Sequential presentation protects working memory from catastrophic interference

Ansgar D. Endress
Department of Psychology, City, University of London, UK
Szilárd Szabó
Department of Mathematics, Budapest University of Technology and Economics, Budapest, Hungary

Draft of February 21, 2020

Abstract

Neural network models of memory are notorious for catastrophic interference: old items are forgotten as new items are memorized (e.g., French, 1999; McCloskey & Cohen, 1989). While Working Memory (WM) in human adults shows severe capacity limitations, these capacity limitations do not reflect neural-network style catastrophic interference. However, our ability to quickly apprehend the numerosity of small sets of objects (i.e., subitizing) does show catastrophic capacity limitations, and this subitizing capacity and WM might reflect a common capacity. Accordingly, computational investigations (Knops, Piazza, Sengupta, Eger, & Melcher, 2014; Sengupta, Surampudi, & Melcher, 2014) suggest that mutual inhibition among neurons can explain both kinds of capacity limitations as well as why our ability to estimate the numerosity of larger sets is limited according to a Weber ratio signature. Based on simulations with a saliency map-like network and mathematical proofs, we provide three results. First, mutual inhibition among neurons leads to catastrophic interference when items are presented simultaneously. The network can remember a limited number of items, but when more items are presented, the network forgets *all* of them. Second, if memory items are presented sequentially rather than simultaneously, the network remembers the most recent items rather than forgetting all of them. Hence, the tendency in WM tasks to sequentially attend even to simultaneously presented items might not only reflect attentional limitations, but an adaptive strategy to avoid catastrophic interference. Third, the mean activation level in the network can be used to estimate the number of items in small sets, but does not accurately reflect the number of items in larger sets. Rather, we suggest that the Weber ratio signature of large number discrimination emerges naturally from the interaction between the limited precision of a numeric estimation system and a multiplicative gain control mechanism.

Keywords: Working memory; Interference; Temporary Memory; Memory Capacity

Introduction

Artificial neural network models are notorious for being susceptible to catastrophic interference: new information can completely wipe out earlier information (e.g., French, 1999; Kirkpatrick et al., 2017; Masse, Grant, & Freedman, 2018; McCloskey & Cohen, 1989; Ratcliff, 1990). This property is widely seen as problematic if these models are meant to be models of human memory and learning performance (e.g., French, 1999; McCloskey & Cohen, 1989), and different authors have sought ways to avoid it (e.g., Kirkpatrick et al., 2017; Kumaran, Hassabis, & McClelland, 2016; Masse et al., 2018; McClelland, McNaughton, & O'Reilly, 1995).

However, at least in infancy, some cognitive abilities such as number processing do show catastrophic performance limitations. For example, while infants reliably choose three over one food items, they are unable to choose between four and one food item; the most common explanation is that the number system used to process small numbers up to three is incompatible with the system used to process large numbers beyond three (e.g., Feigenson, Carey, & Hauser, 2002; Feigenson & Carey, 2005; Zosh & Feigenson, 2015). Further, in line with long-standing theorizing that our cognitive abilities might suffer from a general capacity limit (e.g., Miller, 1956; Cowan, 2005), this small number system might be related or even identical to Working Memory (Cowan, 2005; Piazza, Fumarola, Chinello, & Melcher, 2011), which also has important capacity limitations (e.g., Cowan, 2005; Fukuda, Vogel, Mayr, & Awh, 2010; Luck & Vogel, 1997; Vogel, Woodman, & Luck, 2001). If so, catastrophic interference might be a plausible occurrence in humans at least under some conditions and at least for some mechanisms.[1]

[1]While Working Memory limitations are less catastrophic in adults than in infants, this does not neces-

In line with the view that Working Memory and small number processing rely on a common system, Sengupta et al. (2014) and Knops et al. (2014) recently proposed a neural network model based on a saliency map that, according to these authors, simultaneously accounts not only for Working Memory capacity limitations and small number processing, but also for the processing of numbers that go beyond the limitations of the small number system.

Here, we reanalyze the model, and show that it leads to catastrophic interference when remembering items (albeit, as we will argue below, for a different reason than earlier reports of catastrophic interference). Critically, we also show that a very simple and psychologically interpretable change avoids this type of catastrophic interference: presenting items sequentially rather than simultaneously.

To preview our results, we conclude that a saliency-map-like network is a plausible mechanisms for small number processing, in line with earlier suggestions that the small number system is really an attentional system (e.g., Feigenson et al., 2002; Feigenson & Carey, 2003; Piazza et al., 2011; Trick & Pylyshyn, 1994). However, other mechanisms are likely needed to keep track of numbers beyond the limits of the small number system, again in line with proposals suggesting that small and large numbers are processed by separable mechanisms (e.g., Feigenson, Dehaene, & Spelke, 2004; Izard & Dehaene, 2008; Revkin,

---

sarily reflect memory per se as opposed to auxiliary mechanisms that *also* show important developmental differences and that are critical for memory, including rehearsal, memory strategies, speed of processing, goal maintenance and so forth. For example, if the general processing speed is faster in adults than in infants (e.g., Keating & Bobbitt, 1978), adults can (attentionally) cycle more quickly through the memory representations they need to maintain, which, in turn, allows them to "refresh" the items more efficiently. In contrast, infants might take more time to switch among representations so that the activation of these representations might have decayed by the time they return to them. Alternatively, they might also try to maintain all representations simultaneously, which would then lead to catastrophic interference according to the model analyzed below. As a result, catastrophic failures in infancy might possibly be a "purer" reflection of memory processing constraints than the more sophisticated abilities of adults.

Piazza, Izard, Cohen, & Dehaene, 2008; Trick & Pylyshyn, 1994). In contrast, a saliency-map-like network is only a plausible model of Working Memory (1) if Working Memory relies on attentional pointers and (2) if participants attend items sequentially even when they are presented simultaneously.

Is there a common cognitive capacity limitation for Working Memory and small number processing?

Our ability to retain (and perhaps process) items over brief periods of time is severely limited. This Working Memory (WM) capacity is thought to be limited to about four items (e.g., Miller, 1956; Conway et al., 2005; Cowan, 2001, 2005). However, WM is not our only capacity limitation. When we see fewer than 3 or 4 objects at a time, we know immediately and without counting how many objects we face, a phenomenon that has been called subitizing (e.g., Kaufman & Lord, 1949; Trick & Pylyshyn, 1994). Beyond this limit, we either have to explicitly and sequentially count the objects, or can estimate their number only approximately. The precision of these approximate estimates follows a Weber-law signature and is thus roughly proportional to the quantity the observer tries to estimate. The system for processing large numbers is thus called the *approximate number system* (ANS; e.g., Feigenson et al., 2004; Izard & Dehaene, 2008; Revkin et al., 2008). Finally, a third capacity limitation is related to how many moving objects we can track simultaneously in a display (e.g., Pylyshyn & Storm, 1988; Scholl & Pylyshyn, 1999).

Miller (1956) suggested that these capacity limitations might have a common source. For example, WM, subitizing and parallel individuation might all rely on a system of parallel attention such as the one proposed by Pylyshyn and Storm (1988) (e.g., Cowan, 2001, 2015;

Piazza et al., 2011; Trick & Pylyshyn, 1994).[2] That is, we might be able to remember only 3 or 4 items, because we might remember them by allocating attention to them, using a limited attention capacity of 3 or 4 items (Cowan, 1995).

However, recent evidence would seem problematic at least for the most straightforward versions of such models, because WM-like capacity-limitations in memory arise only under conditions of strong interference among items (Endress & Potter, 2014; Endress & Siddique, 2016; Sands & Wright, 1980), because some WM tasks show only limited interference with attentional tasks (e.g., Fougnie & Marois, 2006; Hollingworth & Maxcey-Richard, 2013; H. Zhang, Xuan, Fu, & Pylyshyn, 2010), and because WM and attention seem to have different properties (Endress, Korjoukov, & Bonatti, 2017). Accordingly, WM might be better described as a continuous resource (e.g., Alvarez & Cavanagh, 2004; Bays & Husain, 2008; van den Berg, Shin, Chou, George, & Ma, 2012; Ma, Husain, & Bays, 2014).

On the other hand, Piazza et al. (2011) showed that individual participants' subitizing range is correlated with their WM capacity as tested in a change detecting experiment (Luck & Vogel, 1997), and that a WM task and a subitizing task mutually interfere with each other when the total number of items is at least 4. However, other investigators did not find interference between WM load and subitizing performance (Shimomura & Kumada, 2011), and the correlation between WM performance and subitizing performance is not observed when WM is measured in complex span tasks (Tuholski, Engle, & Baylis, 2001). The relationship between WM and subitizing thus seems open.

[2]Trick and Pylyshyn (1994) proposed that pre-attentive spatial pointers (i.e., FINSTs) are used to track and enumerate objects, and that WM and subitizing do not rely on a common mechanism. More recently, however, and as discussed below, other authors proposed that subitizing might be linked to WM via attention.

Can a salience map explain both Working Memory and number processing?

Recently, Sengupta et al. (2014) and Knops et al. (2014) revisited the link between WM, small number processing (below 3 and 4), and large number processing. They suggested that all three capacity limitations can be explained based on mutual inhibition among neurons in a saliency map. If so, WM limitations and small number processing would result from the limitations of an attentional mechanism without requiring a special WM system, similar to the theoretical proposals discussed above.

Specifically, Sengupta et al. (2014) and Knops et al. (2014) described a network of neurons coding for the spatial positions of objects. Each neuron excites itself, and inhibits all other neurons. This architecture reflects a saliency map such as those thought to be found in the human posterior parietal cortex (e.g., Bays, Singh-Curry, Gorgoraptis, Driver, & Husain, 2010; Gottlieb, 2007; Roggeman, Fias, & Verguts, 2010).

At the beginning, the network was presented with an input vector coding for the spatial locations of the memory items. After the presentation of this input vector, the network was given the opportunity to find a steady-state activation pattern. Importantly, these authors varied the number of non-zero elements in the input vector, representing the number of memory items. (Hereafter, we will call this number the *set-size* or the *numerosity*.)

Sengupta et al. (2014) and Knops et al. (2014) hypothesized two main response profiles shown in Figure 1. First, according to some models of WM, we have only a fixed number of memory slots (e.g., Cowan, 2001; Luck & Vogel, 1997; Piazza et al., 2011; Rouder et al., 2008; W. Zhang & Luck, 2008); we can retain items only up to the memory capacity

and then do not add any more items to memory. Accordingly, the number of remembered items should increase up to the memory capacity and then plateau; Sengupta et al. (2014) and Knops et al. (2014) operationalize this prediction by positing that the mean network activity should increase up to the memory capacity and then plateau (green dashed line in Figure 1). Second, in an enumeration task, reaction times are relatively constant within the subitizing range, and then increase linearly with the set-size; Sengupta et al. (2014) and Knops et al. (2014) predict the same pattern in the mean network activity (red solid line in Figure 1), though it is not entirely clear why. In fact, we argue below that such a response profile is more consistent with a system processing *large* numbers: it is inactivate for small numbers and then becomes increasingly active as more items are presented.

Sengupta et al. (2014) and Knops et al. (2014) reported simulation results that were in line with these predictions. With relatively high levels of inhibition, the mean activation in the network rose up to a set-size of 4, and then remained constant. They suggested that this network behavior reflects how actual participants behave in WM experiments. Further, Sengupta et al. (2014) and Knops et al. (2014) reported that, at medium levels of inhibition, the mean level of activation in the network was relatively constant for up to 3 items; this pattern of activation thus mirrors the subitizing profile mentioned above. Finally, Knops et al. (2014) measured brain activity while human adults performed a visual WM task or a visual number processing task, and suggested that the posterior parietal cortex can flexibly switch between subserving small-number processing and WM: voxels in the same brain region showed an enumeration profile in an enumeration task, and a WM profile in a WM task.

However, it is not clear to what extent these response profiles reflect WM and subitiz-

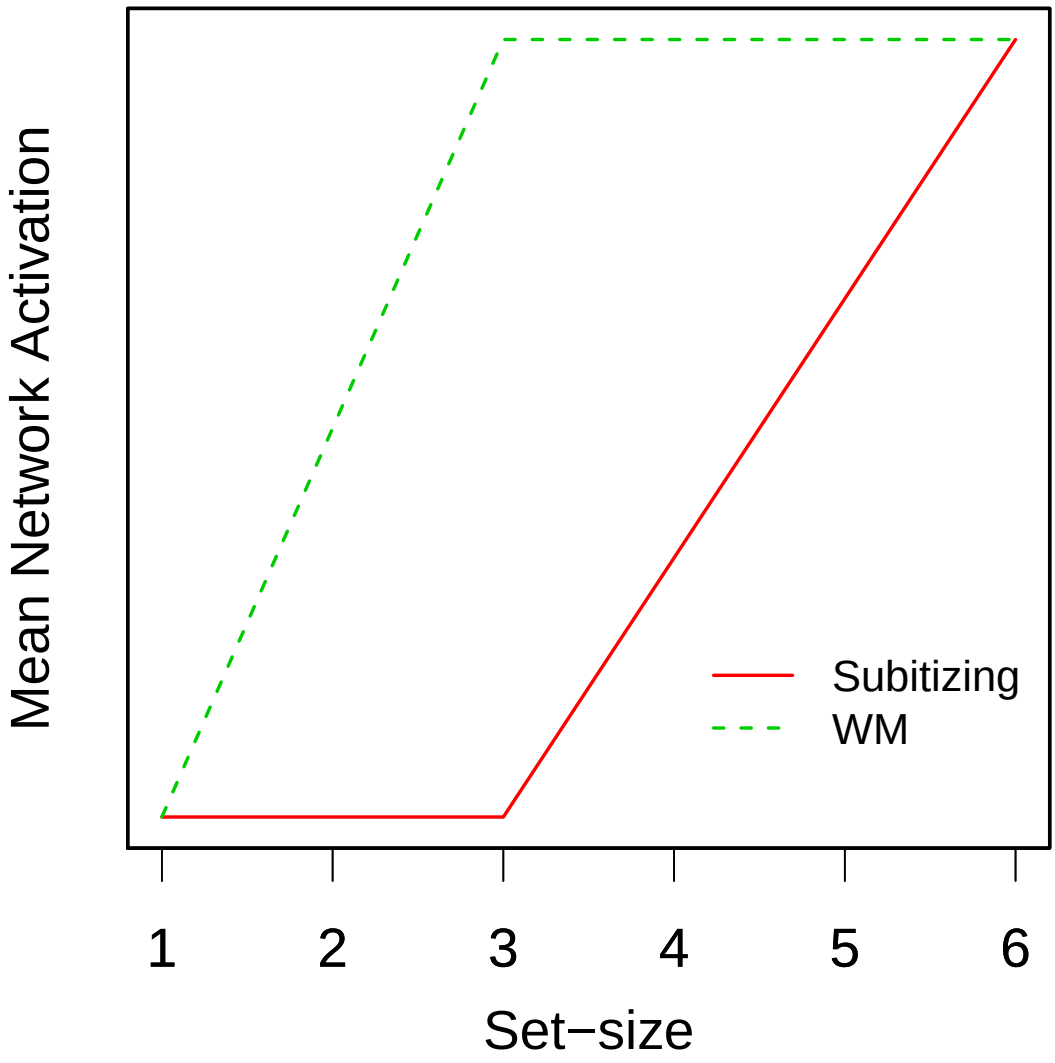

Figure 1. Mean network activation profiles hypothesized by Sengupta et al. (2014) and Knops et al. (2014). According to a subitizing profile, reaction times (and thus, according to Sengupta et al. (2014) and Knops et al. (2014), the mean network activation) should remain constant within the subitizing range, and increase for larger set-sizes. According to a WM profile, the number of remembered items as well as the mean network activation should increase within the WM capacity, and then remain constant.

ing, respectively. With respect to WM, similar response profiles have been reported in numerous other neuroimaging experiments (see Vogel & Machizawa, 2004, for one of the earliest demonstrations). Behaviorally, however, both the memory precision and reaction times deteriorate even within the putative WM capacity (Schneegans & Bays, 2016), and WM-like response profiles do not necessarily reflect memory slots (Bays, Catalao, & Husain, 2009; Endress & Szabó, 2017).

With respect to subitizing, if the subitizing mechanism is specifically involved in the processing of small numerosities, one would expect it to be active for *small* numerosities, and not for large numerosities. This expectation contrasts with the subitizing response profile, where activation is greater for numerosities that are *not* in the subitizing range. As a result, this response profile fits better with the processes involved in attentionally and sequentially "counting" objects beyond the subitizing range. In the subitizing range, these attentional processes are not required, but they become increasingly necessary as we count items beyond the subitizing, which, in turn, would fit the activation in the aforementioned enumeration profile.

This interpretation is in line both with the involvement of the posterior parietal cortex in attentional tasks, and also with experiments with rhesus macaques, where single neurons in a putatively homologous region (the lateral intraparietal area) responded to *large* numbers beyond the subitizing range (Roitman, Brannon, & Platt, 2007), although large number processing dissociates from small number processing (e.g., Feigenson & Carey, 2005; Hauser, Carey, & Hauser, 2000; Izard & Dehaene, 2008; Revkin et al., 2008). It is also in line with Knops et al.'s (2014) data: they showed that a classifier decoding numerosity from brain activity performed better for the numerosities 5 and 6 than for the numerosities

3 and 4, and thus better for numerosities outside the subitizing range. In contrast, other brain imaging studies in humans found the involvement of other regions in small number processing (i.e., the temporo-parietal junction; Ansari, Lyons, van Eimeren, & Xu, 2007; though these differences might reflect the attentional requirements of small vs. large number processing Burr, Turi, & Anobile, 2010; Gliksman, Weinbach, & Henik, 2016).

Be that as it might, while there is considerable evidence suggesting that WM and attention might rely on separable mechanisms, there are also important theoretical and empirical results suggesting a link between WM and (small) number processing — even though this link was traditionally believed to be *via* parallel attention.

Catastrophic interference in neural networks

As mentioned above, artificial neural networks are susceptible to catastrophic interference, in the sense that newly learned information can overwrite and wipe out older information (e.g., French, 1999; Kirkpatrick et al., 2017; Masse et al., 2018; McCloskey & Cohen, 1989; Ratcliff, 1990). At first sight, this does not seem to apply to Sengupta et al.'s (2014) and Knops et al.'s (2014) model, because their model does not involve *learning*. However, Sengupta et al.'s (2014) and Knops et al.'s (2014) model suffers from a different type of interference that occurs during *processing* rather than learning and that arises due to inhibitory interactions among simultaneously active representations. We note that the latter view of processing (rather than learning) interference, where existing representations compete with other representations for accessibility, is critical in psychology, especially for interference-based theories of Working Memory (e.g., D'Esposito, Postle, Jonides, & Smith, 1999; Endress & Szabó, 2017; Oberauer & Lin, 2017; Shipstead & Engle, 2013; Nee, Jonides,

& Berman, 2007). That being said, the behavioral effects of both kinds of interference are similar in that both involve a competition among items for accessibility. In the case of learning interference, the competition involves permanently altering the strength of the representations of some items, or, more likely given that humans have a large memory capacity for individual items (e.g., Brady, Konkle, Alvarez, & Oliva, 2008; Konkle, Brady, Alvarez, & Oliva, 2010; Standing, Conezio, & Haber, 1970; Standing, 1973), the strength of association between these items and relevant retrieval cues; in the case of processing interference, the interference is between simultaneously activated items, possibly through some form of lateral inhibition.

Specifically, in Sengupta et al.'s (2014) and Knops et al.'s (2014) model, catastrophic interference is caused by lateral inhibition among neurons. If a large number of neurons is externally stimulated, lateral inhibition rapidly drives the network activation to zero after the offset of the stimulation, at least for high levels of inhibition. This is because the externally stimulated neurons all exert comparable inhibition on each other, which, in the absence of further stimulation, leads to a complete extinction of the network activation. Below, we refer to the learning-related type of interference as *learning interference* and to the interference among simultaneously active representations as *processing interference*.

While catastrophic interference is usually seen as problematic for modeling the memory performance of actual humans (e.g., French, 1999; Kirkpatrick et al., 2017; Kumaran et al., 2016; Masse et al., 2018; McClelland et al., 1995; McCloskey & Cohen, 1989), we will argue below that catastrophic inhibition is a desirable property in the domain of small number processing (though not for memory processing). Before, however, we will present the results of our simulations and mathematical analyses.

A preview of the simulations

We investigate under which conditions mutual inhibition among items in a simple saliency map-like network provides a plausible model of three aspects of human cognitive performance: WM, small number processing and large number processing.

Regarding WM, we show that, in line with earlier theoretical analyses (Endress & Szabó, 2017), mutual inhibition can — in principle — explain limited WM capacities, because *any* kind of interference or inhibition can lead to limited memory capacities (Endress & Szabó, 2017). However, we also show that mutual inhibition among neurons yields plausible memory capacity limitations only under certain conditions. Specifically, when memory items are presented simultaneously (as in previous simulations), mutual inhibition has catastrophic effects: items are remembered perfectly up to a critical set-size (that depends on the strength of inhibition). When more than this critical number of items are presented, zero items are remembered.

In contrast, when items are presented sequentially one after the other, the network retains a moving window of the most recent items. To the extent that WM capacity limitations are based on mutual inhibition among items, we suggest that these computational results might explain why observers encode items one after the other even when they are presented simultaneously (e.g., Liu & Becker, 2013; Vogel, Woodman, & Luck, 2006; but see Mance, Becker, & Liu, 2012): Rather than reflecting a limitation of attention, sequential encoding might be optimal to avoid catastrophic interference (see also Ihssen, Linden, & Shapiro, 2010, for related results).

Turning to small number processing, we show that the mean activation in the network

is diagnostic of the number of items, but not of the contents of WM, which, in turn, can be decoded from the identity of the activated "neurons." We suggest that the number of items can be read from the network activation for small numbers of items (provided that the brain has appropriate read-out mechanisms). Further, and in contrast to the WM results, we suggest that catastrophic (processing) interference is a desirable property of small number processing in the light of results from developmental psychology (e.g., Feigenson et al., 2002; Feigenson & Carey, 2005; Zosh & Feigenson, 2015).

Finally turning to large number processing, we show that, contrary to earlier claims, the network fails to discriminate larger numbers of items. Rather, we propose that the Weber-law signature of the approximate number system might have the same origin as Weber-fraction signatures in other sensory discrimination: multiplicative gain control mechanisms (e.g., Priebe & Ferster, 2002; Salinas & Thier, 2000).

Materials and method

We use the same network as Sengupta et al. (2014) and Knops et al. (2014). Specifically, the network consists of 70 fully connected neurons (except in some simulations in Appendix F, where we use 400 neurons). The activation of each neuron is determined by decay, self-excitation, inhibition from all other neurons, external input (where applicable), and noise. Specifically, the i^{th} neuron has an activation x_i, whose change is given by

$$\frac{dx_i}{dt} = \dot{x} = -\lambda x_i + \alpha F(x_i) - \beta \sum_{j=1, j\neq i}^{N} F(x_j) + I_i + \text{noise}, \tag{1}$$

where the first term reflects (exponential) decay, the second term self-excitation, the third term inhibition from all other neurons, the fourth term the external input to neuron i, and the fifth term Gaussian noise. $F(x)$ is the activation function. The activation functions we consider are defined in Table 1. Our primary results are based on the rational sigmoid function $F(x) = x/(1+x)$ used by Sengupta et al. (2014) and Knops et al. (2014). However, we show in Appendix B that qualitatively similar results are obtained with other activation functions as long as they are increasing and bounded, with one qualification discussed below. Following Sengupta et al. (2014) and Knops et al. (2014), we use a discrete version of this

Table 1
Activation functions.

Name	$F(x)$
Rational sigmoid	$F(x) = \begin{cases} \frac{x}{1+x} & x \geq 0 \\ 0 & x < 0 \end{cases}$
Threshold	$\begin{cases} 1 & x > \theta \\ 0 & x \leq \theta \end{cases}$
tanh	$\frac{e^x - e^{-x}}{e^x + e^{-x}}$
Sigmoid	$\frac{1}{1+e^{-x}} - \frac{1}{2}$
Linear	γx
Recitfied linear unit	$\max(0, x)$

differential equation for our simulations:

$$x_i(t+1) = x_i(t) - \lambda x_i(t) + \alpha F\left(x_i(t)\right) - \beta \sum_{j=1, j\neq i}^{N} F\left(x_j(t)\right) + I_i + \text{noise} \tag{2}$$

We note that the time step has no unit and is thus arbitrary (see Appendix A for a proof).

As shown in Table 2, we use the same parameters as Sengupta et al. (2014) and Knops et al. (2014). However, given that we confirm our results through mathematical analyses, different parameter sets would generally yield similar results.

Table 2
Summary of network parameters and variables. Variables that are "calculated" are not chosen, but rather calculated during the simulations and analyses.

Symbol	Description	Value
	Free network parameters	
N	Number of neurons	70 (400 in Appendix F)
α	Self-excitation	2.2
β	Inhibition	.01 (low), .1 (medium), .15 (high)
λ	Decay	1
I_i	External excitation	1.0[a] (active), 0 (inactive)
σ_{noise}	Noise standard deviation	.03 or 0
	Other variables	
S	Set-size	varied in simulations
A	Number of active neurons	calculated
m, M	Minimal, maximal activation	calculated

[a]While Sengupta et al. (2014) and Knops et al. (2014) used .33 as the initial activation value, the results are qualitatively similar irrespective of whether 1.0 or .33 is used.

The input is presented in three different ways. In all cases, the input vector comprises S (for set-size) randomly chosen non-zero entries, and $N - S$ zero entries, where N is the total number of neurons in the network. With *simultaneous presentation*, the entire vector is presented to the network for 5 time steps; following this, the network is given 45 time steps to settle into a steady-state pattern. With *continuous sequential presentation*, the non-zero inputs components are presented one after the other, for 5 time steps each; after the presentation of the last non-zero input, the network is again given 45 time steps to settle into a steady-state activation pattern. With *sequential presentation with intermittent consolidation* (presented in Appendix E), each non-zero input is presented for 5 time steps, but, after each input, the network is given 45 time steps to settle into a steady-state activation pattern. In all cases, we analyzed the network output at the end of the simulation.

Following Sengupta et al. (2014) and Knops et al. (2014), we use two primary metrics for evaluating the network performance. First, we use the mean activation in the network as

a measure of *how many* items are retained. The rationale for this metric is related to the fact that the network is fundamentally an attentional model; the question of how many items it can retain thus boils down to the question of how many items can remain active in the attentional system. As a population-level read-out mechanism (e.g., for number processing) does not necessarily take into consideration the specific identify of the neurons, the mean network activation is a good measure of the number of active items.

