Information Geometry of the Retinal Representation Manifold

Xuehao Ding[1], **Dongsoo Lee**[2], **Josh B. Melander**[3], **George Sivulka**[4], **Surya Ganguli**[5], and **Stephen A. Baccus**[6]

[1,5]Department of Applied Physics, Stanford University
[2,3]Neurosciences Phd Program, Stanford University
[4]Department of Electrical Engineering, Stanford University
[5,6]Department of Neurobiology, Stanford University
{xhding, dsnl, melander, gsivulka, sganguli, baccus}@stanford.edu

Abstract

The ability to discriminate visual stimuli is constrained by their retinal representations. Previous studies of visual discriminability were limited to either low-dimensional artificial stimuli or theoretical considerations without a realistic model. Here we propose a novel framework for understanding stimulus discriminability achieved by retinal representations of naturalistic stimuli with the method of information geometry. To model the joint probability distribution of neural responses conditioned on the stimulus, we created a stochastic encoding model of a population of salamander retinal ganglion cells based on a three-layer convolutional neural network model. This model not only accurately captured the mean response to natural scenes but also a variety of second-order statistics. With the model and the proposed theory, we are able to compute the Fisher information metric over stimuli and study the most discriminable stimulus directions. We found that the most discriminable stimulus varied substantially, allowing an examination of the relationship between the most discriminable stimulus and the current stimulus. We found that the most discriminative response mode is often aligned with the most stochastic mode. This finding carries the important implication that under natural scenes noise correlations in the retina are information-limiting rather than aiding in increasing information transmission as has been previously speculated. We observed that sensitivity saturates less in the population than for single cells and also that Fisher information varies less than sensitivity as a function of firing rate. We conclude that under natural scenes, population coding benefits from complementary coding and helps to equalize the information carried by different firing rates, which may facilitate decoding of the stimulus under principles of information maximization.

1 Introduction

Neural populations represent information with their collective activity. In the field of sensory neuroscience, a fundamental problem is to understand how neural populations discriminate between different stimuli. While there is much focus on characterizing the sensitivity of neurons to different stimuli, a quantitative description of discriminability requires precise knowledge of both the sensitivity to stimuli and the structure of noise correlations in the population.

Neural representation manifolds constitute a modern approach to studying neural populations in a geometric framework. On one hand, the recent machine learning literature has seen extensive research on the geometry of representation manifolds of neural networks trained for artificial tasks [1, 2, 3, 4].

Preprint. Under review.

On the other hand, Fisher information has been widely utilized in neuroscience for measuring decoding fidelity, particularly in theoretical studies exploring the effect of noise correlations [5, 6, 7]. In this study, we propose a framework for studying the Fisher information geometry of representation manifolds for natural visual scenes in a population of visual neurons, in particular the retina. In addition, for the first time, we apply such a manifold analysis to the neural representation of natural stimuli for real data.

The discriminability of biological visual systems has long been studied for low-dimensional and discrete artificial stimuli [8, 9, 10]. However, naturalistic visual stimuli in the real world are inherently high-dimensional and discriminability is generally heterogeneous across different types of changes in a stimulus. Recent advances in convolutional neural networks (CNNs) have enabled the accurate prediction of retinal responses to naturalistic stimuli, as well as ethological computations [11, 12, 13]. Nevertheless, previous work has predominantly focused on deterministic trial-averaged neural codes, which is not adequate for computing discriminability. Thus an encoding model of stochastic retinal representations for natural scenes is required, which constitutes one of the primary contributions of this work.

From an analysis of the Fisher information geometry of the retinal representation of stimuli, we find that the most discriminable direction for the retinal ganglion cell population varies greatly across stimuli and is correlated with the mean neural response. We further address a long standing question as to the effect of noise correlations on information. We find that the most discriminative mode in the representation space often aligns with the most stochastic direction, implying that noise correlations in the retina are detrimental to information coding. Finally, we analyze how discriminability depends on firing rate of individual cells and in the population, and discuss the implications in relation to principles of information maximization.

2 Theory

In this section, we describe basic theory of information geometry [14] and its application to sensory neuroscience. Let $\boldsymbol{x}$ denote the stimulus vector and $\boldsymbol{y}$ denote the spike count vector whose dimension is equal to the number of neurons encoding the stimulus. Each stimulus induces a conditional probability distribution of sensory neural responses $P(\boldsymbol{y}|\boldsymbol{x})$ and mean firing rates $\boldsymbol{\mu}(\boldsymbol{x}) \coloneqq E[\boldsymbol{y}|\boldsymbol{x}]$. Along any decoding vector $\boldsymbol{a}$, the sensitivity is defined as $||\nabla_{\boldsymbol{x}} \langle \boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{a} \rangle ||$, while the stochasticity is defined as the standard deviation of $\langle \boldsymbol{y}, \boldsymbol{a} \rangle$ conditioned on $\boldsymbol{x}$ (Figure 1). In our case of retina, $\boldsymbol{\mu}(\boldsymbol{x})$ is modeled by a noiseless CNN and $P(\boldsymbol{y}|\boldsymbol{x})$ is modeled by the stochastic counterpart.

