Nonlinear convergence preserves information

Gabrielle J. Gutierrez[a,b,2], **Fred Rieke**[b,1], **and Eric T. Shea-Brown**[a,1]

[a]University of Washington, Department of Applied Mathematics; [b]University of Washington, Department of Physiology and Biophysics

This manuscript was compiled on May 19, 2020

Neural circuits are structured with layers of converging and diverging connectivity, and selectivity-inducing nonlinearities at neurons and synapses. These components have the potential to hamper an efficient encoding of the circuit inputs. Past computational studies have optimized the nonlinearities of single neurons, or connection weights of networks, to maximize encoded information, but have not grappled with the simultaneous impact of compressive circuit structure and nonlinear response functions for efficient coding. Our approach is to compare model circuits with different combinations of convergence, divergence, and nonlinear neurons to discover how interactions between these components affect coding efficiency. We find that a combination of divergence, convergence, and nonlinear responses retains more information about stimuli than a circuit with linear responses, despite the compressive loss induced by the convergence and the nonlinearities when considered individually. These results show that the combination of selective nonlinearities and a compressive architecture - both elements that induce lossy compression when acting separately - can promote efficient coding.

Neural computation | Efficient Coding | Retina | Sensory Processing | Information Theory

Sensory systems, by necessity, compress a wealth of information gathered by receptors into the smaller amount of information needed to guide behavior. In many systems, this compression occurs via common circuit motifs - namely convergence of multiple inputs into a single neuron and divergence of inputs to multiple parallel pathways (1). Nonlinear circuit elements transform inputs, selecting some parts of the signal while compressing others. Here we investigate how these motifs work together to dictate how much and what information is retained in compressive neural circuits.

These issues are highly relevant to signaling in the retina, because the bottleneck produced by the optic nerve ensures that considerable compression occurs prior to transmission of signals to central targets. This is due to the reduction in the dimensionality of signals as they traverse the retina. In total, signals from $\sim$ 100 million photoreceptors are compressed to modulate the output of $\sim$ 1 million ganglion cells (2, 3).

Receptive field subunits are a key feature of the retina's compressive circuitry. Multiple bipolar cells converge onto a single ganglion cell - forming functional subunits within the receptive field of the ganglion cell (4, 5). Ganglion cell responses can often be modeled as a linear sum of a population of nonlinear subunits. These subunit models have been used to investigate center-surround interactions (6–10) and to explain the nonlinear integration of signals across space (5, 8, 11–13).

While it is clear that subunit coding is inherently compressive, it is not known whether this architecture subserves an efficient code. For decades, information theory (14, 15) has been used to quantify the amount of information that neurons encode (16–25).The efficient coding hypothesis proposes that the distribution of neural responses should be one that is maximally informative about the inputs (19, 20, 26). Take the example of a stimulus variable, such as luminance, where the brightness level is encoded by the number of spikes in the response. An input/output mapping in which most of the possible luminance levels are encoded by the same response, i.e. the same number of spikes, makes bright and dim inputs ambiguous and provides very little information.

Information can be maximized at the level of a single neuron by distributing the responses such that they optimally disambiguate inputs (21). Adaptive rescaling of the response nonlinearity to changes in the input statistics can maintain maximal information in the output (27–29). Alternatively, information can be maximized by optimizing connection weights in the circuit, perhaps in combination with optimizing the nonlinearities (17, 30, 31). This past work, however, has not made explicit how the set of motifs found in most neural circuits, and in the retina in particular, combine to collectively influence coding efficiency.

Our contribution here is to dissect a canonical neural circuit in silico, investigating how much each of its components – some of which are lossy with respect to information on their own – contribute to or detract from the information encoded by the circuit about stimuli. We begin with the simplest motif of converging inputs to single neurons, and analyze the role of rectifying nonlinear subunits applied to each of these multiple inputs. We then add a diverging motif which splits the response into two opposing pathways. We find that rectifying nonlinear subunits mitigate the loss of information from convergence when compared to circuits with linear subunits. This is despite the fact the the nonlinear subunits, considered in isolation,

Significance Statement

Computation in neural circuits relies on a common set of motifs, including divergence of common inputs to parallel paths, convergence of multiple inputs to a single neuron, and nonlinearities that select some signals over others. Convergence and circuit nonlinearities, considered individually, can lead to a loss of information about the inputs, but past work has detailed how to optimize nonlinearities and circuit weights to maximize information. Here, we show that selective nonlinearities, acting together with divergence and convergence, can improve information transmission over a purely linear circuit despite the suboptimality of these components individually. These nonlinearities recode the inputs in a manner that preserves the variance among converged inputs. This extends a broad literature on efficient coding to compressive circuits.

GJG designed experiments, implemented models, performed simulations, wrote manuscript. FR and ETSB conceived experiments, wrote manuscript.

No conflict of interest.

[1]F.R.(Fred Rieke) and E.T.S.B. (Eric T. Shea-Brown) contributed equally to this work as senior co-authors.

[2]To whom correspondence should be addressed. E-mail: ellag9uw.edu

DRAFT

lead to a loss of information. Moreover, this ability of nonlinear subunits to retain information stems from a reformatting of the inputs to encode distinct stimulus features compared with their linear counterparts. Our study contributes to a better understanding of how biologically-inspired circuit structures and neuron properties combine to impact coding efficiency in neural circuits.

Results

We start by quantifying the effect of common circuit motifs, alone and in combination, on coding efficiency. We then explore, geometrically, how nonlinear subunits shape the response distribution to gain intuition as to how they can lead circuits to retain more information. Finally, we explore the implications of nonlinear subunits for encoding stimulus properties. To emphasize the geometrical characterization of the encoding, we use an abstract circuit model without temporal dynamics.

Common circuit components are lossy or inefficient. Our goal is to understand how the combination of divergence of inputs and convergence of nonlinear subunits impacts the retina's ability to efficiently encode spatial inputs. We are particularly interested in the impact of selective nonlinearities on efficient coding. We use Shannon's information to describe the maximum amount of information that a distribution of responses could contain about its inputs (14). We consider deterministic circuits in which the mutual information between the stimulus and response reduces to the entropy of the response. Specifically, we use discrete entropy to compare the information content of continuous distributions of responses generated by different model circuits. The parameters of the discretization were chosen so that distinctions between different distributions could be reasonably interpreted (see Methods).

Many neural circuits are organized in layers that converge and diverge. In the retina (Fig. 1A), this produces a compression and "re-formatting" of a high-dimensional visual input into a lower dimensional neural code that can be interpreted by the brain. In addition, nonlinear responses abound in the neurons that compose these layers. These mechanisms may complicate the ability of the circuit to retain information. For example, two converging inputs can result in ambiguities. With linear convergence, the ability to distinguish the stimulus combinations that sum to the same value is lost and hence this is a form of lossy compression (Fig. 1B). The entropy of the full two-dimensional stimulus (Fig. 1B, top) is 19.85 bits - meaning that a given point in the stimulus space provides 19.85 bits of information about the identity of the stimulus (with our choice of bin size, see Methods). The entropy of the convergent response is smaller (12.50 bits; Fig. 1B, bottom), thus indicating ambiguity in the stimulus identity.

Diverging motifs are another common neural circuit construction. In the example shown in Figure 1C, the divergent responses are identical and the entropy of the 2-dimensional response space (H = 12.01 bits) is the same as the entropy of the 1-dimensional stimulus distribution shown in the top plot (H = 12.01 bits). This demonstrates that divergence of an input into two neurons may produce an inefficient neural architecture by producing redundant (i.e. correlated) signals.