Our second measure is the "faithfulness" of the activations, and indicates to what extent the current activation pattern in the network matches the input pattern the network is supposed to retain.

We also calculate the discriminability between familiar and unfamiliar items. We use $d' = Z(H) - Z(FA)$, where Z is the quantile function of the standard normal distribution and H and FA are the corrected hit and false alarm rates, respectively, using the "log-linear" correction for extreme proportions recommended by Hautus (1995) and Snodgrass and Corwin (1988):

$$\begin{aligned} H &= \frac{\text{Number of Hits} + .5}{S + 1} \\ FA &= \frac{\text{Number of False Alarms} + .5}{N - S + 1} \end{aligned}$$

S and N are the set-size and the number of neurons, respectively.

The code is available at `https://doi.org/10.25383/city.11456106`.

Results

Simultaneous presentation leads to catastrophic interference

Simulations. We first show that the network shows catastrophic interference when memory items are presented simultaneously. The results for the noise-less situation are shown in Figures 2(A) to (C), using three different values for the inhibition parameter β (.01, .1 and .15). Following Sengupta et al. (2014) and Knops et al. (2014), we refer to these regimes as low, medium and high inhibition, respectively.

As shown in Figure 2(A), the mean activation increases up to a maximum, and then decreases again and reaches zero. Beyond some critical set-size that we call $S_{\max}$ below, the mean activation remains at zero. In other words, there is catastrophic (processing) interference: if more than $S_{\max}$ items are presented, *all* of them are forgotten. (For very low inhibition values, the network activation remains positive because, $S_{\max}$ is greater than the total number of neurons in the network; see below for the formulae.)

As shown in Appendix B, other activation functions lead to similar results as long as they are bounded and as long as the slope is not too small at zero. Specifically, when items are presented simultaneously, we find catastrophic (processing) interference for $S > 1 + \frac{\alpha}{\beta}$, while, when items are presented sequentially, the network successfully discriminates seen from unseen items for much larger set-sizes.

In contrast, for activation functions that are unbounded or for which the slope at zero is too small, all neurons — whether they have received stimulation or not — are eventually switched off (or have at least extremely small activation values). In Appendix B.1.2, we show that, if the slope of an activation function is sufficiently small at zero (i.e., $F'(0) <$

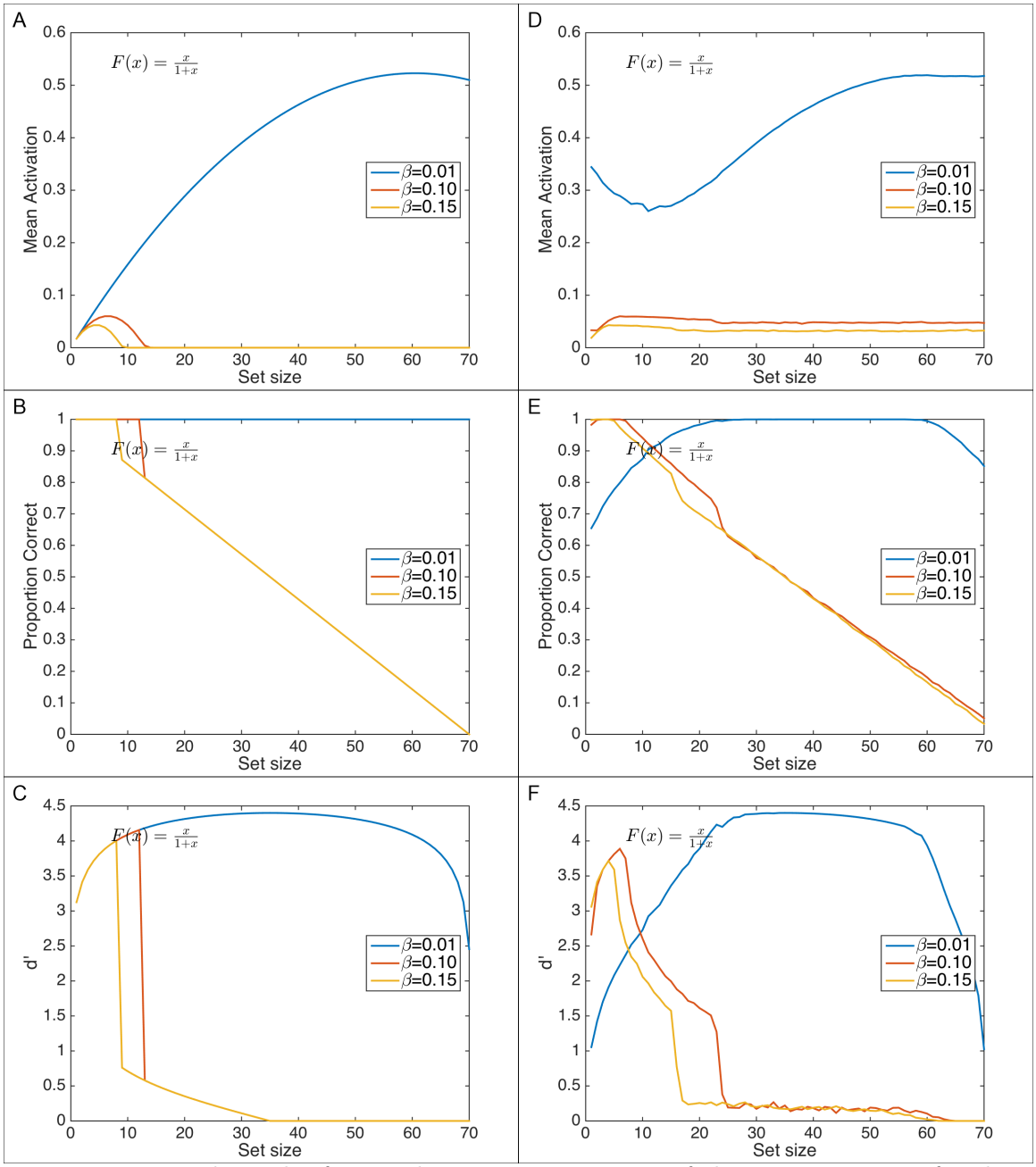

Figure 2. Network results for simultaneous presentation of the memory items, for three different levels of inhibition (low, $\beta = .01$, blue line; medium, $\beta = .1$, red line; high, $\beta = .15$, orange line). The proportion of correct responses is the complement of the Hamming distance between the input and the steady-state activation at end of the simulation. Simulations without noise (A, B, C). (A) The mean activation in the network is a quadratic function with a maximum as $S_{\max}/2$ and a root at $S_{\max}$; beyond $S_{\max}$, the activation is zero. (B) The proportion of correct activations is 1 up to $S_{\max}$, and then linearly decreases to zero, as all output neurons are switched off, and an increasing proportion of input neurons are switched on. (C) Discriminability between stimulated and unstimulated neurons is high below $S_{\max}$, and then drops abruptly. (D, E, F). The simulation results in the presence of noise are similar, except that the network noise maintains a steady-state activation beyond $S_{\max}$.

$\lambda/(\alpha + \beta)$), any difference between activation levels tends to disappear as the network continues to evolve, which, in turn, leads to a complete loss of information irrespective of how many or how few items are presented to the network.

We thus conclude that catastrophic (processing) interference is a fairly general phenomenon, but that there are some activation functions for which sequential presentation does not confer any protection against catastrophic interference.

In Figure 2(B), we show the "faithfulness" of the memories, to estimate the number of items the network can maintain. Following Sengupta et al. (2014), we use the proportion of neurons whose (binary) steady-state activation is identical to that in the corresponding entry of the input vector, after converting the continuous activation values to binary values, using a threshold of .03. That is, we used the complement of the Hamming distance; in our case, this is the proportion of correctly recalled items. The faithfulness is excellent up to $S_{\max}$; beyond a critical set-size $S_{\max}$, the proportion of mistakes grows linearly with the set-size up to 100%.

Finally, we show in Figure 2(C) that the discriminability index d' drops abruptly at $S_{\max}$.

At first sight, this pattern of results seems to fit well with a slot-based WM model: memory is excellent up to the memory capacity, and then gets gradually worse. However, it arises for a very different reason than in WM experiments. Beyond $S_{\max}$, activation is zero for all neurons. As a result, as the set-size grows, an increasing proportion of the input neurons has non-zero activation, and this increasing proportion of non-zero activation drives the increasing difference between the input pattern and the output pattern. In signal detection terms, the network's hit rate approaches zero, while the correct rejection rate

approaches one. However, given that most neurons are inactive, the high correct rejection rate ensures that the percentage of correct responses and d' remain positive, though d' rapidly drops to levels below .5.

These predictions contrast markedly with the behavior of participants in WM experiments, who do not forget all items when presented with a supra-capacity set-size of memory items. In fact, WM capacity is usually calculated from trials that *exceed* a participant's WM capacity (e.g., Cowan, 2001; Rouder et al., 2008), because memory performance is at ceiling while the set-size does not exceed the memory capacity. Hence, it seems that a simple saliency map does not necessarily account for human WM performance.

As shown in Figures 2(D) to (E), the addition of noise prevents the mean activation from decaying to zero, but the network still shows catastrophic interference. It thus seems that the network behavior does not match that of human participants.

Analytic derivation. The simulations so far show that, when items are presented simultaneously, the network will forget *all* of them if more than a critical number of items are presented. For the noiseless case, this network behavior can also be derived from the model equations. We first show that the mean network activation tends to zero for supra-capacity set-sizes, and then calculate the corresponding faithfulness of the network. To do so, we need to assume that

$$\alpha \geq \lambda$$

where α and λ represent the self-excitation and forgetting parameters, respectively. This assumption is necessary, because the assumption $\alpha < \lambda$ would imply exponential decay of the solution of (1), since $\dot{x}_i < (\alpha - \lambda)x_i$.

Further, since $\beta > 0$, the assumption $\alpha \geq \lambda$ also implies

$$\alpha + \beta > \lambda.$$

Steady-state solutions for $\{x_i\}$**.** In the absence of noise, Equation (1) has a steady-state solution. In a steady-state pattern, the time derivative of the activation needs to be zero, indicating no change over time. This yields the following equation:

$$\begin{aligned} 0 &= -\lambda x_i + \alpha F(x_i) - \beta \sum_{j=1, j\neq i}^{N} F(x_j) \\ &= -\lambda x_i + \alpha F(x_i) - \beta(N-1)F(x_j) \end{aligned} \tag{3}$$

In the second step, we used the observation that all active neurons must have the same activation level, given that no neuron is prioritized and all computations are done simultaneously. As shown by Sengupta et al. (2014), expanding F reveals a steady-state solution where the individual activation is given by

$$\hat{x}(S) = \frac{\alpha - (S-1)\beta}{\lambda} - 1, \tag{4}$$

where S is the set-size of memory items and $\hat{x}$ is the steady-state activations of the neurons coding for these memory items. However, in addition to this solution, there is a second solution to Equation (3) as long as $F(0) = 0$: $\hat{x}(S) = 0$ for all S.

Importantly, beyond a critical set-size $S_{\max}$, this trivial solution is the only solution. Specifically, for the steady-state activation to be positive, we have to solve the inequality

$\frac{\alpha-(S-1)\beta}{\lambda} - 1 > 0$. Doing so reveals that there exists no non-zero steady-state solution above a critical set-size

$$S_{\max} = 1 + \frac{\alpha - \lambda}{\beta} \tag{5}$$

For example, with the values from Sengupta et al.'s (2014) high-inhibition situation ($\alpha = 2.2, \beta = .15, \lambda = 1$), $S_{\max} = 9$. Taken together, the steady-state activation in the noiseless situation is thus given by:

$$\hat{x}(S) = \begin{cases} \frac{\alpha-(S-1)\beta}{\lambda} - 1 & S < S_{\max} \\ 0 & S \geq S_{\max} \end{cases} \tag{6}$$

These analyses thus confirm the simulation results above. Below $S_{\max}$, there is network activity, but from $S_{\max}$ onwards, the network forgets all inputs.

Convergence to the steady state solutions. In the following, we will show that the network activation converges to the steady solutions (when they are solutions). For the proofs, we use the discrete version of Equation (1), and rewrite Equation (2) as $x_i(t+1) \equiv g(x_i(t))$ with $g(x) = (1-\lambda)x + (\alpha+\beta)F(x) - S\beta F(x)$; again, we assume that all active neurons have the same activation.

A stationary solution is a fixed point x^* of g such that $g(x^*) = x^*$. According to Banach's fixed point theorem, a sufficient condition for convergence to a unique and stable fixed point is $|g'(x^*)| < 1$, where g' is the derivative $g'(x) = 1 - \lambda + (\alpha+\beta)F'(x) - S\beta F'(x)$.

That is, if $|g'(x^*)| < 1$ holds, we can start with an arbitrary $x(t)$ and subsequent $x(t+T)$ will increasingly approach the fixed point x^*.

In the case of Sengupta et al.'s (2014) and Knops et al.'s (2014) rational sigmoid function, the derivative is given by $F'(x) = 1/(1+x)^2$. We will now show that, when these fixed points exist, there is convergence to them, but that, even when there is no convergence to a fixed point, all neurons will tend to have the same network activation.

Claim 1 (Convergence to 0)**.** *If $x^* = 0$ is a fixed point of Equation* (2) *in the special case that all active neurons have the same activation, there is convergence to that fixed point for $S > S_{max}$.*

Proof. To prove this claim, we need to show that $|g'(0)| < 1$. We thus calculate $|g'(0)|$:

$$\begin{aligned} |g'(0)| &= |1 - \lambda + (\alpha + \beta)F'(0) - S\beta F'(0)| \\ &= |1 - \lambda + (\alpha + \beta) \times 1 - S\beta \times 1| \\ &= |(1 + \alpha + \beta) - (\lambda + S\beta)| \end{aligned} \tag{7}$$

In the second step, we used the fact that $F'(0) = 1$. We thus need to distinguish two cases, depending on whether the term in the absolute value is positive or negative.

The term inside the absolute value is positive for $S < 1 + \frac{\alpha-\lambda}{\beta} + \frac{1}{\beta} = \frac{1}{\beta} + S_{\max}$. Further, this term is smaller than 1 for $S > S_{\max}$. In other words, we have convergence for $S_{\max} < S < S_{\max} + \frac{1}{\beta}$.

The term inside the absolute value is negative for $S > S_{\max} + \frac{1}{\beta}$. Further, it is greater

than -1 for $S < S_{\max} + \frac{2}{\beta}$. Combining these conditions, we find convergence to zero for

$$S_{\max} < S < S_{\max} + \frac{2}{\beta}$$

Does this imply that there is no convergence for $S > S_{\max} + \frac{2}{\beta}$? Equation (1) shows that this is not the case: Any solution to Equation (1) with a larger S will be strictly smaller than a solution with a smaller S. Hence, if there is convergence to zero for $S < S_{\max} + \frac{2}{\beta}$, there will be convergence for $S \geq S_{\max} + \frac{2}{\beta}$ as well. □

Claim 2 (Convergence to $\hat{x}$)**.** *If $\hat{x} = \frac{\alpha-(S-1)\beta}{\lambda} - 1$ is a fixed point of Equation* (2) *in the special case that all active neurons have the same activation, there is convergence to that fixed point for $S < S_{max}$.*

Proof. We first note that $\hat{x}$ is the unique stable fixed point of Equation (2) for $S < S_{\max}$. While $\hat{x}$ is not a stable point of the entire phase space, we show in Appendix C that, in the special case where all activations are identical, $\hat{x}$ is stable, because both the minimum activation and the maximum activation have stable fixed points, which coincide in the special case where all activations are identical. Further, we showed above that the other fixed point at zero is not stable.

To show that $\hat{x}$ is a stable fixed point, we need to calculate $|g'(\hat{x})|$, noting that $F'(\hat{x}) = \frac{\lambda^2}{(\alpha+\beta-S\beta)^2}$. We thus obtain:

$$\begin{aligned} |g'(\hat{x})| &= \left|1 - \lambda + \frac{(\alpha+\beta)\lambda^2}{(\alpha+\beta-S\beta)^2} - \frac{S\beta\lambda^2}{(\alpha+\beta-S\beta)^2}\right| \\ &= \left|\left(1 + \frac{(\alpha+\beta)\lambda^2}{(\alpha+\beta-S\beta)^2}\right) - \left(\lambda + \frac{S\beta\lambda^2}{(\alpha+\beta-S\beta)^2}\right)\right| \end{aligned} \tag{8}$$

The fixed point $\hat{x}$ exists only for $S < S_{\max}$; further, we assume that $\lambda \leq 1$; in this case, the right hand side term in the absolute value is strictly smaller than the left hand side term, and it is sufficient to ask if $g'(\hat{x}) < 1$.

We now rewrite Equation (8), and note that we can express $\alpha + \beta = \beta S_{\max} + \lambda$:

$$\begin{aligned}
|g'(\hat{x})| &= (1-\lambda) + \frac{\lambda^2}{(\alpha+\beta-S\beta)^2}\left((\alpha+\beta)-S\beta\right) \\
&= (1-\lambda) + \frac{\lambda^2}{(\beta S_{\max}+\lambda-S\beta)^2}\left((\beta S_{\max}+\lambda)-S\beta\right) \\
&< (1-\lambda) + \frac{\lambda^2}{(\beta S_{\max}+\lambda-S_{\max}\beta)^2}\left((\beta S+\lambda)-S\beta\right) \\
&= 1-\lambda+\frac{\lambda^3}{\lambda^2} = 1
\end{aligned}$$

□

The results above are specific to Sengupta et al.'s (2014) and Knops et al.'s (2014) rational sigmoid activation function and to the symmetric case where the multivariable non-linear system reduces to a non-linear system of a single variable. In general, it is difficult to provide a precise characterization of the network evolution for arbitrary activation functions. Nonetheless, we show in Appendix B.1.1 that all bounded activation functions admit a fixed point and thus a stationary solution, though we do not know whether these fixed points are convergent or not.

Further, we show in Appendix B.1.2 that, if the slope of the activation function is sufficiently small at zero (more precisely, smaller than $\lambda/(\alpha+\beta)$), all neurons, whether they have been activated or not, will end up having approximately the same level of activation. In other words, for activation functions with this property (e.g., the logistic function), there

is total loss of information.

Stability of the fixed points. Neither of these fixed points is stable. That is, if the network enters a sufficiently small neighborhood of one of the steady-state solutions $x_i = 0$ or $x_i = \hat{x}$, a small perturbation will drive the system away from this neighborhood at the next time-step. This result has an intuitive explanation. First, in the case where all neurons in the network are off, excitation of any single neuron will drive the network away from the switched off state, as this neuron will only receive self-excitation and some forgetting, but no inhibition. This is just to say that, after the network has been swiped clear through catastrophic (processing) interference, it is ready to learn again.

Second, if the network settles at $\hat{x}(S)$ from (6), exciting a single neuron will increase the self-excitation it receives, but also the inhibition it sends to all other neurons, which can increase the separation between the excited neuron and all other neurons. In Appendix C we demonstrate these results more formally.

Critically, however, we also show in Appendix C.2 that, as long as $S < S_{max}$, the minimum activation $\min_{i \leq S}(x_i)$ has a stable upper bound, and the maximum activation $\max_{i \leq S}(x_i)$ has a stable lower bound; both bounds are given by $\hat{x}(S)$. As a result, in the special case where all activations are equal, there is convergence to $\hat{x}(S)$.

Steady state solutions for the average activation. Sengupta et al. (2014) and Knops et al. (2014) used the average activation in the network as a measure of the network's content. To compute the average steady-state activation across all neurons, we have to multiply the individual steady-state activation with S/N, where S is the set-size (and thus the number of active neurons) and N the total number of neurons. After some algebra, the average activation in the network is given by

$$\widehat{\widehat{x}}(S) = \begin{cases} -\frac{\beta}{\lambda N}\left(S^2 - S_{\max}S\right) & S \leq S_{\max} \\ 0 & S > S_{\max} \end{cases} \tag{9}$$

This is a quadratic function with roots at 0 and $S_{\max}$, a maximum at $S_{\max}/2$ and which continues at zero beyond $S_{\max}$. In other words, with simultaneous presentation, the network shows catastrophic (processing) interference. It can remember up to $S_{\max}$ items. However, if more than $S_{\max}$ items are presented, *all* of them are forgotten. As argued above, this network behavior is inconsistent with human behavior in WM experiments.

Faithfulness. Based on the above results, we can also calculate the network faithfulness for the noiseless situation. We call the final steady-state activation of the i^{th}neuron $\widehat{x}_i$, and assume, without restriction of generality, that the first S neurons initially receive external input. The steady-state activation of neuron i is then given by

$$\widehat{x}_i = \begin{cases} \frac{\alpha-\beta(S-1)}{\lambda} - 1 & i \leq S \leq S_{\max} \\ 0 & i > S \\ 0 & S > S_{\max} \end{cases} \tag{10}$$

This is simply because neurons which did not receive external input (i.e., for $i > S$) will remain silent. As a result, the activation of the last $N - S$ neurons necessarily corresponds to their initial input of zero. For the neurons that received external input, the final state activation is set to one if their activation exceeds some threshold ε, that is, if

$$\left|\frac{\alpha - \beta(S-1)}{\lambda} - 1\right| > \varepsilon \tag{11}$$

This leads to two possible cases, depending on whether the term inside the absolute value is positive or negative. It is positive for $S < S_{\max}$, and negative for $S > S_{\max}$. Given that $\widehat{x}_i$ is zero for all i when $S > S_{\max}$, we only need to consider the first case. For $S < S_{\max}$, condition (11) is satisfied if and only if

$$S < S_{\max} - \frac{\lambda}{\beta}\varepsilon \tag{12}$$

For $\varepsilon = .03$ and Knops et al.'s (2014) and Sengupta et al.'s (2014) high inhibition situation, this condition yields $S < 8.8$.

In sum, the faithfulness q of the network output is given by

$$q(S) = \begin{cases} 1 & S < S_{\max} - \frac{\lambda}{\beta}\varepsilon \\ 1 - \frac{S}{N} & \text{otherwise} \end{cases} \tag{13}$$

This function is constant up to $S_{\max} - \frac{\lambda}{\beta}\varepsilon$, then has a jump (i.e., is discontinuous), and then decreases linearly to zero for $S = N$.

Sequential presentation prevents catastrophic (processing) interference and leads to a capacity-limited memory buffer

Simulations. The previous simulations and analyses show that the network shows catastrophic (processing) interference when more than $S_{\max}$ items are presented: the network will forget *all* items. We now show that this catastrophic interference can be prevented by presenting items sequentially, one after the other. Each item was presented for 5 cycles; after all items had been presented, the network was allowed to settle to a steady state for another 45 cycles. Figure 3(A) shows that, in the absence of noise and with strong inhibition, the mean activation of the network increases for small set-sizes up to approximately $S_{\max}/2$, and then reaches an asymptote. Figure 3(B) shows that memory is perfect up to approximately $S_{\max}/2$, and then gradually becomes worse as the set-size increases. Crucially, the proportion of correct activations does not converge to zero, but rather shows an intercept of about $(S_{\max}\text{-}1)/N/2$, suggesting that $(S_{\max}\text{-}1)/2$ neurons remain active. Accordingly, Figure 3(C) shows that the discriminability between old and new items is maintained beyond $S_{\max}$ when items are presented sequentially. Sequential presentation thus protects the network from catastrophic interference in that it enables the network to retain some items even when the set-size massively exceeds $S_{\max}$. That being said, the reason for which the network remembers only the last few items rather than all items is still retroactive interference from the most recent items; critically, however, this interference no longer removes *all* memory traces.

To analyze the content of the memory, we first calculated the number of active neurons at the end of the simulations (see Figure 4(A) for the noiseless case, and Figure 4(D) for

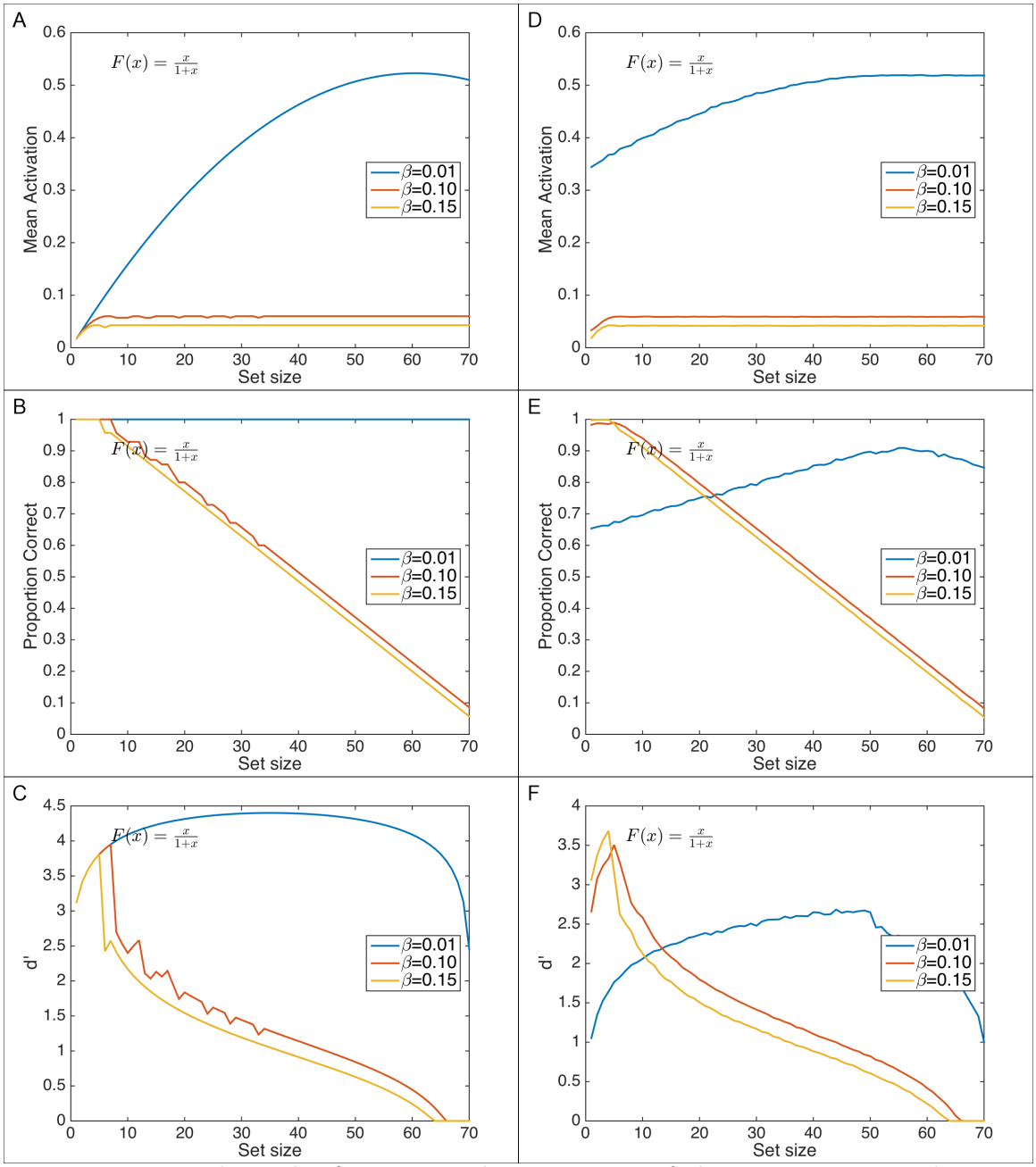

Figure 3. Network results for sequential presentation of the memory items where one memory item is presented after the other, for three different levels of inhibition (low, $\beta = .01$, blue line; medium, $\beta = .1$, red line; high, $\beta = .15$, orange line). The proportion of correct responses is the complement of the Hamming distance between the input and the steady-state activation at end of the simulation. Simulations without noise (A, B, C). (A) The mean activation in the network initially increases, and then remains at an asumptote. (B) The proportion of correct activations is 1 up to $S_{\max}/2$, and then linearly decreases to zero. (C) The discriminability index d' remains positive beyond $S_{\max}$. (D, E, F). The simulations in the presence of noise yield similar results, except that the results are smoothed.

the noisy case). We found that ($S_{\max}$-1)/2 neurons remained active. We then calculated the average activation in the active neurons (see Figure 4(B) for the noiseless case, and Figure 4(E) for the noisy case), and found that it remained relatively high. Finally, and as shown in Figures 4(C) and (F), we calculated (i) the mean activation in the ($S_{\max}$-1)/2 neurons that received the most *recent* external inputs, (ii) the mean activation in a random subset of ($S_{\max}$-1)/2 neurons that received *earlier* inputs, and (iii) the overall mean activation. As shown in Figure 4(C), the activation remained high in the neurons with the most recent inputs, and was almost zero in neurons that have received earlier inputs. In other words, the network maintains a buffer of the last ($S_{\max}$-1)/2 memory items.