The representation manifold is defined as the manifold of mean neural responses induced by stimuli: $\{\boldsymbol{\mu}(\boldsymbol{x})\}$. We also define a statistical manifold of conditional response distributions $\{P(\boldsymbol{y}|\boldsymbol{x})\}$ parameterized by stimuli $\boldsymbol{x}$, which yield a natural coordinate system. To equip the statistical manifold with a Riemannian metric, we consider the Kullback–Leibler divergence between two nearby distributions:

$$\begin{aligned} ds^2(\boldsymbol{x}, \boldsymbol{x}+d\boldsymbol{x}) &\coloneqq 2 \cdot D_{KL}[P(\boldsymbol{y}|\boldsymbol{x})||P(\boldsymbol{y}|\boldsymbol{x}+d\boldsymbol{x})] \\ &\approx \sum_{ij} \int P(\boldsymbol{y}|\boldsymbol{x}) \frac{\partial \ln P(\boldsymbol{y}|\boldsymbol{x})}{\partial \boldsymbol{x_i}} \frac{\partial \ln P(\boldsymbol{y}|\boldsymbol{x})}{\partial \boldsymbol{x_j}} d\boldsymbol{y} d\boldsymbol{x_i} d\boldsymbol{x_j} \\ &= d\boldsymbol{x}^T \cdot \boldsymbol{I}(\boldsymbol{x}) \cdot d\boldsymbol{x}, \end{aligned} \tag{1}$$

where $\boldsymbol{I}(\boldsymbol{x})$ is the Fisher information matrix with respect to the stimulus, and the KL divergence is linearized up to the second order. The Fisher information metric $g^{stim}|_{\boldsymbol{x}}(\boldsymbol{e_i}, \boldsymbol{e_j}) = \boldsymbol{I}(\boldsymbol{x})_{ij}$ naturally characterizes the local discriminability of stimuli near $\boldsymbol{x}$.

When we assume that neural responses follow multivariate Gaussian distributions $\boldsymbol{y}|\boldsymbol{x} \sim N(\boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{\Sigma}(\boldsymbol{x}))$ and the covariance matrix is locally independent of the stimulus, $\frac{d\boldsymbol{\Sigma}(\boldsymbol{x})}{d\boldsymbol{x}} = \boldsymbol{0}$, the Fisher information metric can be expressed as

$$\boldsymbol{I}(\boldsymbol{x}) = \frac{d\boldsymbol{\mu}(\boldsymbol{x})}{d\boldsymbol{x}}^T \cdot \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}) \cdot \frac{d\boldsymbol{\mu}(\boldsymbol{x})}{d\boldsymbol{x}}, \tag{2}$$

where $\frac{d\boldsymbol{\mu}(\boldsymbol{x})}{d\boldsymbol{x}}$ is the Jacobian matrix. The latter assumption is a good approximation in our case because a neural network consisting only of ReLU nonlinearities is locally equivalent to an effective

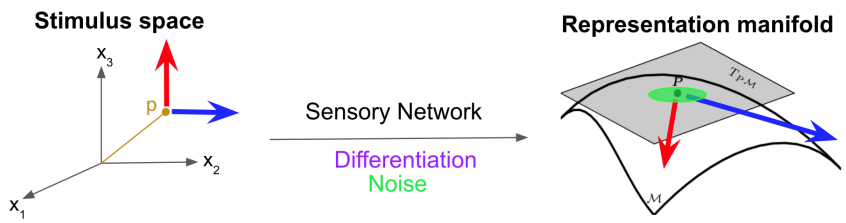

Figure 1: A schematic of basic concepts in a sensory system. The sensory network maps from the stimulus space to the representation manifold. The differentiation operator maps from the stimulus tangent space to the representation tangent space. With noise in the network, each stimulus corresponds to a cloud of points in the representation space. Generally, the sensitivity (the blue and the red arrows) and the stochasticity (the green oval) are both heterogeneous.

linear network [15]. This metric expression is interpretable in that discriminability is proportional to sensitivity and inversely proportional to noise. The length of a geodesic connecting two stimuli under this information metric measures the discriminability between them, when represented by the neural population.

Equivalently, the mean response $\boldsymbol{\mu}(\boldsymbol{x})$ on the representation manifold can also serve as a coordinate system. Then the metric is given by $g^{resp}|_{\boldsymbol{x}}(\boldsymbol{e_i}, \boldsymbol{e_j}) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{x})_{ij}$. The metric matrix transforms between the stimulus space and the representation space as a tensor [1].

