function, will be discussed in subsequent publications.

To introduce the interaction cost, we take an approach similar to the one adopted by Koechlin and Summerfield (2007). In their work, Koechlin and Summerfield (2007) proposed a metric for selecting a single action among multiple alternatives. Here, we will refine this metric to introduce the interaction cost for neural network architectures. Towards this objective, we first leverage the assumptions discussed earlier in
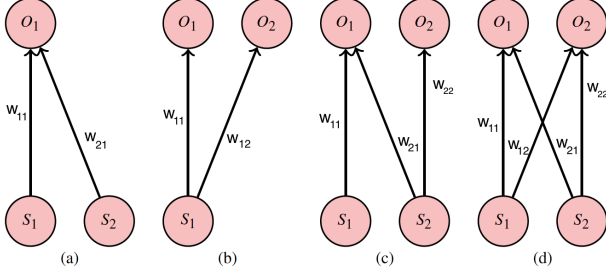
Figure 2: The illustration of convergent, divergent, asymmetric, and symmetric interference.

the section, and abstract out the network configurations presented in Fig. 2 from the network shown in Fig. 1. We also assume that the strength of a task ij from stimulus S_i to output O_j is represented by its non-negative weight $w_{ij} \geq 0$.

Let us first consider the case shown in Fig. 2a. It is obvious that the response in the output component O_j is completely determined by the stimulus if either S_1 or S_2 is activated in the network (executing a single process). However, activating both stimuli S_1 and S_2 simultaneously creates a conflict, since the output can not respond to two distinct stimuli simultaneously (as the activations are linear, the network will always have a response in the output level). In order to measure the level of this competition between stimuli, we define a random variable a_1 associated with the output O_1 such that $a_1 \in \{1, 2\}$ (Fig. 2a). This implies that $a_1 = 1$ or $a_1 = 2$ when the output O_1 is driven completely by S_1 or S_2, respectively. Since a stronger task will have a higher probability of being selected to generate the response, we consider the relative strengths of the task pathways (with associated strengths w_{11} and w_{21}) in order to define the probability of the possible outcomes, i.e. the probability of $a_1 = 1$ and $a_1 = 2$ when both S_1 and S_2 activated. Hence, we compute the probability as

$$P[a_1 = 1] = \frac{w_{11}}{w_{11} + w_{21}}, \quad \text{and,} \quad P[a_1 = 2] = \frac{w_{21}}{w_{11} + w_{21}}.$$

Next, we extend this framework to consider networks with N stimuli and M outputs, wherein an output O_j, $j \in \{1, \ldots, M\}$ responds to a set of stimuli. Let us assume that there are $n \leq N$ incoming edges to a particular output O_j, and each edge is originated from a distinct stimulus S_i, $i = i_1, \ldots, i_n$, where $i_k \in \{1, \ldots, N\}$. Then the probability of the event that output O_j is responding to stimulus S_i is given by,

$$P[a_j = i] = \frac{w_{ij} \mathbb{1}(S_i)}{\sum\limits_{k=1}^{n} w_{i_k j} \mathbb{1}(S_{i_k})}, \tag{1}$$

where $\mathbb{1}(S_i)$ is the indicator function that represents the activation of stimulus S_i such that $\mathbb{1}(S_i) = 1$ if stimulus S_i is active and $\mathbb{1}(S_i) = 0$, otherwise. Then, by building upon the ideas proposed by (Koechlin & Summerfield, 2007), we define the interaction cost as

$$\Psi(a_j = i) = -\log(P[a_j = i]), \tag{2}$$

where the logarithm is with respect to base 2.

Equation 1 implies the $P[a_j = i] = 1$ when only the relevant stimulus S_i associated with task ij is activated in the network. Hence the interaction cost is computed as $\Psi(a_j = i) = 0$ which implies that there is no interaction cost. Moreover, when multiple processes are competing due to the activation of multiple stimuli, $P[a_j = i] \to 0$ as the competition increases, and as a consequence the interaction cost $\Psi(a_j = i) \to \infty$.

We further extend equation 1, to encapsulate the probability associated with parallel processing of task pathways in the network. Thus, we introduce the joint probability of distinct output components responding to a set of stimuli. For instance, let us consider the parallel processing of tasks with strength w_{11} and w_{21} in Fig. 2a, and calculate $P[a_1 = 1, a_1 = 2]$. This is the probability of output component O_1 responding to both S_1 and S_2, and by definition we know that this probability is zero. For the case illustrated in Fig. 2b, the joint probability $P[a_1 = 1, a_2 = 1] = 1$ since activation of S_1 will activate both processes with strengths w_{11} and w_{12}, and there is no competition in outputs O_1 and O_2. This result is parallel to the observation made by (Musslick et al., 2016), who stated that divergent interference is not actually an interference but a dependency on the stimuli.

