Reviewer #2 :

The sparseness of neurons, or equivalently tuning bandwidth, has motivated several theoretical models that attempt to describe why neurons have certain response properties (e.g. why simple cells are Gabor-like). This theoretical paper takes a fresh perspective on sparseness, by testing how different levels of sparseness impact neural coding (discrimination and generalization). The authors find that optimal sparseness values fall within the range of values measured experimentally. Thus this paper provides a compelling explanation for "why" different brain areas might have the sparseness values that they do. As such, I believe it will be of general interest to the J Neuroscience audience.

Comments: There are a number of instances in which I found the figures confusing and in some cases, the figures did not directly align with the legends and/or text. I'm also including suggestions that might help make this (highly theoretical) paper more accessible to the general reader. Finally, I have a few suggestions for the discussion.

Fig 1:
Figure 1 is a great intuitive introduction for what follows.

The axes in B and D should be labeled or described.

When I first read the paper, my thought was that one of the papers main points could be more explicitly captured in Fig 1D by plotting a third function under the generalization and discrimination ramps that shows performance of the linear readout (an inverted U). I see that this information is captured later in Fig 9. Still, the authors might consider foreshadowing.

Fig 2:
The title of this figure should read "Segregated representations are NOT linearly separable" (while it is true that they are non-linearly separable, that is not the point)

D, E) The x-axis label should read "Number of inputs TO BE classified" (as is, it is more naturally read as the number of inputs that are classified)

I am confused about how the authors are measuring "classification". It could mean each input should be classified differently (N inputs result in an N-way classification problem) or it could mean that N inputs are categorized into a 2-way classification problem. The text reads as if the problem was the former. Panel E has a lower performance bound of 0.5, suggesting the latter (given that chance for an N-way classification is 1/N).

Fig 3:

What is labeled in Fig 2D as "Num. Dimensions" appears to have transformed into "Dimensionality" in Fig 3. Unless these two concepts are different, they should have consistent labels.

f should be defined in the figure legend and in the text when panels 3A and B are discussed

In subpanel A, it's called "f"; in subpanel C, it's called "Coding level".

The explanation for panel C is indecipherable. First, it explains that it is a plot of the "the black markers of panels A,B" which are described earlier as "the point where 95% of the patterns could be classified correctly". But then it describes that the "ordinate denotes the minimum number of RCNs required to classify 80% of the patterns normalized by ..."; 80% and 95% are inconsistent. Finally, the label on the ordinate is "Nrcn/p" and it's mysterious how that quantity relates to 80% performance as described.

Crit. coding should be defined in the legend and/or text and the term "f crit" should be used consistently.

Fig 4:

p20: I believe the phrase "(obtained by flipping a random fraction n of the source neurons' activities)" is meant to read "(obtained by flipping a random subset of n% of neurons' activities)"

For comparison with figure 5, what is the noise level used in Fig 4? Or if this somehow does not depend on noise level, please explain.

Fig 5:
The y-axis labels for panels C & F should reflect what is plotted (Nrcn required ...)

Panel E and F: p somehow became P

Measuring discrimination and generalization: given that this is a linear discrimination, how are these factors related to the components of the commonly used measure d-prime/ROC, or population-based measures thereof (i.e. aren't they just the numerator and the denominator of these measures)? If these quantities could be related to measures that the general reader is familiar with, that would go along way toward demystifying these quantities.

Discussion:

Simulations are performed under the assumption that all possible input-output functions are of interest. There are many cases in which only a subset of things are of interest (or ever observed in the natural world). In these cases (which I suspect are a majority of cases), I assume that randomly connected networks are probably not the best solution (but rather specifically connected mixtures). How do these results translate to those scenarios?

This paper emphasizes the "average" coding level but one consistent experimental observation is that neurons in any given brain area take on a broad range of sparseness values. The authors might consider speculating on why this might be the case, based on their results.

Reviewer #1 :

"The sparseness of mixed selectivity neurons controls the generalization-discrimination trade off" studies how randomly connected neural networks can reduce correlation due to representing both stimuli and context in order to produce a representation that can be easily read out by a linear classifier. It also shows that a sparseness of ~0.1 for each neuron in the randomly connected network achieves roughly optimal performance. This is a very interesting and insightful paper the builds on recent work by the Fusi lab, which I think is highly creative and innovative. Its models are very reminiscent of cortical data, especially in the frontal cortex, and its prediction for the sparseness of neural firing does correspond roughly with experimental data in many parts of the cortex. Overall, I am very enthusiastic about publication, but have several comments to address below.

In particular, there are several places in the manuscript that seem overly jargony, as the authors seem quite familiar with notation and interpretive language which is not widely used or known. They should go through the entire manuscript with an eye towards reducing jargon and improving clarity for those who have not read their last couple papers.

Major:

• The Introduction is non-standard. This material is appropriate to begin the Results section. A proper Introduction, laying out a broader background to the research and citing relevant papers, needs to be written.

• It is very important to consider the generality of the current result. At least three major questions/issues come to mind. While these questions will be outside of the scope of the technical results of this paper, the authors should consider these issues and try to address them in some fashion in Discussion:

1) The authors consider neural activity patterns that are correlated in a particular form, namely stimulus information in some (half) the neurons combined with context information in the other half of the neurons. If stimuli can be (A,B) and context (C,D), then the network activity patterns can be: AC, AD, BC, or BD. The following patterns are not possible: AA, AB, BB, CC, CD, or DD. This is an interesting and important case, and one that seems particularly apt at in the frontal cortex. However, many other patterns of correlation, many of them inspired by properties of lower stages in the cortical hierarchy, have been considered in the literature: i) in retina and V1, correlation generally refers to spatial correlation due to the properties of natural visual stimuli; decorrelation and/or sparseness does leads to a more efficient neural code rather than changing what information can be read out; ii) in V2, V4, and IT cortex, a common form of correlation comes from sensory
stimuli from the same object having some common, invariant features; here the goal seems to be to classify objects using their invariant features while ignoring information about irrelevant details.