The eigenvector of $\boldsymbol{I}(\boldsymbol{x})$ associated with the largest eigenvalue is the most discriminable direction in the stimulus tangent space (MDS) at stimulus $\boldsymbol{x}$ and the square root of the eigenvalue is the corresponding discriminability. The MDS in the stimulus space can be pushed forward to the most discriminative response mode (MDR) in the representation tangent space via the Jacobian matrix: MDR $\propto \frac{d\boldsymbol{\mu}(\boldsymbol{x})}{d\boldsymbol{x}} \cdot$ MDS. Similarly, the most sensitive direction in the stimulus tangent space (MSS) can be obtained as the principal eigenvector of $\frac{d\boldsymbol{\mu}(\boldsymbol{x})}{d\boldsymbol{x}}^T \cdot \frac{d\boldsymbol{\mu}(\boldsymbol{x})}{d\boldsymbol{x}}$ and the associated most sensitive response mode (MSR) obeys MSR $\propto \frac{d\boldsymbol{\mu}(\boldsymbol{x})}{d\boldsymbol{x}} \cdot$ MSS. Additionally, the most stochastic response mode (MNR) is defined as the principal eigenvector of the noise covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{x})$. For a ReLU neural network with small stochasticity, the stimulus dependence of the aforementioned tangent space features is solely derived from the stimulus dependence of the effective linear network [15].

One practical challenge of the above analysis with natural visual stimuli is the extremely high dimensionality of the stimulus space. Fortunately, we only need to compute the Fisher information matrix in a subspace defined by the recorded cell population. Define the instantaneous receptive field for each cell as the direction of the greatest sensitivity to the stimulus at one stimulus point, which is the gradient of that cell's response, $\frac{d\boldsymbol{\mu_i}(\boldsymbol{x})}{d\boldsymbol{x}}$ [12]. We only need to compute Fisher information in the subspace spanned by instantaneous receptive fields, as justified by the following theorem.

Theorem: the top N most discriminable directions are linear combinations of the instantaneous receptive fields, where N is the number of output neurons.

Proof: Let $IR := \text{span}(\{\frac{d\boldsymbol{\mu_i}(\boldsymbol{x})}{d\boldsymbol{x}}\})$ represent the subspace spanned by instantaneous receptive fields. Suppose that the most discriminable direction $\boldsymbol{v_1}$ is not in IR, then we can decompose $\boldsymbol{v_1} = \boldsymbol{v_1^o} + \alpha \boldsymbol{v_1^p}$, where $\boldsymbol{v_1^o}$ is the orthogonal component in $IR^{\perp}$ and $\alpha \boldsymbol{v_1}^p$ is the parallel component in IR, $||\boldsymbol{v_1}|| = ||\boldsymbol{v_1^p}|| = 1$, $|\alpha| < 1$. Then following Eq. 2 we have

$$\begin{aligned} g^{stim}(\boldsymbol{v_1}, \boldsymbol{v_1}) &= \alpha^2 g^{stim}(\boldsymbol{v_1^p}, \boldsymbol{v_1^p}) \\ &< g^{stim}(\boldsymbol{v_1^p}, \boldsymbol{v_1^p}), \end{aligned} \tag{3}$$

which contradicts with the fact that $\boldsymbol{v_1}$ is the most discriminable direction. Therefore, $\boldsymbol{v_1}$ is in IR.

Replacing IR with $IR \cap \{\boldsymbol{v_1}\}^{\perp}$, with the same analysis we can prove that the second most discriminable direction $\boldsymbol{v_2}$ is also in IR. Subsequently, the whole theorem can be proven.

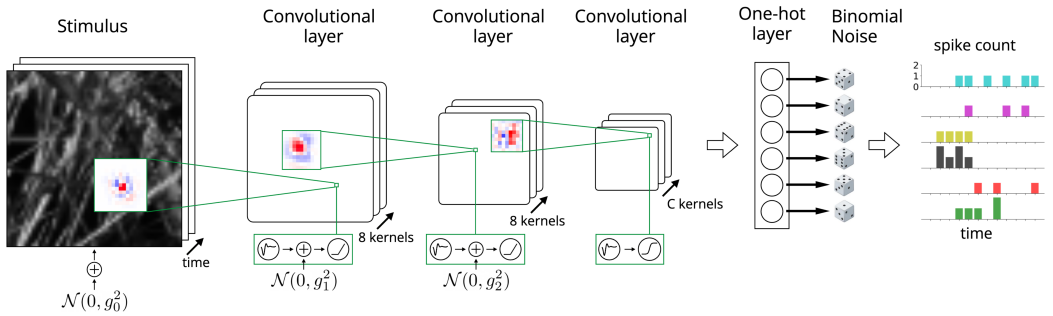

Figure 2: Model architecture. The first layer consists of eight 15×15 spatiotemporal convolutional filters. The second layer consists of eight 11×11 filters. The final layer consists of C 9×9 filters, where C represents the number of cell types. Batch normalization is applied after each convolutional filter. The nonlinearities in the first two layers are ReLU, while in the last layer, the nonlinearity consists of the SoftPlus function followed by a parametric $\tanh$ function. Gaussian white noise is added to the stimulus and the first two convolutional layers. Independent binomial noises are applied to the final nonlinearity. The one-hot layer selects the location of the unit in the final convolutional layer to match the recorded neurons. Further details are explained in the main text.