Nonlinearities are abundant in neural circuits, and firing rates generally have a nonlinear relationship to inputs. On a more granular level, neural spike generation mechanisms are nonlinear and can be characterized by a thresholded function. The rectified linear nonlinearity is a tractable representation that captures key features of neural nonlinearities, including the selectivity for some inputs over others. For the subunits and output units in our model, we interpret these nonlinear response functions in terms of the relationship between input and excitatory synaptic output.

Similarly to convergence, nonlinear transformations can lead to loss of information by introducing ambiguities. Take the example of a rectified-linear transformation that is thresholded at zero and is therefore selective for positive inputs (Fig. 1D). It is a non-invertible nonlinearity where half of the stimulus distribution is encoded faithfully and half is mapped to an output of 0 by the thresholded response. Therefore, this nonlinearity induces lossy compression: the information that would distinguish these thresholded stimuli has been irretrievably discarded. Correspondingly, the entropy of the rectified-linear (ReLU) response (H = 6.50 bits) is nearly half

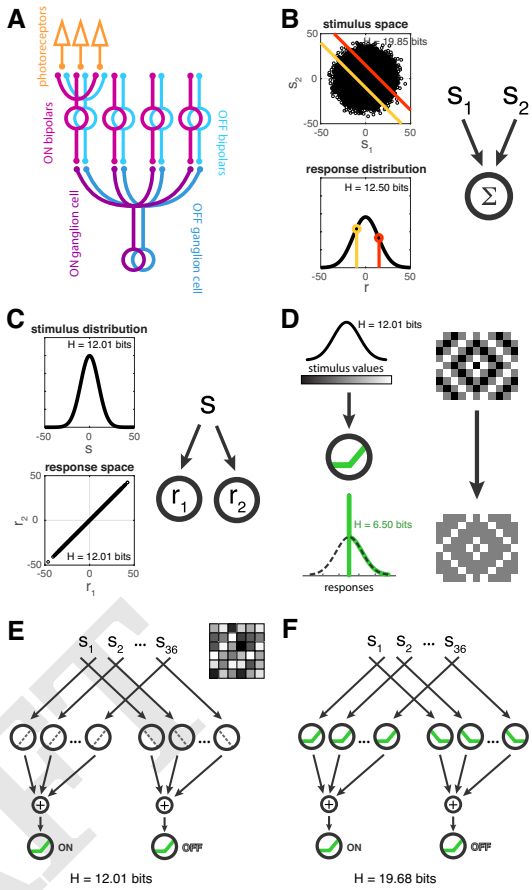

Fig. 1. Neural circuits are composed of inherently lossy components. (A) schematic of retina circuit with its convergent and divergent structure. (B) Converging two inputs results in ambiguities. A 2D stimulus space is reduced to a 1D response space in which one response (bottom: yellow and orange points) represents all stimuli along an isoline (top: yellow and orange lines) where $s_1 + s_2 =$ constant. All entropy values shown are based on a discrete entropy computation (see Methods). (C) Diverging to two outputs can produce redundancies. (D) Nonlinear transformation of a gaussian distributed input with a ReLU (rectified linear unit) can distort the distribution, producing a compressed response in which some portion of the stimulus information is discarded. (E-F) Convergent, divergent circuits with (E) linear subunits, or (F) nonlinear subunits. Subunit responses are weighted by $1/\sqrt{36}$. Example stimulus image is shown.

DRAFT

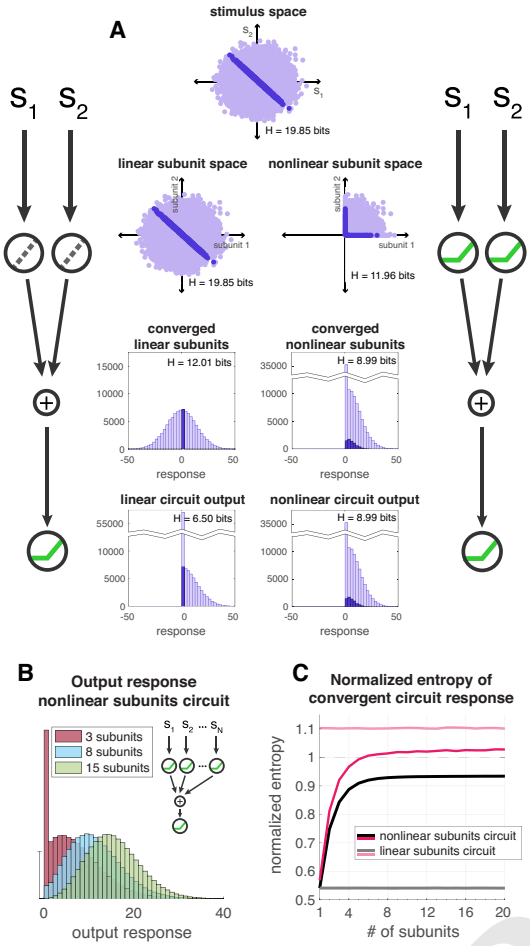

Fig. 2. (A) The encoding of the stimulus space (top) within each layer of a 2-subunit circuit configuration without divergence. Subunits (2nd row); summed subunits response distributions (3rd row); nonlinear output response distributions (4th row). Left, linear subunits circuit; right, nonlinear subunits circuit. The output nonlinearity does not have an additional effect on the summed nonlinear subunits without noise. (B) Compression with increasing numbers of converging subunits. Histograms of the output response for the nonlinear subunits circuit are shown for configurations with 3, 8, and 15 subunits.The subunit responses are normalized so that each subunit is weighted by $1/\sqrt{N}$. The inputs $[s1, s2, \ldots, sN]$ are independently drawn from a gaussian distribution (see Methods). (C) Normalized entropy of output response as a function of number of subunits where subunits are normalized as in B and the circuit entropy is normalized by the entropy of the summed linear subunits (see Methods). Black curve, nonlinear subunits circuit ; gray, linear subunits circuit; dark pink, nonlinear subunits with optimized sigmoidal output nonlinearity; light pink, linear subunits with optimized sigmoidal output nonlinearity. Error bars are too small to show, standard deviation of entropy over 10 runs is on the order of between 10^{-4} and 10^{-2} bits.

of that for the stimulus distribution (H = 12.01 bits).

Each of the common circuit motifs described above is inefficient or discards information when considered in isolation (Figs. 1A-D). How much information can a neural circuit with all of these components retain? We constructed a model circuit that compresses a high-dimensional spatial input into a low-dimensional output. It has a N-dimensional input structure that diverges along two pathways, an ON and an OFF pathway, each culminating in a single output neuron. The inputs to each output neuron come from a layer of subunits - the building blocks for the receptive field structure of the output neuron. Each subunit receives input from one of the N stimulus inputs that compose a stimulus image, and each pixel of the stimulus is independently drawn from a gaussian distribution. Within each pathway, the normalized subunit responses linearly sum at the output neuron and are then rectified.

The ON and OFF output responses lie in a 2-dimensional space, and form a low-dimensional representation of the N-dimensional input. We compute the entropy of the 2-dimensional output response after showing many stimulus samples to the circuit. For a 36-dimensional input space, the circuit with linear subunits has 12.01 bits of entropy (Fig. 1E), while the circuit with nonlinear subunits has 19.68 bits of entropy (Fig. 1F). The greater entropy of the nonlinear subunit circuit is counterintuitive because the nonlinear elements considered in isolation lead to a loss of information (Fig. 1D).

No processing operation can increase the information content of an input signal (Data Processing Inequality, 15). Despite the nonlinear subunits circuit retaining more information than the linear subunits circuit, neither retains the full amount of information in the 36-dimensional input signal which has a much higher entropy (H = 432.46 bits, see Methods) than the 2-dimensional outputs produced by either circuit. The convergence of the inputs necessarily limits the information in the output (15).

