Adaptation of retinal discriminability to natural scenes

Xuehao Ding[a], Dongsoo Lee[b], Joshua B. Melander[b], Surya Ganguli[a], Stephen A. Baccus[c]
[a]Department of Applied Physics, Stanford University
[b]Neurosciences PhD Program, Stanford University
[c]Department of Neurobiology, Stanford University

Abstract. Sensory systems discriminate stimuli to direct behavioral choices, a process governed by two distinct properties – neural sensitivity to specific stimuli, and stochastic properties that importantly include neural correlations. Two questions that have received extensive investigation and debate are whether visual systems are optimized for natural scenes, and whether noise correlations contribute to this optimization. However, the lack of sufficient computational models has made these questions inaccessible in the context of the normal function of the visual system, which is to discriminate between natural stimuli. Here we take a direct approach to analyze discriminability under natural scenes for a population of salamander retinal ganglion cells using a model of the retinal neural code that captures both sensitivity and stochasticity. Using methods of information geometry and generative machine learning, we analyzed the manifolds of natural stimuli and neural responses, finding that discriminability in the ganglion cell population adapts to enhance information transmission about natural scenes, in particular about localized motion. Contrary to previous proposals, noise correlations reduce information transmission and arise simply as a natural consequence of the shared circuitry that generates changing spatiotemporal visual sensitivity. These results address a long-standing debate as to the role of retinal correlations in the encoding of natural stimuli and reveal how the highly nonlinear receptive fields of the retina adapt dynamically to increase information transmission under natural scenes by performing the important ethological function of local motion discrimination.

1 Introduction

A central question of sensory neuroscience is how the performance of neural circuits is adapted to the statistics of natural stimuli.[1] Normative approaches, including efficient coding theory, use theoretical methods to derive how optimal sensory circuits should function.[2–4] Although such theories have revealed that simple receptive field properties such as linear stimulus sensitivity or thresholds match the expected optimum,[5–8] they rely on strong and inaccurate assumptions such as Gaussian stimulus statistics and independent neural noise, and have not accounted for complex nonlinear encoding properties.

A second approach has been to measure information transmission under natural and artificial statistics, either in the brain or in behavior.[9–13] However, such analyses are limited to the stimuli presented in finite experiments, and also give no direct insight into the internal computational

properties or mechanisms that generate this increased efficiency.

Here we take a new approach using a mechanistically interpretable network model of the neural code of the retinal ganglion cell population that accurately captures both stimulus sensitivity and stochastic correlations during the presentation of natural scenes. We further model the distribution of natural stimuli using methods from generative machine learning. From these models, we calculate how stimuli can be discriminated given any specific stimulus. This approach not only allows us to assess whether natural stimuli are encoded with greater information, but how the properties of sensory tuning and correlated neural noise adapt to contribute to information transmission for each specific stimulus, thus allowing a direct answer to the long-debated question of the role of noise correlations.[14]

We analyzed the manifolds of natural scenes and ganglion cell responses using information geometry to compute the Fisher information, which accounts for distinguishable stimulus changes as embodied by the "just noticeable difference" that is frequently measured psychophysically in humans. This discriminability limits all behaviors, including subcortical reflexive behaviors and perceptual behaviors distinguishing stimuli in distinct categories. We show that the ganglion cell population adapts dynamically to transmit greater information about the manifold of natural scenes than for random changes in the stimulus. Furthermore, we show that a key property of nonlinear visual processing that contributes to this optimization is the discrimination of local motion. We further find that this optimization arises from the dynamic receptive field structure of the ganglion cell population whereas noise correlations reduce, not enhance, information transmission. These results show that the nonlinear visual sensitivity of the retina adapts dynamically to perform an important behavioral function, to discriminate local motion in natural scenes.

2 Results

2.1 A stochastic model of the retinal code

Discriminability depends on both sensitivity, defined here as the incremental change in the average population response caused by a change in the stimulus, and the stochasticity, defined here as the joint distribution of the population response when the stimulus is fixed. Estimating sensitivity requires a differentiable model that maps the stimulus space to the representation manifold (Fig. 1a), which is composed of the mean activity. To analyze both sensitivity and stochasticity, we created a stochastic encoding model of the retina that captured both the average response to natural scenes and the correlated noise.

We began with a three-layer CNN adopted from Ref.,[15–17] previously shown to capture the visual sensitivity of the retina under natural scenes and other stimuli (Fig. 1b). Stimuli were images from a natural image database[18] jittered with the statistics of fixational eye movements. In addition to this model predicting the average natural scene response nearly to within the variability of ganglion cells, the model trained only on natural scenes generalizes to a wide range of ethological computations of motion processing, detection of stimulus patterns and adaptation. Furthermore, the model is mechanistically interpretable in that the activity of internal units is correlated with retinal interneuron responses recorded separately and not used in training. This model was then extended to extrapolate from recorded neurons to entire populations of recorded cell types by adding a third convolutional layer that was then sampled by each of the recorded neurons (see Methods and Ref.[19]).

We extended this deterministic model of retinal sensitivity to capture the stochastic properties of the ganglion cell population. While freezing all deterministic parameters, stochastic parame-

ters in the model were optimized to fit the single cell variability and pairwise correlations of the recorded ganglion cell population. The only noise sources were independent Gaussian noise added to each cell within the model, with a different noise amplitude for each layer, and binomial noise for each ganglion cell. Simply by adding these few parameters defining independent noise, the full model after optimization successfully captured a variety of commonly used first and second-order statistics including the mean activity, single-cell variability and correlations between cells (Fig. 1c, Fig. S1).

2.2 Retinal discriminability for natural scenes

A complete knowledge of both the deterministic neural code, or sensory tuning, and a model of the population noise for each stimulus determines the geometry of the manifold of neural responses, as well as the transmitted information and discrimination of stimuli.[20] Using principles of Information Geometry,[21] we developed a theory and protocol to analyze the Fisher information of the network in order to determine the discriminability of any changes in the stimulus. For a given stimulus, the network can be considered as a fixed spatiotemporal linear network given that a model consisting only of rectified linear (ReLU) nonlinearities is locally equivalent to an effective linear network,[22] although this linear approximation changes with each stimulus. The stimulus sensitivity (or tuning) of such a linear network is represented by the Jacobian matrix, the rows of which are the instantaneous receptive fields (IRFs) of each ganglion cell. The IRFs change dynamically over timescales of tens to hundreds of ms reflecting adaptation to the specific stimulus sequence[16] (Fig. S9). We estimated the noise covariance matrix of the neural response for each stimulus empirically by presenting the stimulus repeatedly to the model. We then calculated the Fisher Information Matrix (FIM) for each stimulus using the Jacobian matrix to capture sensitivity and the covariance

matrix to capture stochasticity[21] (see Methods).

For each stimulus, we directly computed the stimuli that were most effectively discriminated given the neural code and noise. The most discriminable input (MDI) is the change in the stimulus that the network conveys the most information about in the current state, and was computed as the largest eigenvector of the Fisher Information Matrix. Additional MDIs could be computed as the 2nd and smaller eigenvectors of the FIM. In addition to the MDI, we computed the response induced by the MDI, which is the most discriminative response (MDR), or the change in the response that would convey the most information to the higher brain about the MDI.