3 Methods

The spiking activity of a population of tiger salamander retinal ganglion cells was recorded in response to a sequence of jittering natural images. We created a stochastic encoding model capturing $P(\boldsymbol{y}|\boldsymbol{x})$ with the CNN architecture that included independent noise in the stimulus and in each unit of the CNN. The CNN parameters were trained using backpropagation to fit the mean firing rates $\boldsymbol{\mu}(\boldsymbol{x})$ to create a deterministic model, and then these parameters were frozen. Noise parameters included one parameter for each layer to set the standard deviation of the independent noise, and one parameter for binomial noise in each ganglion cell. These were then optimized by maximizing the log-likelihood of the empirical data and exploring the parameter space to match a few second-order statistics of the neural responses. Once the parameter optimization was completed, we computed the Fisher information metric with the model numerically for each stimulus and analyzed the most discriminable direction.

3.1 Stochastic model

The model architecture is shown in Figure 2. The model takes a spatiotemporal visual stimulus as input and its output is a set of spike counts, one for each neuron in each time bin. The skeleton of the model is a three-layer CNN similar to the architecture in Ref. [11]. Every convolutional filter in our model was implemented by linearly stacking a sequence of 3×3 small filters, which outperforms the traditional method [16]. A parametric $\tanh$ nonlinearity was attached to the last convolutional layer for the purpose of enforcing the refractory period constraint:

$$f(x; r_{max}, a, b) = \frac{r_{max}}{2} \cdot [\tanh(a(x-b)) + 1]\,, \tag{4}$$

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 fitted 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.

Formally, the one-hot layer is defined as

$$\begin{aligned} &\text{out}(k) = \sum_{ij} w_{ij}^k \cdot \text{input}(C(k), i, j),\ \forall\, k \\ &\sum_{ij} w_{ij}^k = 1,\ \forall\, k \\ &w_{ij}^k \geq 0,\ \forall\, k, i, j \end{aligned} \tag{5}$$

Each w^k will converge to a one-hot array after model training with our one-hot loss function

$$\begin{aligned} &\mathcal{L}_{\text{one-hot}}(w^k) \propto -\ln \sum_{pq} |\prod_{ij} (w_{ij}^k + \delta_{pq,ij} - 1)|, \\ &\delta_{pq,ij} = \begin{cases} 1 & p = i, q = j \\ 0 & \text{otherwise} \end{cases}, \end{aligned} \tag{6}$$

which reflects the locations of recorded neurons. Each output channel represents a ganglion cell type and the neuron-to-channel map $C(k)$ was determined using hierarchical clustering of output channels (see the Supplementary Material for details).

Gaussian white noise was added to the stimulus to simulate stochasticity in photoreceptors, and to pre-threshold signals in the first two layers to simulate stochasticity in bipolar cells and amacrine cells. The stochasticity of ganglion cells was modeled as a parametric binomial probability mass function:

$$p^B(n; r, k, N) \propto \frac{\Gamma(k \cdot N + 1)}{\Gamma(k \cdot (N - n) + 1) \cdot \Gamma(k \cdot n + 1)} r^{k \cdot n} \cdot (1 - r)^{k \cdot (N - n)}, \quad n = 0, 1, ..., N \tag{7}$$

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, and N is the cell-dependent upper bound of spike counts in one time bin reflecting the refractory period of that cell. 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., $\mu(r) = E_{p(n;r,k,N)}(n)$. Although we have explored a variety of probability mass functions to fit the sub-Poisson statistics of retinal ganglion cells [17, 18], the binomial noise Eq. 7 performs the best among them (see the Supplementary Material) and also is the most natural choice considering the refractory period constraint.

3.2 Optimization

CNN parameters and one-hot parameters $\{w^k\}$ were optimized together with the 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 [19] via Pytorch [20] on NVIDIA 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. 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.

3.3 Computation of Fisher information metric

For each stimulus, the noise covariance matrix $\mathbf{\Sigma}$ was estimated by repeatedly simulating the stochastic model thousands of times. We then eigendecomposed $\mathbf{\Sigma}$ obtaining local information of the representation manifold: $\mathbf{\Sigma} = \sum_i w_i \boldsymbol{v_i} \boldsymbol{v_i}^T$. We applied the Gram–Schmidt orthogonalization to instantaneous receptive fields to obtain a basis of IR which is a subspace of the stimulus tangent space. Then following Eq. 2, the Fisher information matrix with respect to the stimulus can be computed as

$$\boldsymbol{I} = \sum_i \frac{d\,\langle \boldsymbol{\mu}(\boldsymbol{z}), \boldsymbol{v_i} \rangle}{d\boldsymbol{z}}^T w_i^{-1} \frac{d\,\langle \boldsymbol{\mu}(\boldsymbol{z}), \boldsymbol{v_i} \rangle}{d\boldsymbol{z}}, \tag{8}$$

where $\boldsymbol{z}$ is the stimulus coordinate under the IR basis. In practice, we summed only over hundreds of the most stochastic modes, as the sensitivity and stochasticity of the other modes are both close to zero indicating that those modes are orthogonal to the tangent space of the representation manifold. The gradient with respect to the stimulus was computed using the noiseless model.