This unexpected higher information with nonlinear subunits motivated us to consider how each component or structure interacts with the other components to determine the encoded information. Rather than optimizing the circuit weights or input biases, our goal is to explore the extent to which the information retention of a circuit can benefit from selective nonlinearities that operate independently on signals in each subunit. We next investigate how convergence interacts with subunit nonlinearities.

Lossy nonlinear subunits benefit from convergence. To understand the joint impact of nonlinear subunits and convergent connectivity on encoded information, we examined circuit configurations with a single pathway, i.e. without divergence (Fig. 2). Pathways with two subunits permitted visualization of the input and response spaces. Stimuli that sum to the same value (example highlighted with dark purple in the top plot of Fig. 2A) elicit the same response in the circuit pathway with linear subunits because the subunits do not transform the inputs (Fig. 2A, left, 3rd and 4th rows). The nonlinear subunits transform the stimulus space such that all points are compressed into a single quadrant (Fig. 2A right, 2nd row). Summing the nonlinear subunits (Fig. 2A, right, 3rd row) allows the ambiguous stimuli to have a more distributed representation in the output response - meaning that they are represented more distinctly by the nonlinear subunits pathway than the pathway with linear subunits.

For a configuration with a single subunit, the linear and nonlinear subunits circuits would have identical output responses so long as there remained an output nonlinearity. The 2-subunit circuit (Figure 2A) showed improved information transmission with nonlinear subunits, and this prompted us to ask whether there would be a continued improvement with additional nonlinear subunits. We computed the entropy of the output responses for the linear and nonlinear subunit configurations that converge to a single output for a range of subunit numbers (Fig. 2B,C). With increasing numbers of subunits, more subunit responses are compressed into the output response. To observe a relative change in entropy as the

DRAFT

number of subunits is increased, the subunits were normalized (see Methods).

The distribution of output responses for the nonlinear subunits pathway qualitatively changes with the number of subunits (Fig. 2B). With few subunits, the output response distribution resembles the truncated gaussian seen for the rectified output response in Figures 1D and 2A. With increasing numbers of subunits, the output response distribution approximates a gaussian (due to the central limit theorem) with a mean that shifts towards more positive values (Fig. 2B; also see Fig. S1).

The entropy for the nonlinear subunits pathway increases with increasing subunit dimension (Fig. 2C, black line). It saturates near a normalized value of 0.9, before ever reaching the entropy of the converged linear subunits (where normalized entropy is 1); thus, although increasing convergence improves the information retention of nonlinear subunits, the entropy of the converged nonlinear subunits is apparently bounded by the entropy of the converged linear subunits. The nonlinear subunits only encode positive inputs whereas the linear subunits encode positive and negative inputs. However, when the summation of the linear subunits is followed by a nonlinear rectification at the output, the response entropy is reduced (H = 6.50 bits, Fig. 2C, grey line, normalized H = 0.54) and does not increase beyond that regardless of the number of convergent subunits.

The output nonlinearity reduces the entropy of the linear subunits circuit whereas in the nonlinear subunits circuit, the output nonlinearity does not impact the entropy of the summed nonlinear subunits since the responses have already been rectified. The summed linear subunits produce a gaussian distribution, and the summed nonlinear subunits approach a gaussian distribution as greater numbers of subunits are converged. The entropy of either circuit could be maximized by replacing the output nonlinearity with a sigmoidally shaped nonlinearity that is the cumulative gaussian of the distribution of summed subunit inputs bounded by the maximum and minimum values of that distribution (21, see Methods). Doing so benefits the linear subunits motif more than the rectified subunits motif (compare dark and light pink curves, Fig. 2C) because the variance of the full distribution of summed linear subunits is greater than that for the distribution of summed nonlinear subunits.

Figure 2 shows that convergence reduces the impact of the subunit nonlinearity on the entropy of the circuit. This is because with increasing convergence, the summed nonlinear subunit responses shift so that the output nonlinearity will not truncate them. But the reduced effect on entropy does not mean that the nonlinear subunits cease to affect the circuit outputs; instead, they continue to differentiate stimuli that are made ambiguous by linear convergence (i.e. that sum to the same value). We wondered whether this effect of nonlinear convergence was sufficient to explain why the divergent nonlinear subunits circuit in Figure 1F has higher entropy than the divergent linear subunits circuit (Fig. 1E). We next explore the impact of divergence on information coding with nonlinear subunits.

Divergent circuit structure leverages selectivity of nonlinear subunits. To understand the combined impact of divergence, convergence, and nonlinearities, we present a geometrical exploration of the transformations that take place in the different layers of the circuit with either linear or nonlinear subunits. Our demonstration uses circuits with two input dimensions to facilitate visualization of the stimulus and subunit spaces (Fig. 3).

To determine the optimal nonlinear thresholds, we swept through a range of thresholds for ON and OFF subunits in a divergent, convergent circuit with two inputs and computed the response entropy for each combination of threshold values. Very low thresholds approximate linear functions while high thresholds are extremely rectifying. We found that the optimal combination of ON/OFF subunit thresholds cross at zero (Fig. S2). These zero-crossing nonlinearities are used for all other figures in the main text. Furthermore, when output responses are considered abstractly as static firing rates, this position of thresholds maintains reasonably low output responses (Fig. S3).

The linear ON subunit space (Fig. 3A, 2nd row, left) is identical to the stimulus space because no transformation or compression has taken place through the linear subunits. The OFF subunits receive a negative copy of the same stimulus that the ON subunits receive which reflects the stimuli about the diagonal (Fig. 3A, 2nd row, right). When the linear subunits converge within their respective pathways, the ON and OFF responses are compressed onto a diagonal line because they are anti-correlated (Fig. 3B). This emphasizes the fact that the ON and OFF linear subunits do not have stimulus selectivities in the strictest sense. When the output nonlinearities are applied, this linear manifold is folded into an L-shape (Fig. 3C).

The entropy for the output response of the linear subunits circuit with diverging pathways (H = 12.01 bits, normalized H = 1, Fig. 3C,F, grey) is higher than it was with just a single pathway (H = 6.50 bits, normalized H = 0.54 bits, Fig. 2C, grey). However, it is only increased enough to match the entropy of a single pathway response without any nonlinearities in either the subunits or the output (H = 12.01 bits, normalized H = 1). In other words, the OFF pathway in the linear subunits circuit with output nonlinearities (Fig. 3C) encodes the information discarded by the output nonlinearity in the ON pathway, but it does not enable the divergent linear subunits circuit in Figure 3C to do any better than the convergence of only ON linear subunits (Fig. 2C, dashed line, normalized H = 1). This is because the linear subunits do not select for anything specific and nothing is lost to selectivity; instead the loss of entropy (relative to the entropy of the stimuli) occurs from convergence. Furthermore, when the convergence of the ON linear subunits is followed by a nonlinearity, the positive-summing stimuli are selected. A divergent OFF pathway selects the negative-summing stimuli that the ON pathway discards. Visually, one can see that nothing is lost by folding the linear response space into an L. Thus, the divergent linear subunits circuit recovers what is lost by the output nonlinearities, but not what is lost by convergence.