We further examined the relationship between the local computation of MDIs, and discriminability between stimuli separated by larger distances in pixel space (Fig. S6). We found that the ordering of MDIs is preserved across distance, indicating that the local measurement of discriminability is representative for larger stimulus changes. We also computed how different cell types contributed to discriminability, either when considered together or separately (Fig. S7). This analysis showed that different parts of the stimulus contributed to the MDRs for different cell types, indicating that different cell types convey information about salient stimulus features in different regions of the image.

2.3 The retina is adapted to natural scenes

We then analyzed whether the retina's ability to discriminate stimuli adapted to the statistics of natural scenes using a generative model of natural images. The set of natural images lies on a manifold (natural image manifold, NIM) with a much smaller dimensionality than the set of all possible images.[1,23] For one natural image represented by a point on this manifold, a change in that image along a direction tangent to this manifold represents a natural change of the image,

whereas a change with components in an orthogonal direction signifies an unnatural change. We tested whether the retinal code is optimized to discriminate stimulus changes within the natural image manifold.

We parameterized the natural image manifold using a generative adversarial network (GAN) trained on natural images, where the latent code input to the generator serves as a coordinate system of the manifold. This approach has been widely used in image editing research primarily because of the discriminative nature of the GAN's latent space and the high quality of generated images[24,25] (Fig. 2a).

For each stimulus, we used the GAN to determine the natural tangent space (NTS), the tangent space of the natural image manifold (NIM), which permits only natural changes. The NTS is a subspace of the full tangent space (FTS), which is the tangent space of the full pixel space. For each image, we generated stimulus frames in the NTS or the FTS that were then jittered with the same trajectory as the original 400 ms stimulus sequence in order to compute discriminability. We found that the most discriminable stimuli (MDI) solved in the full tangent space changed little when it was projected into the natural tangent space (Fig. 2b, Fig. 2c). This was remarkable considering that the dimension of the FTS is substantially higher than that of the NTS, indicating that the neural code is adapted to discriminate changes in natural scenes. Examining the structure of the MDIs, we observed many spatially biphasic MDIs split by edges in the stimulus, indicating that discriminability was tuned to local translational motion (Fig. 2b, Fig. S4).

From the Fisher Information Matrix, we can compute the discriminability of any vector representing a change in the stimulus as the norm of vector defined by the metric of the FIM (see Methods), and so compared the discriminability of a large sample of natural and random changes in the stimulus. We found the discriminability of vectors in the NTS was significantly higher than

those in a randomly selected space of the same dimensionality (random tangent space, RTS) (Fig. 2d), and that constraining the MDI to mixtures of the natural and random tangent spaces steadily increased discriminability as the space became more similar to the NTS (Fig. 2e). Discriminability under natural scenes also decreased with increasing mean intensity consistent with Weber adaptation,[26] and also decreased with increasing standard deviation of the stimulus (Fig. S8).

As an alternative simpler method, we instead altered the Fourier spectra of images, as numerous studies have identified the power-law form of Fourier amplitude spectra as a characteristic feature of natural scenes.[1,27,28] An approach often used in psychophysical studies of discriminability is to alter the phases while preserving the power spectrum to either distort a natural image or to blend two natural images naturally.[11,13] We approximated natural image changes by fixing the power spectrum of natural images but randomly shifting the phases of Fourier components. This approximation was only applied to changes in the image (the natural tangent space), and the full natural image manifold was used to sample the base stimulus. The results from this approximation were very similar to those using the GAN to model the natural tangent space, confirming that the neural code was optimized to discriminate natural scenes (Fig. 2).

2.4 Nonlinear neural code and dynamic adaptation in the retina

We then examined which properties of the neural code caused the retina to discriminate natural stimuli. We created a simplified version of the CNN model by removing the intermediate nonlinearities while keeping other model parameters and noise levels untouched. This model was the commonly used linear-nonlinear (LN) model, having fixed receptive fields that do not adapt to the stimulus. (Fig. S9). For the LN model, the MDI across the population is stimulus-dependent because different cells are activated by different stimuli. To quantify the adaptive properties of the

MDI to natural stimuli for the CNN model compared to the LN model, we defined for each stimulus the MDI selectivity index as the ratio between the discriminability of the MDI of the current stimulus and the average discriminability for many random natural stimuli. We found that compared to the LN model, the nonlinearities in the retina make the MDI more selective. Therefore, the MDI of the full model changes more dramatically with the stimulus than the LN model (Fig. 3a), indicating that the nonlinear retinal neural code dynamically adapts to natural scenes.

We then examined which specific stimulus features were enhanced or suppressed by the more complex nonlinearities of the retina. To characterize the differences in how the CNN and LN models discriminated natural stimuli, we defined the Fisher information difference (FID) as the element-wise difference between the Fisher information matrices computed using the full model and the LN counterpart. Eigenvectors of the FID with positive eigenvalues are stimulus changes whose discriminabilities are enhanced by the nonlinearities, while negative eigenvectors are stimulus changes that are suppressed by those nonlinearities (Fig. 3b). We then examined the structure of the eigenvectors of the FID. In the field of computer vision, it has been proposed extensively that motion estimation in natural scenes can be accomplished effectively using the phase of localized sinusoidal functions known as wavelets.[29–32] As the spatial scale of a wavelet filter increases, the corresponding motion becomes more and more global. We found surprisingly that for most natural stimuli the most enhanced and the most suppressed discriminable stimulus changes can be well approximated by Gabor filters, a type of wavelet (Fig. 3c). The scales of the most suppressed stimulus changes are significantly larger than those of the most enhanced stimulus changes (Fig. 3d), indicating that the nonlinearities in the retina enhance discriminability for local motion but suppress discriminability for global motion. Earlier studies have identified Object Motion Sensitive (OMS) ganglion cells that are sensitive to differential motion as might arise from object motion

and insensitive to global motion as might occur from eye movements.[33] Although these earlier studies used artificial gratings and measured only sensitivity rather than discriminability, our current analysis of the retinal neural code overall points to object motion sensitivity as a key function that drives the dynamic adaptation of discriminability for natural scenes.

2.5 Retinal noise correlations limit information

We then tested which properties of the neural code – sensitivity created by dynamic receptive fields, or stochasticity created by noise correlations – contributed to its adaptation for natural scenes by solving for the most discriminable inputs and most discriminative responses when constrained to either natural or random changes in the stimulus. We found that the property of sensitivity was higher for natural than for random changes (Fig. 4a). However, we found that stochasticity was not different when constrained to the natural or random tangent spaces. This finding demonstrated that optimization of the neural code to natural scenes is driven by adapting receptive fields. This further prompted us to systematically investigate the role of retinal stochasticity, specifically the structure of RGC noise correlations in the processing of information within natural images — a subject of longstanding debate in the field. Multiple scenarios have been advanced where noise correlations are detrimental to information coding.[34,35] However, others suggest that noise correlations under certain circumstances may not limit information coding and can even be beneficial.[36–38] More specifically, theoretical studies have proposed that compared to independent noise, neural correlations driven by noise decrease information when their covariance is aligned with neural correlations driven by signal in the stimulus, whereas noise correlations can increase information when they are structured to be orthogonal to stimulus-driven correlations[14] (Fig. 4b).