In contrast, when inhibition is low, the network storage capacity is fundamentally unbounded (see Figures 3(B) and 4(A)), in line with earlier analyses (Endress & Szabó, 2017) and results (Endress & Potter, 2014).

In Appendix B, we show that other activation functions than Sengupta et al.'s (2014) and Knops et al.'s (2014) rational sigmoid function yield qualitatively similar results in that sequential presentation protects the network from catastrophic (processing) interference — as long as the activation function is bounded and its slope sufficiently strong at zero.

In Appendix E, we also consider a different kind of sequential presentation, where the network is given 45 time steps to reach a stable activation pattern after *each* input. The results are similar to those below, except that the network behavior becomes oscillatory.

Analytic derivation. An analytic solution to (1) is more difficult to find for the general case where items are presented sequentially. Nevertheless, we can show that the difference between the maximal activation and the minimal activation remains positive (if it ever was positive). As a result, if the maximal activation is greater than the minimal

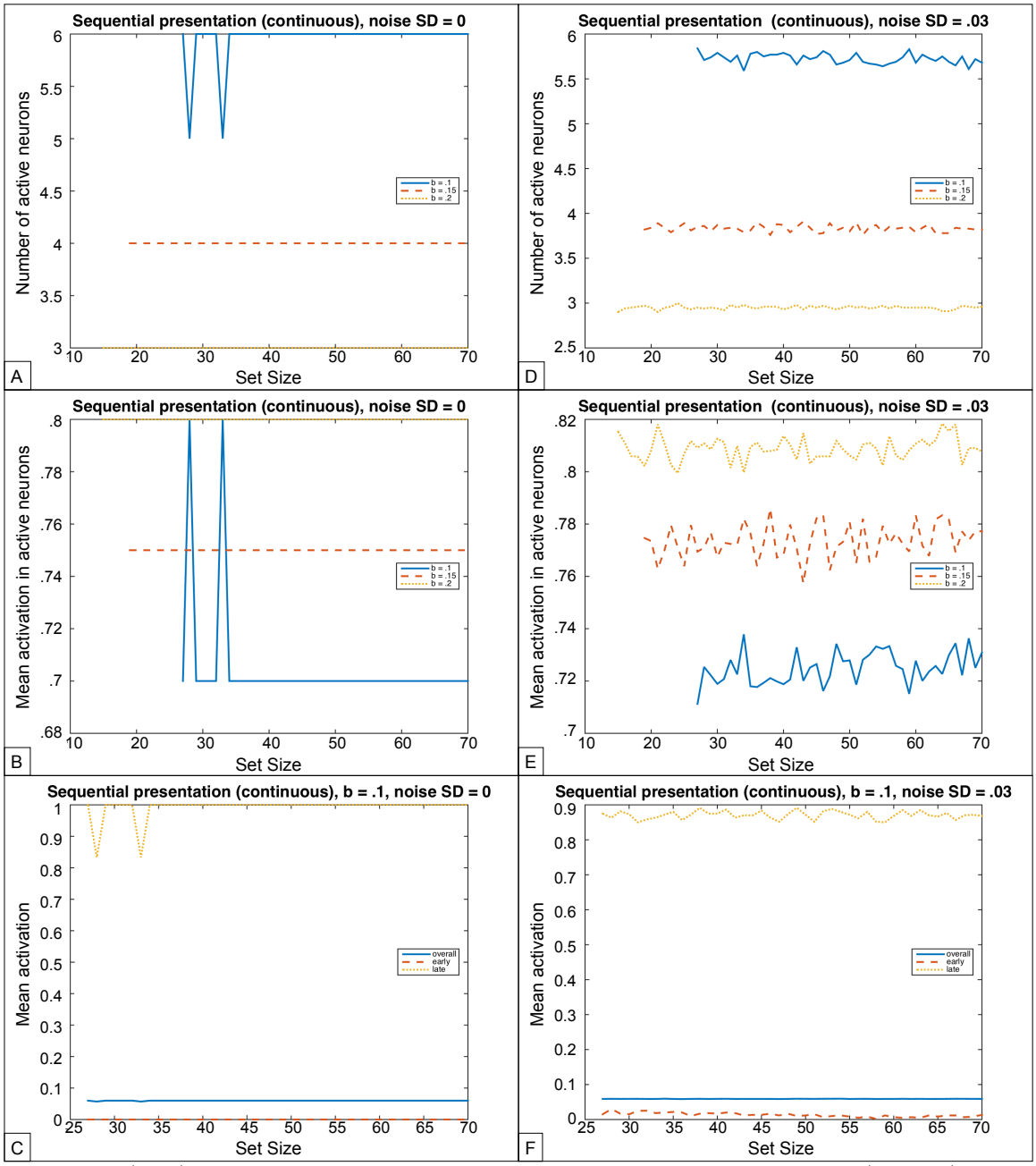

Figure 4. (Top) Number of active neurons at the end of the simulations. (Middle) Mean activation in the active neurons at the end of the simulations. (Bottom) Average activation in the most recent ($S_{\max}$-1)/2 target neurons compared to a random subset of ($S_{\max}$-1)/2 target neurons presented earlier during the input. When each input is presented immediately after the previous one, there is steady activation for the last inputs, and zero activation for the earlier targets. (Left) Without noise. (Right). With noise.

activation, at least the maximal activation must remain positive.

Specifically, we show that, even if the difference between the minimum and the maximum decreases, as soon as the maximal activation decreases below $\sqrt{(\alpha+\beta)/\lambda}-1$, the difference between the maximum and the minimum starts increasing again. Given that the minimum cannot be smaller than zero, the maximum must be at least as large as the difference between the maximum and the minimum. As a result, if the difference stays positive, at least some neurons must remain active.

In Appendix D, we will show that this difference is strictly greater than a function that increases if the maximum descends below $\sqrt{(\alpha+\beta)/\lambda}-1$. As a result, the network activation cannot disappear as long as there was some initial activation; further, $\sqrt{(\alpha+\beta)/\lambda}-1$ is a lower bound for this activation.

The role of network topology

So far we have shown that the network suffers from catastrophic interference if all items are presented simultaneously: it remembers items perfectly up to a set-size $S_{\max}$, and forgets *all* items if more than $S_{\max}$ items are presented. In contrast, if items are presented one after the other, the network maintains a trace of the last few items. We will now explore the consequences of the network structure.

Knops et al.'s (2014) and Sengupta et al.'s (2014) network is fully connected. To assess the importance of this choice, we limited the number of inhibitory connections. That is, we arranged the 70 neurons on a 10×7 grid, and manipulated the distance within which inhibition took place. For example, if inhibition takes place only among nearest neighbors, all neurons providing inhibitory input can be reached by going up or down and/or left or

right by one step on the grid, and are thus contained within a sphere of radius $\sqrt{2}$. More generally, if interacting neurons can be reached in d_s (horizontal and/or vertical) steps (so that diagonal displacement is allowed as well), they are contained in a sphere of radius $\sqrt{2}d_s$. This sphere comprises $(2d_s+1)^2-1$ inhibiting neurons. In other words, the distance in which interacting neurons can be reached is simply a measure of the number of effectively interacting neighbors.

In Figure 5, we plot the mean network activation and the faithfulness as a function of the set-size and the maximal number of steps in which interacting neurons must be reached. The results show that memory performance is somewhat better when inhibition takes place between nearest neighbors only. This is because restricting the distance at which inhibitory interactions can take place also restricts the number of neurons that can interfere with each other. In a local neighborhood, the number of effectively interacting neurons is thus more likely to remain below $S_{\max}$. In line with this interpretation, as soon as interactions are allowed to take place between neurons that are two or three steps apart, the pattern of result becomes very similar to that of the fully connected network.

In Appendix F, we repeat these simulations with a larger network of 400 neurons, arranged on a 20×20 grid, with a maximum set-size of 70. As a result, the activation in this network is much sparser than in the 10×7 network. Unsurprisingly, when long-range inhibitory interactions are allowed, results are similar to the 10×7 network; in contrast, when only short-range inhibitory interactions are allowed, performance remains excellent for larger set-sizes as well, because the effective inhibitory input of each neuron is lower, and because the network has a large storage capacity for low levels of inhibition.

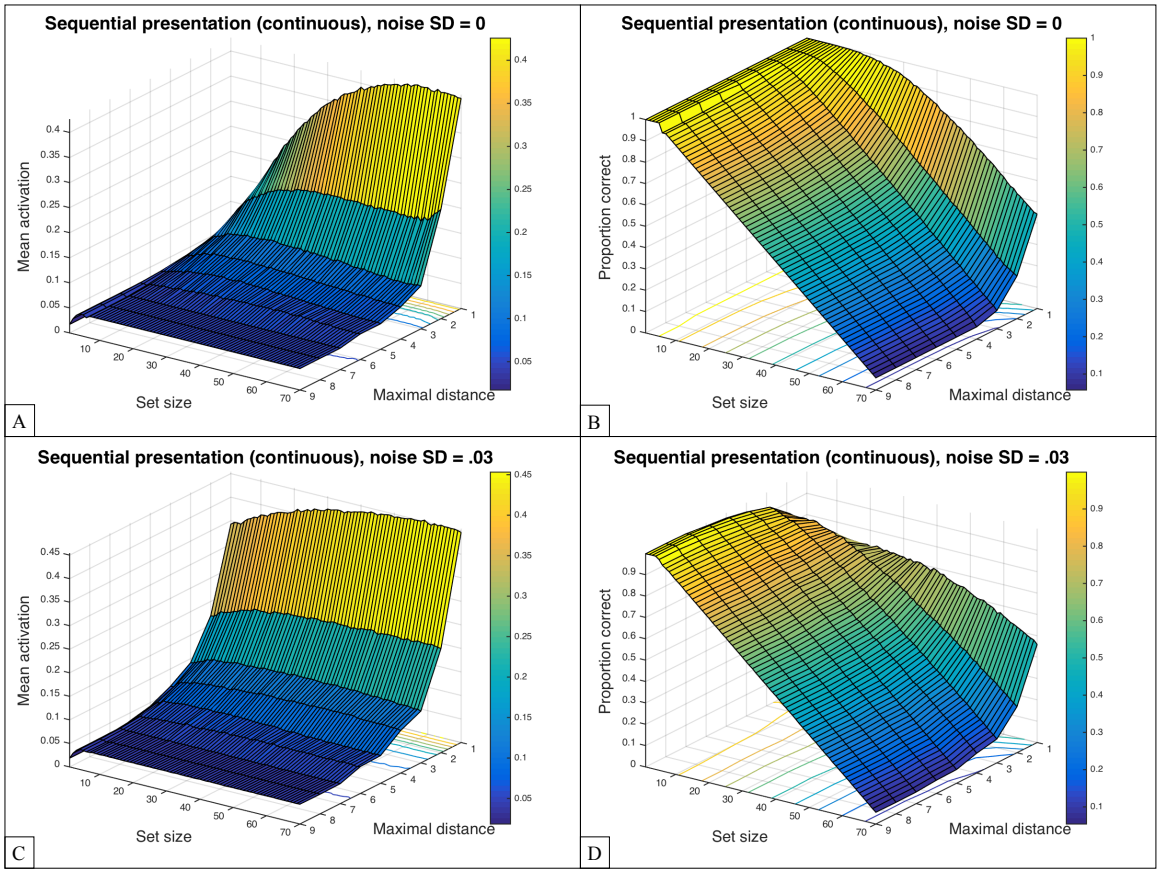

Figure 5. Network results for sequential presentation of the memory items where one memory item is presented after the other. The inhibition parameter is set to $\beta = .15$. The maximal distance is the maximal number of steps on the grid in which mutually inhibiting neurons can be reached. The proportion of correct responses is the complement of the Hamming distance between the input and the steady-state activation at end of the simulation. (A) Mean network activation in the absence of noise. (B) Faithfulness in the absence of noise. (C) Mean network activation in the presence of noise. (D) Faithfulness in the presence of noise. When inhibition is allowed between neurons that are 2 or 3 steps apart (in the horizontal and/or vertical direction, so that diagonal displacements are allowed as well), the network behavior is similar to the fully connected situation.

Mutual inhibition and small number processing

The results so far suggest that the network can remember only up to a certain number of items when the items are presented simultaneously and shows catastrophic interference beyond this critical set-size. Further, we have shown that this critical set-size depends on the effective connectivity between neurons. In contrast, when items are presented sequentially, the network retains a buffer of the last few items. However, this network was also proposed to account for the processing of both small and larger numbers. We now turn to this issue.

As Knops et al. (2014) and Sengupta et al. (2014), we find that, for medium and high levels of inhibition, the mean activation of the network increases within the subitizing range as a function of the set-size, for both simultaneous and sequential presentation. As a result, if organisms are endowed with a readout mechanisms that can interpret the mean activation over the entire saliency map (such as the summation units in the neural network models by Dehaene & Changeux, 1993, and Verguts & Fias, 2004), this information might be used for number identification in the subitizing range.

Mutual inhibition and large number processing

Knops et al. (2014) and Sengupta et al. (2014) suggested that the mean activation in the network at low levels of inhibition might be used to estimate numerosities in the approximate number systems range (i.e., for numerosities above 4). However, for the mean network activation to be diagnostic of the number of items, the relationship between the mean activation and the numerosity must be monotonic, and Figure 2 shows that this is not the case, especially in the presence of noise.

To illustrate this issue, we simulated a numerosity comparison task using the same

strategy as Sengupta et al. (2014). Specifically, the network had to "decide" whether or not a given test numerosity was larger than a reference numerosity (in our case 16). To simulate this task, we calculated the average mean network activation for the reference numerosity across 100 simulations, providing us with a reference mean network activation. For all test numerosities, we then counted the proportion of simulations (out of 100 simulations) where the mean network activation (i) was smaller than the reference activation by a margin of .0001 (receiving a score of 0), (ii) larger than the reference activation by a margin of .0001 (receiving a score of 1), or (iii) equal to the reference activation (that is, within these margins; receiving a score of .5). We then averaged these scores across the 100 simulations for each numerosity. Based on human data, we would expect these scores to describe a sigmoid function of the ratio between the test numerosity and the reference numerosity, as large number discrimination follows a Weber law (e.g., Izard & Dehaene, 2008; Whalen, Gallistel, & Gelman, 1999; Halberda, Ly, Wilmer, Naiman, & Germine, 2012).

As shown in Figure 6, medium and high inhibition levels clearly do not show a sigmoid profile; we thus focus on the low inhibition situation. Figure 6(A) shows that, in the absence of noise, and when the reference numerosity is 16, the network judges lower numerosities to be smaller than 16, and higher numersities to be greater. However, given that the mean network activation is a quadratic function of the numerosity, this is not the case when the reference numerosity is close to the maximum. As shown in Figure 6(B), when choosing 59 as the reference numerosity (because, for the low inhibition situation modeled here, it is close to the maximum of the mean network activation of 60.5), the only numerosities judged greater than 59 are 60 and 61, but all numerosities larger than 62 items are judged *smaller* than 59. The network thus does not successfully discriminate large numbers.

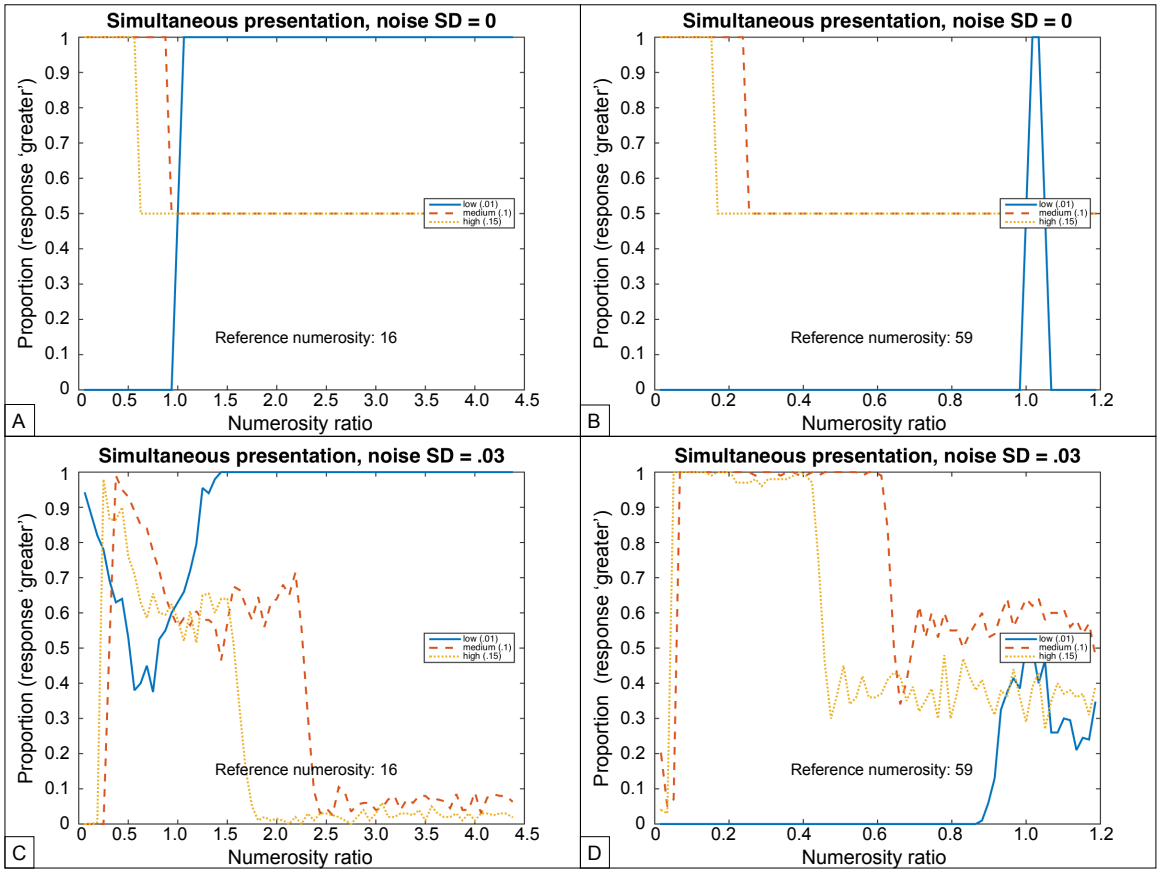

Figure 6. Number comparison performance for the reference numerosity of 16 (left) or 59 (right), in the absence (top) presence (bottom) of noise. The inputs are presented simultaneously. The y-axis shows the proportion of simulations for which the mean network activation was larger than that for the reference numerosity.

As shown in Figures 6(C) and (D), the network does not succeed in discriminating large numbers in the presence of noise either. As a result, at least with simultaneous presentation, mutual inhibition does not seem to explain the Weber signature of the approximate number system.

Figure 7 shows that the comparison performance is not sigmoid with sequential presentation either, reflecting again the fact that the relationship between the mean network activation and the numerosity is not monotonic.

There is also a more general reason for which mutual inhibition is unlikely to account

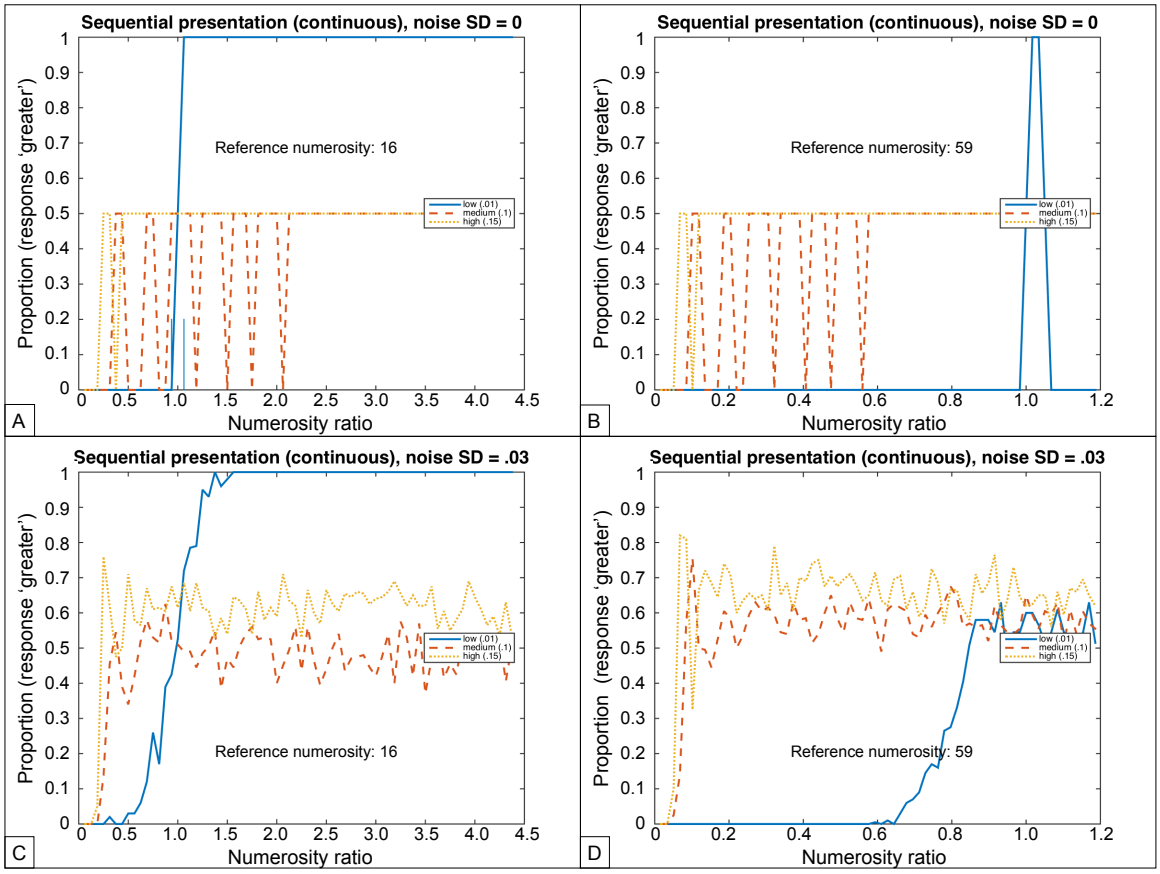

Figure 7. Number comparison performance for the reference numerosity of 16 (left) or 59 (right), in the absence (top) or presence (bottom) of noise. The inputs are presented sequentially, with no consolidation intervals between presentations. The y-axis shows the proportion of simulations for which the mean network activation was larger than that for the reference numerosity.

for large number processing in humans. In the current model, noise is independent of the set-size, while, for large number discriminations, noise must scale with the set-size to account for the Weber signature (e.g., Gallistel & Gelman, 2000; Izard & Dehaene, 2008). Critically, however, in the General Discussion, we suggest that this property might not require any specific explanation, as it is expected in any system that combines a limited precision with a gain control mechanism.

General Discussion

In the current analyses, we asked to what extent a simple saliency map based on mutual inhibition among neurons can explain human performance in Working Memory (WM), small number processing and large number processing. We found four crucial results. First, mutual inhibition among neurons leads to catastrophic interference if memory items are presented simultaneously. Items are remembered up to the memory capacity; when more items are presented, the network forgets *all* of them. For activation functions with $F'(0) < \frac{\lambda}{\alpha+\beta}$, there is even total loss of information irrespective of how many items are presented.

Second, sequential presentation protects the network from catastrophic interference; with sequential presentation, the network retains the most recent items as long as the activation function is bounded and has a sufficiently strong slope at zero.

Third, we confirm earlier results that the mean activation level in the network might in principle be used to enumerate small sets of items (i.e., in the subitizing range). Fourth, we showed that, in contrast to the small number situation, mutual inhibition does not account for number discrimination in the range of the approximate number system. Rather, we suggest below that the Weber signature of the large number system might follow from the combination of a limited precision and a multiplicative gain control mechanism.

We will now discuss to what extent the model behavior matches human performance in the three domains investigated — WM, small number processing and large number processing. Following this, we will raise the more general question of whether these domains of processing have a common processing capacity.

Does a saliency map based on mutual inhibition exhibit human-like behavior in Working Memory?

When items are presented simultaneously, Sengupta et al.'s (2014) and Knops et al.'s (2014) model shows catastrophic (processing) interference: when more than $S_{\max}$ items are presented, all of them are forgotten. In contrast, when items are presented sequentially one after the other, the network maintains a buffer of the last few items.