4 Results

4.1 Model evaluation

Early stochastic encoding models of the retina were primarily developed for artificial stimuli [21, 22, 23], and their results cannot be directly generalized to natural scenes due to the stimulus-dependent nature of noise in the retina. Here our fitted model was evaluated by comparing with neural data various commonly used second-order statistics including correlation measures and single-cell variability measures (Figure 3). Surprisingly, the model prediction matches the neural data remarkably well despite that the model only has a few noise parameters. This result suggests that because noise in the retina propagates through the same network pathways as the signal, they are both captured by the CNN structure. This indicates a strikingly simple origin for noise correlations between all ganglion cell pairs: independent noise that starts in individual upstream cells, and then propagates through the same network that creates mean stimulus sensitivity. Statistics for preparations with longer test sequences tend to be more reliable and consequently our model performs better on these.

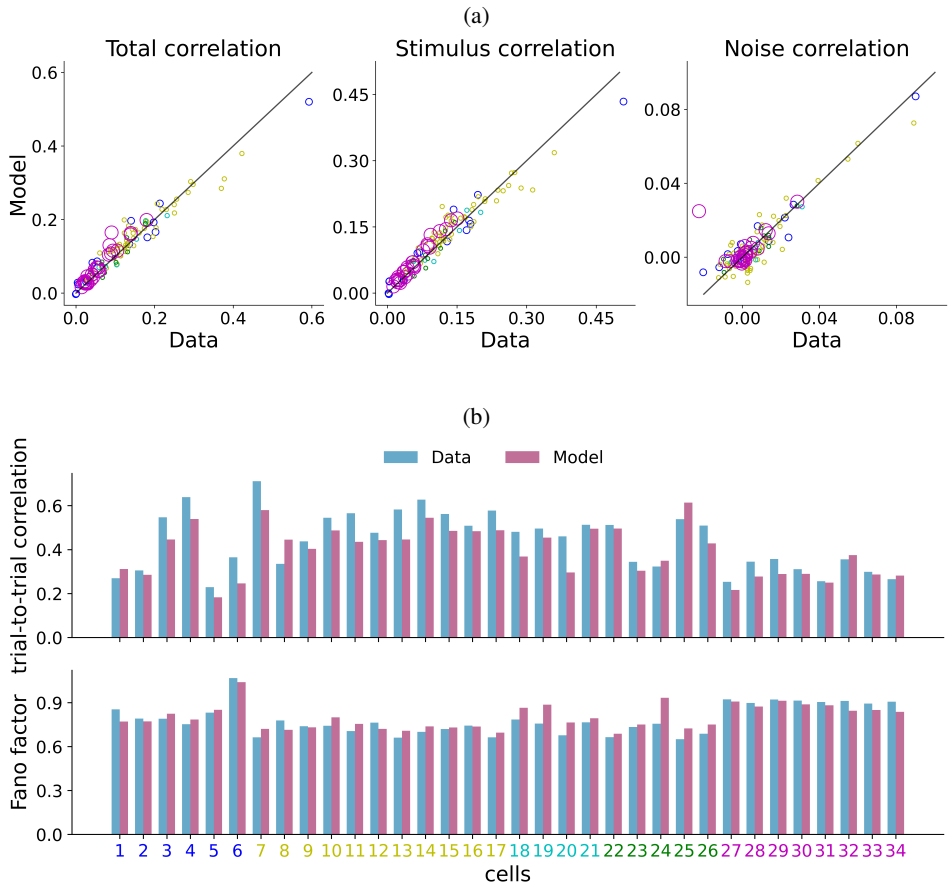

Figure 3: The model captures a variety of second-order statistics of neural responses (see the Supplementary Material for definitions). (a) Comparison of total pairwise correlation, stimulus correlation, and noise correlation between neural data and model prediction. Black lines represent perfect agreement. Each circle corresponds to a pair of neurons in an experimental preparation coded by its color and the circle radius is proportional to the square root of the total length of the test set. (b) Comparison of single-cell variability measures including the Fano factor and trial-to-trial correlation between neural data and model prediction. Neuron indices share the same color code with (a).

4.2 The most discriminable directions

Some examples of the most discriminable stimulus direction (MDS) and the most discriminative response mode (MDR) are shown in Figure 4a (see the Supplementary Material for more examples). Often the top eigenvalues were similar, and so to avoid over-interpreting the importance of individual eigenmodes, we found the sparsest direction (with the least L1 norm) in the space spanned by the top 5 MDSs for the purpose of visualization. We found that the MDS is strongly stimulus dependent, and is typically localized in a region where the sensitivity is high, an effect that arises from the locality of convolutional connections. The MDS tends to appear around the central region (red box in Figure 4a) because pixels within this area affect more output units than pixels near the edges or corners.