We tested these two hypotheses for the retina by computing cosine similarities between the

most discriminative response (MDR) and the most noisy response (MNR), which was computed as the largest eigenvector of the covariance matrix. We found that the most discriminative response mode was aligned with the most stochastic mode and orthogonal to modes with low noise (Fig. 4c), implying that noise correlations in the retina are information-limiting correlations. We also performed a more direct test of this conclusion by analyzing trial-shuffled responses, which have independent noise while preserving the single-cell variability. We found that removing noise correlations in this way created responses that were significantly more discriminative than the original model outputs with correlated noise for every population size (Fig. 4d). These observations, in conjunction with our model's architecture, show that noise correlations do not improve information transmission under natural scenes and may simply arise as a natural consequence of shared circuitry in the retina.

3 Discussion

These results show that the receptive fields of ganglion cells dynamically adapt to enhance discriminability to natural stimulus environments, in particular by enhancing discriminability for local motion. The generalizable approach we have created enables an investigation of stimulus discriminability for a broad range of specific stimulus classes and cell types, allowing a direct examination of the relationship between the neural code and information transmission under any set of stimuli. Given the success of machine learning methods at capturing the sensory neural code[16] and stimulus statistics,[40] we expect that this approach will reveal similar insight into information transmitted about natural signals in other sensory systems.

Our stochastic model was a simple extension from the deterministic model, created by adding independent noise with no additional effort needed to refit the model's synaptic weights to capture

the correlated noise. We therefore consider the possibility that noise correlations may arise as a natural consequence of independent neural noise and synaptic weights constrained by evolution, rather than as a mechanism to optimize information transmission as previously proposed.[36–38] Although our analysis indicates that across a large stimulus ensemble and the full ganglion cell population noise correlations limit information, there remains the possibility that noise correlations increase information under some stimulus conditions and for some cell types.

The results in this work are grounded in the examination of infinitesimal stimulus changes. Although this local discriminability ultimately serves as the limit to discriminate more distinct stimuli, accurately predicting discrimination across larger stimulus changes will require an analysis of the geodesics under the information metric to navigate between different stimuli.[41]

Visual behaviors of any type are similarly limited by the discriminability of the retina. As additional data is gathered from behavioral and retinal experiments under parallel conditions, a direct analysis of the relationship between information transmission and the neural code will allow an assessment of how the visual properties of different cell types are optimized either alone or in larger populations to discriminate different classes of stimuli in order to support diverse visual behaviors.

4 Methods

4.1 Retinal recordings and data preparation

A video monitor projected visual stimuli at 30 Hz controlled by Matlab (Mathworks), using Psychophysics Toolbox.[42,43] Stimuli had a constant mean intensity of $8.3\ mW/m^2$. Images were presented in a 50×50 grid with a square size of $55\ \mu m$ on the retina. We presented a natural scene stimulus that was a sequence of jittered images sampled from a natural image database.[18]

The responses of tiger salamander retinal ganglion cells from 5 animals were recorded using a 60 channel multielectrode array. Further experimental details are described in Ref.[44]

Spiking responses were binned in 10 ms bins and for the training set were further smoothed using a 10 ms Gaussian filter. Each spatiotemporal stimulus had a duration of 400 ms as we focused on shorter timescales of integration. For each preparation, the training dataset of 60 minutes in total was divided according to a $90\%/10\%$ train/validation split, and the test dataset consisted of repeated trials to novel stimuli of 60 seconds natural scene.

4.2 Details of the stochastic encoding model and optimization

The model architecture is shown in Figure 1b. The model takes a spatiotemporal visual stimulus as an input and its output is a set of spike counts, one for each neuron in each time bin. Every convolutional filter in our model was implemented by linearly stacking a sequence of 3×3 small filters, which outperforms the traditional method.[45] A parametric $\tanh$ nonlinearity was attached to the last convolutional layer for the purpose of enforcing the refractory period constraint:

$$f(x; r_{max}, a, b) = \frac{r_{max}}{2} \cdot [\tanh(a(x-b)) + 1]\,, \tag{1}$$

where a, b are cell-dependent model parameters, and r_{max} represents the cell-dependent maximal firing rate.

The essential module that enables us to study retinal population coding is the one-hot layer. Model parameters are fit with the one-hot layer, which bijectively maps a few output units of the final convolutional layer to recorded neurons by the end of the training. The one-hot layer is then removed after parameter optimization so that we can analyze the representation of the full

population, typically consisting of thousands of units. The rationale behind this method is the mosaic organization of each retinal cell type, in which all ganglion cells of a given type tile the retina with their dendrites.[44] Formally, the one-hot layer is defined as

$$\begin{aligned} &\text{out}(k) = \sum_{ij} w_{ij}^k \cdot \text{input}(C(k), i, j),\ \forall\, k \\ &\sum_{ij} w_{ij}^k = 1,\ \forall\, k \\ &w_{ij}^k \geq 0,\ \forall\, k, i, j \end{aligned} \tag{2}$$

where the input is the output of the last convolutional layer, k is the index of recorded neurons, i, j (and p, q below) are location indices, w^k is the linear combination weight vector, and $C(k)$ is the neuron-to-channel map. Each w^k will converge to a one-hot array after model training with our one-hot loss function, which is adapted from the semantic loss in Ref.[46]

$$\begin{aligned} &\mathcal{L}_{\text{one-hot}}(w^k) \propto -\ln \sum_{pq} |\prod_{ij} (w_{ij}^k + \delta_{pq,ij} - 1)|, \\ &\delta_{pq,ij} = \begin{cases} 1 & p = i, q = j \\ 0 & \text{otherwise} \end{cases}, \end{aligned} \tag{3}$$

and which reflects the locations of recorded neurons. In order to determine the number of channels in the last convolutional layer and the neuron-to-channel map $C(k)$, we first fit a model for each preparation assigning a separate channel to each neuron, i.e., $C(k) = k$. We then computed the cosine similarity between the response vectors of every pair of different channels averaged over stimuli. A standard hierarchical clustering was then applied using the cosine similarity matrix as the affinity matrix. Neurons that were clustered into the same group then shared the same cell-type

channel in the final model. The number of clusters as a hyperparameter was explored and the lowest number that did not limit the model performance significantly was chosen.