At first sight, these results seem to suggest that the model does not account for human performance; we clearly do not forget all items in WM when a single additional item is presented. However, the finding that catastrophic (processing) interference is preventable through sequential presentation also raises the possibility that participants behave optimally by sequentially allocating attention to simultaneously presented items in WM tasks. Specifically, participants attend objects in WM experiments one after another, even when few objects are presented, and even when they are presented simultaneously (e.g., Liu & Becker, 2013; Vogel et al., 2006; but see Mance et al., 2012). While sequentially attending items has been interpreted as an attentional limitation, such an attentional strategy might well be adaptive, if sequentially attending to items avoids catastrophic interference (see also Ihssen et al., 2010, for related results). If so, items might be maintained in WM by attentionally cycling through the memory items, similarly to how people sequentially cycle through verbal memory items when they (sub-vocally) rehearse and refresh them.

Does a saliency map based on mutual inhibition exhibit human-like behavior in small number processing?

While catastrophic interference does not seem to be a feature of WM (at least in adults), catastrophic capacity limitations are well documented in number processing in infants (e.g., Feigenson et al., 2002; Feigenson & Carey, 2005; Zosh & Feigenson, 2015). As mentioned above, infants reliably choose two or three food items over a single food item, but they have no preference for four items over a single food item. The traditional explanation is that, in infants, the small number system can process only numerosities up to 3; beyond 3, the approximate number system must be used. As the two number systems are mutually incompatible (see e.g. Izard & Dehaene, 2008; Revkin et al., 2008, for evidence in adults that these systems have different properties), the infant has no way to compare the numerosities of numbers below 3 and above 3. Further, given that the small number system might really be an attentional system (e.g., an object file system; Feigenson et al., 2002; Feigenson & Carey, 2003; Piazza et al., 2011), it is particularly plausible that an attentional device such as a saliency map might account for small number processing.

Does a saliency map based on mutual inhibition exhibit human-like behavior in large number processing?

While the saliency map-like architecture might account for small number processing, the model fails to account for large number processing, because the mean network activation is not a monotonic function of the number of items presented to the network. There are also two more theoretical reasons suggestion that a saliency-based model is unlikely to account for large number processing.

First, the mutual inhibition model is fundamentally spatial, while even newborns can recognize large numerosities across modalities (e.g., Izard, Sann, Spelke, & Streri, 2009). For example, when familiarized with sequences of 4 sounds, newborns look longer to displays with 4 items than with 12 items, and vice-versa. It is unclear how a system modeled after spatial attention would account for such results.

Second, and crucially, given that participants can estimate 100 dots on a screen within 100 ms (e.g., Izard & Dehaene, 2008), and that a saliency map is fundamentally an attention device, we believe that it is simply implausible that participants can (covertly) orient to the so many dots so quickly. This also fit with the respective attentional demands of small vs. large number processing. While small number processing requires attention, large number processing does not (e.g., Burr et al., 2010; Gliksman et al., 2016; see also Railo, Koivisto, Revonsuo, & Hannula, 2008). As a result, an attentional mechanisms is unlikely to be at the root of large number processing by the approximate number system (though such a mechanism is likely involved in sequentially *counting* items).

Do the properties of large number processing result from gain control in a system with a limited precision?

While an attention-based mechanisms is unlikely to provide an explanation for large number processing, we suggest that, to the extent that large number discriminations operate similarly to other sensory systems, no explanation for its Weber signature is required as it follows naturally from a phenomenon that is as ubiquitous as Weber signatures: multiplicative gain control (e.g., Abbott, Varela, Sen, & Nelson, 1997; Priebe & Ferster, 2002; Rudd & Brown, 1997; Salinas & Thier, 2000; Willmore, Cooke, & King, 2014).

The basic fact to explain is that the difficulty of discriminating two numbers depends on their ratio rather than their absolute difference. For example, it is easier to discriminate 15 from 10 (ratio 1.5) than 600 from 500 (ratio 1.2) even though the absolute difference is much larger in the second case. Levi (1969) proposed that such Weber ratio signatures follow from the presence of multiplicative gain control, and that gain control also explained the fact that Weber's law is a more accurate approximation at higher rather than lower stimulus intensities. Further, Cicchini, Anobile, and Burr (2014) argued that, outside the domain of number processing, gain control seems to be the usual explanation for Weber signatures. However, as this model is rarely made explicit, and many authors explain Weber ratio signatures based on assumptions about the format and distribution of the mental representations of quantities that are not always independently motivated (e.g., Dehaene, Izard, Spelke, & Pica, 2008; Feigenson et al., 2004; Izard & Dehaene, 2008; Nieder, 2017; Pardo-Vazquez et al., 2019; Piazza, Izard, Pinel, Bihan, & Dehaene, 2004; Whalen et al., 1999), we present the gain control model in some detail in its purest and simplest form.

The idea is analogous to the precision of the display of a pocket calculator. Specifically, if the display has three digits, we can represent numbers up to 999 exactly, which allows us to discriminate 100 from, say, 101. However, the representation of 1,000 would be 100e01; hence, we can discriminate 100e01 from 101e01 (i.e., 1,000 from 1,010), but not 1,000 from 1,001 because only the first three digits are represented. Likewise, the representation of 10,000 would be 100e02; hence, we can discriminate 100e02 from 101e02 (i.e., 10,000 from 10,100), but not 10,000 from 10,010. More generally, for a number of magnitude 10^{3+a}, there is uncertainty about the last a digits, leading to an average error of $< 0 \ldots (10^a - 1) > \approx 5 \times 10^{a-1}$. As a result, for a system representing a quantity with a fixed precision and

multiplicative gain control, the uncertainty about the quantity is necessarily proportional to the quantity, with a ratio of $\frac{5\times10^{a-1}}{10^{3+a}} = .05\%$. Given the ubiquity of multiplicative gain control in the brain (e.g., Priebe & Ferster, 2002; Salinas & Thier, 2000), we thus suggest that the Weber signature of the approximate number system might be an inevitable consequence of the fact that we have a system that discriminates large quantities.

Gain control also explains why quantities seem to be represented logarithmically (e.g., Dehaene et al., 2008; Izard & Dehaene, 2008; Piazza et al., 2004; Whalen et al., 1999). For example, the difference in pitch of any two adjacent (white or black) keys on a (well tempered) piano is perceived as being the same, but the pitches are really separated by the same frequency *ratio* rather than by the same frequency difference. The frequency difference between the middle C and the next semitone is 15.55 Hz, while, one octave higher, the frequency difference between C and the next semitone is 31.1 Hz; in both cases, however, the *relative* difference is about 6%.

The fact that constant relative differences leads to the perception of equally spaced stimuli follows naturally from multiplicative gain control. If we try to discriminate two pairs of *physical* quantities x_1 and x_2 (e.g., the frequency of the middle C and of the next semitone) and z_1 and z_2 (e.g., the frequency of these tones shifted by one octave), the *internal representations* of these quantities will be scaled by a factor that is roughly proportional to these quantities, that is $\xi_1 \approx \frac{x_1}{x_1} = 1$; $\xi_2 \approx \frac{x_2}{x_1}$; $\zeta_1 \approx \frac{z_1}{z_1} = 1$; $\zeta_2 \approx \frac{z_2}{z_1}$. The difference in terms of the internal representations is thus simply the ratio of the physical quantities. Semitones (or more generally quantities with constant ratios) are therefore perceived as equidistant, which is just to say that the quantities appear to be represented logarithmically.

For the same reason, noise in the internal representations of quantities might appear

to be log-normally distributed (e.g., Izard & Dehaene, 2008). As the internal representations are scaled through gain control, normally distributed noise in the internal representations appears to be log-normally distributed if the internal representations are transformed back to the original physical quantities.

In line with this view, Cicchini et al. (2014) showed that, in an estimation task (where observers had to estimate rather than compare quantities), quantities on previous trials affected estimates on later trials, and suggested that observers might use information from previous trials to adjust the "gain" for the representations.

A simple illustration of this model is given in Figure 8; the code for the model is available at `https://doi.org/10.25383/city.10751186`. In this model, we ask an observer to compare two numbers. The first number establishes the gain; we model this by first scaling the number to fit between 0 and 10, and then centering it to five; the same scaling and centering is then applied to the second number. The limited precision is implemented by rounding the numbers to 2 significant digits, and by restricting the possible range of number to between 0 and 10. However, the results are largely identical when the representations are not rounded, as the presence of noise (see below) makes the last few digits uninformative in any case.

We then add Gaussian noise (with a mean of zero) to both numbers. As mentioned above, this assumption is equivalent to previous assumptions that the noise in a representation is proportional to its magnitude (e.g., Izard & Dehaene, 2008; Whalen et al., 1999). Finally, we ask whether the internal representation is greater for the first or the second number. We ran 500 simulations for every noise-level/ratio combination. Of course, biologically systems do not necessarily operate on a decimal basis, but the choice of the basis just

affects the quantitative predictions and needs to be determined empirically.

The results are shown in Figure 8, for when the model needs to compare numbers to 5 (a) or to 500 (b). Performance is better for more discriminable ratios, and follows the expected psychometric function (e.g., Halberda, Mazzocco, & Feigenson, 2008; Halberda et al., 2012), whose slope reflects the strength of the internal noise. Further, in Appendix G, we show that there is a very simple mathematical reason why gain control *necessarily* implies a sigmoid response profile. Critically, the results are virtually identical irrespective of whether the reference number is 5 or 500; performance is exclusively driven by the ratio of the numbers.

Multiplicative gain control thus provides a natural explanation of Weber's law, the apparent logarithmic representation of quantities, the apparently log-normal distribution of noise in the representation of quantities and the sigmoid dependency between the ratio of two quantities and discrimination performance. The generality of Weber's law might thus reflect the ubiquity of multiplicative gain control in the brain.

Interestingly, it has been suggested that multiplicative gain control relies on the balance between excitation and inhibition, though the underlying mechanisms tend to be synaptic, and not due to the number of neurons providing inhibitory input as in the current model (e.g., Ayaz & Chance, 2009; Chance, Abbott, & Reyes, 2002; Prescott & De Koninck, 2003; Rothman, Cathala, Steuber, & Silver, 2009).

That is, to detect the numerosity of a set, one would still need neurons that sum the total network activity as in the models by Dehaene and Changeux (1993) and Verguts and Fias (2004). However, the logarithmic scaling of numerosities characteristic of the Weber signature might be a consequence of multiplicative gain control mechanisms, and not require

an intrinsically logarithmic representation.

Is there a general cognitive capacity?

Our results show that a saliency-map like model shows catastrophic (processing) interference unless items are presented sequentially. This is potentially problematic for a model of WM, but it is desirable for a model of small number processing. In contrast, the properties of large number processing are a necessary consequence of a system with a limited precision and gain control.

Based on these results, we will consider Sengupta et al.'s (2014) and Knops et al.'s (2014) critical question: do WM and small number processing rely on a common system? The evidence for this proposal is somewhat mixed. On the one hand, the most commonly accepted reason for this possibility seems inconsistent with recent evidence. As mentioned above, WM and small number processing might be linked through a system of parallel attention (e.g., Cowan, 2001, 2015; Piazza et al., 2011). However, some WM tasks show only limited interference with attentional tasks (e.g., Fougnie & Marois, 2006; Hollingworth & Maxcey-Richard, 2013; H. Zhang et al., 2010), and WM and attention seem to have different properties (Endress et al., 2017). One the other hand, Piazza et al. (2011) showed that individuals' WM capacity predicts their subitizing range (at least when WM is assessed through change detection paradigms, but see Shimomura & Kumada, 2011; Tuholski et al., 2001).[3] To better understand to what extent these computations are part of a common system or just rely on the same auxiliary systems (e.g., both might require object-based

[3] As mentioned above, Knops et al. (2014) also suggested that WM and small number processing might rely on the same brain regions. However, an alternative interpretation of Knops et al.'s (2014) enumeration profile is that it might select voxels that are *not* involved in small number processing, and only become involved beyond the subitizing range.

attention, but then perform different computations on these representations), we thus need a psychological theories of *how* these computations might be related.

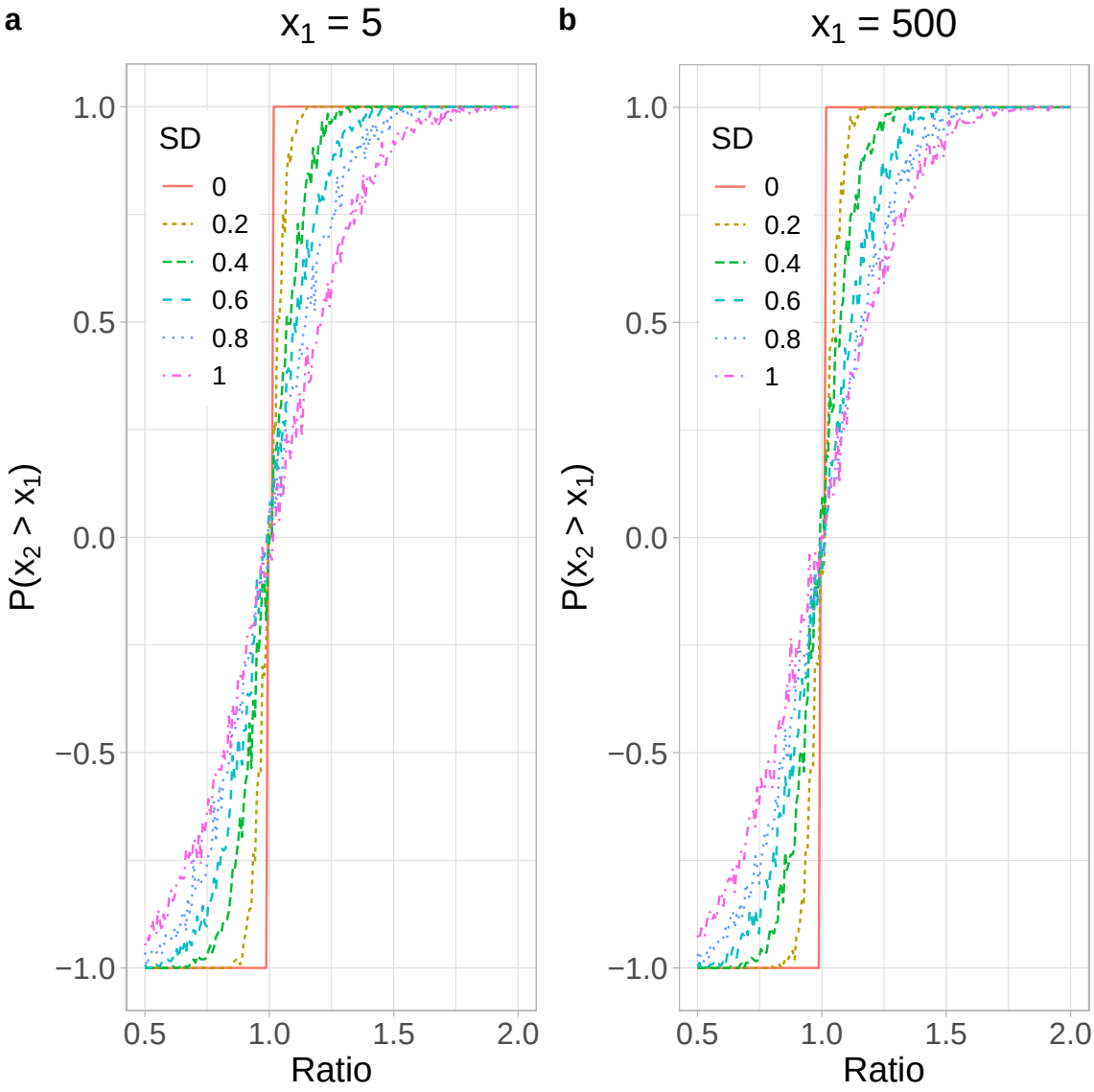

Figure 8. Average decisions whether x_2 is greater than x_1 in a quantity comparison task between two quantities x_1 and x_2 based on the pocket calculator model. Each simulation received a score of 1 when the internal representation of x_2 was greater than that of x_1 and a score of –1 when the internal representation of x_1 was greater than that of x_2. In line with behavioral data, the average decision shows a sigmoid dependency on the ratio between the quantities. The shape of the response profile is indistinguishable irrespective of whether the reference quantity is 5 (a) or 500 (b). In Appendix G, we show that this dependency is necessarily sigmoid.

References

Abbott, L. F., Varela, J. A., Sen, K., & Nelson, S. B. (1997). Synaptic depression and cortical gain control. *Science*, *275*, 220–224.

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*(2), 106–111.

Ansari, D., Lyons, I. M., van Eimeren, L., & Xu, F. (2007). Linking visual attention and number processing in the brain: The role of the temporo-parietal junction in small and large symbolic and nonsymbolic number comparison. *Journal of Cognitive Neuroscience*, *19*(11), 1845–1853.

Ayaz, A., & Chance, F. S. (2009). Gain modulation of neuronal responses by subtractive and divisive mechanisms of inhibition. *Journal of Neurophysiology*, *101*, 958–968. doi: 10.1152/jn.90547.2008

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), 7.1–711. doi: 10.1167/9.10.7

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*(5890), 851–854. doi: 10.1126/science.1158023

Bays, P. M., Singh-Curry, V., Gorgoraptis, N., Driver, J., & Husain, M. (2010). Integration of goal- and stimulus-related visual signals revealed by damage to human parietal cortex. *Journal of Neuroscience*, *30*, 5968–5978. doi: 10.1523/JNEUROSCI.0997-10.2010

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(38), 14325–14329. doi: 10.1073/pnas.0803390105

Burr, D. C., Turi, M., & Anobile, G. (2010). Subitizing but not estimation of numerosity requires attentional resources. *Journal of Vision*, *10*(6), 20. doi: 10.1167/10.6.20

Chance, F. S., Abbott, L. F., & Reyes, A. D. (2002). Gain modulation from background synaptic input. *Neuron*, *35*, 773–782.

Cicchini, G. M., Anobile, G., & Burr, D. C. (2014). Compressive mapping of number to space reflects dynamic encoding mechanisms, not static logarithmic transform. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(21), 7867–7872. doi: 10.1073/pnas.1402785111

Conway, A. R., Kane, M., Bunting, M., Hambrick, D., Wilhelm, O., & Engle, R. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin and Review*, *12*, 769-786.

Cowan, N. (1995). *Attention and memory: An integrated framework.* Oxford, UK: Oxford University Press.

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114.

Cowan, N. (2005). *Working memory capacity.* Hove, UK: Psychology Press.

Cowan, N. (2015). George Miller's magical number of immediate memory in retrospect: Observations on the faltering progression of science. *Psychological Review*, *122*(3), 536–541. doi: 10.1037/a0039035

Dehaene, S., & Changeux, J. P. (1993). Development of elementary numerical abilities: a neuronal model. *Journal of Cognitive Neuroscience*, *5*(4), 390–407. doi: 10.1162/jocn.1993.5.4.390

Dehaene, S., Izard, V., Spelke, E., & Pica, P. (2008). Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. *Science*, *320*(5880), 1217–1220. doi: 10.1126/science.1156540

D'Esposito, M., Postle, B. R., Jonides, J., & Smith, E. E. (1999). The neural substrate and temporal dynamics of interference effects in working memory as revealed by event-related functional MRI. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(13), 7514–7519.

Endress, A. D., Korjoukov, I., & Bonatti, L. L. (2017). Category-based grouping in working memory and multiple object tracking. *Visual Cognition*. doi: 10.1080/13506285.2017.1349229

Endress, A. D., & Potter, M. C. (2014). Large capacity temporary visual memory. *Journal of Experimental Psychology. General*, *143*(2), 548–65. doi: 10.1037/a0033934

Endress, A. D., & Siddique, A. (2016). The cost of proactive interference is constant across presentation conditions. *Acta Psychologica*, *170*, 186 - 194. doi: dx.doi.org/10.1016/j.actpsy.2016.08.001

Endress, A. D., & Szabó, S. (2017). Interference and memory capacity limitations. *Psychological Review*, *124*(5), 551–571. doi: 10.1037/rev0000071

Feigenson, L., & Carey, S. (2003). Tracking individuals via object-files: evidence from infants' manual search. *Devel Sci*, *6*(5), 568-584. doi: 10.1111/1467-7687.00313

Feigenson, L., & Carey, S. (2005). On the limits of infants' quantification of small object arrays. *Cognition*, *97*(3), 295-313. doi: 10.1016/j.cognition.2004.09.010

Feigenson, L., Carey, S., & Hauser, M. (2002). The representations underlying infants' choice of more: Object files versus analog magnitudes. *Psychological Science*, *13*(2), 150-6.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307-14. doi: 10.1016/j.tics.2004.05.002

Fougnie, D., & Marois, R. (2006). Distinct capacity limits for attention and working memory: Evidence from attentive tracking and visual working memory paradigms. *Psychological Science*, *17*(6), 526–534. doi: 10.1111/j.1467-9280.2006.01739.x

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, *3*(4), 128-135.

Fukuda, K., Vogel, E., Mayr, U., & Awh, E. (2010). Quantity, not quality: the relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin and Review*, *17*(5), 673–679. doi: 10.3758/17.5.673

Gallistel, C. R., & Gelman, R. (2000). Non-verbal numerical cognition: From reals to integers. *Trends in cognitive sciences*, *4*(2), 59–65.

Gliksman, Y., Weinbach, N., & Henik, A. (2016). Alerting cues enhance the subitizing process. *Acta*

psychologica, *170*, 139–145. doi: 10.1016/j.actpsy.2016.06.013

Gottlieb, J. (2007). From thought to action: the parietal cortex as a bridge between perception, action, and cognition. *Neuron*, *53*, 9–16.

Gray, R. (1971). *Toeplitz and circulant matrices: a review* (Tech. Rep.). Palo Alto, CA: Information Systems Laboratory Technical Report, Stanford University.

Halberda, J., Ly, R., Wilmer, J. B., Naiman, D. Q., & Germine, L. (2012). Number sense across the lifespan as revealed by a massive internet-based sample. *Proceedings of the National Academy of Sciences*, *109*(28), 11116-11120. doi: 10.1073/pnas.1200196109

Halberda, J., Mazzocco, M. M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*, 665–668. doi: 10.1038/nature07246

Hauser, M. D., Carey, S., & Hauser, L. B. (2000). Spontaneous number representation in semi-free-ranging rhesus monkeys. *Proceedings. Biological Sciences*, *267*(1445), 829-833. doi: 10.1098/rspb.2000.1078

Hautus, M. (1995). Corrections for extreme proportions and their biasing effects on estimated values of *d'*. *Behavior Research Methods*, *27*, 46-51.

Hollingworth, A., & Maxcey-Richard, A. M. (2013). Selective maintenance in visual working memory does not require sustained visual attention. *Journal of Experimental Psychology. Human Perception and Performance*, *39*(4), 1047–1058. doi: 10.1037/a0030238

Ihssen, N., Linden, D. E. J., & Shapiro, K. L. (2010). Improving visual short-term memory by sequencing the stimulus array. *Psychonomic Bulletin and Review*, *17*(5), 680–686. doi: 10.3758/PBR.17.5.680

Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition*, *106*(3), 1221–1247. doi: 10.1016/j.cognition.2007.06.004

Izard, V., Sann, C., Spelke, E. S., & Streri, A. (2009). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 10382–

10385. doi: 10.1073/pnas.0812142106

Kaufman, E. L., & Lord, M. W. (1949). The discrimination of visual number. *The American Journal of Psychology*, *62*, 498–525.

Keating, D. P., & Bobbitt, B. L. (1978). Individual and developmental differences in cognitive-processing components of mental ability. *Child Development*, *49*(1), 155. doi: 10.2307/1128604

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, *114*, 3521–3526. doi: 10.1073/pnas.1611835114

Knops, A., Piazza, M., Sengupta, R., Eger, E., & Melcher, D. (2014). A shared, flexible neural map architecture reflects capacity limits in both visual short-term memory and enumeration. *Journal of Neuroscience*, *34*(30), 9857–9866. doi: 10.1523/JNEUROSCI.2758-13.2014

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Scene memory is more detailed than you think: the role of categories in visual long-term memory. *Psychological Science*, *21*(11), 1551–1556. doi: 10.1177/0956797610385359

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What learning systems do intelligent agents need? Complementary Learning Systems Theory updated. *Trends in Cognitive Sciences*, *20*, 512–534. doi: 10.1016/j.tics.2016.05.004

Levi, L. (1969). Automatic gain control model for vision. *Nature*, *223*, 396–397.

Liu, T., & Becker, M. W. (2013). Serial consolidation of orientation information into visual short-term memory. *Psychological Science*, *24*(6), 1044–50. doi: 10.1177/0956797612464381

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281. doi: 10.1038/36846

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356. doi: 10.1038/nn.3655

Mance, I., Becker, M. W., & Liu, T. (2012). Parallel consolidation of simple features into visual short-term memory. *Journal of Experimental Psychology. Human Perception and Performance*, *38*(2), 429–438. doi: 10.1037/a0023925

Masse, N. Y., Grant, G. D., & Freedman, D. J. (2018). Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *Proceedings of the National Academy of Sciences of the United States of America*, *115*, E10467–E10475. doi: 10.1073/pnas.1803839115

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419–457. doi: 10.1037/0033-295X.102.3.419

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 24, pp. 109–164). Academic Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

Nee, D. E., Jonides, J., & Berman, M. G. (2007). Neural mechanisms of proactive interference-resolution. *Neuroimage*, *38*(4), 740–751. doi: 10.1016/j.neuroimage.2007.07.066

Nieder, A. (2017). Evolution of cognitive and neural solutions enabling numerosity judgements: lessons from primates and corvids. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, *373*. doi: 10.1098/rstb.2016.0514

Oberauer, K., & Lin, H.-Y. (2017). An interference model of visual working memory. *Psychological Review*, *124*, 21–59. doi: 10.1037/rev0000044

Pardo-Vazquez, J. L., Castiñeiras-de Saa, J. R., Valente, M., Damião, I., Costa, T., Vicente, M. I., ... Renart, A. (2019). The mechanistic foundation of weber's law. *Nature neuroscience*, *22*, 1493–1502. doi: 10.1038/s41593-019-0439-7

Piazza, M., Fumarola, A., Chinello, A., & Melcher, D. (2011). Subitizing reflects visuo-spatial object individuation capacity. *Cognition*, *121*(1), 147–153. doi: 10.1016/j.cognition.2011.05.007

Piazza, M., Izard, V., Pinel, P., Bihan, D. L., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, *44*(3), 547-55. doi: 10.1016/j.neuron.2004.10.014

Prescott, S. A., & De Koninck, Y. (2003). Gain control of firing rate by shunting inhibition: roles of synaptic noise and dendritic saturation. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 2076–2081.