Now let us consider the case introduced in Fig. 2c which can be thought of as the composition of the two cases presented in Fig. 2a-b. We compute the interaction cost of parallel processing the tasks with strengths w_{11} and w_{22}. This requires simultaneous activation of S_1 and S_2, which indirectly engages the task with strength w_{21}, and initiates a competition in the output O_1. Thus, the interaction cost of parallelism between tasks represented by w_{11} and w_{22} is given by

$$\Psi_1(a_1 = 1, a_2 = 2) = -\log(P[a_2 = 2] \cdot P[a_1 = 1 | a_2 = 2])$$
$$= -\log\left(1 \cdot \frac{w_{11}}{w_{11} + w_{21}}\right).$$

Here $P[a_2 = 2] = 1$ since task with weight w_{22} is not competing with any other process in the output O_2. The competition, however, takes place in O_1, and the interaction cost associated with w_{11} for this case has already been computed when we discussed the case in Fig. 2a.

In a similar way, we can compute the interaction cost of parallelism between tasks represented by w_{11} and w_{22} in Fig. 2d, and we have

$$\Psi_2(a_1 = 1, a_2 = 2) = -\log(P[a_2 = 2] \cdot P[a_1 = 1 | a_2 = 2])$$
$$= -\log\left(\frac{w_{11}}{w_{11} + w_{21}} \cdot \frac{w_{22}}{w_{22} + w_{12}}\right).$$

In this case, simultaneous activation of S_1 and S_2 causes competition in both outputs O_1 and O_2. Thus, by revealing further insight about the strength and directionality of interference, the interaction cost serves as an extension of the interference definition presented by (Musslick et al., 2016). For instance, for the same values of $w_{11}, w_{21}, w_{22} \geq 0$ in both configurations in Fig. 2c-d, and given $w_{12} \geq 0$ for the configuration in

Fig. 2d, it is obvious that

$$\Psi_1(a_1 = 1, a_2 = 2) \leq \Psi_2(a_1 = 1, a_2 = 2).$$

Neural Network Simulation

In order to investigate the effect of directionality for the third indirect interference during parallel processing, we implemented a synthetic neural network simulation[3] identical to our earlier work (Musslick et al., 2016). The neural network used for this simulation maps stimulus input encoded at a stimulus layer via a non-linear associative layer to non-linear response layer. A separate task input layer encodes the current task to be performed with respect to that stimulus and projects to both the associative layer and response layer. Units in the stimulus layer were grouped into six stimulus dimensions with three units per dimension. Similarly, units in the response layer was grouped into six response dimensions with three units per dimension. The network was trained on 12 tasks, where each task corresponds to a one-to-one mapping between a subset of three input features in a stimulus layer to a subset of three response units in an output layer.

We then used the methods described in Musslick et al. (2016) to extract a bipartite task graph from single representations encoded at the associative layer. The representations associated with each task can be characterized by calculating, for each unit in the associative and output layers, the mean of its activity over all of the stimuli for a given task; this mean pattern of activity can then be used as a representation of the task. Correlating these patterns of activity across tasks yields a task similarity matrix that can be examined separately for the associative and output layers of the network. This can then be used to assess the extent to which different tasks rely on similar or different representation within each layer of the network. Tasks that have similar representations over the associative layer can be inferred to rely on the same input dimension that is, they share an input component in the bipartite graph representation of the network and tasks that are similar at the output layer can be inferred to share an output component. Accordingly, a bipartite graph can be constructed by measuring the patterns of activity observed in the network while it performs each individual task.

The extracted bipartite graph can be used to extract interference patterns between pairs of tasks (cf. Fig. 2). We use this to extract all possible learned task-pairs involving no interference case (Fig. 3a), and two distinct cases of interference as shown in Fig. 3b and Fig. 3c from single-task representations. Fig. 3 shows the activation patterns of the output units for the simultaneous execution of two tasks, averaged across the patterns of all task pairs for a given interference structure. That is, no interference (Fig. 3a) leads to very accurate response patterns (i.e. the current activation shown in orange) is very close to the desired activation pattern shown in grey). For the case in Fig. 3b, the response pattern of task $(S_1 : O_1)$ is primarily impaired due to the interference aris-

---
[3]Simulation details are omitted due to space constraints.

ing from task $(S_2 : O_1)$. However, for the case of interference illustrated in Fig. 3c, the response patterns for both tasks $(S_1 : O_1)$ and $(S_2 : O_2)$ are impaired as observed by the activation patterns. These simulation results reflect the influence of the directionality of interference between tasks as predicted by the proposed interaction cost.
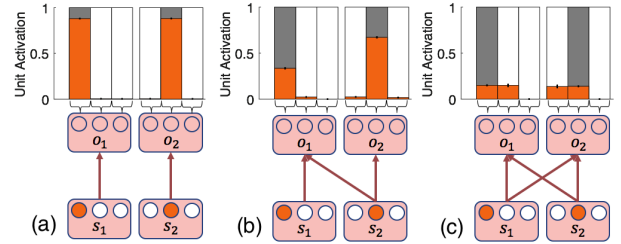

Figure 3: This figure illustrates the performance of a task-pair for a given interference pattern. Each tasks maps a subset of three stimulus input units onto three response units (see text). The orange color in the bar plots indicates unit activation of response units relevant to the depicted tasks, while gray indicates desired response pattern of those units.