Gaussian white noise was added to the stimulus to simulate stochasticity in photoreceptors, and was also added to pre-threshold signals in the first two layers to simulate stochasticity in bipolar cells and amacrine cells. We found that Gaussian noise outperforms Poisson noise in these layers, which can presumably be attributed to the noise generation mechanism in the retina and the central limit theorem. The stochasticity of ganglion cells was modeled as a parametric binomial probability mass function (Fig. S2):

$$p^B(n; r, k, N) \propto \frac{\Gamma(k \cdot N + 1)}{\Gamma(k \cdot (N - n) + 1) \cdot \Gamma(k \cdot n + 1)} r^{k \cdot n} \cdot (1 - r)^{k \cdot (N - n)}, \quad n = 0, 1, ..., N \quad (4)$$

where n denotes the spike count in a time bin, r is the rate parameter having a one-to-one correspondence with the mean firing rate μ, k is the variability parameter to be optimized, and N is the cell-dependent upper bound of spike counts in one time bin reflecting the refractory period of that cell, which one can directly determine from data. In practice, for given k and N we can interpolate the $r - \mu$ mapping and use it to determine the rate parameter producing the desired mean firing rate, i.e., $\mu(r) = E_{p(n;r,k,N)}(n)$. We stress that the binomial noise on ganglion cells is independent noise, and thus will not qualitatively affect the geometry and the functional implication of noise correlations.

Since the second order statistics are functions of the entire dataset and cannot be decomposed into terms depending on the individual stimulus, we optimized deterministic parameters and stochastic parameters separately. CNN parameters and one-hot parameters $\{w^k\}$ were optimized together with the standard Poisson loss function and $\mathcal{L}_{\text{one-hot}}$ to fit mean firing rates smoothed

using a $10\ ms$ Gaussian filter. The optimization was performed using ADAM[47] via Pytorch[48] on NVIDIA TITAN Xp, GeForce GTX TITAN X, GeForce RTX 3090, TITAN RTX, TITAN V, and GeForce RTX 2080 Ti GPUs. The network was regularized with an L2 weight penalty at each layer and an L1 penalty on the model output. Then $\tanh$ parameters of the final nonlinearity were fitted with the least square regression. Then we optimized stochastic parameters while freezing all deterministic parameters. Variability parameters k were optimized by maximizing the log-likelihood of recorded spike counts based on empirical firing rates. Standard deviations of Gaussian noise $\{g_0, g_1, g_2\}$ in each layer were optimized through grid searches in order to best fit empirical noise correlations and stimulus correlations by minimizing the smoothed and weighted average of mean squared errors.

4.3 Generative adversarial network

The generator of a trained GAN acts as a mapping from latent space to pixel space, denoted as $G : \mathcal{Z} \to \mathcal{I}$. Consequently, the differentiation of the generator, dG, is a linear transformation between the tangent spaces of these two spaces. We approximated the NTS as the image of dG for each generated image.

We follow the same architecture of deep convolutional generative adversarial networks (DCGAN)[24,49] with the image size $1 \times 128 \times 128$. We trained the network using images sourced from the same natural image database that provided the natural stimuli for the retinal recordings.[18] To prevent mode collapse — a common issue where the network generates a limited variety of outputs — we implemented minibatch discrimination techniques.[50] We experimented with a range of latent dimensions from 100 to 512, finding that our results were consistent regardless of the specific dimensionality chosen.

4.4 Theory of Information Geometry

Let $\boldsymbol{x}$ denote the stimulus vector and $\boldsymbol{y}$ denote the spike count vector. Each stimulus induces a conditional probability distribution of sensory neural responses $P(\boldsymbol{y}|\boldsymbol{x})$ and mean firing rates $\boldsymbol{\mu}(\boldsymbol{x}) \coloneqq E[\boldsymbol{y}|\boldsymbol{x}]$. The representation manifold is defined as the manifold of mean neural responses induced by stimuli: $\{\boldsymbol{\mu}(\boldsymbol{x})\}$. In our case, $\boldsymbol{\mu}(\boldsymbol{x})$ is modeled by the noiseless CNN and $P(\boldsymbol{y}|\boldsymbol{x})$ is modeled by the full stochastic model.

We here define several important matrices. The sensitivity structure is characterized by the Jacobian matrix $\frac{d\boldsymbol{\mu}(\boldsymbol{x})}{d\boldsymbol{x}}$. The stochasticity structure is characterized by the noise covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{x}) \coloneqq E[(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{x})) \cdot (\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{x}))^T|\boldsymbol{x}]$. Assuming Gaussian statistics, the discriminability structure is characterized by the linear Fisher information matrix

$$\boldsymbol{I}(\boldsymbol{x}) \coloneqq \frac{d\boldsymbol{\mu}(\boldsymbol{x})}{d\boldsymbol{x}}^T \cdot \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}) \cdot \frac{d\boldsymbol{\mu}(\boldsymbol{x})}{d\boldsymbol{x}}, \tag{5}$$

which as an information-geometric Riemannian metric measures the Kullback–Leibler divergence between response distributions induced by infinitesimally close stimuli.[21] Consequently, the length of a geodesic connecting two stimuli under this information metric measures the discriminability between them, when represented by the neural population.

The rationale for using linear Fisher information rather than the full metric is that (1) this is a good approximation for our model because a neural network consisting only of ReLU nonlinearities is locally equivalent to an effective linear network[22] and the covariance matrix of a linear network is a constant matrix; and (2) linear Fisher information bounds the performance of unbiased linear decoders which may be more biologically plausible than nonlinear decoders.[51–53]

For any vector $\boldsymbol{\xi}$ in the stimulus tangent space of stimulus $\boldsymbol{x}$ such that $||\boldsymbol{\xi}|| = 1$, the sensitivity

and discriminability of it are given by $||\frac{d\boldsymbol{\mu}(\boldsymbol{x})}{d\boldsymbol{x}} \cdot \boldsymbol{\xi}||$ and $\sqrt{\boldsymbol{\xi}^T \cdot \boldsymbol{I}(\boldsymbol{x}) \cdot \boldsymbol{\xi}}$, respectively. For any response mode $\boldsymbol{a}$ such that $||\boldsymbol{a}|| = 1$ with stimulus $\boldsymbol{x}$, its sensitivity and stochasticity are given by $||\frac{d\boldsymbol{\mu}(\boldsymbol{x})}{d\boldsymbol{x}}^T \cdot \boldsymbol{a}||$ and $\sqrt{\boldsymbol{a}^T \cdot \boldsymbol{\Sigma}(\boldsymbol{x}) \cdot \boldsymbol{a}}$, respectively. Furthermore, MDI and MNR can be obtained as the principal eigenvectors of $\boldsymbol{I}(\boldsymbol{x})$ and $\boldsymbol{\Sigma}(\boldsymbol{x})$, respectively. And MDR is defined as the pushforward of MDI by $\boldsymbol{\mu}(\boldsymbol{x})$: MDR $\propto \frac{d\boldsymbol{\mu}(\boldsymbol{x})}{d\boldsymbol{x}} \cdot$ MDI.