Priebe, N. J., & Ferster, D. (2002). A new mechanism for neuronal gain control (or how the gain in brains has mainly been explained). *Neuron*, *35*, 602–604.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vision*, *3*(3), 179–197.

Railo, H., Koivisto, M., Revonsuo, A., & Hannula, M. M. (2008). The role of attention in subitizing. *Cognition*, *107*(1), 82–104. doi: 10.1016/j.cognition.2007.08.004

Ratcliff, R. (1990). Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, *97*, 285–308.

Revkin, S. K., Piazza, M., Izard, V., Cohen, L., & Dehaene, S. (2008). Does subitizing reflect numerical estimation? *Psychological Science*, *19*(6), 607–614. doi: 10.1111/j.1467-9280.2008.02130.x

Roggeman, C., Fias, W., & Verguts, T. (2010). Salience maps in parietal cortex: imaging and computational modeling. *NeuroImage*, *52*, 1005–1014. doi: 10.1016/j.neuroimage.2010.01.060

Roitman, J. D., Brannon, E. M., & Platt, M. L. (2007). Monotonic coding of numerosity in macaque lateral intraparietal area. *PLoS Biology*, *5*(8), 1-11. doi: 10.1371/journal.pbio.0050208

Rothman, J. S., Cathala, L., Steuber, V., & Silver, R. A. (2009). Synaptic depression enables neuronal gain control. *Nature*, *457*, 1015–1018. doi: 10.1038/nature07604

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008).

An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(16), 5975–5979. doi: 10.1073/pnas.0711295105

Rudd, M. E., & Brown, L. G. (1997). A model of Weber and noise gain control in the retina of the toad bufo marinus. *Vision Research*, *37*, 2433–2453.

Salinas, E., & Thier, P. (2000). Gain modulation: a major computational principle of the central nervous system. *Neuron*, *27*, 15–21.

Sands, S. F., & Wright, A. A. (1980). Primate memory: Retention of serial list items by a rhesus monkey. *Science*, *209*(4459), 938–940.

Schneegans, S., & Bays, P. M. (2016). No fixed item limit in visuospatial working memory. *Cortex*, *83*, 181–193. doi: 10.1016/j.cortex.2016.07.021

Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: Clues to visual objecthood. *Cognitive Psychology*, *38*(2), 259-90.

Sengupta, R., Surampudi, B. R., & Melcher, D. (2014). A visual sense of number emerges from the dynamics of a recurrent on-center off-surround neural network. *Brain Research*, *1582*, 114–124. doi: 10.1016/j.brainres.2014.03.014

Shimomura, T., & Kumada, T. (2011). Spatial working memory load affects counting but not subitizing in enumeration. *Attention, Perception and Psychophysics*, *73*(6), 1694–1709. doi: 10.3758/s13414-011-0135-5

Shipstead, Z., & Engle, R. W. (2013). Interference within the focus of attention: working memory tasks reflect more than temporary maintenance. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*, 277–289. doi: 10.1037/a0028467

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology. General*, *117*(1), 34–50.

Standing, L. (1973). Learning 10,000 pictures. *The Quarterly Journal of Experimental Psychology*, *25*(2), 207–222.

Standing, L., Conezio, J., & Haber, R. (1970). Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. *Psychonomic Science*, *19*(2), 73-74.

Trick, L., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review*, *101*(1), 80-102.

Tuholski, S. W., Engle, R. W., & Baylis, G. C. (2001). Individual differences in working memory capacity and enumeration. *Memory and Cognition*, *29*(3), 484–492.

van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*(22), 8780–8785. doi: 10.1073/pnas.1117465109

Verguts, T., & Fias, W. (2004). Representation of number in animals and humans: a neural model. *Journal of Cognitive Neuroscience*, *16*, 1493–1504. doi: 10.1162/0898929042568497

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature*, *428*, 748–751. doi: 10.1038/nature02447

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions and objects in visual working memory. *Journal of Experimental Psychology. Human Perception and Performance*, *27*(1), 92–114.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2006). The time course of consolidation in visual working memory. *Journal of Experimental Psychology. Human Perception and Performance*, *32*(6), 1436–1451. doi: 10.1037/0096-1523.32.6.1436

Whalen, J., Gallistel, C., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science*, *10*(2), 130-137. doi: 10.1111/1467-9280.00120

Willmore, B. D. B., Cooke, J. E., & King, A. J. (2014). Hearing in noisy environments: noise invariance and contrast gain control. *Journal of Physiology*, *592*, 3371–3381. doi: 10.1113/jphysiol.2014.274886

Zhang, H., Xuan, Y., Fu, X., & Pylyshyn, Z. W. (2010). Do objects in working memory compete with objects in perception? *Visual Cognition*, *18*(4), 617-640. doi: 10.1080/13506280903211142

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235. doi: 10.1038/nature06860

Zosh, J. M., & Feigenson, L. (2015). Array heterogeneity prevents catastrophic forgetting in infants. *Cognition*, *136*, 365–380. doi: 10.1016/j.cognition.2014.11.042

Appendix A

Arbitrariness of the time step in the discrete version of the model

We will now show that the time step in the discrete version of the model has no unit and is arbitrary.

Let $F : \mathbf{R}_+ \to [0, 1]$ be any continuous function and $N \geq 2$ an integer. We consider the system of first-order nonlinear ordinary differential equations

$$\dot{x_i} = -\lambda x_i + \alpha F(x_i) - \beta \sum_{j=1, j\neq i}^{N} F(x_j)$$

where $x_1(t), \ldots, x_N(t)$ are real-valued functions on $\mathbf{R}_+$. Fixing an arbitrary $h > 0$, the discrete-time version of this system is the following system of equations:

$$\frac{x_i(t+h) - x_i(t)}{h} = -\lambda x_i(t) + \alpha F(x_i(t)) - \beta \sum_{j=1, j\neq i}^{N} F(x_j(t)).$$

After rearranging, this becomes

$$x_i(t+h) = x_i(t) - h\lambda x_i(t) + h\alpha F(x_i(t)) - h\beta \sum_{j=1, j\neq i}^{N} F(x_j(t)). \tag{14}$$

Now, introducing

$$t = ht', \quad \lambda' = h\lambda, \quad \alpha' = h\alpha, \quad \beta' = h\beta \text{ and } X_i(t') = x_i(ht')$$

this system reads as

$$X_i(t'+1) = X_i(t') - \lambda' X_i(t') + \alpha' F(X_i(t')) - \beta' \sum_{j=1, j\neq i}^{N} F(X_j(t')).$$

Thus, in order to study the discretized system it is sufficient to consider (14) with $h = 1$.

Appendix B

Generalization to other activation functions

B.1 Analytic derivation

To find a stationary solution, we need to find a solution to Equation (3) so that the time derivative vanishes:

$$0 = \dot{x} = -\lambda x + (\alpha - \beta(S-1))F(x) \tag{3}$$

Here we use again the observation that the activation in all active neurons should be equal. This allows us to prove the following claim:

Claim 3. *For activation functions that satisfy $F(0) = 0$ and $F(x) > 0$ for all positive x, $x(t) = 0$ is always a solution to Equation (3). It is also the only constant solution for a sufficiently large S such that $S > 1 + \frac{\alpha}{\beta}$.*

Proof. To verify this claim, we assume that there is another solution and rewrite Equation (3) as

$$F(x) = \frac{\lambda}{\alpha - (S-1)\beta} x$$

for $\alpha \neq \beta(S-1)$ (in which case the equation describes exponential decay). The term on the

right hand side is negative for $S > 1 + \frac{\alpha}{\beta}$, which contradicts the assumption that $F(x) > 0$ for positive x. Hence, there is no non-zero constant solution for $S > 1 + \frac{\alpha}{\beta}$. □

B.1.1 Existence of fixed points. In general, we cannot say much about the evolution of the network activity with arbitrary activation functions. However, we can show that any bounded activation function admits a fixed point.

For the proof, assume that we have N neurons in total, and that for some $1 \leq S \leq N$ the levels of activation of the first S neurons is identical, and that the levels of activation of the remaining $N - S$ neurons is identical as well (and potentially different from that of the first S neurons):

$$x_1(0) = \cdots = x_S(0)$$

$$x_{S+1}(0) = \cdots = x_N(0).$$

For reasons of symmetry, for all $t \geq 0$ such that the external input I_i vanishes on $[0, t]$ for all i, we equally have

$$x_1(t) = \cdots = x_S(t)$$

$$x_{S+1}(t) = \cdots = x_N(t).$$

In what follows, we assume that this holds for any t. Let us denote by

$$x(t) = x_1(t), \quad y(t) = x_{S+1}(t)$$

the two different levels of activation. Let us set

$$z^+ = \max(z, 0),$$

and introduce the function

$$\begin{aligned} f : \mathbf{R}_{\geq 0} \times \mathbf{R}_{\geq 0} &\to \mathbf{R}_{\geq 0} \times \mathbf{R}_{\geq 0} \\ \begin{pmatrix} x \\ y \end{pmatrix} &\mapsto \begin{pmatrix} [(1-\lambda)x + \alpha F(x) - \beta(S-1)F(x) - \beta(N-S)F(y)]^+ \\ [(1-\lambda)y + \alpha F(y) - \beta S F(x) - \beta(N-S-1)F(y)]^+ \end{pmatrix}. \end{aligned} \tag{15}$$

As the sum and maximum of two continuous functions is continuous, so is f.

Furthermore, we have

$$\begin{pmatrix} x(t+1) \\ y(t+1) \end{pmatrix} = f \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}.$$

A stationary point of the process without noise and external activation is precisely a fixed point of the function f.

Claim 4. *For any continuous function $F : \mathbf{R}_+ \to [0, 1]$, and integers $1 \leq S \leq N$, the function f introduced in* (15) *above admits at least one fixed point.*

Proof. Let us fix any $R > \left(\frac{\alpha}{\lambda}\right)^+$ and let x, y satisfy $x \leq R, y \leq R$. Then, we have the

estimate

$$\begin{aligned}(1-\lambda)x + \alpha F(x) - \beta(S-1)F(x) - \beta(N-S)F(y) &\leq (1-\lambda)x + \alpha F(x) \\ &\leq (1-\lambda)R + \alpha \\ &< (1-\lambda)\frac{\alpha}{\lambda} + \alpha = \frac{\alpha}{\lambda} < R,\end{aligned}$$

and similarly we find

$$(1-\lambda)y + \alpha F(y) - \beta S F(x) - \beta(N-S-1)F(y) < R.$$

This shows that f preserves the square

$$D = [0, R] \times [0, R].$$

As D is homeomorphic to the closed unit disc and f is a continuous self-map of D, Brouwer's fixed point theorem implies that f admits a fixed point. □

We note that in general we can not say anything about the convergence of $(x(t), y(t))$ to the stationary point, unless we have information about the Jacobian of f at the stationary point.

B.1.2 Convergence towards similar activation values. While it is difficult to say something general about the evolution of the network activity, we can show that, under certain conditions, all neurons in the network, whether they have been activated or not, will end up having a similar activation value. As a result, all information about the

stimuli that have been presented is lost.

To prove this claim, we note that, initially, the S neurons that have been externally excited with have some activation $x_1(t)$, while the $N-S$ neurons that have not been excited will have some other activation $x_2(t)$. We then consider the difference

$$\Delta(t+1) = x_1(t+1) - x_2(t+1)$$

and show that it converges towards zero as long as $(\alpha+\beta)F'(0) < 1$, without completely reaching it. That is, all neurons will end up having a similar, and practically identical, activation.

Claim 5 (Disappearance of activation differences)**.** *For* $F'(0) < \frac{\lambda}{\alpha+\beta}$, $\Delta(t+1) = x_1(t+1) - x_2(t+1)$ *converges to zero, but stays strictly positive.*

Proof. Using Equation (2), we can calculate $\Delta(t+1)$:

$$\begin{aligned}
\Delta(t+1) &= x_1(t+1) - x_2(t+1) \\
&= \{(1-\lambda)x_1(t) + \alpha F(x_1(t)) - \\
&\quad \beta(S-1)F(x_1(t)) - \beta(N-S)F(x_2(t))\} - \\
&\quad \{(1-\lambda)x_2(t) + \alpha F(x_2(t)) - \\
&\quad \beta S F(x_1(t)) - \beta(N-S-1)F(x_2(t))\}
\end{aligned} \tag{16}$$

$$\Delta(t+1) = (1-\lambda)(x_1(t) - x_2(t)) + (\alpha+\beta)(F(x_1(t)) - F(x_2(t)) \tag{17}$$

We can now rewrite Equation (17) using the Taylor expansion around zero of the activation

function $F(x) = \sum_{n=0}^{\infty} \frac{F^{(n)}(0)}{n!} x^n$:

$$\begin{aligned}
\Delta(t+1) &= (1-\lambda)(x_1(t) - x_2(t)) + (\alpha+\beta)\left\{\sum_{n=0}^{\infty} \frac{F^{(n)}(0)}{n!} x_1^n(t) - \sum_{n=0}^{\infty} \frac{F^{(n)}(0)}{n!} x_2^n(t)\right\} \\
&= (1-\lambda)\Delta(t) + (\alpha+\beta)\left\{\sum_{n=0}^{\infty} \frac{F^{(n)}(0)}{n!}(x_1^n(t) - x_2^n(t))\right\} \\
&\equiv g(\Delta(t)) \qquad (18)
\end{aligned}$$

Using Banach's fixed point theorem, we now show that $\Delta^* = 0$ is a convergent fixed point of g as defined in Equation (18) above. $\Delta^* = 0$ is clearly a fixed point, as $g(0) = 0$. It thus remains to be shown that $|g'(\Delta^* = 0)| < 1$. To see why this is the case, we note that $\frac{\partial F(x(\Delta))}{\partial \Delta} = F'(x(\Delta))\frac{\partial x(\Delta)}{\partial \Delta}$, and that, since $\Delta = x_1 - x_2$, we have $\frac{\partial x_1}{\partial \Delta} - \frac{\partial x_2}{\partial \Delta} = 1$. As a result, we find the following expansion for g':

$$\begin{aligned}
g'(\Delta) &= (1-\lambda) + (\alpha+\beta)F'(0) + (\alpha+\beta)\left\{\sum_{n=2}^{\infty} \frac{F^{(n)}(0)}{n!} n\left(x_1^{n-1}\frac{\partial x_1}{\partial \Delta} - x_2^{n-1}\frac{\partial x_2}{\partial \Delta}\right)\right\} \\
&= (1-\lambda) + (\alpha+\beta)F'(0) + (\alpha+\beta)\left\{\sum_{n=2}^{\infty} \frac{F^{(n)}(0)}{(n-1)!}\left(\left(x_1^{n-1} - x_2^{n-1}\right)\frac{\partial x_1}{\partial \Delta} + x_2^{n-1}\right)\right\}
\end{aligned}$$

Since $x_1 = x_2$ at $\Delta = 0$, we obtain:

$$\begin{aligned}
g'(0) &= (1-\lambda) + (\alpha+\beta)F'(0) + (\alpha+\beta)\left\{\sum_{n=2}^{\infty} \frac{F^{(n)}(0)}{(n-1)!} x_2^{n-1}\right\} \\
&= (1-\lambda) + (\alpha+\beta)F'(0) + (\alpha+\beta)\left\{\sum_{n=0}^{\infty} \frac{F^{(n+1)}(0)}{n!} x_2^n - F'(0)\right\} \\
&= (1-\lambda) + (\alpha+\beta)F'(x_2)
\end{aligned}$$

Activation functions that go into saturation are concave such that $F'(x) \leq F'(0)$ for $x > 0$. Hence, for such activation functions, a sufficient condition for the difference to

converge to zero is

$$g'(\Delta^* = 0) \quad \leq \quad (1 - \lambda) + (\alpha + \beta)F'(0) < 1 \tag{19}$$

Further, for *all* activation functions, the first order Taylor expansion of Equation (17) is given by

$$\Delta(t + 1) \approx \{(1 - \lambda) + (\alpha + \beta)F'(0)\}\, \Delta(t) \tag{20}$$

Hence, Equation (19) is an approximate sufficient condition for convergence to $\Delta^* = 0$ for all activation functions. In other words, we have convergence for

$$F'(0) < \frac{\lambda}{\alpha + \beta}$$

In other words, if the slope of the activation function at zero is smaller than $\frac{\lambda}{\alpha+\beta}$, the activation function will eliminate activation differences between neurons. That being said, Equation (18) shows that any solution $\Delta(t)$ will be strictly larger than an exponentially decaying solution; since an exponentially decaying function will converge to zero without ever reaching it, the differences between the x_i will also converge to zero without ever reaching it (though they might reach it within the precision of numerical simulations). □

B.2 Simulation results with different activation functions

We now confirm these results by running simulations with other activation functions, namely a threshold function, the tanh function, the logistic (sigmoid) function (shifted by .5 so that $F(0) = 0$), a linear function, and a rectified linear function; the equations are

given in Table 1 above.

For each function, we report four sets of simulations, reflecting the simultaneous vs. sequential presentation of the stimulus items, in the presence or the absence of noise. For each simulation, we report both the average activation in the network as well as the accuracy of the activations.

The results support the predictions above: For all activation functions, there is catastrophic interference when items are presented simultaneously; if the set-size grows too large, all neurons are switched off. For activation functions whose slope at zero is not too small (i.e., $F'(0) > \frac{\lambda}{\alpha+\beta}$, and thus for all functions except for the logistic function) and that are not unbounded, sequential presentation protects the network from catastrophic interference in that successful discrimination between seen and unseen items is maintained for larger set-sizes.

When the slope at zero is too small (i.e., $F'(0) < \frac{\lambda}{\alpha+\beta}$, and thus for the logistic function), *all* neurons end up having the same activation, irrespective of whether they were part of the stimulus or not. As a result, the network activity is no longer diagnostic of the stimuli, and there is thus complete loss of information. In Section B.3, we show this result analytically.

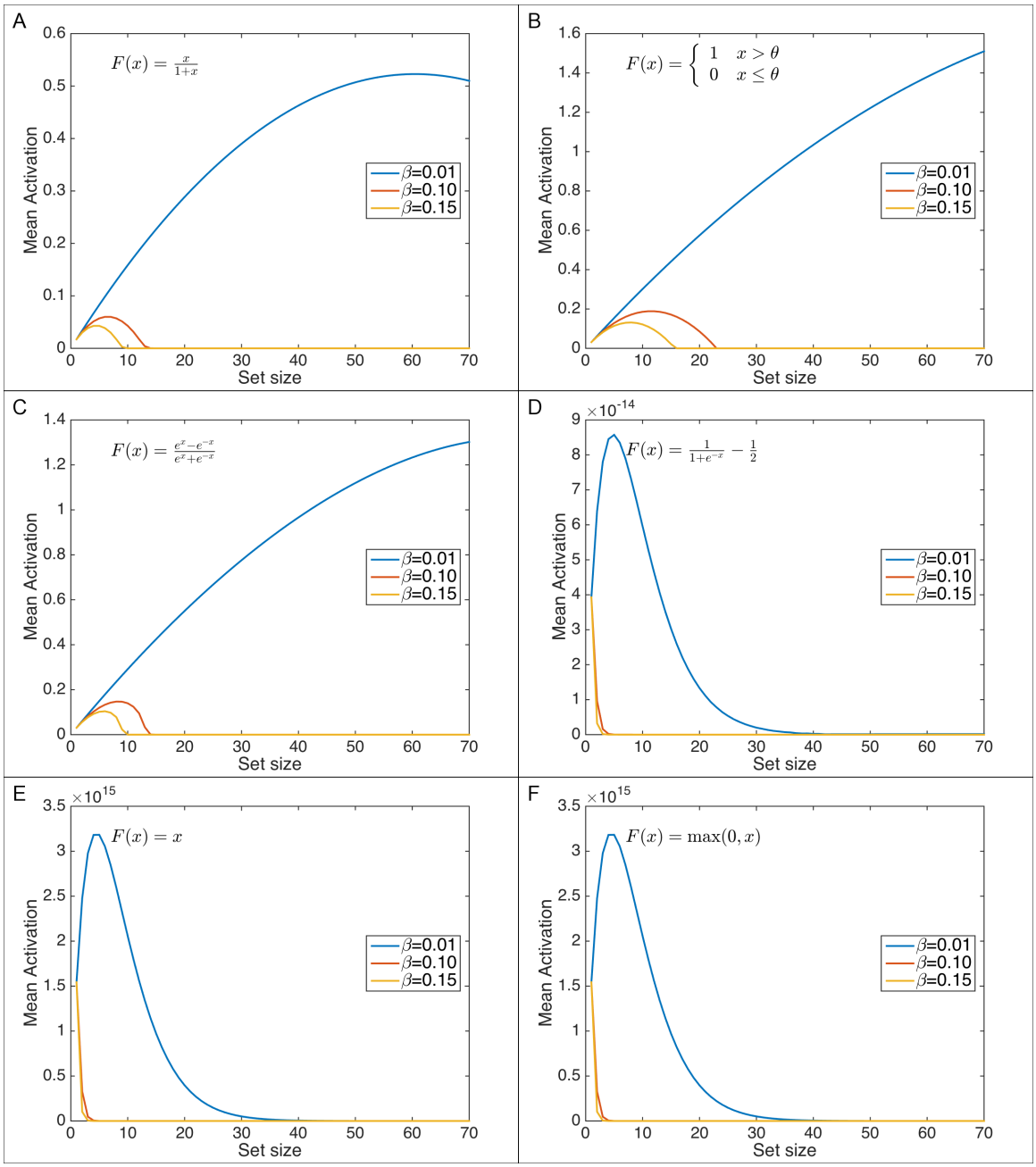

Figure B1. Mean activation level when items are presented simultaneously and in the absence of noise, for (A) Knops et al.'s (2014) rational sigmoid activation function, (B) a threshold function, (C) the hyperbolic tangent function, (D) the logistic function, (E) a linear function and (F) a rectified linear function. The network shows catastrophic interference with all activation functions in that, beyond a critical set-size, and neurons are switched off.

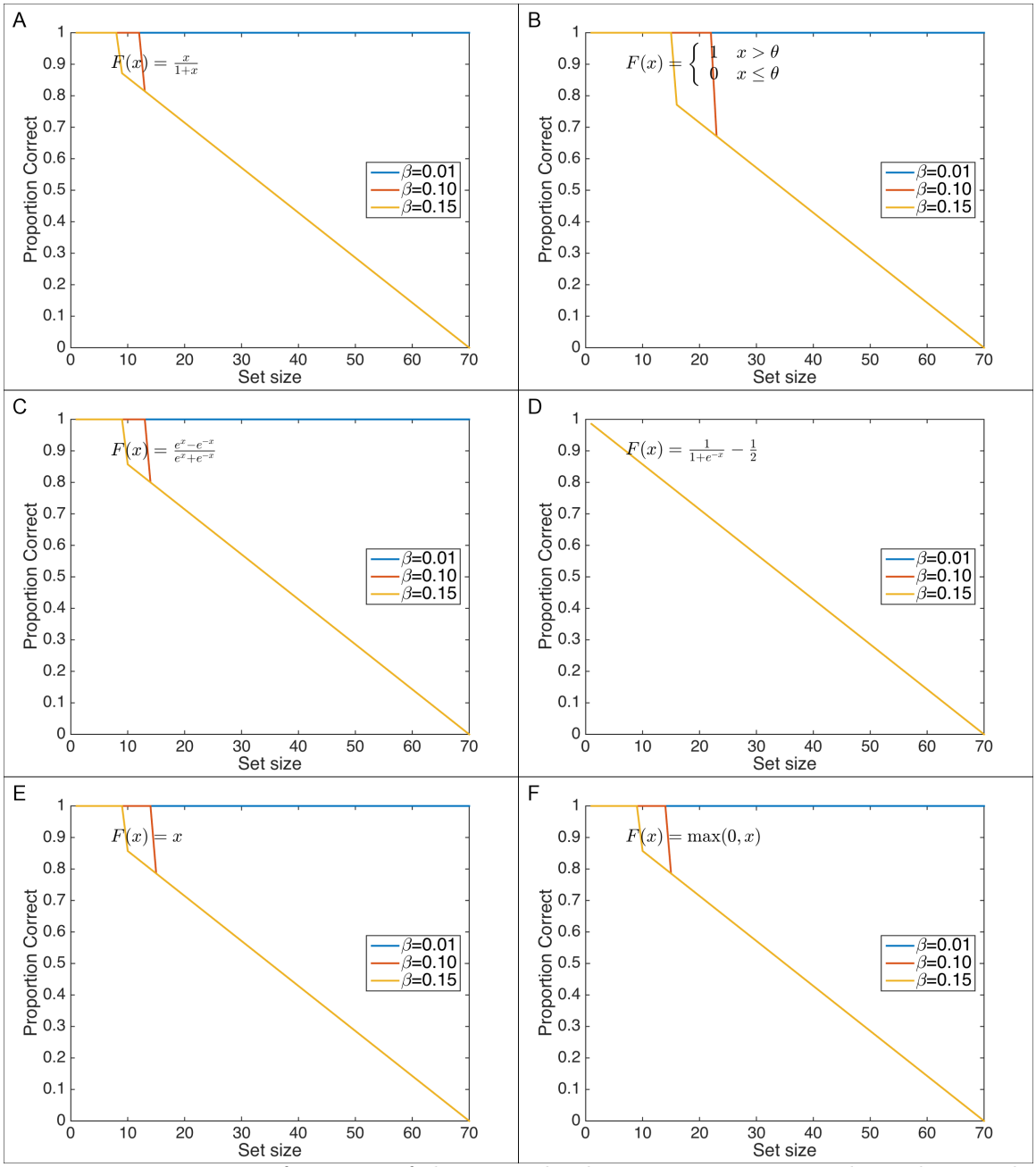

Figure B2. Memory performance of the network when items are presented simultaneously and in the absence of noise, for (A) Knops et al.'s (2014) rational sigmoid activation function, (B) a threshold function, (C) the hyperbolic tangent function, (D) the logistic function, (E) a linear function and (F) a rectified linear function. (A,B,C,E,F) All functions with $F'(0) \geq \frac{\lambda}{\alpha+\beta}$ show catastrophic interference. As a result, they selectively activate the "correct" input neurons as long as they are within $S_{\max}$, and show a 100% miss rate beyond $S_{\max}$, reflecting a decrease in the proportion of correct responses as a function of the set-size. (D) For the logistic function, neurons have extremely low activation levels that converge to zero, as any differences in activation across neurons eventually disappear and all neurons are switched off. This leads to a high number of misses for all set-sizes.