General Discussion and Conclusion

In this study, we have introduced two new measures to determine costs associated with intensity and interaction for the demand on control. First, we quantify the intensity cost as a function of the amount of control bias that is supplementary to stimulus-specific processing in order to achieve a desired response from the network. Doing so, we formalize the probability of achieving a desired task given the stimulus, weights and biases infused to the network. Since the stimuli and weights are considered as network properties, the intensity cost to achieve desired response is defined as the amount (value) of control biases required to be injected to the input and output components of the network. The detailed analysis of intensity cost revealed an interesting optimization problem to maximize the probability of surpassing a specified activation for a given budget of resources (i.e. an upper bound on the control biases). The existence of a solution of this optimization problem implicitly reveals whether the desired objective is feasible. However, as it can be foreseen that under certain circumstances the solution does not exist due to interference between the involved processes. Such boundary conditions motivated the second metric introduced in our paper in which we formalize the interaction cost to measure the level of interactions/interference between processes by means of their type of connections and weights.

With the introduced characterization of intensity and interaction costs, it is possible to formally define whether a process is considered a reflex, automatic or controlled. Concretely, a process is considered a reflex if the underlying weight guarantees a successful execution. In other words, a reflex can be successfully executed without any intensity or interaction costs. We assume that the execution of both controlled and automatic processes carries with it an intensity cost as some amount of control bias is needed to elicit a response. However, unlike the former, controlled processes are

subject to interference and thus yield interaction costs large than zero.

The metrics proposed here can also be used towards further understanding cognitive effort as well as synthetic neural networks designed to achieve goal-driven tasks. By using the intensity cost, which reveals the interrelationship between control bias and probability of achieving a desired objective, we will investigate the limitations of any given neural network architecture by allocating a budget of control bias. The intensity cost can also be used to investigate the feasibility of achieving a desired objective defined by the set of processes of interest in a network.

In the interaction cost analysis, we have assumed that there exist a response in the output for any stimulus activation and this may not be the case for a nonlinear activation in the output components. Hence, one major research direction is the detailed analysis for the classification of processes with nonlinear activation in output. Another possible direction for future work is to further analyze the interaction cost in order to capture the properties of the overall network (not only a subset of tasks of interest). This will allow one to use the interaction cost as an objective function for network training. Another possible direction is to explore the interrelationship between intensity and interaction cost. In our work (Musslick et al., 2016), we noticed a fundamental trade-off between shared representations in a network and its parallel processing capability (separated representations). Intuitively, we envision this separation will decrease interaction cost while increasing the likelihood of successful execution for a given budget of control bias.

Acknowledgements

References

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765.

Botvinick, M. M. (2007). Conflict monitoring and decision making: Reconciling two perspectives on anterior cingulate function. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(4), 356–366.

Botvinick, M. M., & Braver, T. (2015). Motivation and Cognitive Control: From Behavior to Neural Mechanism. *Annual Review of Psychology*, *66*(1), 83–113.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652.

Botvinick, M. M., & Cohen, J. D. (2014). The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cognitive Science*, *38*(6), 1249–1285.

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends in Cognitive Sciences*, *8*(12), 539–546.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the stroop effect. *Psychological Review*, *97*(3), 332–361.

Feng, S. F., Schwemmer, M., Gershman, S. J., & Cohen, J. D. (2014). Multitasking vs. Multiplexing: Toward a normative account of limitations in the simultaneous execution of control-demanding behaviors. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(1), 129-146.

Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, *11*(6), 229–235.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *The Behavioral and Brain Sciences*, *36*(6), 661–679.

Muraven, M., Tice, D. M., & Baumeister, R. F. (1998). Self-control as a limited resource: Regulatory depletion patterns. *Journal of Personality and Social Psychology*, *74*(3), 774-789.

Musslick, S., Dey, B., Özcimder, K., Patwary, M. M. A., Willke, T. L., & Cohen, J. D. (2016). Controlled vs. Automatic Processing: A Graph-Theoretic Approach to the Analysis of Serial vs. Parallel Processing in Neural Network Architectures. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1547–1552). Philadelphia, PA.

Ortega, P. A., & Braun, D. A. (2013). Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of Royal Society A*, *469*(2153), 20120683.

Posner, M. I., & Snyder, C. R. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information Processing and Cognition: The Loyola Symposium* (p. 55-85). Lawrence Erlbaum.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). *Toward a rational and mechanistic account of mental effort.* (submitted to Annual Reviews of Neuroscience)

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127-190.