One can also evaluate the Fisher information matrix and MDI constrained in a subspace of the full stimulus tangent space by

$$\begin{aligned} \widetilde{\boldsymbol{I}}(\boldsymbol{x}) &= \frac{d\boldsymbol{z}}{d\boldsymbol{x}}^T \cdot \frac{d\boldsymbol{z}}{d\boldsymbol{x}} \cdot \boldsymbol{I}(\boldsymbol{x}) \cdot \frac{d\boldsymbol{z}}{d\boldsymbol{x}}^T \cdot \frac{d\boldsymbol{z}}{d\boldsymbol{x}} \\ &= \frac{d\boldsymbol{z}}{d\boldsymbol{x}}^T \cdot \frac{d\boldsymbol{\mu}[\boldsymbol{x}(\boldsymbol{z})]}{d\boldsymbol{z}}^T \cdot \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}) \cdot \frac{d\boldsymbol{\mu}[\boldsymbol{x}(\boldsymbol{z})]}{d\boldsymbol{z}} \cdot \frac{d\boldsymbol{z}}{d\boldsymbol{x}}, \end{aligned} \tag{6}$$

where $\boldsymbol{z}$ denotes the coordinate with respect to an orthonormal basis of the subspace. Since $\frac{d\boldsymbol{z}}{d\boldsymbol{x}}$ is a semi-orthogonal matrix, eigendecomposing $\widetilde{\boldsymbol{I}}(\boldsymbol{x})$ simplifies to eigendecomposing $\frac{d\boldsymbol{\mu}[\boldsymbol{x}(\boldsymbol{z})]}{d\boldsymbol{z}}^T \cdot \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}) \cdot \frac{d\boldsymbol{\mu}[\boldsymbol{x}(\boldsymbol{z})]}{d\boldsymbol{z}}$, which is more feasible computationally due to the lower dimensionality of the subspace. This method facilitated the computation of MDIs constrained in subspaces discussed in the main text. Additionally, even for computing MDIs in the spatio-temporal full stimulus tangent space, this method was also employed as we proved elsewhere that all MDIs must reside in the subspace spanned by instantaneous receptive fields, i.e., span$(\{\frac{d\boldsymbol{\mu_i}(\boldsymbol{x})}{d\boldsymbol{x}}\})$.[19]

4.5 Statistical measures

Let the neural response be represented by the tensor $R_{tsc} \in \mathbb{N}$ denoting the spike count of the c-th cell during the s-th time bin in the t-th trial. Statistical measures used in the main text are defined as follows.

Total pairwise correlation:

$$C_{i,j}^{total} := \langle \frac{\langle (R_{tsi} - \langle R_{ts'i} \rangle_{s'}) \cdot (R_{tsj} - \langle R_{ts'j} \rangle_{s'}) \rangle_s}{\sqrt{\langle (R_{tsi} - \langle R_{ts'i} \rangle_{s'})^2 \rangle_s \cdot \langle (R_{tsj} - \langle R_{ts'j} \rangle_{s'})^2 \rangle_s}} \rangle_t. \tag{7}$$

Stimulus correlation:

$$C_{i,j}^{stim} := \frac{\langle (\langle R_{tsi} \rangle_t - \langle R_{ts'i} \rangle_{ts'}) \cdot (\langle R_{tsj} \rangle_t - \langle R_{ts'j} \rangle_{ts'}) \rangle_s}{\sqrt{\langle (R_{tsi} - \langle R_{t's'i} \rangle_{t's'})^2 \rangle_{ts} \cdot \langle (R_{tsj} - \langle R_{t's'j} \rangle_{t's'})^2 \rangle_{ts}}}. \tag{8}$$

Noise correlation:

$$C_{i,j}^{noise} := \frac{\langle (R_{tsi} - \langle R_{t'si} \rangle_{t'}) \cdot (R_{tsj} - \langle R_{t'sj} \rangle_{t'}) \rangle_{ts}}{\sqrt{\langle (R_{tsi} - \langle R_{t's'i} \rangle_{t's'})^2 \rangle_{ts} \cdot \langle (R_{tsj} - \langle R_{t's'j} \rangle_{t's'})^2 \rangle_{ts}}}. \tag{9}$$

Trial-to-trial correlation:

$$C_i^{trial} := \langle \frac{\langle (R_{asi} - \langle R_{as'i} \rangle_{s'}) \cdot (R_{bsi} - \langle R_{bs'i} \rangle_{s'}) \rangle_s}{\sqrt{\langle (R_{asi} - \langle R_{as'i} \rangle_{s'})^2 \rangle_s \cdot \langle (R_{bsi} - \langle R_{bs'i} \rangle_{s'})^2 \rangle_s}} \rangle_{a \neq b}. \tag{10}$$

Fano factor:

$$F_i := \langle \frac{\langle (R_{tsi} - \langle R_{t'si} \rangle_{t'})^2 \rangle_t}{\langle R_{tsi} \rangle_t} \rangle_s. \tag{11}$$

In some previous work, the stimulus correlation is alternatively defined as the average correlation between the neural responses of two cells in different trials. Correspondingly, the noise correlation is defined as the difference between the total correlation and the stimulus correlation. In practice, this set of definitions always produced very similar values to our definitions.

4.6 Gabor filter

The Gabor filters we used take the following form:

$$\psi(x, y; x_0, y_0, a, b, \theta, k, \phi) \propto \exp(-\frac{x'^2}{2a^2} - \frac{y'^2}{2b^2}) \cdot \cos(k \cdot x' + \phi),$$
$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \cdot \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix}, \tag{12}$$

where $\{x_0, y_0, a, b, \theta, k, \phi\}$ is the set of parameters. We optimized these parameters to maximize the cosine similarity. The area of each Gabor filter is defined as the area of the ellipse contour: πab.

Acknowledgments

This work was supported by grants from the NEI, R01EY022933, R01EY025087 and P30EY026877 (SAB).

References

1 Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

2 Simon Laughlin. A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung c*, 36(9-10):910–912, 1981.

3 Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

4 William E Vinje and Jack L Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287:1273–1276, 2000.

5 J Hans Van Hateren. Spatiotemporal contrast sensitivity of early vision. *Vision research*, 33(2):257–267, 1993.

6 Xaq Pitkow and Markus Meister. Decorrelation and efficient coding by retinal ganglion cells. *Nature neuroscience*, 15(4):628–635, 2012.

7 Michael S Lewicki. Efficient coding of natural sounds. *Nature neuroscience*, 5(4):356–363, 2002.

8 Eizaburo Doi, Jeffrey L Gauthier, Greg D Field, Jonathon Shlens, Alexander Sher, Martin Greschner, Timothy A Machado, Lauren H Jepson, Keith Mathieson, Deborah E Gunning, et al. Efficient coding of spatial information in the primate retina. *Journal of Neuroscience*, 32(46):16256–16264, 2012.

9 Yang Dan, Joseph J Atick, and R Clay Reid. Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *Journal of neuroscience*, 16(10):3351–3362, 1996.

10 Fred Rieke, DA Bodnar, and William Bialek. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 262(1365):259–265, 1995.

11 David J Tolhurst and Yoav Tadmor. Discrimination of spectrally blended natural images: Optimisation of the human visual system for encoding natural images. *Perception*, 29(9):1087–1100, 2000.

12 C Alejandro Parraga, Tom Troscianko, and David J Tolhurst. The human visual system is optimised for processing the spatial information in natural visual images. *Current Biology*, 10(1):35–38, 2000.

13 Ingo Fruend and Elee Stalker. Human sensitivity to perturbations constrained by a model of the natural image manifold. *Journal of Vision*, 18(11):20–20, 2018.

14 Bruno B Averbeck, Peter E Latham, and Alexandre Pouget. Neural correlations, population coding and computation. *Nature reviews neuroscience*, 7(5):358–366, 2006.