Unlike the linear subunits, the stimulus undergoes a transformation within the nonlinear subunits layer (Fig. 3D), producing a complimentary compression for the ON and OFF pathways. When these subunits converge in their respective pathways (Fig. 3E), the output response has some similarities to that for the linear subunits circuit (Fig. 3C). The L-shaped manifold is still present, but the points representing

DRAFT

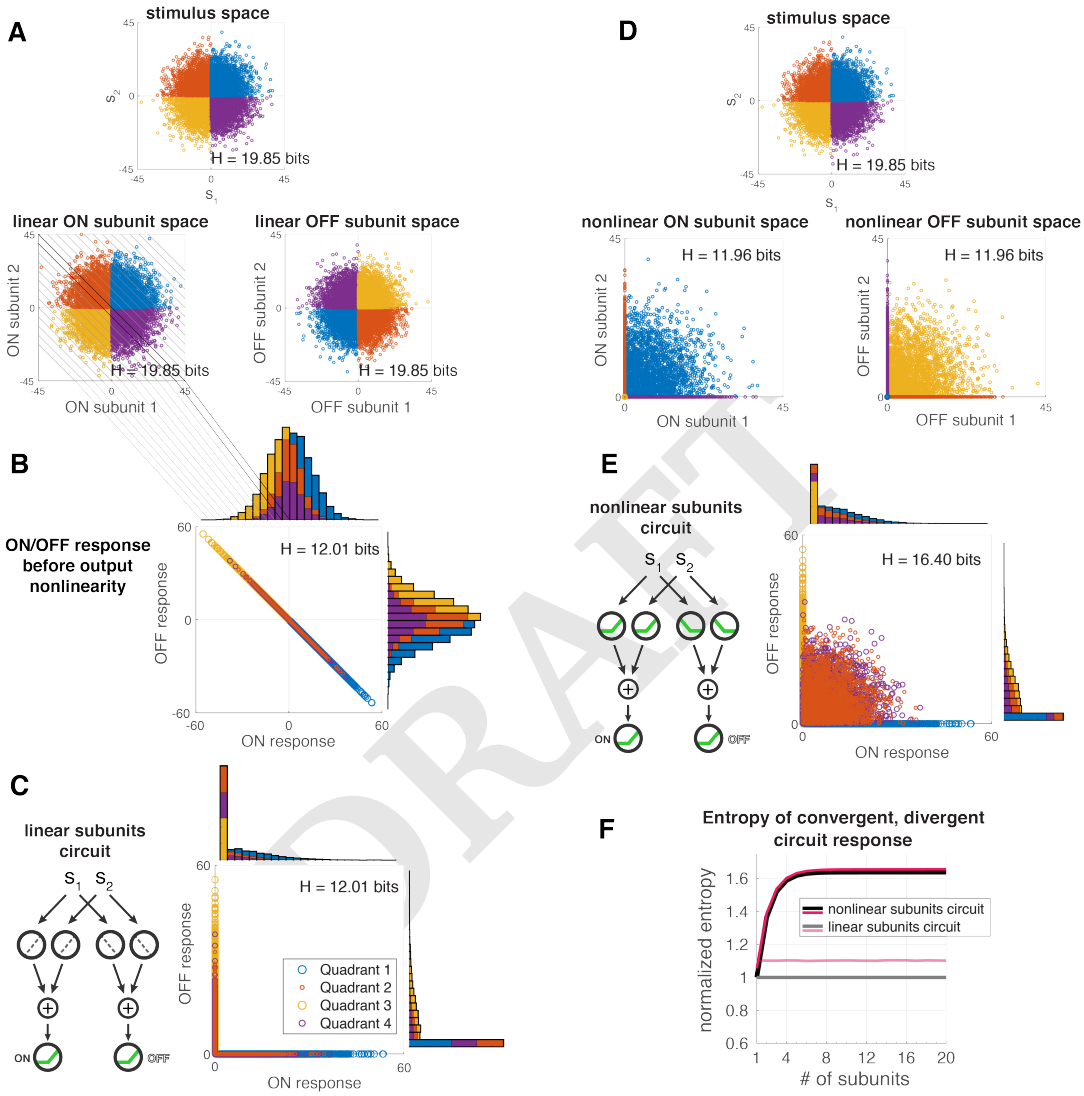

Fig. 3. Visualization of stimulus and response mappings at each level of a convergent, divergent circuit with two inputs, two subunits for each pathway (ON and OFF pathways), and a nonlinear output neuron for each pathway. The points in all subsequent plots are color-coded by the stimulus quadrant from which they originate. (A) The spaces are color-coded according to the 4 stimulus quadrants (top). The 2D stimulus space maps onto a 2D linear subunit space for each pathway (second row, left: ON; right: OFF). The subunit spaces are shown before subunit normalization. (B) The response space is shown for the linear sum of subunits before the nonlinearity is applied and (C) after the nonlinear output response. (D) The 2D stimulus space maps onto a 2D **nonlinear** subunit space for each pathway (second row, left: ON; right: OFF). (E) The output response space for the nonlinear subunits circuit. Note that the output response before the output nonlinearity is applied is identical to the output response after the output nonlinearity is applied for the circuit with nonlinear subunits. (F) Normalized entropy of the output response for convergent, divergent circuits with increasing input and subunit dimension (subunits are normalized as before). The circuit entropy is normalized by the entropy of the summed linear subunits. Gray, linear subunits circuit in C; black, nonlinear subunits circuit in E; light pink, linear subunits circuit with optimal sigmoidal output nonlinearity; dark pink, nonlinear subunits circuit with optimal sigmoidal output nonlinearity.

DRAFT

the stimulus inputs with mixed sign have been projected off it. By virtue of having these points leave the manifold and fill out the response space, entropy is increased. In fact, as more nonlinear subunits converge in a divergent circuit, a greater fraction of points are projected off the manifold along the axes, and as a result the entropy continues to increase until saturation (Fig. 3F, black curve).

The nonlinear subunits circuit does nothing to save the dually positive (blue quadrant) or dually negative (yellow quadrant) stimuli from information loss by convergence. Those are ultimately encoded in the same way as by the linear subunits circuit. In fact, the circuit entropy is not very sensitive to the subunit linearities when the stimuli corresponding to different subunits are correlated (i.e. between s_1 and s_2) and is most sensitive when the stimuli are anti-correlated (Fig. S4). The advantage conferred by the divergent nonlinear subunits is to preserve the variance among the mixed sign stimuli, not only within a single pathway, but also across ON and OFF pathways (this is why adding a bias to the summed linear subunits to evade the output nonlinearity will not give the same results as the nonlinear subunits circuit). As the stimulus dimension is increased, the mixed sign stimuli make up a larger and larger proportion of all stimuli, resulting in the increasing advantage of the nonlinear subunits circuit and its saturation.

To show that the nonlinear subunits themselves confer a unique advantage, we once again replace the output nonlinearities with optimal sigmoidal nonlinearities. The entropy of the linear subunits circuit is increased (Fig. 3F, light pink), however, it is not increased beyond the entropy of the nonlinear subunits circuit with (Fig. 3F, dark pink) or without (Fig. 3F, black) optimal output nonlinearities. This demonstrates that the entropy of the convergent, divergent circuit can be increased beyond an optimization of the output nonlinearities by implementing selective nonlinear subunits.

The rectified output nonlinearities have the effect of decorrelating the ON and OFF output responses in the linear subunits circuit, while for the nonlinear subunits circuit, it is the nonlinear subunits themselves that decorrelate the output response (correlation coefficients: linear response = -1, Fig. 3B; linear subunits, nonlinear output circuit = -0.4670, Fig. 3C; nonlinear subunits circuit = -0.4669, Fig. 3E). Indeed, although the output nonlinearity decorrelates the ON and OFF outputs of the linear subunits circuit, this decorrelation does not produce any gains in entropy relative to the linear subunits circuit before output nonlinearities are applied. Furthermore, the ON and OFF responses of the nonlinear subunits circuit are as decorrelated as for the linear subunits circuit, but unlike the linear subunits circuit, it experiences an entropy gain over the converged linear subunits alone. Complementing the geometrical explanations above, Appendix II presents an analytic argument for why the nonlinear subunits circuit has greater entropy than the linear subunits circuit using the fact that the summed subunit distributions in both circuits are gaussian in the limit of large N subunits (Supplemental Material).