So the central question is: Do the results of this paper - for instance, the effectiveness of random networks and the benefit of sparse activity in this network - generalize to these other kinds of correlation induced by other tasks performed in different parts of the cortex?

2) A striking and obviously important feature of cortical circuitry is the massive, ongoing levels of synaptic plasticity. Random connectivity is attractive as a very weak assumption about the network. But are the properties described here going to be stable under known plasticity rules, like STDP? Alternatively, can synaptic plasticity build on the performance achieved by random networks and give rise to even better performance?

3) The operation described here happens in a single step: strongly correlated neural activity patterns in the input network are successfully decorrelated in a single synaptic transfer within the random network (RCN). But sensory-motor pathways in the cortex are characterized by multiple hierarchical levels. If the

operation studied in this paper is centrally important to cortical computation, then why does the cortex seem to need/use well many hierarchical levels of circuitry?

• Methods, Approximation of the test error:
- this section could use a better transition saying that the paper presents exact numeric simulations of the model but that analytic approximations are useful to understand better various limiting behaviors, etc.; currently, the transition seems quite abrupt and the reader isn't sure if exact calculations will appear.
- this section lists a number of analytic results made; these results need to be derived, otherwise the reader has no way of understanding the approximation that was made. This is particularly true starting with Eqn. 9; there should be either derivations or references.

Moderately important:

• I find the notation in Methods to be reasonably confusing.
- x: you should state that it's values run from 1,...,m_1, as you state for a.
- why combine "x" and "a" in this fashion? one usually thinks of these variables as having completely different roles, yet here they are analogous
- the variable xi uses a superscript mu in most places, but "xa" in its first definition; better to always use mu and then define it in the same sentence
- Eqn. 1: why is there no threshold? this is standard for linear classifiers.

• Figure 6: very confusing
- is this real data or simulated data?
- if these neurons are supposed to be RCNs, then how can one know the actual distance between patterns in the input space (x-axis of the inset)? Wouldn't this require recording from the input neurons, which might be in a different cortical area?
- the arrows illustrating delta_1 and delta_2 seem to be incorrect; delta_2 should link AC with BD as well as AD and BC, right?

Minor:

• Page 5-6: "Unfortunately,transformations that decorrelate (good discrimination) tend to destroy the information about the relative distances in the original space, making it more diffi&#x000E;cult for the readout neurons to generalize (i.e. generate the same output when the inputs are sufficiently similar)."

While this concept might seem obvious to the authors, it is not widely accepted. Authors should add a reference and/or foreshadowing that the manuscript will demonstrate this in Results.

• Page 13: "Figure 2B shows a simple example that illustrates the problem. The four possible configurations of the two populations of N input neurons are four points in a 2N dimensional space. Four points span at most a 3D space (i.e. a solid), but, in our example, these patterns are all on a 2D plane because of their correlations (Figure 2C)."

This notion of correlation is really not very clear. So the point is that the patterns AA, AB, BB, CC, CD, and DD never appear? But if each N neuron activity pattern is randomly generated, then can't you actually get every possible 2N neuron pattern in the population with equal frequency?

• Figure 2 caption, edit for clarity: (p = 25, 100, or 225 possible patterns).

• Figure 3: panel B should be on the upper right and panel C on the lower left.

• Page 15, 17: "As the total number p of input patterns increases and the space spanned by the inputs grows, the RCNs become progressively more efficient at increasing the dimensionality because

they have more chances to be activated (panel B)."
  - isn't another reason the fact that 2m differs more from m^2 as the number of patterns increases?

• Figure 5: just to be clear, these are a discrete set of numeric results, but with so many calculations that one cannot see the discretization on the x-axes? If so, it would help to mention that these are exact numeric calculations.
  - also: the ticks on the x-axes should be bigger and should include coding level = 0.1, as the paper makes a big point about the importance of this value

• Page 23: "Indeed, for a linear transformation (gamma = 0) the dimensionality of the original input space does not increase, and the neural representations would remain non linearly separable."
 - Why does gamma = 0 correspond to a linear transformation?

• Discussion: "Generalization is defined here as the ability to respond in the same way to several noisy
variants of the same input."
  - This notion of generalization is quite non-standard. Often one refers to responding the same way to the same object presented slightly differently to the sensory periphery. Or another connotation is pattern completion of a memory in a Hopfield network.

• Discussion, add a comma: "(as in the case of semantic memories, Hinton, 1981)."

• Discussion, page 29: "Mixed representations arising from dendritic integration would be analogous to the transformation we describe performed by randomly connected neurons. These dendritic representations, however, would not be directly observable in extracellular recordings. In contrast, we know that neurons with non-linear mixed selectivity are widely observed in all areas of the brain (see e.g. Asaad et al., 1998; Rigotti et al., 2010; Warden and Miller, 2010)."
  - The logic here is not clear. If dendrites are extensively nonlinear, would that cause the spiking patterns of the neurons not to have mixed selectivity? It doesn't seem like this evidence rules out highly nonlinear dendritic processing.

• Discussion, page 30: "This is the same scaling as in the case in which the synaptic weights are carefully chosen with an e&#x000E;fficient algorithm, while, surprisingly, RCNs do not require any training. "
  - please add a reference about scaling vs. N for the efficient algorithm.