15 Lane McIntosh, Niru Maheswaranathan, Aran Nayebi, Surya Ganguli, and Stephen Baccus. Deep learning models of the retinal response to natural scenes. *Advances in neural information processing systems*, 29, 2016.

16 Niru Maheswaranathan, Lane T McIntosh, Hidenori Tanaka, Satchel Grant, David B Kastner, Joshua B Melander, Aran Nayebi, Luke E Brezovec, Julia H Wang, Surya Ganguli, et al. Interpreting the retinal neural code for natural scenes: From computations to neurons. *Neuron*, 111(17):2742–2755, 2023.

17 Hidenori Tanaka, Aran Nayebi, Niru Maheswaranathan, Lane McIntosh, Stephen Baccus, and Surya Ganguli. From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction. *Advances in neural information processing systems*, 32, 2019.

18 Gašper Tkačik, Patrick Garrigan, Charles Ratliff, Grega Milčinski, Jennifer M Klein, Lucia H Seyfarth, Peter Sterling, David H Brainard, and Vijay Balasubramanian. Natural images from the birthplace of the human eye. *PLoS one*, 6(6):e20409, 2011.

19 Xuehao Ding, Dongsoo Lee, Joshua Melander, George Sivulka, Surya Ganguli, and Stephen Baccus. Information geometry of the retinal representation manifold. *Advances in Neural Information Processing Systems*, 36, 2024.

20 Nikolaus Kriegeskorte and Xue-Xin Wei. Neural tuning and representational geometry. *Nature Reviews Neuroscience*, 22(11):703–718, 2021.

21 Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

22 Haozhe Shan and Haim Sompolinsky. Minimum perturbation theory of deep perceptual learning. *Physical Review E*, 106(6):064406, 2022.

23 Javier Portilla and Eero P Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40:49–70, 2000.

24 Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 597–613. Springer, 2016.

25 Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

26 Jingyang Zhou, Lyndon R Duong, and Eero P Simoncelli. A common framework for discriminability and perceived intensity of sensory stimuli. *bioRxiv*, pages 2022–04, 2022.

27 David J Tolhurst, Yoav Tadmor, and Tang Chao. Amplitude spectra of natural images. *Ophthalmic and Physiological Optics*, 12(2):229–232, 1992.

28 Daniel L Ruderman and William Bialek. Statistics of natural images: Scaling in the woods. *Physical review letters*, 73(6):814, 1994.

29 David J Fleet and Allan D Jepson. Computation of component image velocity from local phase information. *International journal of computer vision*, 5:77–104, 1990.

30 Temujin Gautama and MA Van Hulle. A phase-based approach to the estimation of the optical

flow field using spatial filtering. *IEEE transactions on neural networks*, 13(5):1127–1136, 2002.

31 Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T Freeman. Phase-based video motion processing. *ACM Transactions on Graphics (ToG)*, 32(4):1–10, 2013.

32 Pierre-Étienne Fiquet and Eero Simoncelli. A polar prediction model for learning to represent visual transformations. *Advances in Neural Information Processing Systems*, 36, 2024.

33 Bence P Ölveczky, Stephen A Baccus, and Markus Meister. Segregation of object and background motion in the retina. *Nature*, 423(6938):401–408, 2003.

34 Ingmar Kanitscheider, Ruben Coen-Cagli, and Alexandre Pouget. Origin of information-limiting noise correlations. *Proceedings of the National Academy of Sciences*, 112(50):E6973–E6982, 2015.

35 Rava Azeredo da Silveira and Fred Rieke. The geometry of information coding in correlated neural populations. *Annual Review of Neuroscience*, 44:403–424, 2021.

36 Jon Cafaro and Fred Rieke. Noise correlations improve response fidelity and stimulus encoding. *Nature*, 468(7326):964–967, 2010.

37 Rubén Moreno-Bote, Jeffrey Beck, Ingmar Kanitscheider, Xaq Pitkow, Peter Latham, and Alexandre Pouget. Information-limiting correlations. *Nature neuroscience*, 17(10):1410–1417, 2014.

38 Oleg I Rumyantsev, Jérôme A Lecoq, Oscar Hernandez, Yanping Zhang, Joan Savall, Radosław Chrapkiewicz, Jane Li, Hongkui Zeng, Surya Ganguli, and Mark J Schnitzer. Fundamental bounds on the fidelity of sensory cortical coding. *Nature*, 580(7801):100–105, 2020.

39 Elad Schneidman, Michael J Berry, Ronen Segev, and William Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, 2006.

40 Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

41 Olivier J Hénaff and Eero P Simoncelli. Geodesics of learned representations. In *4th International Conference on Learning Representations, ICLR 2016*, 2016.

42 Brainard and D. H. The psychophysics toolbox. *Spatial vision*, 10(4):433–436, 1997.

43 Denis G. Pelli. The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision*, 10:437–442, 1997.

44 David B Kastner and Stephen A Baccus. Coordinated dynamic encoding in the retina using opposing forms of plasticity. *Nature neuroscience*, 14(10):1317–1322, 2011.

45 Xuehao Ding, Dongsoo Lee, Satchel Grant, Heike Stein, Lane McIntosh, Niru Maheswaranathan, and Stephen Baccus. A mechanistically interpretable model of the retinal neural code for natural scenes with multiscale adaptive dynamics. In *2021 55th Asilomar Conference on Signals, Systems, and Computers*, pages 287–291. IEEE, 2021.

46 Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Broeck. A semantic loss function for deep learning with symbolic knowledge. In *International conference on machine learning*, pages 5502–5511. PMLR, 2018.

47 Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

48 Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

49 Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

50 Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

51 Jeffrey Beck, Vikranth R Bejjanki, and Alexandre Pouget. Insights from a simple expression for linear fisher information in a recurrently connected population of spiking neurons. *Neural computation*, 23(6):1484–1502, 2011.

52 Ingmar Kanitscheider, Ruben Coen-Cagli, Adam Kohn, and Alexandre Pouget. Measuring fisher information accurately in correlated neural populations. *PLoS computational biology*, 11(6):e1004218, 2015.

53 Adam Kohn, Ruben Coen-Cagli, Ingmar Kanitscheider, and Alexandre Pouget. Correlations and neuronal population information. *Annual review of neuroscience*, 39:237–256, 2016.