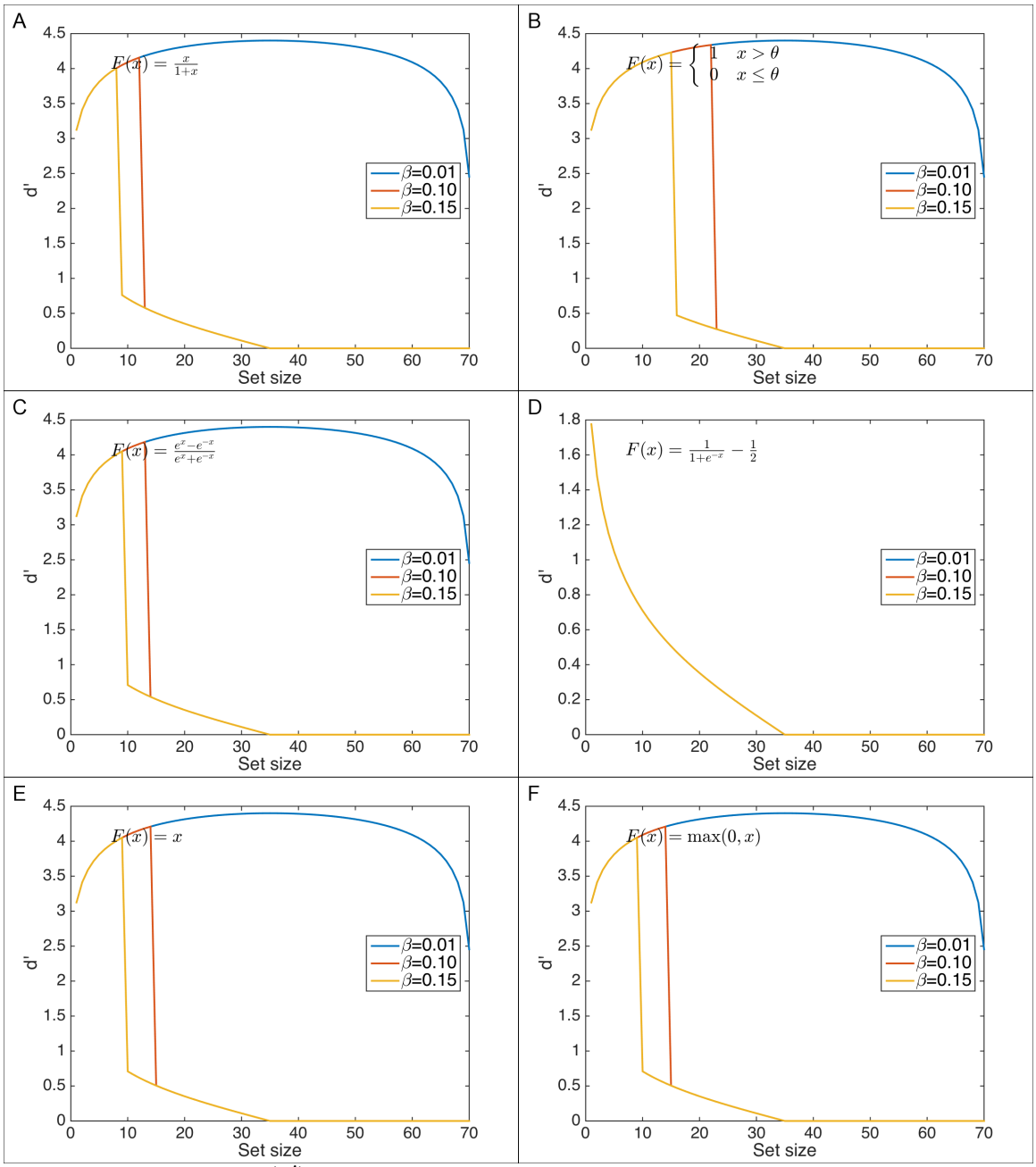

Figure B3. Sensitivity (d') when items are presented simultaneously and in the absence of noise, for (A) Knops et al.'s (2014) rational sigmoid activation function, (B) a threshold function, (C) the hyperbolic tangent function, (D) the logistic function, (E) a linear function and (F) a rectified linear function. (A,B,C,E,F) All functions with $F'(0) \geq \frac{\lambda}{\alpha+\beta}$ show catastrophic interference in that d' drops abruptly beyond some critical set-size. (D) For the logistic function, neurons have extremely low activation levels that converge to zero, as any differences in activation across neurons eventually disappear such that all neurons are switched off. As a result, discriminability is poor even for small set-sizes.

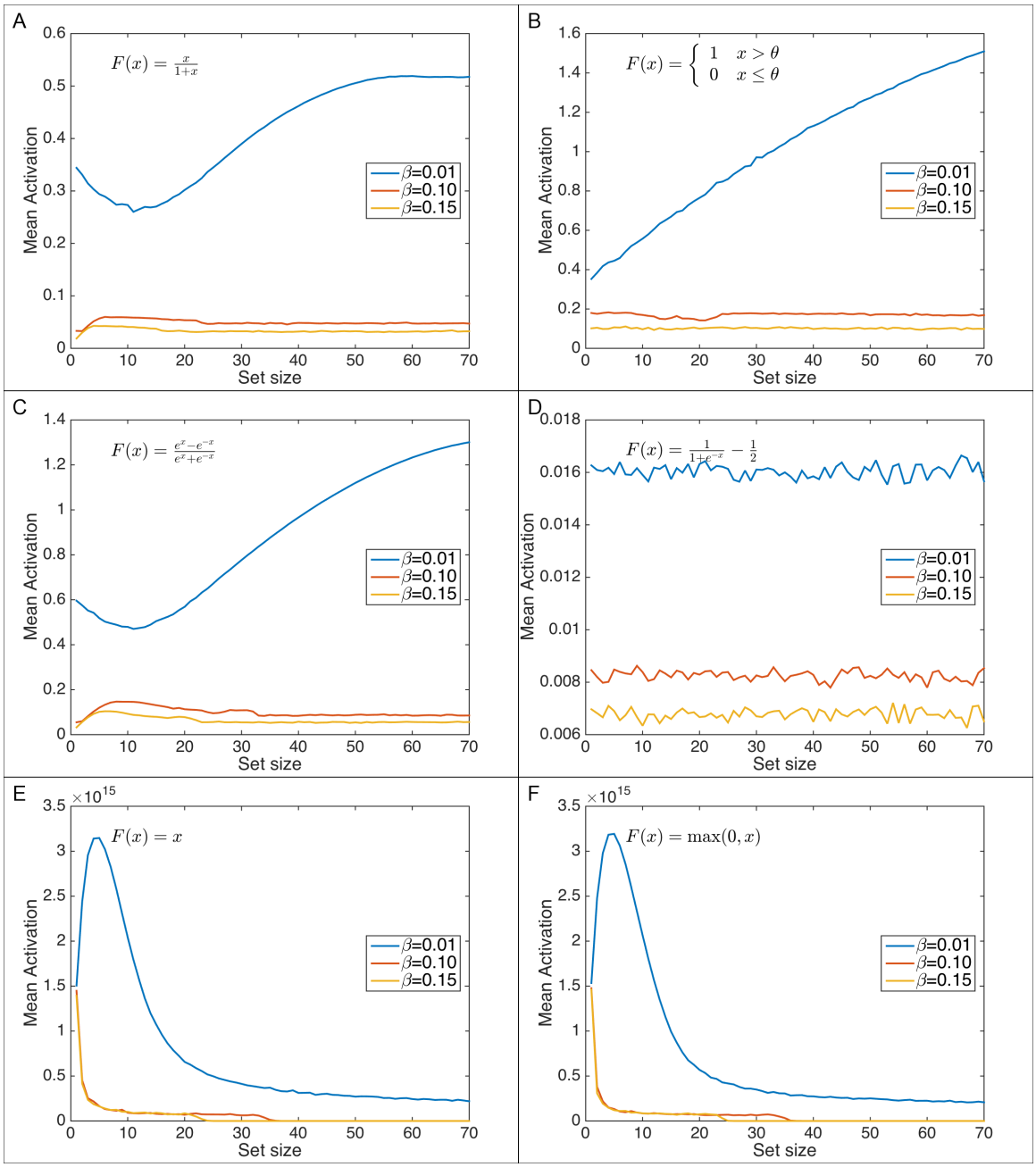

Figure B4. Mean activation level when items are presented simultaneously and when the noise standard deviation is set to .03, for (A) Knops et al.'s (2014) rational sigmoid activation function, (B) a threshold function, (C) the hyperbolic tangent function, (D) the logistic function, (E) a linear function and (F) a rectified linear function. (A,B,C,E,F) All functions with $F'(0) \geq \frac{\lambda}{\alpha+\beta}$ show catastrophic interference. (D) For the logistic function, *all* neurons end up having the same, low level of activation irrespective of the number of input neurons when inhibition levels are low.

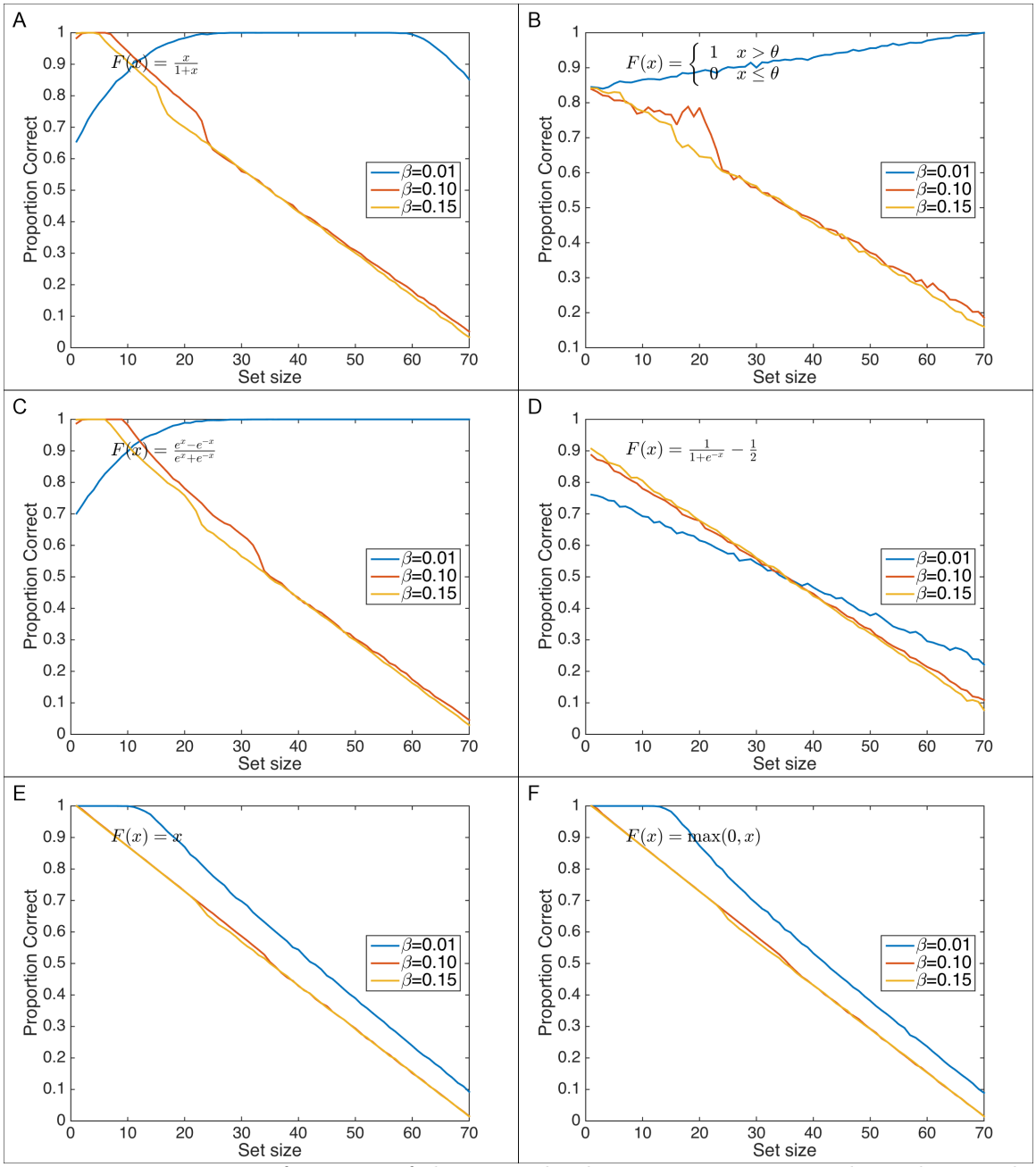

Figure B5. Memory performance of the network when items are presented simultaneously and the noise standard deviation is set to .03, for (A) Knops et al.'s (2014) rational sigmoid activation function, (B) a threshold function, (C) the hyperbolic tangent function, (D) the logistic function, (E) a linear function and (F) a rectified linear function. (A,B,C,E,F) All functions with $F'(0) \geq \frac{\lambda}{\alpha+\beta}$ show catastrophic interference. As a result, they selectively activate the "correct" input neurons as long as they are within $S_{\max}$, and show a 100% miss rate beyond $S_{\max}$, reflecting a decrease in the proportion of correct responses as a function of the set-size. (D) For the logistic function, neurons have extremely low activation levels that converge to zero, as any differences in activation across neurons eventually disappear such that all neurons are switched off. As a result, discriminability is poor even for small set-sizes.

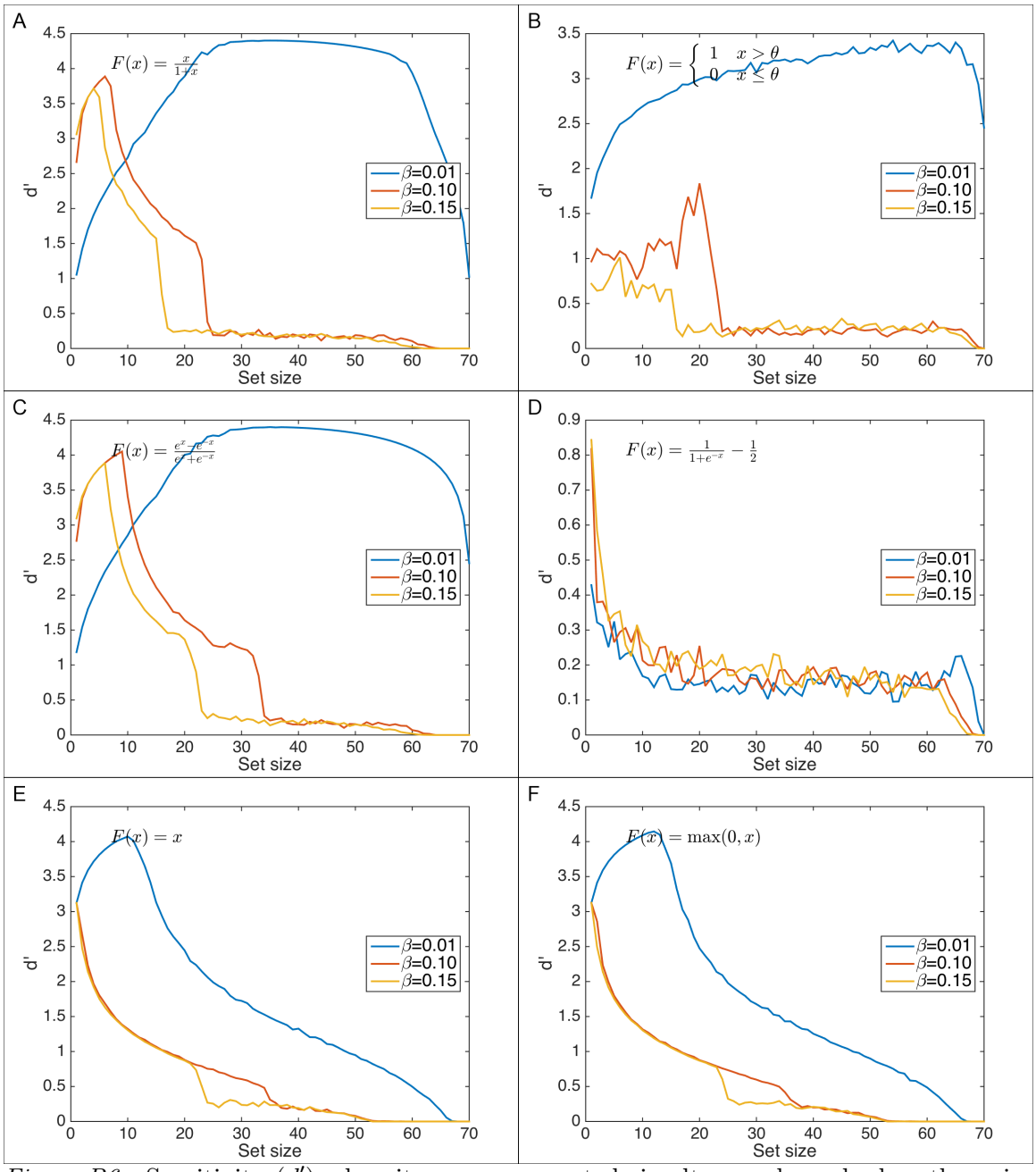

Figure B6. Sensitivity (d') when items are presented simultaneously and when the noise standard deviation is set to .03, for (A) Knops et al.'s (2014) rational sigmoid activation function, (B) a threshold function, (C) the hyperbolic tangent function, (D) the logistic function, (E) a linear function and (F) a rectified linear function. (A,B,C,E,F) All functions with $F'(0) \geq \frac{\lambda}{\alpha+\beta}$ show catastrophic interference. (D) For the logistic function, neurons have extremely low activation levels that converge to zero, as any differences in activation across neurons eventually disappear such that all neurons are switched off. As a result, discriminability is poor even for small set-sizes.

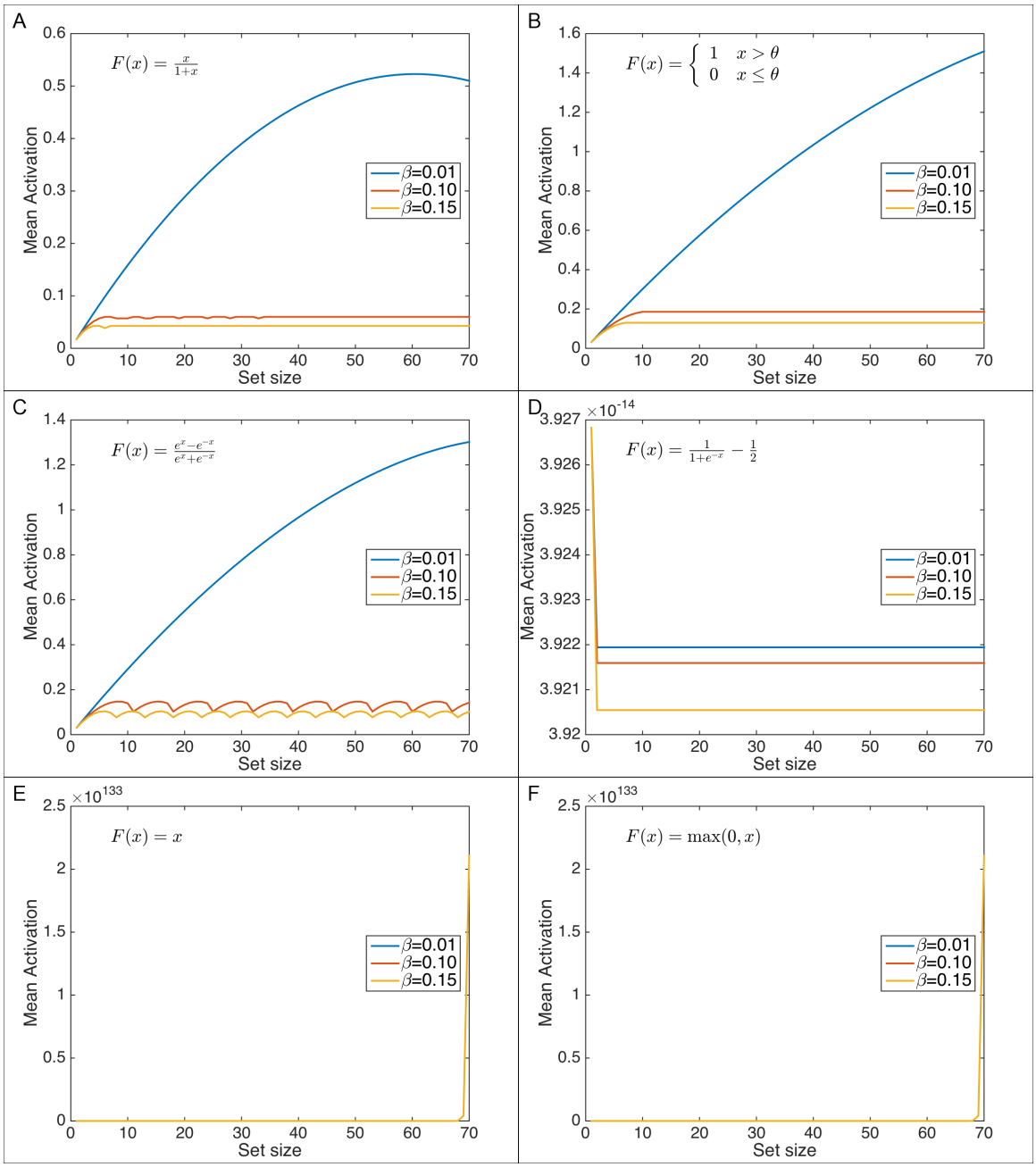

Figure B7. Mean activation level when items are presented sequentially and in the absence of noise, for (A) Knops et al.'s (2014) rational sigmoid activation function, (B) a threshold function, (C) the hyperbolic tangent function, (D) the logistic function, (E) a linear function and (F) a rectified linear function. (A, B, C) For bounded activation functions with $F'(0) \geq \frac{\lambda}{\alpha+\beta}$, sequential presentation protects the network from catastrophic interference. (D, E, F) For unbounded functions or functions with $F'(0) < \frac{\lambda}{\alpha+\beta}$, the network shows similar behavior as in the simultaneous case: *All* neurons end up being switched off irrespective of the set-size.

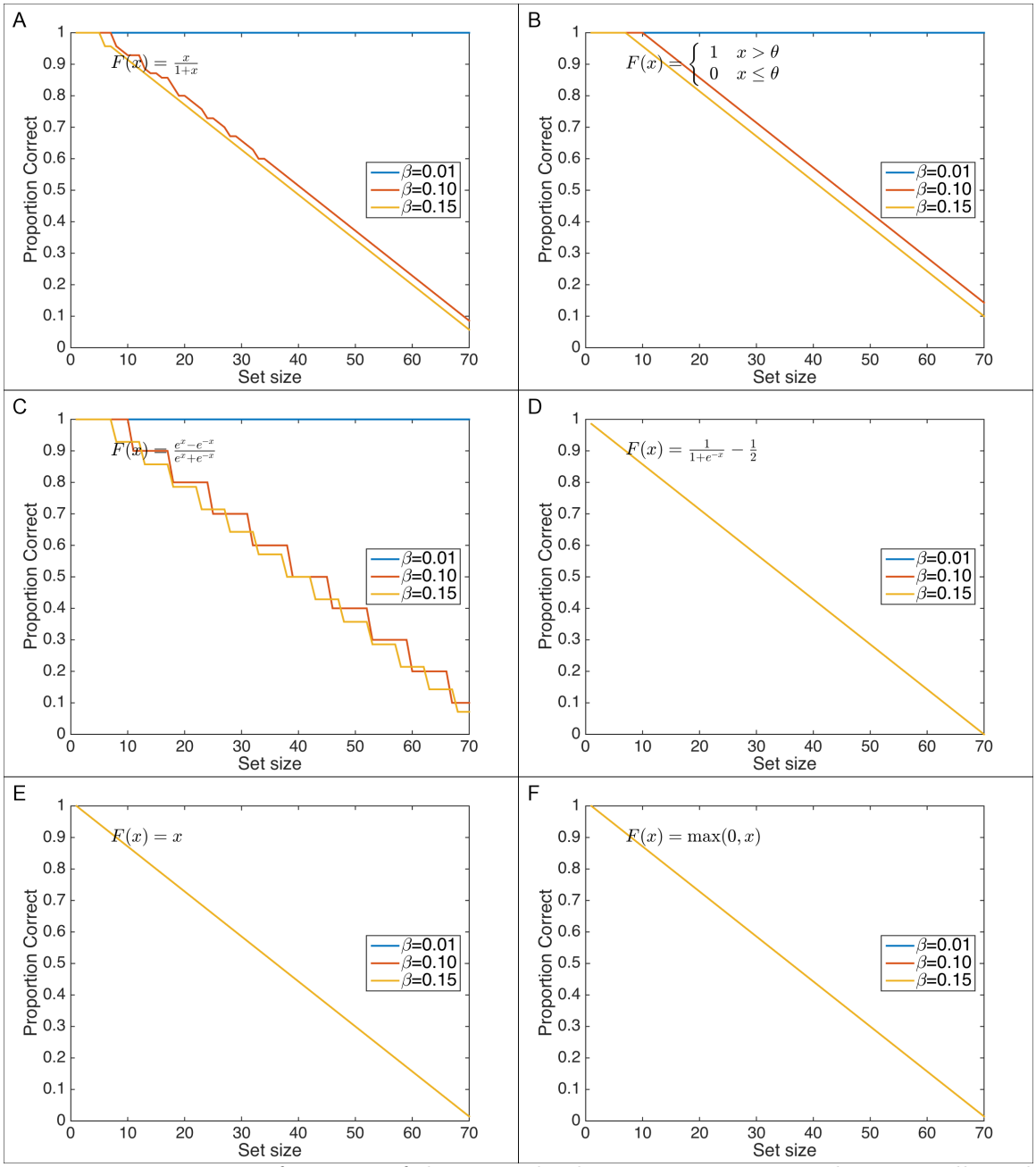

Figure B8. Memory performance of the network when items are presented sequentially and in the absence of noise, for (A) Knops et al.'s (2014) rational sigmoid activation function, (B) a threshold function, (C) the hyperbolic tangent function, (D) the logistic function, (E) a linear function and (F) a rectified linear function. (A, B, C) For all bounded activation functions with $F'(0) \geq \frac{\lambda}{\alpha+\beta}$, sequential presentation protects the network from catastrophic interference, so that it can retain a limited number of items. As a result, they selectively activate the "correct" input neurons as long as they are within the capacity of this buffer, and show a 100% miss rate beyond, reflecting a decrease in the proportion of correct responses as a function of the set-size. (D, E, F) For unbounded functions or functions with $F'(0) < \frac{\lambda}{\alpha+\beta}$, the logistic function, *all* neurons are switched off irrespective of the set-size, leading to a high number of misses, and thus to a decreasing proportion of correct responses as the set size increase.