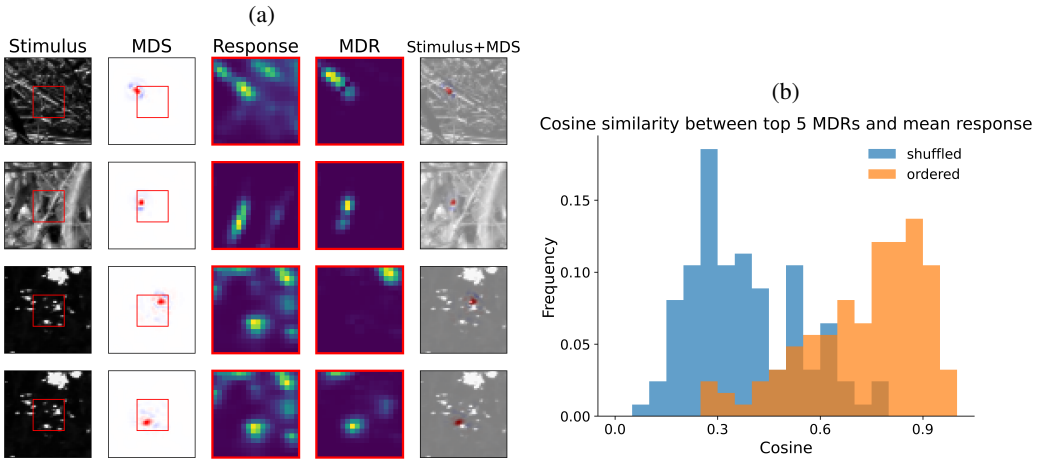

Figure 4: (a) A few examples of the most discriminable direction for different stimuli. The first column shows a representative frame of the spatiotemporal stimulus. The second column shows the spatial component of the spatiotemporal MDS obtained from the singular value decomposition. Red boxes correspond to the unit array in the final layer. The third column shows the mean response averaged over cell-type channels. The fourth column shows the absolute values of MDR averaged over channels. The last column shows the superposition of MDS and stimulus. (b) Histogram of the cosine similarity between the mean response vector and its projection onto the space spanned by the top 5 MDRs. Similarities between shuffled pairs are also shown for baseline comparison.

One general observation is that MDS varies greatly across stimuli. From a mathematical point of view, this is to say that certain features of the tangent space depend on the base point. With respect to neural function, this variability potentially indicates the presence of adaptation to the stimulus. We found that the most discriminable direction is related to the base point through the representation space such that MDR and the mean response are correlated (Figure 4b). This correlation can be attributed to the fact that neurons with intermediate firing rates exhibit the highest single-cell sensitivity and discriminability (Figure 6a). As a result, MDS sometimes captures the most salient feature of the stimulus (the first row of Figure 4a), although this is not always the case (the second row of Figure 4a). Similar stimuli and mean responses likely produce distinct MDSs especially when their mean response pattern has multiple spatial components that are candidates to yield a localized MDR (the third and fourth rows of Figure 4a). Moreover, the correlation between MDR and mean response provides a mechanism for higher brain regions to estimate future response changes of retinal ganglion cells based on the current response.

4.3 Noise correlations in the retina limit information coding

The role of noise correlations in population coding has long been debated. Multiple scenarios have been advanced where noise correlations are detrimental to information coding [24, 9]. However, others suggest that noise correlations under certain circumstances may not limit information coding and can even be beneficial [25, 5, 8]. More specifically, it has been proposed that in response space,

noise correlations decrease information when the differences in the signal are aligned with the noise direction, and increase information when they are orthogonal [26] (Figure 5a).

Here we tested these two hypotheses for the retina by computing cosine similarities between MDRs and MNRs. It turns out that the most discriminative mode is quite aligned with the most stochastic mode and orthogonal to modes with small variability (Figure 5b), implying that noise correlations in the retina are information-limiting correlations. To understand the mechanism, we analyzed the eigenspectrum of the Fisher information matrix (Figure 5c). Both the sensitivity and the stochasticity increase with the discriminability of MDR, and their ratio reaches a maximum at the noisiest mode. Therefore, we propose that the fundamental reason for such information-limiting correlation is that noise propagates through the same network as the signal, causing the stochasticity ellipsoid to align closely with, and be rounder than, the sensitivity ellipsoid.