The additional entropy conferred by divergence for the nonlinear subunits circuit is due to *how* the nonlinear subunits decorrelate the ON and OFF pathways, and not merely the fact that those outputs have been decorrelated. It is this subunit processing step that pulls responses off the linear manifold in the output response space leading to an increase in response entropy. The space of the responses in the linear case can be expanded by manipulating the linear subunit weights; however, we find that no rotation of the linear subunit weights can cause the entropy of the linear subunits circuit to surpass that of the nonlinear subunits circuit (Figs. S5 and S6).

To determine whether the increase in entropy for the nonlinear subunits circuit is due to a "synergistic" effect whereby the ON and OFF output responses convey more information together than the sum of the information that each output contains individually (3, 32, 33), we computed the synergy metric ($syn(R_1, R_2) = I(S; R_1, R_2) - I(S; R_1) - I(S; R_2)$) for the different circuit configurations and for a range of subunit numbers (Fig. S7). Positive values of this metric indicate synergy while negative values indicate redundancy. None of the circuits have synergy, and the nonlinear subunits circuit has greater redundancy than the linear subunits circuit. Thus, the nonlinear subunits circuit has higher redundancy *and* higher entropy than the linear subunits circuit.

Increased response entropy could reflect an increased precision in encoding the same stimulus features or the encoding of new stimulus features. We next explore how the processing of mixed sign stimuli by nonlinear subunits creates sensitivity to stimulus features that are not encoded with linear subunits.

Nonlinear subunits circuit encodes both mean and contrast information. To determine whether the higher entropy of the nonlinear subunits circuit accompanies an encoding of additional stimulus features, we visualized the stimulus and response spaces for the linear and nonlinear circuit configurations. The stimulus inputs are assumed to represent luminance values and the distributions are the same as before. We chose two basic features of visual stimuli to investigate: mean luminance and contrast. In Figure 4A, the stimulus space is color-coded by bands of mean luminance levels. In both of the response spaces in Figure 4A, a banded structure is preserved, indicating that there is a separation of the mean luminance levels within the response spaces for the circuits with linear subunits and with nonlinear subunits.

The bottom row shows how contrast is encoded by the different circuits (Fig. 4B). The stimulus space is color-coded for three contrast levels (Fig. 4B, left). The highest contrast areas of the space are in the mixed sign quadrants. The representations for low, medium, and high contrast stimuli overlap each other in the output response space of the linear subunits circuits (Fig. 4B, middle). However, there is separation of these contrast levels in the output response space of the nonlinear subunits circuit (Fig. 4B, right). As in Figure 3F, the encoding of contrast (and mixed sign subunit responses) will increase with increasing numbers of converging subunits. This is reinforced by the result that the nonlinear subunits circuit is more sensitive to anti-correlated stimuli than the linear subunits circuit (Fig. S4, right panel). Thus, the nonlinear subunits circuit encodes both mean and contrast information whereas the linear subunits circuit only encodes mean luminance.

Discussion

In a circuit like the retina, inputs diverge to distinct cell types while neurons receive converging inputs from many presynaptic neurons. This combination of divergence and convergence reorganizes and compresses visual inputs. To determine the

DRAFT

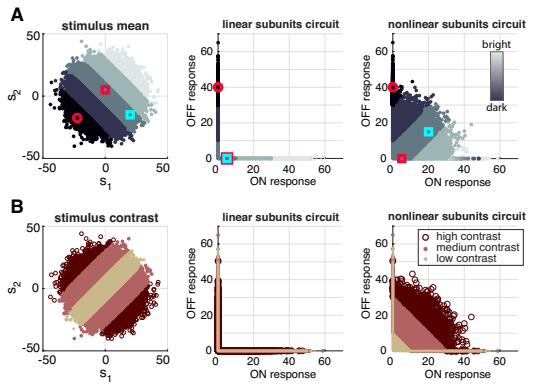

Fig. 4. Mean and contrast encoding of convergent, divergent circuits from Figure 3 (2 inputs, ON and OFF outputs). (A) Visualization of the stimulus mean and response spaces. For example, the bright mean stimulus band contains the 2-pixel image samples that have the highest mean luminance. The red square is an arbitrary reference point. In the stimulus space, the cyan square has the same mean luminance as the red square but a different contrast, while the red circle has the same contrast as the red square but a different mean luminance. (B) Visualization of the stimulus contrast and response spaces. The high contrast stimulus bands contain 2-pixel image samples that have high contrast, whereas the low contrast band contains 2-pixel image samples where the pixel luminance is more correlated.

impact of these common circuit properties on information encoding, we built a circuit model and compared the response entropies of linear and nonlinear subunit configurations. Divergence, convergence, and non-invertible nonlinear signal transformations each have a negative impact on efficiency, or information, individually. However, when arranged together they can mitigate the loss of information that is imposed by the reduction in dimension from inputs to outputs.

The advance made by our study is to demonstrate that rectified nonlinearities can increase the response entropy in a circuit with convergence and divergence, not merely by decorrelating inputs, but by reformatting them. We predict that the information encoded by neurons is maximized by circuit mechanisms that exploit such nonlinearities before inputs converge. This complements known mechanisms, such as adaptation, that enhance coding efficiency by providing a good match of input stimuli and responses at the level of the output neuron (21, 28–30, 34).

Transforming convergent inputs maximizes circuit information. For a single neuron receiving a single input with a known distribution, classical and influential studies prescribe how transmitted information can be maximized by matching the response function to that distribution (21, 28, 29). We consider a complementary question here: when there are multiple input components converging to a neuron, how should those inputs be transformed to maximize the information that a neuron – or that multiple neurons within a divergent output population – can transmit?

The weights with which inputs are combined is a key factor in the information encoded by a circuit and has been studied extensively (17, 31). Here, we highlight a distinct factor: selective "subunit" transformations that are applied to each input component separately before they are combined. We chose non-invertible nonlinearities that exhibit generic selectivity (ON or OFF) and that, individually, induce lossy compression of stimuli with no inherent spatial statistical redundancy to exploit. Despite these properties, we found that, within a convergent, divergent circuit architecture, rectified subunit nonlinearities preserved more information than linear subunits.

This increase comes from a reformatting of the stimulus distribution in a manner that reduces the ambiguities produced by the convergence of multiple inputs (Fig. 2). In our linear subunits circuit, it was possible to spread out the circuit responses by tuning the subunit weights (Fig. S5) such that the ON subunits could be made independent of the OFF subunits. After applying the output nonlinearities, the response space for the linear subunits circuit resembles that for the nonlinear subunits circuit. However, this does not result in an increase in entropy for the linear subunits circuit (see Figs. S5 and S6) because it does not reformat the mixed sign inputs as the nonlinear subunits circuit does (Fig. S5). This reformatting facilitates the encoding of multiple stimulus features (mean luminance and contrast) in Figure 4. Thus, in the circuits we study here, efficient coding can be achieved with selective nonlinear components. We note that even invertible nonlinearities, when followed by noise, will become difficult to invert and will thus behave like the non-invertible nonlinearity studied here.

Reduction in correlations and redundancies is not necessary to increase information. Maximizing information is often seen as equivalent to reducing redundancy (23, 26, 32, 35, 36), this is not the case for the role of subunit nonlinearities in the circuit we study here. The responses from the nonlinear subunits circuit in Figure 3 have more information than those from the linear subunits circuit despite having greater redundancy (Fig. S7) and the same degree of correlation. The neural code in the retina is highly redundant (3, 32), as the degree to which neighboring ganglion cells share information has been estimated as roughly ten-fold (35). Our results indicate that this redundancy need not come at the expense of information about the stimuli.