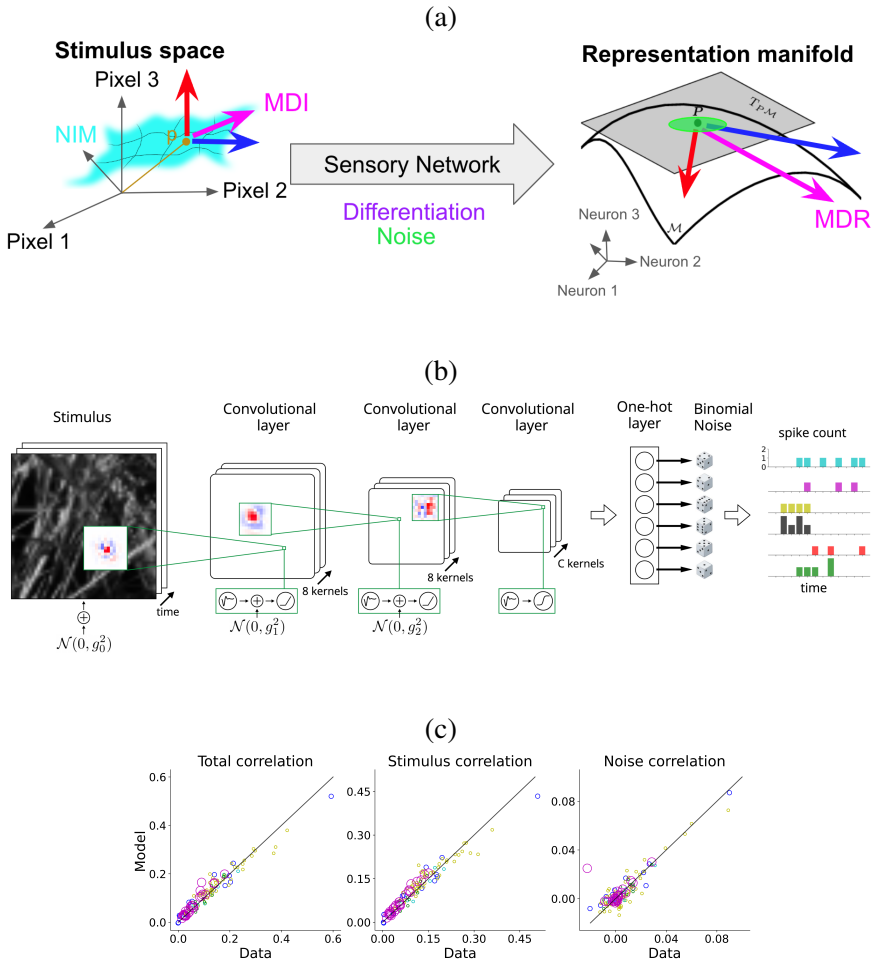

Fig 1: A stochastic model of the retina. (a) Schematic of a general sensory system. With noise in the network, each stimulus corresponds to a cloud of points in the representation space. The stimulus sensitivity (blue and the red arrows) and the stochasticity (the green oval) both vary with the specific stimulus. (b) Architecture of the retinal convolutional neural network model. Layer 1: 8 cell types with 15×15 convolutional filters. Layer 2: 8 cell types with 11×11 filters. Layer 3: C 9×9 filters, where C is the number of empirically determined cell types. Gaussian white noise is added to the stimulus and the first two convolutional layers, and independent binomial noise is applied after the final nonlinearity. The one-hot layer selects the location of the unit in the final convolutional layer to match that of the recorded neuron. (c) Comparison of total pairwise correlation, stimulus correlation, and noise correlation between neural data and model prediction. Diagonal line is the identity. Each circle corresponds to a pair of recorded neurons in an experimental preparation coded by its color whose radius is proportional to the square root of the test set duration.

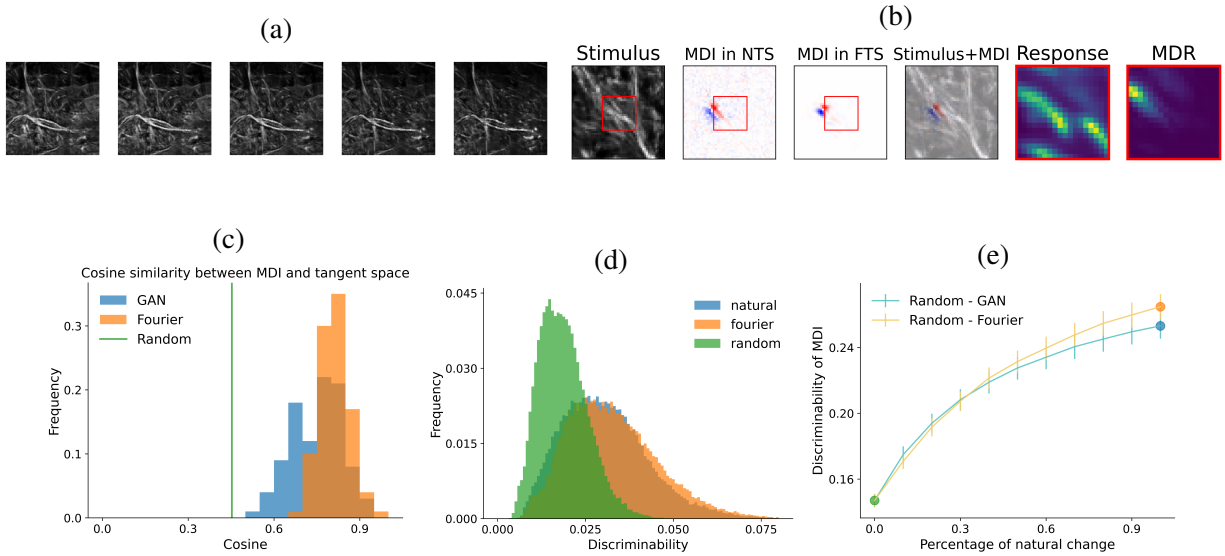

Fig 2: The retina is adapted to encode natural scenes (a) Example of a sequence of images generated by GAN with latent codes changing gradually along a straight path. (b) Example of the stimulus alongside the corresponding most discriminable input (MDI). Left to right, Stimulus frame, the MDIs constrained in the natural tangent space (NTS) , the full tangent space (FTS), and an overlay of the MDI constrained in the NTS on the stimulus. Only the first frames of these spatio-temporal vectors are shown given the jittering nature of the stimulus (see the main text). The last two columns show the mean response and the absolute values of the most discriminative response (MDR), averaged over cell-type channels. Red boxes correspond to the unit array in the final layer. (c) Histograms of the cosine similarity between the MDI computed in the full tangent space and its projection onto the natural tangent space, as estimated by both GAN and Fourier analysis across various stimuli. The theoretical value of the expected cosine similarity for a random space (RTS) is shown for comparison. (d) Histograms of discriminabilities of vectors randomly drawn from the NTS (as estimated by both GAN and Fourier analysis) and the RTS, across different stimuli. (e) The discriminability of the MDI constrained to the mixed space spanned by basis randomly drawn from the NTS and the RTS with different ratios.