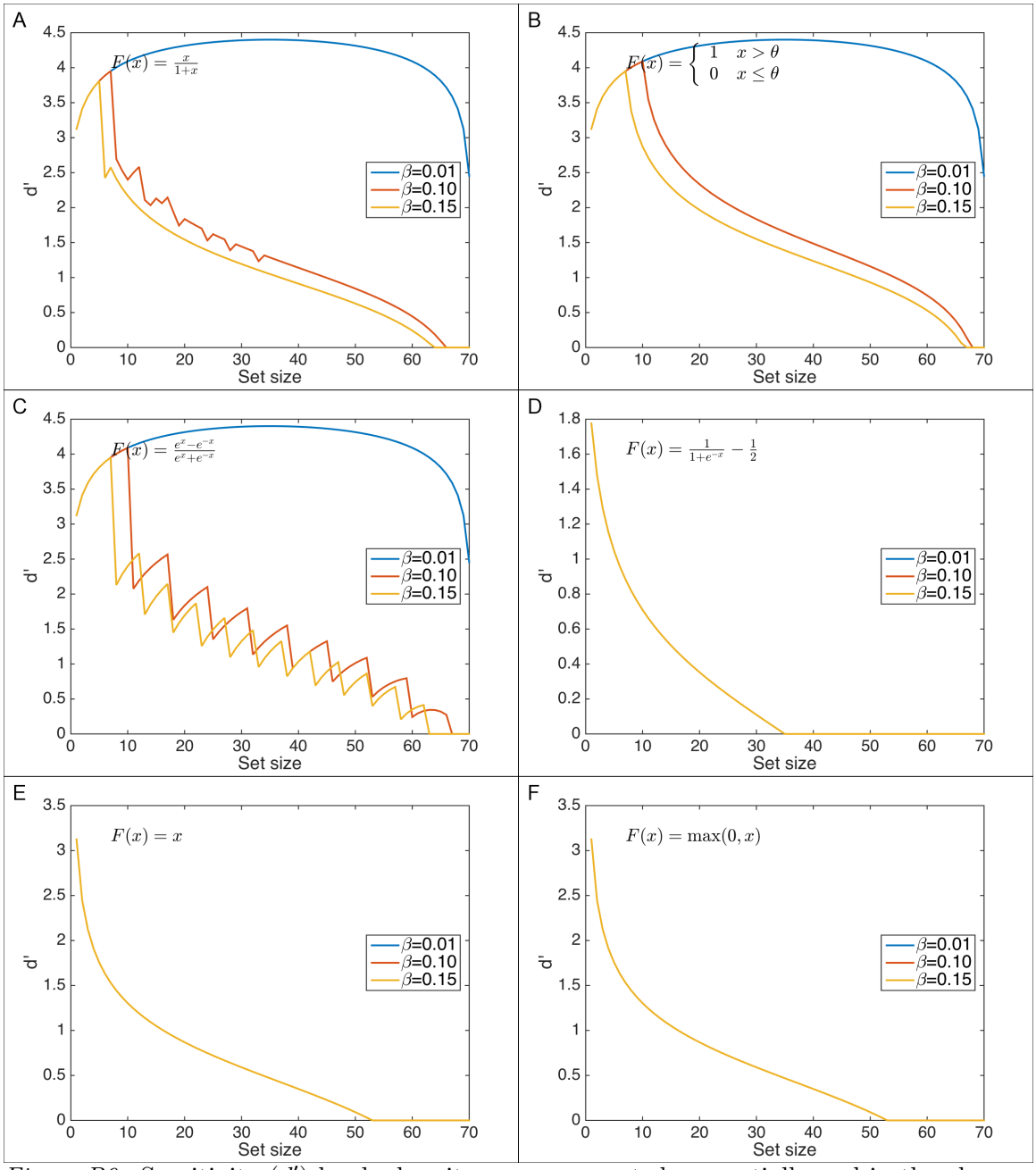

Figure B9. Sensitivity (d') level when items are presented sequentially and in the absence of noise, for (A) Knops et al.'s (2014) rational sigmoid activation function, (B) a threshold function, (C) the hyperbolic tangent function, (D) the logistic function, (E) a linear function and (F) a rectified linear function. (A, B, C) For all bounded activation functions with $F'(0) \geq \frac{\lambda}{\alpha+\beta}$, sequential presentation protects the network from catastrophic interference. (D, E, F) For unbounded functions or functions with $F'(0) < \frac{\lambda}{\alpha+\beta}$, *all* neurons are switched off, leading to poor discrimination even for small set-sizes.

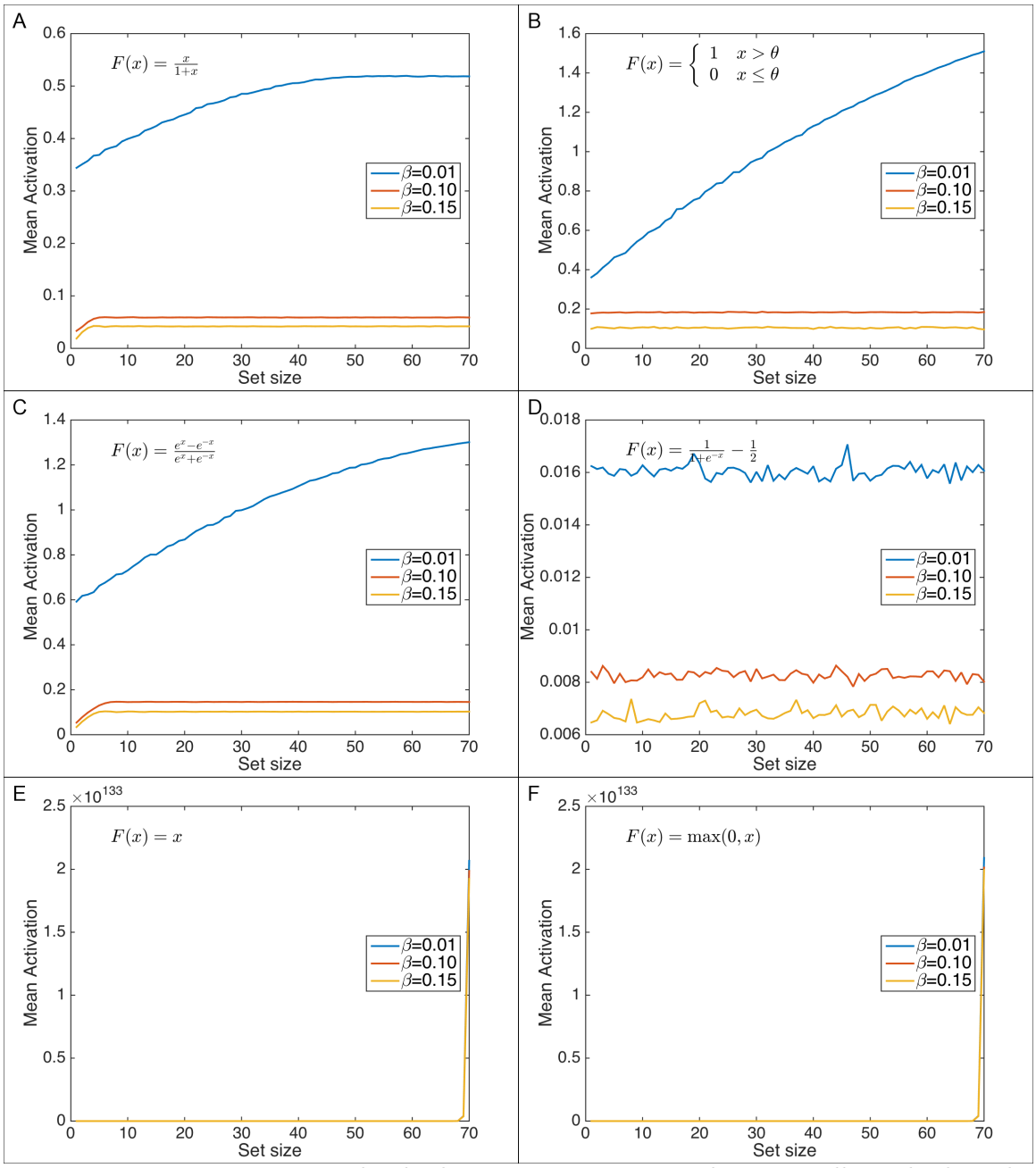

Figure B10. Mean activation level when items are presented sequentially and when the standard deviation for the noise is set to .03, for (A) Knops et al.'s (2014) rational sigmoid activation function, (B) a threshold function, (C) the hyperbolic tangent function, (D) the logistic function, (E) a linear function and (F) a rectified linear function. (A, B, C) For all bounded activation functions with $F'(0) \geq \frac{\lambda}{\alpha+\beta}$, sequential presentation protects the network from catastrophic interference. (D, E, F) For the activation functions that are unbounded or for which $F'(0) < \frac{\lambda}{\alpha+\beta}$, the network shows similar behavior as in the simultaneous case: *all* neurons end up having activations close to zero.

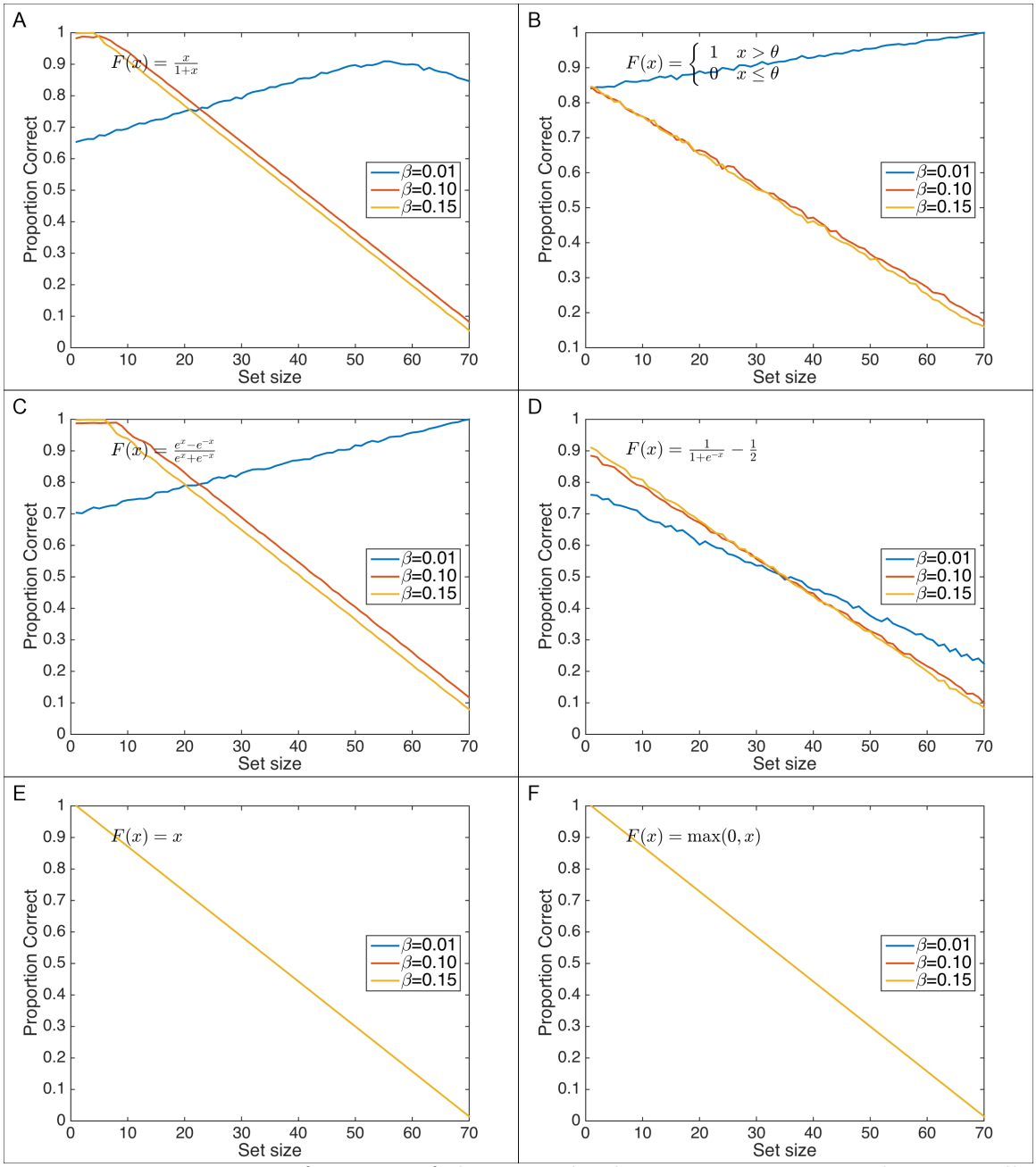

Figure B11. Memory performance of the network when items are presented sequentially and when the noise standard deviation is set to .03, for (A) Knops et al.'s (2014) rational sigmoid activation function, (B) a threshold function, (C) the hyperbolic tangent function, (D) the logistic function, (E) a linear function and (F) a rectified linear function. (A, B, C) For all bounded activation functions with $F'(0) \geq \frac{\lambda}{\alpha+\beta}$, sequential presentation protects the network from catastrophic interference, so that it can retain a limited number of items. As a result, they selectively activate the "correct" input neurons as long as they are within the capacity of this buffer, and show a 100% miss rate beyond, reflecting a decrease in the proportion of correct responses as a function of the set-size. (D, E, F) For activation functions that are unbounded or for which $F'(0) < \frac{\lambda}{\alpha+\beta}$, *all* neurons are switched off, leading to a high number of misses, and thus to a decreasing proportion of correct responses as the set size increase.

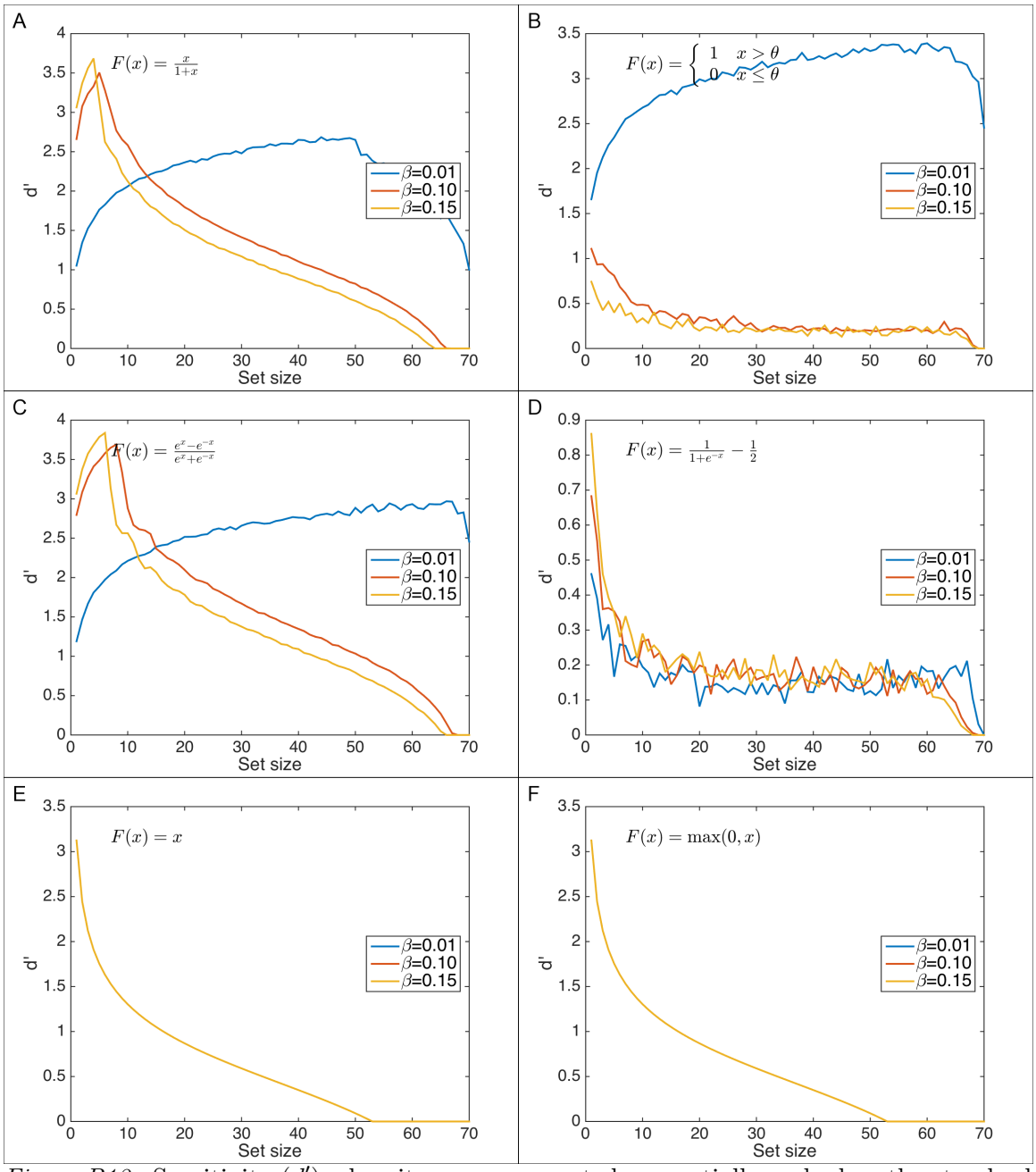

Figure B12. Sensitivity (d') when items are presented sequentially and when the standard deviation for the noise is set to .03, for (A) Knops et al.'s (2014) rational sigmoid activation function, (B) a threshold function, (C) the hyperbolic tangent function, (D) the logistic function, (E) a linear function and (F) a rectified linear function. (A, B, C) For bounded functions with $F'(0) \geq \frac{\lambda}{\alpha+\beta}$, sequential presentation protects the network from catastrophic interference in that successful discrimination is maintained far beyond $S_{\max}$. (D, E, F) For activation functions that are unbounded or with $F'(0) < \frac{\lambda}{\alpha+\beta}$, the network does not maintain any discrimination ability beyond small set-sizes.

B.3 Analytic derivation of the behavior of the sigmoid function

We first show that, with a sigmoid activation function, all neurons in the network will end up having the same activation. Initially, the S neurons that have been externally excited with have some activation $x_1(t)$, while the $N-S$ neurons that have not been excited will have some other activation $x_2(t)$. We now consider the difference:

$$\Delta(t+1) = x_1(t+1) - x_2(t+1) \tag{21}$$

As shown in Claim 5 in Appendix B.1.2, such differences disappear if $F'(0) < \frac{\lambda}{\alpha+\beta}$. For the logistic function, $F'(x) = \frac{e^x}{(1+e^x)^2}$, and thus $F'(0) = 1/4$. Δ thus converges to zero as long as $\alpha < 4\lambda - \beta$, which is the case in all of our simulations.

To calculate the steady state activation, we thus need to solve the equation

$$\begin{aligned} 0 &= \dot{x} \\ &= -\lambda x + (\alpha + \beta - \beta N)\, F(x) \\ &\approx -\lambda x + (\alpha + \beta - \beta N)\, \frac{x}{4} \end{aligned} \tag{22}$$

In the last step, we used the first order Taylor expansion of the logistic function. After some algebra, the solution to this equation is given by

$$\hat{x} = 2\frac{\alpha + \beta(1-N)}{4\lambda - \alpha - \beta(1-N)} \tag{23}$$

Unless $\lambda = \frac{\alpha-\beta(N-1)}{4}$, $x(t) = 0$ is the only solution. With the logistic activation

function, all neurons will thus be switched off.

Appendix C

Stability of the steady-state solutions

C.1 Stability of the steady-state solutions for simultaneous presentations

When all items are presented simultaneously, Equation (1) has two steady-state solutions: $\hat{x}(S) = \frac{\alpha-(S-1)\beta}{\lambda} - 1$, and $\hat{x}(S) = 0$. We will now show that neither solution is stable. However, we show in Section C.2 below that there are stable solutions for the maximal and the minimal activation levels.

Claim 6. *The constant solutions of Equation* (1) $\hat{x}(S) = \frac{\alpha-(S-1)\beta}{\lambda} - 1$ *and* $\hat{x}(S) = 0$ *are not stable.*

The stable submanifold at both stationary points is either the 1-dimensional linear subspace $x_1 = \cdots = x_N$ *or empty, depending on whether the inequality* $S > S_{max}$ *holds (for* $\hat{x} = 0$*) or similarly whether the inequality* $S > 1 + \frac{-\lambda+Q\alpha}{Q\beta}$ *holds (for* $\hat{x}(S) = \frac{\alpha-(S-1)\beta}{\lambda} - 1$*), with* $Q = \frac{\lambda^2}{(\alpha-(S-1)\beta)^2}$.

Proof. Let $\dot{x_i} = g(x) = -\lambda x_i + \alpha F(x_i) - \beta \sum_{j \neq i} F(x_j)$. A fixed point is stable if the Jacobian matrix of partial derivatives $\partial g_i / \partial x_j$ has only negative Eigenvalues. The Jacobian is given by

$$J = \begin{pmatrix} -\lambda + \alpha F'(x_1) & -\beta F'(x_2) & \cdots \\ -\beta F'(x_1) & -\lambda + \alpha F'(x_2) & \cdots \\ \vdots & & \ddots & \vdots \\ -\beta F'(x_1) & \cdots & -\lambda + \alpha F'(x_N) \end{pmatrix} \tag{24}$$

Given that all x_i are equal at the fixed points, this reduces to

$$J = \begin{pmatrix} -\lambda + \alpha F'(\hat{x}) & -\beta F'(\hat{x}) & \cdots & \\ -\beta F'(\hat{x}) & -\lambda + \alpha F'(\hat{x}) & \cdots & \\ \vdots & & \ddots & \vdots \\ -\beta F'(\hat{x}) & \cdots & & -\lambda + \alpha F'(\hat{x}) \end{pmatrix} \tag{25}$$

We now need to show that the Jacobian has only negative Eigenvalues. The Jacobian is circulant. For a circulant matrix with identical diagonal elements a and identical non-diagonal elements b, the Eigenvalues η_i are given by

$$\eta_1 = a + (N-1)b \tag{26}$$

$$\eta_{i>1} = a - b \tag{27}$$

(see e.g. Gray, 1971). With $F(x) = x/(1+x)$, we have $F'(x) = 1/(1+x)^2$, and thus $F'(0) = 1$ and $F'(\hat{x}) = \lambda^2/(\alpha - (S-1)\beta)^2$. We thus calculate the Jacobians at both fixed

points:

$$J^0 = \begin{pmatrix} -\lambda+\alpha & -\beta & \cdots \\ -\beta & -\lambda+\alpha & \cdots \\ \vdots & & \ddots \quad \vdots \\ -\beta & \cdots & -\lambda+\alpha \end{pmatrix} \tag{28}$$

$$\eta_1^0 = -\lambda+\alpha-(N-1)\beta \tag{29}$$

$$\eta_{i>1}^0 = -\lambda+\alpha+\beta \tag{30}$$

$$J^{\hat{x}} = \begin{pmatrix} -\lambda+\alpha Q & -\beta Q & \cdots \\ -\beta Q & -\lambda+\alpha Q & \cdots \\ \vdots & & \ddots \quad \vdots \\ -\beta Q & \cdots & -\lambda+\alpha Q \end{pmatrix} \tag{31}$$

$$\eta_1^{\hat{x}} = -\lambda-Q(\beta(N-1)-\alpha) \tag{32}$$

$$\eta_{i>1}^{\hat{x}} = -\lambda+Q(\alpha+\beta) \tag{33}$$

where $Q = \frac{\lambda^2}{(\alpha-(S-1)\beta)^2}$.

As mentioned above, for the fixed points to be stable, all Eigenvalues of the Jacobian must be negative. For the fixed point at zero, this yields the conditions $\alpha-(N-1)\beta<\lambda$ (for the first Eigenvalue), and $\alpha+\beta<\lambda$ (for the other Eigenvalues). Further, the second condition implies the first one. Critically, given that we assume that $\alpha+\beta>\lambda$ (otherwise, the solution to (1) would show exponential decay), the second condition is not satisfied; as a result, this fixed point is not stable.

For the fixed point at $\hat{x}$, the conditions are $(\alpha-(N-1)\beta)/(\alpha-(S-1)\beta)^2<1/\lambda$

(for the first Eigenvalue), and $(\alpha+\beta)/(\alpha-(S-1)\beta)^2 < 1/\lambda$ (for the other Eigenvalues). Since the left hand side of the first condition is strictly smaller than the left hand side of the second condition, it is sufficient to consider the second condition. For low inhibition values, the fixed point is stable for $S < 72.3$ or $S > 396.6$; for medium inhibition values, the fixed point is stable for $S < 7.8$ and $S > 38.2$; for high inhibition values, the fixed point is stable for $S < 1.04$ and $S > 2934.3$. As a result, neither fixed point is generally stable.

That being said, there is a stable submanifold at both stationary points when the first Eigenvalue is negative. This submanifold is either the 1-dimensional linear subspace $x_1 = \cdots = x_N$ or empty. For the stationary point at $\hat{x} = 0$, this condition is met if $\eta_1^0 < 0$, or, equivalently, if and only if

$$S > S_{\max} = 1 + \frac{-\lambda+\alpha}{\beta}.$$

For the stationary point at $\hat{x}(S) = \frac{\alpha-(S-1)\beta}{\lambda} - 1$, this condition is met if $\eta_1^{\hat{x}} = -\lambda + Q(\alpha-(N-1)\beta) < 0$, with Q as defined above, or equivalently if and only if

$$S > 1 + \frac{-\lambda+Q\alpha}{Q\beta}.$$

In either case, the Eigendirection for the first Eigenvalue is clearly spanned by the vector $(1,\ldots,1)$. □

C.2 Convergence of the minimum and the maximum of the x_i

While the steady state solutions of Equation (1) for simultaneous presentation are not generally stable, we now show that there are stable, non-trivial solutions for the maximum

and the minimum of activity among the neurons as long as $S < S_{\max}$.

C.2.1 Bounds on the maximum. We will now show that there exists a stable lower bound for the maximum of the x_i, $M = \max(x_i)$. Below, S (for set-size) will refer to the number of active neurons.