We find that the efficiency of divergent circuits can be enhanced by nonlinearities that decorrelate the outputs, as others have found (30, 37). Indeed, our findings show that diverging ON and OFF pathways resulted in efficiency gains for both the linear and nonlinear subunits circuits (compare the entropies for the single pathway configurations in Figure 2C to those for the corresponding divergent circuits in Figure 3F). Pitkow and Meister (38) showed that nonlinear responses in ganglion cells have more of an effect on decorrelating their responses than their center-surround receptive field properties. However, as they point out, weak correlation is not necessarily weak dependence. In the divergent, convergent circuits in Figure 3, nonlinearities located either in the output neurons or in the subunits decorrelate the outputs to a similar extent. However, only subunit nonlinearities produce an increase in entropy relative to a fully linear circuit.

Bell and Sejnowski (17) showed that nonlinearities have the effect of reducing redundancy between output neurons by separating statistically independent parts of the inputs. Following that, it was shown that the efficient encoding of natural signals is facilitated by a nonlinear decomposition whose implementation is similar to the nonlinear behaviors observed in neural circuits through divisive normalization (39). Our study contributes to this body of work by showing how nonlinear processing contributes to a more informative, compressed representation without relying on redundancy reduction or decorrelation.

DRAFT

Reconciling selectivity with efficiency. Nonlinearities can have different functional consequences for neurons. Nonlinear transformations can induce selectivity in that they can cause a neuron to encode a very particular aspect of the stimulus or its inputs (40, 41). Nonlinearities can otherwise optimize efficiency by maximizing the entropy of the response distribution (21). The rectified nonlinearity that we used does not maximize the response entropy of the individual neuron that receives gaussian-distributed inputs, but it does enforce a strict selectivity for inputs above threshold. Selectivity would appear to be in conflict with efficient coding in that discarding information is a poor way to maximize it. Our results reveal how selectivity can work in concert with a circuit structure of parallel pathways to produce an efficient encoding of inputs.

The selective coding of features is often conflated with redundancy reduction, but it is important to make a distinction in the context of efficient coding - where a redundancy reducing code is reversible and is expected to maximize information about the stimulus (36). Selectivity indicates that some stimulus information will be irreversibly discarded. The existence of selective cell types that compute different aspects of the visual scene appears to confound an efficient coding framework (38). Yet, properties of selectivity are crucial to the functions of a diverse array of cell types, such as object-selective cells in medial temporal lobe (42), face-selective cells in the inferior temporal cortex (43, 44), and direction-selective cells, orientation-selective cells, and edge detector cells in the retina (45). Furthermore, many cell types in the retina and other circuits have both an ON and an OFF variant, indicating that this kind of ON/OFF selectivity is beneficial to sensory information processing (18, 30).

Implications for artificial neural networks. Although mean and contrast are elementary features of visual inputs, the striations seen in the response space in Figure 4 (nonlinear subunits circuit, right plots) reflect the concept that hidden nonlinear neural units can facilitate the categorization of stimulus features (46). Indeed, in our study simply inserting nonlinear subunits with uniform weights immediately produced a representation that qualitatively enables linear classification or decoding of the mean and contrast levels of the input.

Feedforward artificial neural networks (ANNs) were inspired by the layered organization of biological neural networks. Neural units have activation functions, or static nonlinearities, that transform inputs. Rectified Linear Units (ReLU) such as those used in our nonlinear neural units, enforce a strict selectivity for inputs above threshold; whereas smooth nonlinearities implement a less rigid selectivity, if at all. In both cases, selectivity is dependent on the bias and weight parameters, which can be adjusted by learning, to offset the nonlinearity such that it truncates the input distribution to various degrees. The ReLU frequently has the best performance among other nonlinear activation functions (47, 48) in tasks ranging from the discrimination of handwritten digits to restricted Boltzmann machines (49). The findings presented here of the information preserving capabilities of a selectivity-inducing nonlinear activation within an architecture that is reminiscent of a feedforward ANN complement our knowledge of the ReLU's favorable performance in machine learning and the remarkable classification capabilities of ANNs.

Future directions. Our study did not include noise explicitly, however, the discretization of the response distributions mimics the effects of noise because stimuli that fall into the same discrete bin are ambiguated. Brinkman et al (34) investigated the dependence of mutual information on the source and magnitude of noise in their divergent circuit. They found that for low noise conditions, the optimal sigmoidal nonlinearities cross at their "lower bend," similar to the crossing at threshold for the thresholded nonlinearities in our study. In their study, noise had the effect of changing the optimal parameters for their nonlinearities, but it is unclear whether including or modifying convergence would be more effective for maximizing information than changing the properties of the nonlinearity. Our future studies will investigate how noise impacts the efficiency of a circuit with rectified nonlinearities.

Compressed Sensing is a coding paradigm that has been used to explain the structure of circuit motifs, in olfactory systems in particular. In this framework, signals that are sparse in some basis are efficiently reconstructed in low dimensional representations. Compressed Sensing can be implemented in feedforward networks with converging and diverging architectures and nonlinear neurons (50–52). While the inputs in our study are high dimensional, they are not sparse. Also, in contrast to Compressed Sensing studies, we did not utilize a sensing matrix that determines the optimal connection weights. Instead, we employed uniform weights so as to highlight the contribution of the rectifying nonlinear subunits to the entropy of the circuit responses. Whether convergent subunit nonlinearities may contribute to coding in networks with optimal sensing weights is an interesting question that is still open for investigation.

Finally, our model did not include temporal dynamics. We opted for a granular, geometrical analysis of the set of all possible responses to a fixed and finite set of stimuli so that we could clearly ascertain the counterintuitive finding that rectified nonlinearities allow a convergent, divergent circuit to preserve more information. Despite the lack of temporal dynamics, we compared the effects of different output nonlinearities which abstractly approximate different spike generating mechanisms. Future studies will explicitly include time-dependence to investigate how adapting subunit nonlinearities impact the efficient encoding of inputs with changing stimulus statistics.

Materials and Methods

All simulations and analyses were done in Matlab using custom-written scripts. These can be found on the corresponding author's Github page: https://github.com/gabrielle9/nonlinear-convergence-info-entropy-retention.

We used Shannon's information (14) to quantify the information retention of our model circuits because it quantifies how many distinct neural responses are possible given a particular stimulus distribution, and this relates to the specificity of encoding even though it does not indicate which specific stimulus features are encoded. Since there was no noise anywhere in the circuit, the mutual information between the stimulus and the response reduces to the entropy of the response. Mutual information is defined as (15):

$$MI = H[r] - \langle H[r|s] \rangle \qquad [1]$$

In our study, there is a deterministic relationship between the response and the stimulus due to the lack of noise. The second term of the MI goes to zero and one is left with the entropy of the response. We were careful to avoid any effects that could distort the interpretation of the entropy. The ReLU was ideal here

DRAFT

because it compresses the input signal without necessarily scaling it. In our model, the compression was entirely derived from the non-invertibility of the nonlinearity rather than a linear gain factor.

The convergent structure of the retina reduces the dimension of the high-resolution visual input it receives, placing an upper bound on the amount of information that can possibly be transmitted through the optic nerve. In general, the data compression implemented by the circuit architecture may perform lossless or lossy compression or some combination, depending on the statistics of the inputs. In this study, we focus on lossy compression. By using input images of uncorrelated gaussian random pixels (i.e. no redundant structure), we place the inputs into a regime where lossless compression is impossible or assumed to have already taken place. Therefore, the circuit configuration that experiences less lossy compression has a higher entropy than that which experiences more lossy compression. We thus consider higher entropy to be an indication of better performance.