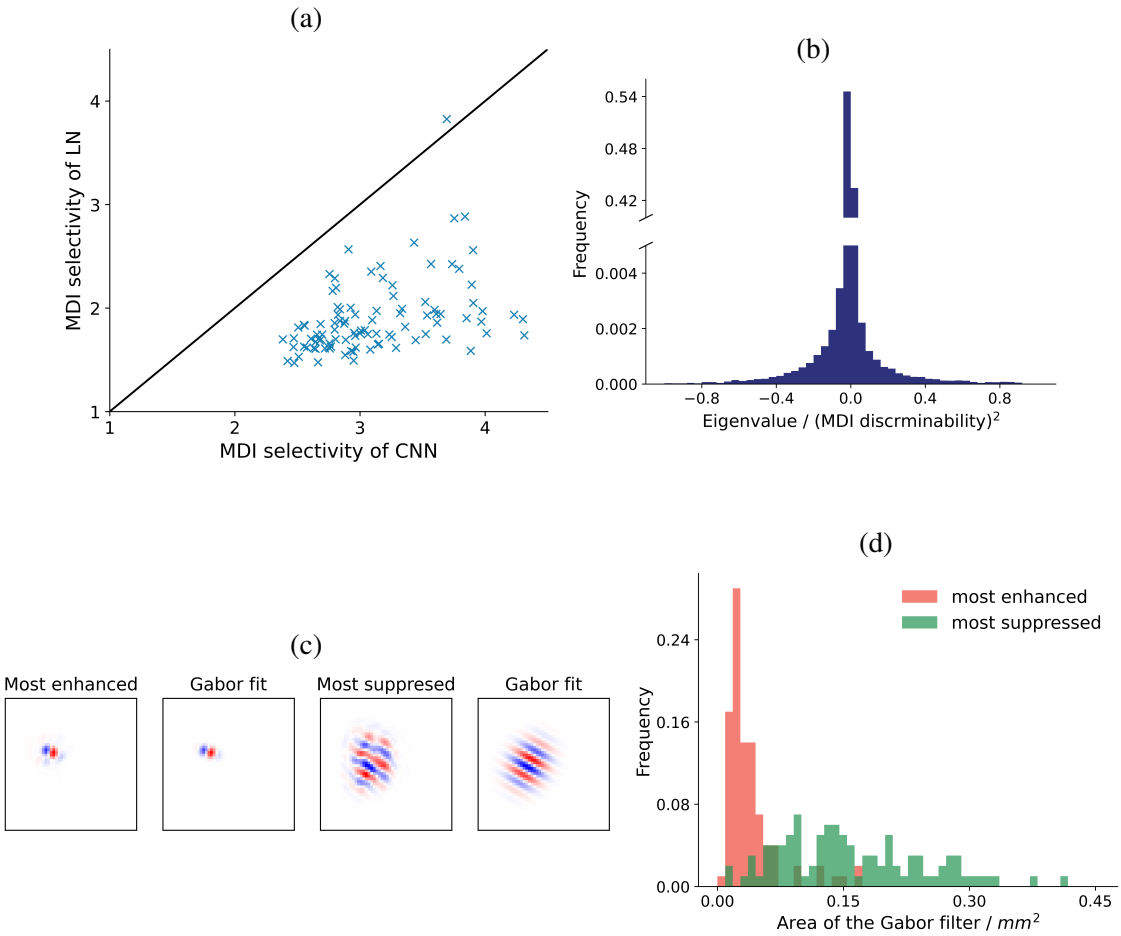

Fig 3: Nonlinearity in the retina enhances the discriminability of local motion but suppresses global motion. (a) MDI selectivity index of the LN model versus the CNN model for different stimuli (see the main text). The black line indicates the identity. (b) Eigenvalue spectrum of the Fisher information difference (FID) normalized by the square of the MDI discriminability over stimuli. (c) Example of the most positive (enhanced) and the most negative (suppressed) eigenvectors of the FID and the Gabor filters fit to them. (d) Histograms of areas of the Gabor filters fit to the most enhanced and the most suppressed stimulus changes for different stimuli (see Methods).

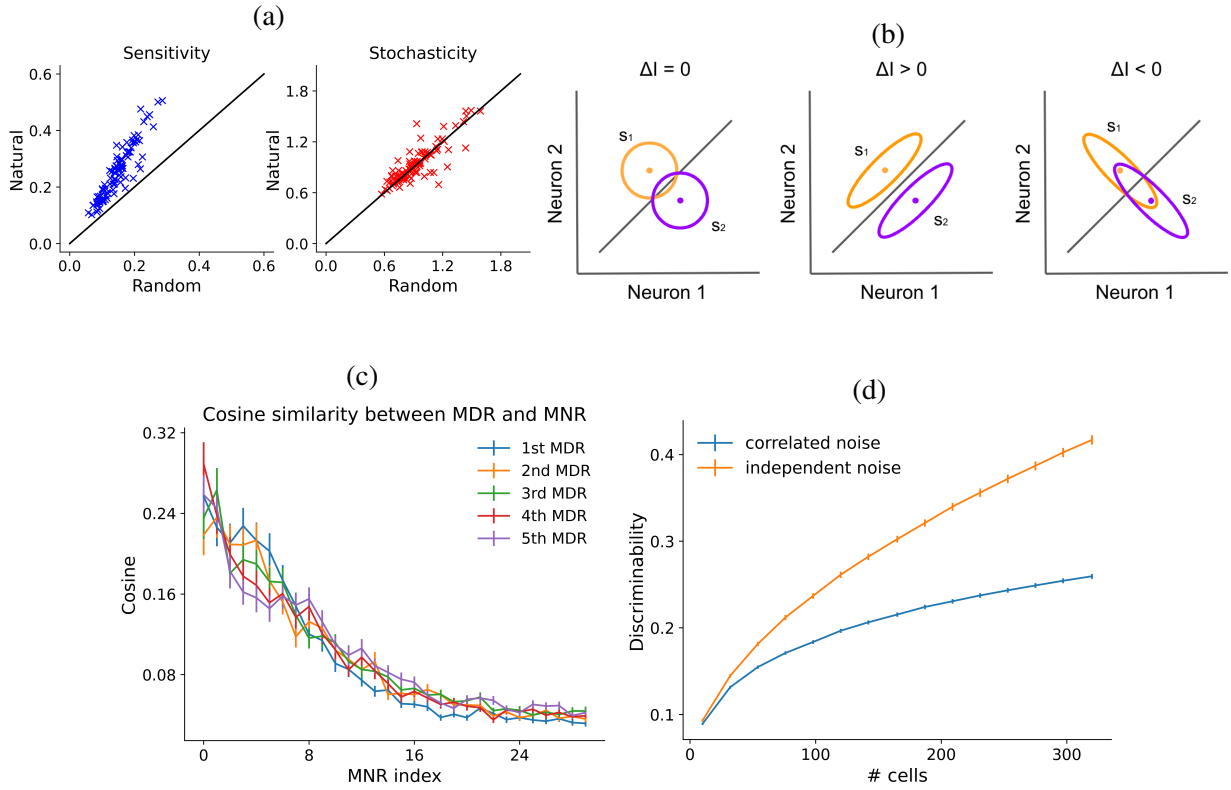

Fig 4: Noise correlations in the retina are detrimental to information coding. (a) Sensitivity and stochasticity of most discriminable inputs constrained to natural and random tangent spaces. (b) Compared to independent noise (left), noise correlations can either increase information (middle) or decrease information (right).[14] (c) Cosine similarities between the top 5 most discriminative responses (MDRs) and the top 30 most noisy responses (MNRs) averaged over stimuli. (d) Discriminability of MDI for different numbers of output units. The plot for trial-shuffled responses with independent noise is shown for comparison. Output units were randomly selected from the central $500\mu m \times 500\mu m$ region.[39] MDIs in (c) and (d) were computed in the NTS, similar results are seen for the full tangent space (Fig. S5).