Claim 7. *According to Equation* (1)*, the maximum activation among neurons,* $M(t)$*, has a stable lower bound given by*

$$M_S = \frac{\alpha - (S-1)\beta}{\lambda} - 1$$

unless $M(t=0) = 0$*. The maximal activation level thus converges to* M_S*.*

Proof. As $F(x)$ is increasing and $\beta > 0$, we see by replacing each $F(x_j)$ in the sum by $F(\max(x_i))$ that the right-hand side of Equation (1) is bounded from below by

$$(\alpha - (S-1)\beta)\frac{\max(x_i)}{1+\max(x_i)} - \lambda \max(x_i).$$

As a result, the function $M = \max(x_i)$ can be bounded from below by the solution of the differential equation

$$\dot{M} = (\alpha - (S-1)\beta)\frac{M}{1+M} - \lambda M, \quad M(0) = \max(x_i(0)). \tag{34}$$

The stationary solutions of this differential equation are given by

$$M_0 = 0, \quad M_S = \frac{\alpha - (S-1)\beta}{\lambda} - 1. \tag{35}$$

The non-trivial stationary solution M_S is positive if and only if

$$S < S_{\max} = \frac{\alpha - \lambda}{\beta} + 1;$$

equivalently,

$$0 < \lambda < \alpha - \beta(S-1). \tag{36}$$

From now on, we assume (36) holds. Setting

$$f(M) = (\alpha - (S-1)\beta)\frac{M}{1+M} - \lambda M,$$

we see that

$$f'(M) = (\alpha - (S-1)\beta)\frac{1}{(1+M)^2} - \lambda.$$

Plugging M_0 and M_S into this formula yields

$$f'(M_0) = (\alpha - (S-1)\beta) - \lambda$$
$$f'(M_S) = \lambda\left(\frac{\lambda}{\alpha - (S-1)\beta} - 1\right).$$

Taking into account (36), these in turn imply

$$f'(M_S) < 0 < f'(M_0).$$

We infer that the critical point M_0 is unstable, while M_S is a stable one. As a result, if $M(t=0) > 0$ the solution to (34) converges to M_S. Therefore, denoting by $M(t) = \max(x_i)$,

we get

$$\lim_{t\to\infty} M(t) \geq M_S$$

unless $M(0) = 0$. □

C.2.2 Bounds on the minimum. In this section, we demonstrate a similar claim about the stability of the minimal activation level $m(t) = \min(x_i)$ in any solution to Equation (1).

Claim 8. *According to Equation* (1)*, the minimum activation among neurons,* $m(t)$*, has a stable lower bound given by*

$$m_S = \frac{\alpha - (S-1)\beta}{\lambda} - 1 = M_S$$

unless $m(t = 0) = 0$*. The maximal activation level thus converges to* m_S*.*

Proof. Considering now $m(t) = \min(x_i)$ over the active neurons $1 \leq i \leq S$, we easily find that it is bounded from above by the solution of

$$\dot{m} = (\alpha - (S-1)\beta)\frac{m}{1+m} - \lambda m, \quad m(0) = \min(m_i(0)).$$

Therefore, according to the analysis in Section C.2.1 above, if the limit x of $\min(x_i)$ exists, $m(t)$ has two fixed points $m_0 = 0$ and $m_S = \frac{\alpha-(S-1)\beta}{\lambda} - 1$, and only m_S is stable. As a result, we find that

$$\lim_{t\to\infty} m(t) \leq m_S$$

unless $m(0) = 0$. □

Appendix D

Persistence of activation differences

Claim 9. *If neurons have different activations at some time, their activations remain different; the difference between the maximal and the minimal activation increases as soon as the maximal activation drops below* $\sqrt{(\alpha+\beta)/\lambda}-1$. *Further, if a stationary solution exists,* $\sqrt{\frac{\alpha+\beta}{\lambda}}-1$ *is a lower bound for the maximal activation.*

Proof. Below, we have set $M(t) = \max(x_i(t))$ and $m(t) = \min(x_i(t))$. Clearly, we have $M(t) \geq m(t)$. Equation (1) implies

$$\dot{M} = -\lambda M + (\alpha+\beta)F(M) - \beta\sum_{j=1}^{N} F(x_j) \tag{37}$$

$$\dot{m} = -\lambda m + (\alpha+\beta)F(m) - \beta\sum_{j=1}^{N} F(x_j). \tag{38}$$

We deduce

$$\dot{M} - \dot{m} = -\lambda(M-m) + (\alpha+\beta)(F(M) - F(m)). \tag{39}$$

By letting $\Delta = M - m$, we can rewrite (39) as

$$\begin{aligned}
\dot{\Delta} &= -\lambda\Delta + (\alpha+\beta)(F(M) - (F(M-\Delta)) \\
&= -\lambda\Delta + (\alpha+\beta)\frac{\Delta}{(1+M)(1+M-\Delta)} \\
&= -\lambda\Delta + \frac{\alpha+\beta}{1+M}\sum_{k=1}^{\infty}\frac{\Delta^k}{(1+M)^k}
\end{aligned}$$

In the last step, we used the Taylor expansion of the right-hand side around zero. If the network activation converges to zero, the difference Δ also converges to zero. In this case,

we can use the first Taylor term of the right hand side of (39).

$$\dot{\Delta} \approx -\lambda\Delta + \frac{\alpha+\beta}{(1+M)^2}\Delta$$

Note that any solution to (39) is strictly larger than the approximate solution, since the higher-order Taylor terms are strictly positive. Given that $\alpha+\beta > \lambda$, the difference between the minimum and the maximum is increasing when M goes to zero. As a result, as long as there is an initial positive difference between the maximum and the minimum, this difference remains positive. More generally, the difference Δ increases for $M < \sqrt{(\alpha+\beta)/\lambda} - 1$, decreases for $M > \sqrt{(\alpha+\beta)/\lambda} - 1$, and is stationary otherwise. Given that the maximal activation is at least as large as the difference between the minimum and the maximum, at least one neuron thus maintains a positive activation.

Further, $\sqrt{(\alpha+\beta)/\lambda}-1$ is also a lower bound for M (that is, the maximal activation). (39) shows that a stationary point (m_0, M_0) such that the difference between the maximum and the minimum is stationary as well must satisfy

$$\frac{F(M_0) - F(m_0)}{M_0 - m_0} = \frac{\lambda}{\alpha+\beta}, \tag{40}$$

Geometrically, the condition for such a stationary point amounts to saying that the function F has a chord of slope

$$\frac{\lambda}{\alpha+\beta} \tag{41}$$

defined by its values over the points m_0 and M_0. Notice that the derivative of F

$$F'(y) = \frac{1}{(1+y)^2}$$

is strictly decreasing, i.e. F is strictly concave, with $F'(0) = 1$ and $\lim_{y\to\infty} F'(y) = 0$. Further, we assumed that $\lambda/(\alpha+\beta) < 1$ (otherwise, the network would show exponential decay). Hence, $\lambda/(\alpha+\beta)$ is between the values of F' at 0 and infinity, and a chord with this slope thus exists. Further, let y_0 be the unique positive solution of

$$F'(y_0) = \frac{\lambda}{\alpha+\beta}.$$

That is, we identify the argument for which F takes the critical slope that a stationary point (m_0, M_0) must take. This argument is given by

$$y_0 = \sqrt{\frac{\alpha+\beta}{\lambda}} - 1.$$

Since F is concave, it follows that M_0 lies to right of y_0:

$$y_0 < M_0$$

Hence, if a stationary point (m_0, M_0) exists, y_0 is a lower bound for M_0.

In sum, we showed that (1) if neurons have different activations at some time, their activations remain different; that (2) differences in activation do not converge to zero, which implies that at least one neuron must remain active; and that, (3) if a stationary solution

exists, we can identify a lower bound for the maximal activation. □

Appendix E

The role of memory consolidation between input presentations

The simulations with sequential presentation in the main text presented one input after the other, with no consolidation time between inputs. In this section, we describe results of simulations where the network was allowed to settle to a steady-state activity pattern after *each* input presentation. Specifically, the inputs were presented one-by-one, as in the sequential presentation regime above. After each input presentation, however, the network was given 45 time steps to find a steady-state activation.

As shown in Figures E1, E2, and F1, the results are very similar to the sequential results reported above, except that the network performance appears more step-wise and shows oscillatory behavior.

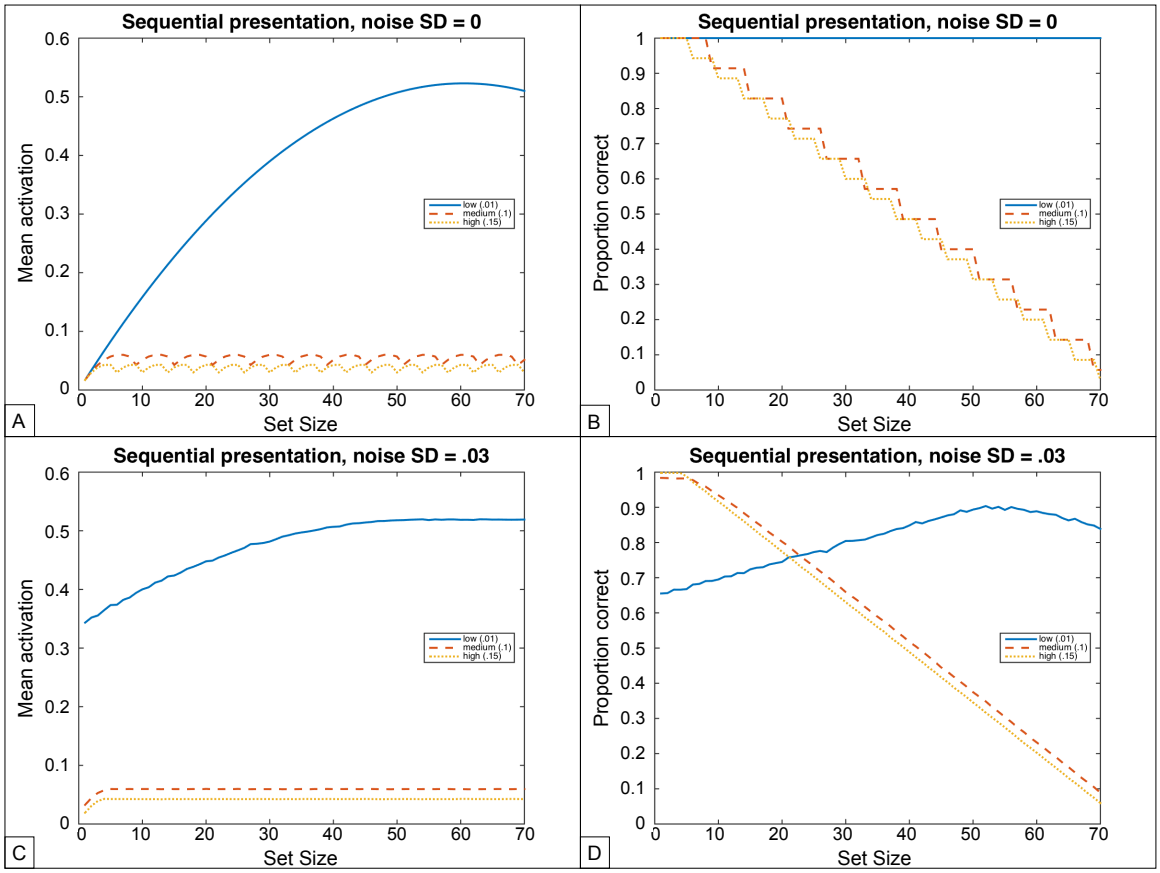

Figure E1. Network results for sequential presentation of the memory items where the network was allowed to consolidate memories after the presentation of each item, for three different levels of inhibition (low, $\beta = .01$, solid blue line; medium, $\beta = .1$, dashed green line; high, $\beta = .15$, dotted red line). The proportion of correct responses is the complement of the Hamming distance between the input and the steady-state activation at end of the simulation. Simulations without noise (A, B). (A) The mean activation in the network is a periodic function. (B) The proportion of correct activations is 1 up to $S_{\max}$, and then linearly decreases to zero in a step-wise fashion. (C,D). The simulation results in the presence of noise are similar, except that they are smoother.

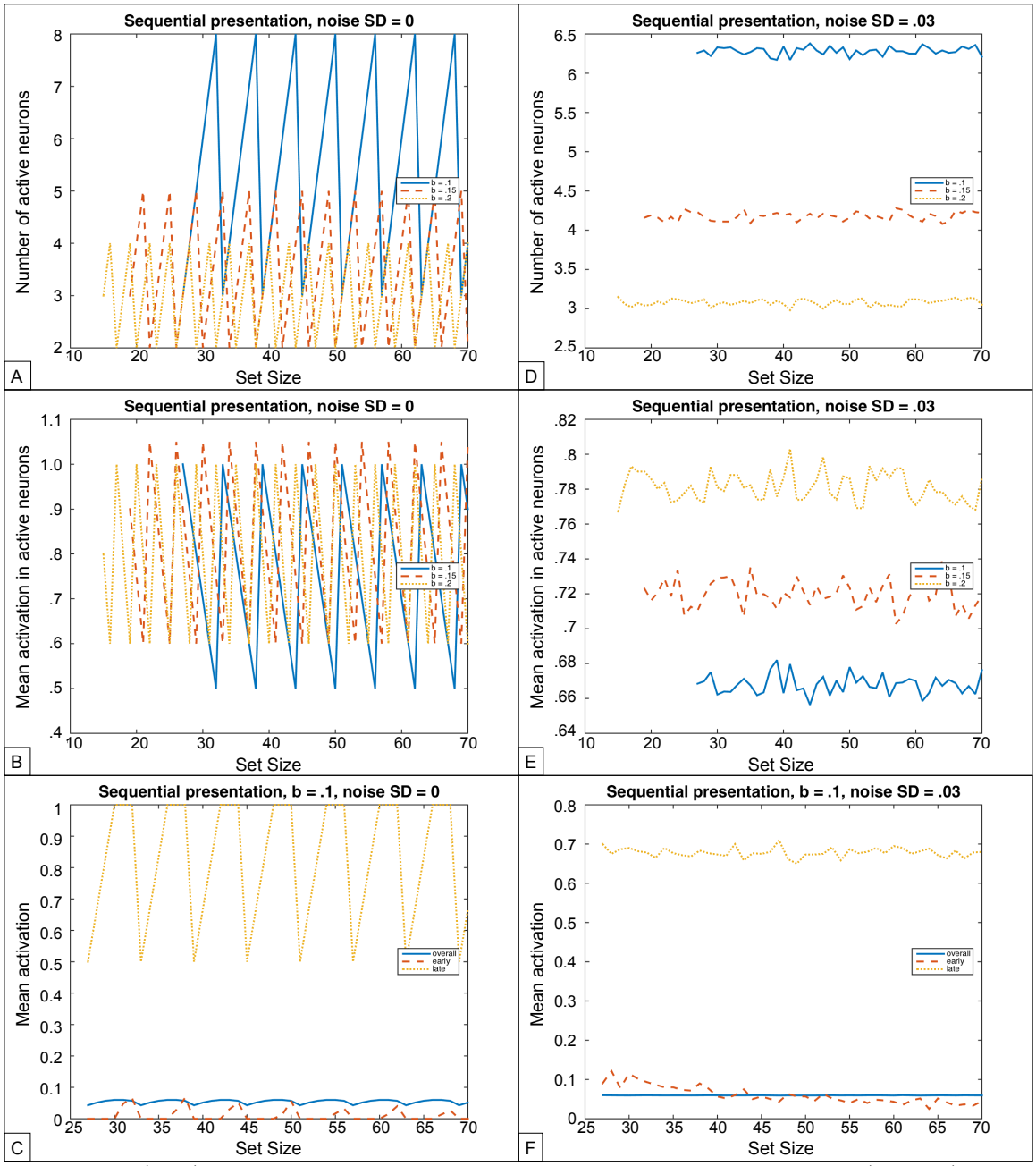

Figure E2. (Top) Number of active neurons at the end of the simulations. (Middle) Mean activation in the active neurons at the end of the simulations. (Bottom) Average activation in the most recent ($S_{\max}$-1)/2 target neurons compared to a random subset of ($S_{\max}$-1)/2 target neurons presented earlier during the input. When the network is given 45 cycles to find a steady-state after each input, there is steady activation for the last inputs, and zero activation for the earlier inputs, though the network behavior becomes oscillatory. (Left) Without noise. (Right). With noise.

Appendix F

The role of network topology in a larger network

In the main simulations above, the network comprised only 70 neurons. As a result, with a maximal set-size of 70, all neurons received external input.

To investigate how the sparsity of the input and the inhibitory interactions would affect the simulation results, we repeated the simulations above with a network with 400 neurons, arranged on a 20×20 grid. We varied the set-size from 1 to 70, and the maximal distance within which inhibitory interactions took place from 1 to 19 (in terms of horizontal and/or vertical steps). As mentioned above, the distance in which interacting neurons can be reached is simply a measure of the number of effectively interacting neighbors.

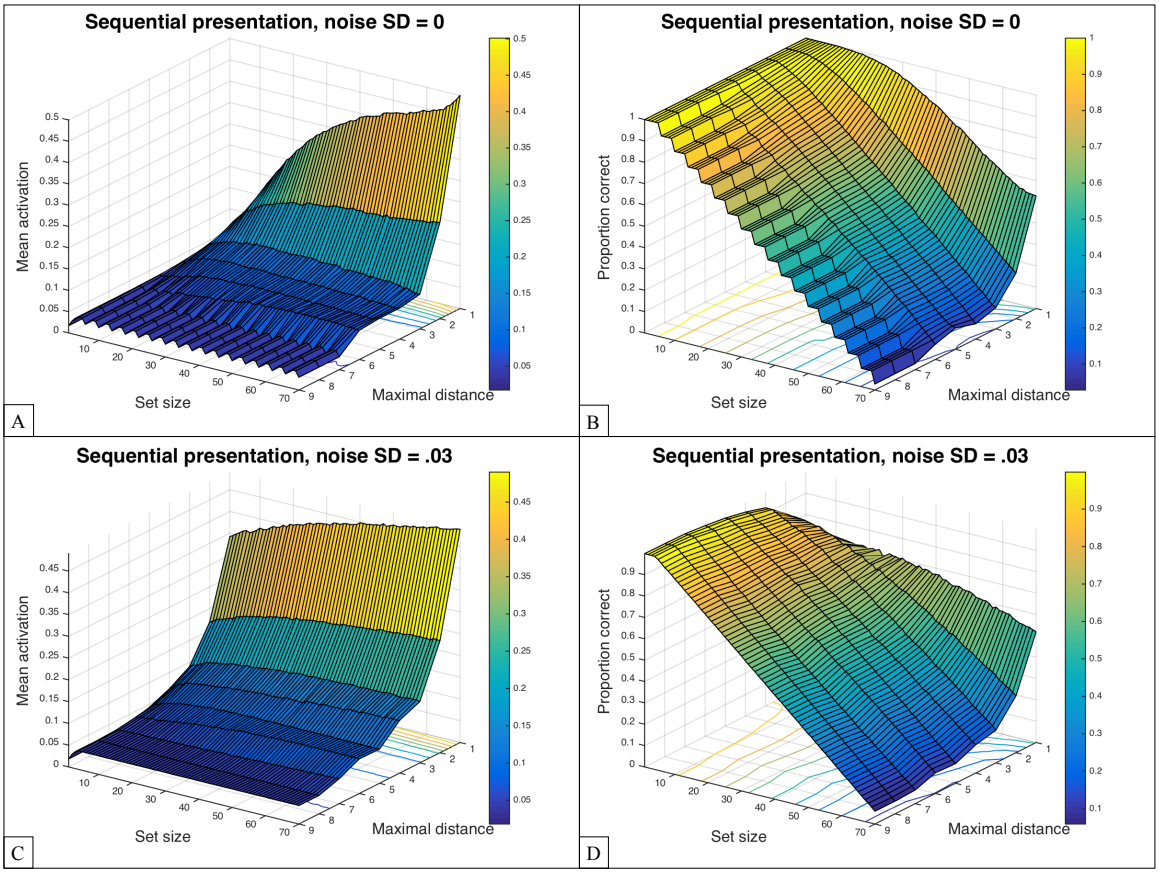

Figure F1. Network results for sequential presentation of the memory items where the network was allowed to consolidate memories after the presentation of each item. The inhibition parameter is set to $\beta = .15$. The maximal distance is the maximal number of (horizontal and/or vertical) steps on the grid in which mutually inhibiting neurons can be reached. The proportion of correct responses is complement of the Hamming distance between the input and the steady-state activation at end of the simulation. (A) Mean network activation in the absence of noise. (B) Faithfulness in the absence of noise. (C) Mean network activation in the presence of noise. (D) Faithfulness in the presence of noise. When inhibition is allowed between neurons that are 2 or 3 steps apart, the network behavior is similar to the fully connected situation.

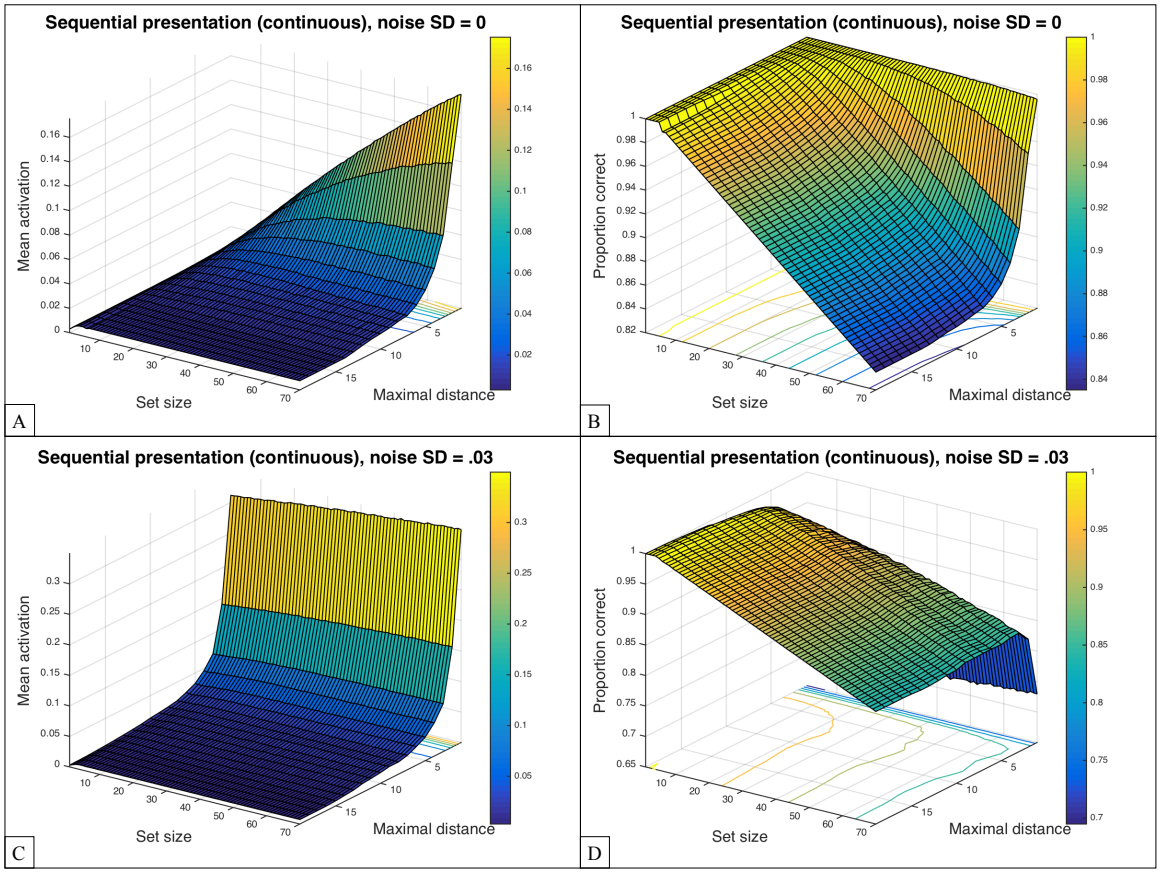

Figure F2. Network results for sequential presentation of the memory items where one memory item is presented after the other, in a network with 400 neurons arranged on a 20×20 grid. The inhibition parameter is set to $\beta = .15$. The maximal distance is the maximal number of (horizontal and/or vertical) steps on the grid in which mutually inhibiting neurons can be reached. The proportion of correct responses is the complement of the Hamming distance between the input and the steady-state activation at end of the simulation. (A) Mean network activation in the absence of noise. (B) Faithfulness in the absence of noise. (C) Mean network activation in the presence of noise. (D) Faithfulness in the presence of noise.

The results are shown in Figure F2. When long-range inhibitory interactions are allowed, the results are very similar to the main simulations reported above. That said, Figures F2(B) and (D) show that the proportion of correct outputs remains at about 80% even for large set-sizes and long-range inhibitory connections. However, this high value is misleading, because the external input to at least $400 - 70 = 330$ neurons was zero. As

a result, even if all neurons in the network are inactive, the proportion of correct outputs would still be at least 330/400 = .825. As shown in Figures F2(A) and (C), the average network activation is very low for large set-sizes and long-range interactions. As a result, the relatively high proportion of correct outputs simply reflects the sparsity of the external input.

Appendix G

Gain control necessarily leads to a sigmoid response profile in a number comparison task

We will now show that there is simple mathematical reason for which performance in a comparison task is necessarily a sigmoid function of the ratio of the comparison stimuli.

In a comparison between two stimuli x_1 and x_2 (with $x_2 > x_1$), the probability of a correct decision is simply the probability that the *internal* representation of x_2 (ξ_2) is greater than the internal representation of x_1 (ξ_1):

$$\text{accuracy} = P\left(\xi_2 - \xi_1 > 0\right) = P\left(\frac{x_2 - x_1}{K} + \epsilon_{\sqrt{2}\sigma} > 0\right),$$

where K is the gain control scaling factor and ϵ is Gaussian noise with a mean of zero and a standard deviation of $\sqrt{2}\sigma$. (The factor $\sqrt{2}$ is due to the fact that we need to add the variances of ξ_1 and ξ_2.)

As mentioned above, the scaling factor K will generally be proportional to x_1. $\frac{x_2 - x_1}{K}$ becomes $\frac{x_2 - x_1}{\alpha x_1}$ and thus $\frac{1}{\alpha}(R-1)$, where α is the proportionality factor. The accuracy can thus be expressed in terms of the ratio of the quantities and the noise level:

$$\text{accuracy} = P\left(\frac{1}{\alpha}(R-1) + \epsilon_{\sqrt{2}\sigma} > 0\right) = P\left(\epsilon_{\sqrt{2}\sigma} > -\frac{1}{\alpha}(R-1)\right)$$
$$= P\left(\epsilon_{\sqrt{2}\alpha\sigma} < R-1\right)$$

This, however, is just an expression for the error function (scaled for a standard deviation of $\sqrt{2}\alpha\sigma$) and thus a sigmoid function.