Model simulations and visualizations. All simulations, visualizations, and entropy computations were done in Matlab. The dimension of the stimulus always matches the dimension of the subunits within a pathway, and a stimulus consists of N stimulus inputs. For example, if there are 5 subunits in each of the ON and OFF pathways, then the stimulus has 5 stimulus inputs (sometimes referred to as pixels). Each stimulus input was independently drawn from a gaussian distribution with arbitrary units ($\mu = 0, \sigma^2 = 10$). Each subunit receives input from one stimulus input. For all figures in this paper, linear subunits did not transform stimulus inputs and therefore the ON linear subunit response was equivalent to the stimulus input and the OFF linear subunit response was the negative of the stimulus input.

All weights were uniform with unit weights from stimulus inputs to subunits and normalized weights from subunits to outputs. The subunits were normalized so that the variance of the linear sum of subunits is maintained. With N subunits, each subunit weight is $1/\sqrt{N}$. This normalization facilitated a comparison between circuit configurations with linear and nonlinear subunits and varying numbers of convergent subunits. All circuit configurations are subject to the same uniform weighting and subunit normalization throughout the paper except where noted in the Supplemental Information and Figures.

Each rectified nonlinear unit has unit slope (slope $= \pm 1$) and applies a threshold to the stimulus input - effectively a positive-pass filter for ON subunits and a negative-pass filter for OFF subunits.

The output neuron linearly sums the subunit responses in its pathway and then applies the output nonlinearity. The output response to a given stimulus was a single value that represents a steady state response, as this model did not have temporal dynamics.

$$R_{ON\ nonlinear\ subunit}(s) = \begin{cases} s, & \text{if } s > 0 \\ 0, & \text{otherwise} \end{cases} \quad [2]$$

$$R_{OFF\ nonlinear\ subunit}(s) = \begin{cases} -s, & \text{if } s < 0 \\ 0, & \text{otherwise} \end{cases} \quad [3]$$

For the summed subunit response before or without an output nonlinearity in a single pathway,

$$R_{output} = \sum_{i}^{N} \frac{1}{\sqrt{N}} R_{subunit\ i} \quad [4]$$

And with an output nonlinearity, the single pathway output response is

$$R_{output} = \begin{cases} \sum_{i}^{N} \frac{1}{\sqrt{N}} R_{subunit\ i}, & \text{if } R_{output} > 0 \\ 0, & \text{othewise} \end{cases} \quad [5]$$

The optimal sigmoidal output nonlinearity was computed as the cumulative gaussian of the neural input distribution, bounded by the maximum and minimum input values:

$$R_{CG\ output}(x) = (max(x) - min(x))CDF(x) + min(x) \quad [6]$$

$$CDF(x) = \frac{1}{2}[1 + erf(\frac{x-\mu}{\sigma\sqrt{2}})] \quad [7]$$

Visualizations in stimulus, subunit, and response spaces. Each quadrant was color-coded such that: $s_1 > 0, s_2 > 0 : blue; s_1 < 0, s_2 > 0 : orange; s_1 < 0, s_2 < 0 : yellow; s_1 > 0, s_2 < 0$ Output response histograms in Figure 3 are also color-coded in this way to show which response bins represent which stimuli. For mean luminance and contrast visualization, spaces were color-coded to indicate bands of mean stimulus luminances, M, and stimulus contrasts, Λ. Each stimulus image, Υ, consists of N stimulus inputs, $\Upsilon = [s_1, s_2, ..., s_N]$. In Figures 3A-E and 4, $N = 2$.

$$M(\Upsilon) = \frac{s_1 + s_2 + ... + s_N}{N} \quad [8]$$

$$\Lambda(\Upsilon) = \left| \sqrt{\frac{1}{2} \sum_{i} (s_i - \langle s \rangle)^2} \right| \quad [9]$$

Entropy calculations. Entropy computations were done by simulating the circuit responses to batches of 10^5 stimulus samples. Discrete entropy was used to quantify continuous stimulus and response distributions. Distributions of stimuli and responses were binned and probabilities were computed from the binned distributions. These binned probability distributions were used to calculate the entropy of the responses. Information entropy is defined as

$$H = -\sum P[r] log_2 P[r] \quad [10]$$

where P[r] is a discrete probability distribution. The entropy quantities presented are the average over 10 batches of 10^5 samples per batch. The normalized entropy was computed by dividing the entropy of the circuit by the entropy of the sum of linear subunits. Thus,

$$H_{normalized} = H_{output}/H_{summed\ linear\ subunits} \quad [11]$$

The entropy of the summed linear subunits was the same for the single pathway as it was for the divergent pathways ($H_{summed\ linear\ subunits} = 12.01$ bits) since the OFF linear subunits are perfectly anti-correlated with the ON linear subunits. This normalization facilitated a comparison between the entropies of the circuit configurations in Figures 2 and 3.

A Freedman-Diaconis histogram bin approximation was used to determine an appropriate bin width for the stimulus and response distributions (53). A consistent bin width of 0.01 was used for all entropy calculations to facilitate comparison. This bin width was used for all dimensions. For example, in a 2-dimensional response space, bins would be boxes that are 0.01 x 0.01. Discretization parameters were chosen carefully to ensure that the bins were sufficiently small to capture the shape of the distributions, but not so small that the log(N) bound was reached.

To determine confidence in the entropy calculation, the sample size at which the entropy converged was determined by computing the entropy for different batch sizes and finding the number of samples for which the entropy converged to a saturation point. These convergence tests were performed for different distributions and a lower bound on sample size was found at 10^5 samples which were used for all of the entropy analyses in this paper.

The 36-dimensional input space used in Figure 1E,F was too large for a numerical computation of the entropy. Its discrete entropy was estimated analytically from its continuous entropy with a bin-correction term as in equation 12 where m = 36, bin width b = 0.01, and K is the covariance matrix of the stimuli.

$$H(s) = h(s) - mlog2(b) = \frac{1}{2} log_2[(2\pi e)^m |K|] - mlog_2(b) \quad [12]$$

ACKNOWLEDGMENTS. We thank Joel Zilberberg, Stephen Baccus, Max Turner, Adree Songco-Aguas, Iris Jianghong Shi, and Matthew Farrell for their helpful feedback and comments on the manuscript and Leenoy Meshulam for their helpful discussions.

Funding sources: GJG is supported by NIH NINDS K22 1K22NS104187-01A1 and WRF UWIN postdoc fellowship, ETSB

DRAFT

was supported by the Boeing professorship in Applied Mathematics at UW and the UW Center for Sensorimotor Neural Engineering. FR is supported by EY028111.