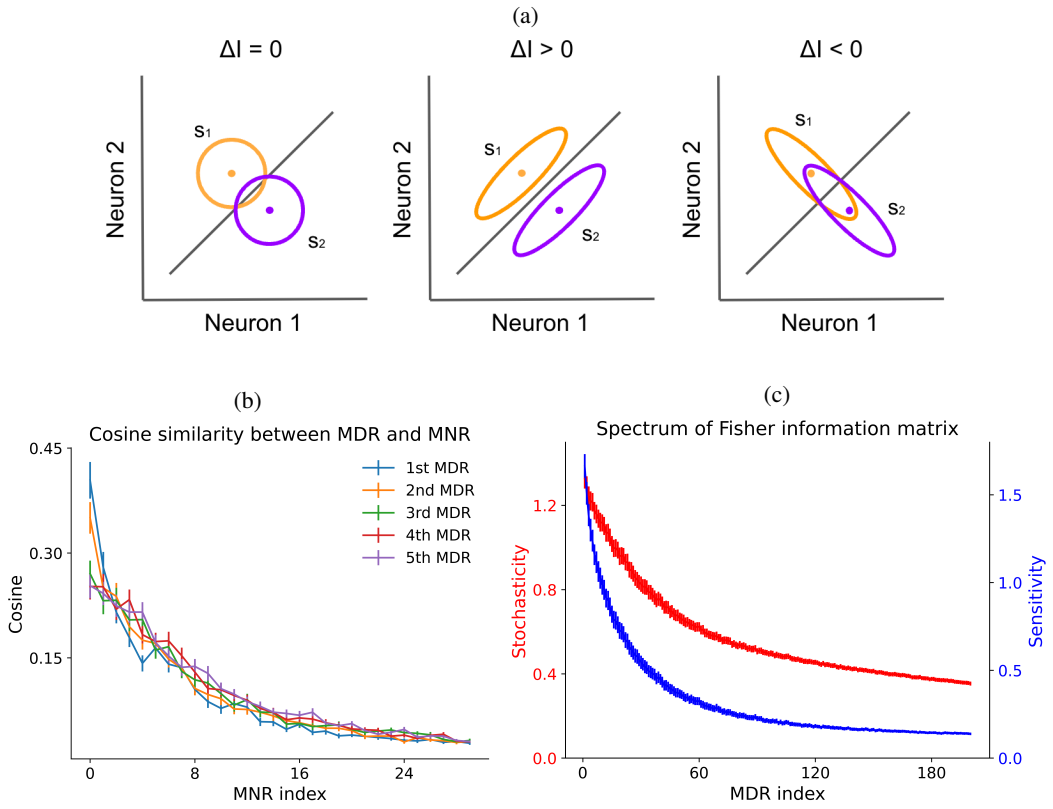

Figure 5: Noise correlations in the retina are detrimental to information coding. (a) Compared to independent noise (left), noise correlation can either increase information (middle) or decrease information (right). (b) Cosine similarities between the top 5 MDRs and the top 30 MNRs averaged over stimuli. (c) Sensitivity and stochasticity of the top 200 MDRs averaged over stimuli.

4.4 Firing rate dependency

The firing rate dependencies of stochasticity, sensitivity, and discriminability for individual neurons and the whole population are shown in Figure 6. For individual neurons, both the sensitivity and the stochasticity peak around intermediate firing rates and decay at high firing rates (a consequence of the refractory period). Thus, the single-cell discriminability also saturates around the intermediate firing rate. In contrast, population coding does not exhibit such high-rate decay, implying that neurons can encode information complementarily, which is a benefit of population coding. In other words, even when a fraction of neurons get saturated by the stimulus and become uninformative at high firing rates, other neurons can still provide discriminative responses to stimulus changes. Although the population sensitivity and stochasticity significantly increase with the mean firing rate, the firing rate

dependency of the population discriminability remains relatively flat, which is potentially beneficial for information transmission assuming that stimuli are uniformly distributed [27, 28].