1. JM Jeanne, RI Wilson, Convergence, Divergence, and Reconvergence in a Feedforward Network Improves Neural Speed and Accuracy. *Neuron* **88**, 1014–1026 (2015).
2. L Zhaoping, Theoretical understanding of the early visual processes by data compression and data selection. *Network: Comput. Neural Syst.* **17**, 301–334 (2006).
3. S Nirenberg, SM Carcieri, AL Jacobs, PE Latham, Retinal ganglion cells act largely as independent encoders. *Nature* **411**, 698–701 (2001).
4. JB Demb, JH Singer, Functional Circuitry of the Retina. *Annu. review vision science* **1**, 263–289 (2015).
5. C Enroth-Cugell, JG Robson, The contrast sensitivity of retinal ganglion cells of the cat. *The J. physiology* **187**, 517–552 (1966).
6. C Enroth-Cugell, AW Freeman, The receptive-field spatial structure of cat retinal Y cells. *The J. Physiol.* **384**, 49–79 (1987).
7. S Hochstein, RM Shapley, Linear and Nonlinear Spatial Subunits in Y Cat Retinal Ganglion-Cells. *The J. physiology* **262**, 265–284 (1976).
8. HB Barlow, Summation and inhibition in the frog's retina. *The J. Physiol.* **119**, 69–88 (1953).
9. MH Turner, GW Schwartz, F Rieke, Receptive field center-surround interactions mediate context-dependent spatial contrast encoding in the retina. *eLife* **7**, e38841 (2018).
10. JB Demb, L Haarsma, MA Freed, P Sterling, Functional circuitry of the retinal ganglion cell's nonlinear receptive field. *J Neurosci* **19**, 9756–9767 (1999).
11. MH Turner, F Rieke, Synaptic Rectification Controls Nonlinear Spatial Integration of Natural Visual Inputs. *Neuron* **90**, 1257–1271 (2016).
12. HK Hartline, The effects of spatial summation in the retina on the excitation of the fibers of the optic nerve. *Am. J. Physiol. Content* **130**, 700–711 (1940).
13. MA Freed, P Sterling, The ON-alpha ganglion cell of the cat retina and its presynaptic cell types. *J. Neurosci.* **8**, 2303–2320 (1988).
14. CE Shannon, W Weaver, *The mathematical theory of communication*. (Univ. of Illinois Press, Urbana), (1949) OCLC: 246600266.
15. TM Cover, JA Thomas, *Elements of information theory*. (Wiley-Interscience, Hoboken, N.J), 2nd ed edition, (2006) OCLC: ocm59879802.
16. P Reinagel, Information theory in the brain. *Curr. Biol.* **10**, R542–R544 (2000).
17. AJ Bell, TJ Sejnowski, An Information Maximization Approach to Blind Separation and Blind Deconvolution. *Neural computation* **7**, 1129–1159 (1995).
18. JJ Atick, Could information theory provide an ecological theory of sensory processing? *Network: Comput. Neural Syst.* **3**, 213–251 (1992).
19. F Attneave, Some informational aspects of visual perception. *Psychol. Rev.* **61**, 183–193 (1954).
20. HB Barlow, Possible principles underlying the transformation of sensory messages in *Sensory communication*, ed. WA Rosenblith. (MIT Press), pp. 217–234 (1961).
21. S Laughlin, A Simple Coding Procedure Enhances a Neurons Information Capacity. *Zeitschrift Fur Naturforschung C-a J. Biosci.* **36**, 910–912 (1981).
22. SE Palmer, O Marre, MJ Berry, W Bialek, Predictive information in a sensory population. *Proc. Natl. Acad. Sci. United States Am.* **112**, 6908–6913 (2015).
23. F Rieke, DK Warland, RR de Ruyter van Steveninck, W Bialek, *Spikes: exploring the neural code*, Computational neuroscience. (MIT Press, Cambridge, Mass), (1997).
24. P Sterling, S Laughlin, *Principles of neural design*. (2015).
25. V Balasubramanian, P Sterling, Receptive fields and functional architecture in the retina. *The J. Physiol.* **587**, 2753–2767 (2009).
26. JJ Atick, AN Redlich, Towards a Theory of Early Visual Processing. *Neural Comput.* **2**, 308–320 (1990).
27. N Brenner, W Bialek, RD van Steveninck, Adaptive rescaling maximizes information transmission. *Neuron* **26**, 695–702 (2000).
28. AL Fairhall, GD Lewen, W Bialek, R van Steveninck, Efficiency and ambiguity in an adaptive neural code. *Nature* **412**, 787–792 (2001).
29. B Wark, BN Lundstrom, A Fairhall, Sensory adaptation. *Curr. Opin. Neurobiol.* **17**, 423–429 (2007).
30. J Gjorgjieva, H Sompolinsky, M Meister, Benefits of pathway splitting in sensory coding. *J Neurosci* **34**, 12127–12144 (2014).
31. TO Sharpee, JA Berkowitz, Linking neural responses to behavior with information-preserving population vectors. *Curr. Opin. Behav. Sci.* **29**, 37–44 (2019).
32. E Schneidman, W Bialek, MJ Berry, Synergy, Redundancy, and Independence in Population Codes. *J. Neurosci.* **23**, 11539–11553 (2003).
33. N Brenner, SP Strong, R Koberle, W Bialek, RR de Ruyter van Steveninck, Synergy in a Neural Code. *Neural Comput.* **12**, 1531–1552 (2000).
34. BAW Brinkman, AI Weber, F Rieke, E Shea-Brown, How Do Efficient Coding Strategies Depend on Origins of Noise in Neural Circuits? *PLoS comp bio* **12**, e1005150 (2016).
35. JL Puchalla, E Schneidman, RA Harris, MJ Berry, Redundancy in the Population Code of the Retina. *Neuron* **46**, 493–504 (2005).
36. H Barlow, Redundancy reduction revisited. *Network: Comput. Neural Syst.* **12**, 241–253 (2001).
37. DB Kastner, SA Baccus, TO Sharpee, Critical and maximally informative encoding between neural populations in the retina. *Proc. Natl. Acad. Sci. United States Am.* **112**, 2533–2538 (2015).
38. X Pitkow, M Meister, Decorrelation and efficient coding by retinal ganglion cells. *Nat. Neurosci.* **15**, 628–635 (2012).
39. O Schwartz, EP Simoncelli, Natural signal statistics and sensory gain control. *Nat Neurosci* **4**, 819–825 (2001).
40. T Gollisch, Features and functions of nonlinear spatial integration by retinal ganglion cells. *J. physiology, Paris* **107**, 338–348 (2013).
41. T Gollisch, M Meister, Eye Smarter than Scientists Believed: Neural Computations in Circuits of the Retina. *Neuron* **65**, 150–164 (2010).
42. MJ Ison, et al., Selectivity of pyramidal cells and interneurons in the human medial temporal lobe. *J Neurophys* **106**, 1713–1721 (2011).
43. S Eifuku, WC De Souza, R Tamura, H Nishijo, T Ono, Neuronal Correlates of Face Identification in the Monkey Anterior Temporal Cortical Areas. *J. Neurophysiol.* **91**, 358–371 (2004).
44. ME Hasselmo, ET Rolls, GC Baylis, The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behav. Brain Res.* **32**, 203–218 (1989).
45. JR Sanes, RH Masland, The types of retinal ganglion cells: current status and implications for neuronal classification. *Annu. review neuroscience* **38**, 221–246 (2015).
46. J DiCarlo, D Zoccolan, N Rust, How Does the Brain Solve Visual Object Recognition? *Neuron* **73**, 415–434 (2012).
47. X Glorot, A Bordes, Y Bengio, Deep sparse rectifier neural networks in *Journal of Machine Learning Research*. (Universite de Technologie de Compiègne, Compiegne, France), pp. 315–323 (2011).
48. Y LeCun, Y Bengio, G Hinton, Deep learning. *Nature* **521**, 436–444 (2015).
49. V Nair, GE Hinton, Rectified linear units improve Restricted Boltzmann machines in *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*. (University of Toronto, Toronto, Canada), pp. 807–814 (2010).
50. Y Zhang, TO Sharpee, A Robust Feedforward Model of the Olfactory System. *PLoS Comput. Biol* **12**, e1004850 (2016).
51. CF Stevens, What the fly's nose tells the fly's brain. *PNAS* **112**, 9460–9465 (2015).
52. S Qin, Q Li, C Tang, Y Tu, Optimal compressed sensing strategies for an array of nonlinear olfactory receptor neurons with and without spontaneous activity. *Proc Natl Acad Sci USA* **116**, 20286–20295 (2019).
53. D Freedman, P Diaconis, On the histogram as a density estimator:L 2 theory. *Z. Wahrscheinlichkeitstheorie verw Gebiete* **57**, 453–476 (1981).

DRAFT