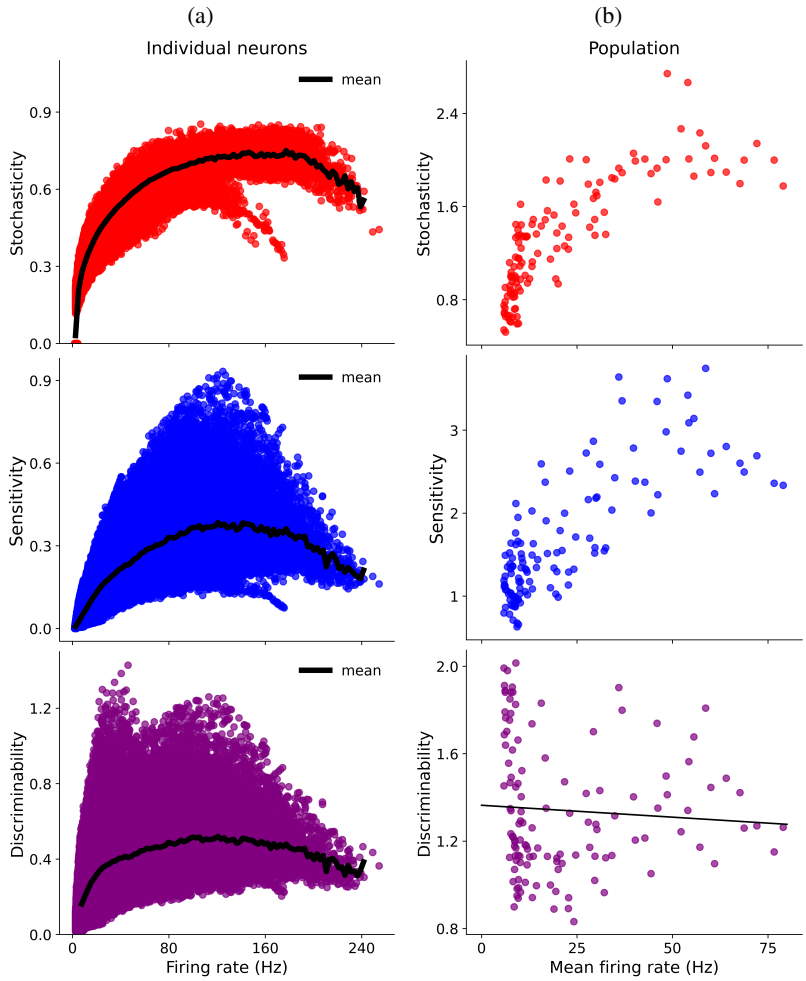

Figure 6: Firing rate dependency. (a) Firing rate dependencies of stochasticity, sensitivity, and discriminability for individual neurons. Mean values were plotted with thick black lines. (b) Same plots as (a) for the first MDR and averaged firing rates over the population. The linear regression of the populational discriminability is shown.

5 Discussion

In summary, we established a novel framework to understand the high dimensional representation manifold of the retinal population for natural scenes using an information geometric approach. However, there still exist remaining questions and further work that can be pursued within this framework.

First, our analysis primarily focused on examining the local metric tensor which characterizes the effects of infinitesimal stimulus changes. It would be interesting to extend the analysis to finite stimulus changes to explore more global geometry of the representation manifold. For example, one exciting direction is to find geodesics between two stimuli under the information metric and observe how one stimulus transitions into another with the least amount of discriminable changes for the retinal population.

Second, we found that noise correlations in the retina are information-limiting correlations. Nevertheless, it remains unclear whether this is a general result for all feedforward networks with independent noise applied to their inputs and intermediate layers. If this is indeed the case, noise correlations in other brain areas that do not limit information coding must arise from different mechanisms.

Finally, the Fisher information metric in this work was computed in the full stimulus space. However, most points in the stimulus space represent meaningless noise images that no animal has ever encountered. Therefore, an intriguing avenue for future research would involve incorporating the structure of natural stimuli and computing the metric with stimuli constrained to a natural-scene manifold.

Acknowledgements

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

References

[1] Michael Hauser and Asok Ray. Principles of riemannian geometry in neural networks. *Advances in neural information processing systems*, 30, 2017.

[2] Uri Cohen, SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. Separability and geometry of object manifolds in deep neural networks. *Nature communications*, 11(1):746, 2020.

[3] Ben Sorscher, Surya Ganguli, and Haim Sompolinsky. Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, 119(43):e2200800119, 2022.

[4] Binxu Wang and Carlos R Ponce. A geometric analysis of deep generative image models and its applications. In *International Conference on Learning Representations*, 2021.

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

[6] Larry F Abbott and Peter Dayan. The effect of correlated variability on the accuracy of a population code. *Neural computation*, 11(1):91–101, 1999.

[7] Bruno B Averbeck and Daeyeol Lee. Effects of noise correlations on information encoding and decoding. *Journal of neurophysiology*, 95(6):3633–3644, 2006.

[8] 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.

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

[10] 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.

[11] 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.

[12] Niru Maheswaranathan, Lane T McIntosh, David B Kastner, Josh Melander, Luke Brezovec, Aran Nayebi, Julia Wang, Surya Ganguli, and Stephen A Baccus. Deep learning models reveal internal structure and diverse computations in the retina under natural scenes. biorxiv. *URL: https://www. biorxiv. org/content/early/2018/06/14/340943. http://dx. doi. org/10.1101/340943. arXiv: https://www. biorxiv. org/content/early/2018/06/14/340943. full. pdf*, 2018.

[13] 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.

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

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

[16] 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.

[17] Michael J Berry, David K Warland, and Markus Meister. The structure and precision of retinal spike trains. *Proceedings of the National Academy of Sciences*, 94(10):5411–5416, 1997.

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

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

[20] 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.

[21] EJ Chichilnisky. A simple white noise analysis of neuronal light responses. *Network: computation in neural systems*, 12(2):199, 2001.

[22] Justin Keat, Pamela Reinagel, R Clay Reid, and Markus Meister. Predicting every spike: a model for the responses of visual neurons. *Neuron*, 30(3):803–817, 2001.

[23] Jonathan W Pillow, Liam Paninski, Valerie J Uzzell, Eero P Simoncelli, and EJ Chichilnisky. Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model. *Journal of Neuroscience*, 25(47):11003–11013, 2005.

[24] 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.

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

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

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

[28] Xue-Xin Wei and Alan A Stocker. Mutual information, fisher information, and efficient coding. *Neural computation*, 28(2):305–326, 2016.