Learning Predictive Filters

Lane McIntosh*
Neurosciences Graduate Program, Stanford University
Stanford, CA 94305.
(Dated: December 13, 2013)

We examine how a system intent on only keeping information maximally predictive of the future would filter data with temporal correlations. We find analytical solutions for the optimal predictive filters when data is Gaussian distributed and numerically find the optimal filters for non-Gaussian data using the Broyden-Fletcher-Goldfarb-Shanno algorithm. We then test the hypothesis that biological organisms process visual information to preserve only predictive information by comparing these optimal filters with measured filters from tiger salamander retinas. These results suggest that neural systems optimally capture information about the future, as well as give insights about how we should pre-process image data for deep learning networks.

Keywords: information bottleneck, predictive information, retinal coding, image pre-processing

INTRODUCTION

Neural systems and machine learning algorithms face a common challenge: to learn the simplest possible models from past data that generalize well to future, unseen data.

This problem is particularly acute in the early visual processing of the retina, where bits about the visual world are encoded via action potentials that cost 400×10^6 ATP apiece [1]. While the strict energy budget of the retina has led to theories of efficient coding, where information between the visual world and the retina's output is maximized by stripping away redundancy in the visual signal, relatively little attention has been paid to the need for retinal output to be informative about the future. From the moment light hits the eye, over a hundred milliseconds elapse before our first perception of the signal - meaning that we are constantly living in the past. Since organisms must react to the future given data collected in the past, one conceivable function of the retina would be to only transmit information that will be useful for the near future.

Such predictive processing is by no means beyond the purview of the retina. The last several decades of research on the retina have demonstrated that the eye acts as far more than a simple camera capturing light intensity from the external world [2]. In just a few layers of cells, the retina detects object motion [3], dynamically predicts arbitrarily complex temporal and spatial visual patterns [4], reduces the temporal and spatial redundancy inherent in natural scenes [5, 6], and transmits this compressed representation efficiently [7, 8].

Are these retinal properties the result of individual "bug detectors" crafted over the course of evolution, analogous to hand-selected feature creation, or do they fall out of general optimization principles? The theory of efficient coding suggests that maximizing mutual information between stimuli and the neural response provides such an optimization principle [6, 7, 9]. By examining correlations between the output cells of the retina, recent results have suggested that instead of trying to encode all information present in the stimulus, neural systems instead only seek to efficiently encode information that is predictive about the future stimulus [10].

One elegant idea that emerges from this work is that the retina's selectivity for specific features like object motion could arise not for its own sake, but rather because encoding object motion provides an efficient way of representing the future in a world with inertia [10]. In other words, maximizing predictive information in a neural network may be sufficient for capturing important information from the environment, rather than resorting to handcrafted feature detectors. This is important not only for a low-dimensional understanding of how the brain processes sensory information, but also for establishing the computations artificial systems need to implement in order to perform "intelligent" tasks like computer vision.

Here we derive the filters that maximally predict the system's future input in the context of Gaussian distributed stimuli with temporal correlations. We then compare these optimal filters to measured filters in tiger salamander retinas, and conclude that the retina is indeed optimally capturing information from the past that is predictive of the future.

PREDICTIVE INFORMATION BOTTLENECK

Problem Statement

Since the information processing inequality states that we will maximize the amount of predictive information by just keeping all information about the past, we avoid this trivial solution and incorporate a cost to encoding more bits by penalizing information about the past. A model-agnostic objective function that selectively keeps information about the future while penalizing information about the past is the information bottleneck problem

$$\min I(\overleftarrow{X}; \overleftarrow{Y}) - \beta I(\overleftarrow{Y}; \overrightarrow{X}), \tag{1}$$

where $\overleftarrow{X}$ and $\overrightarrow{X}$ are the past and future inputs to our system, $\overleftarrow{Y}$ is the past output of the system, and β pa-

rameterizes how much information we are willing to keep about the past.

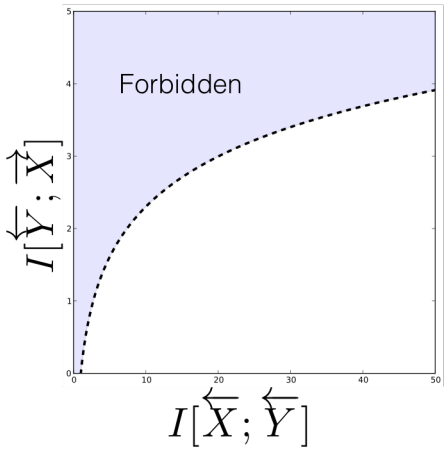

FIG. 1: An example information curve given a stimulus with exponentially decaying temporal correlations. The dotted line represents the maximal number of predictive bits that can be extracted given $I(\overleftarrow{X};\overleftarrow{Y})$ stored bits about the past. The solution to the information bottleneck problem is one point along this curve that depends on the particular value of β.

System Definition

We assume that our stimuli $X \sim \mathcal{N}(\mathbf{0}, \Sigma)$ where Σ has arbitrary correlations, and that our system linearly combines past stimuli values to form its output. In particular, we have

$$\overleftarrow{X} = \begin{bmatrix} x_t \\ x_{t-1} \\ \vdots \\ x_{t-k+1} \end{bmatrix}, \quad \overrightarrow{X} = \begin{bmatrix} x_{t+1} \\ \vdots \\ x_{t+k} \end{bmatrix}, \quad \text{and} \quad \overleftarrow{Y} = \begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-k+1} \end{bmatrix},$$

where

$$\begin{aligned} x_{t+1} &= Ax_t + \eta \\ y_{t+1} &= C_\beta x_{t+1} + D_\beta y_t + \xi \end{aligned}$$

and $\eta \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$, $\xi \sim \mathcal{N}(\mathbf{0}, \Sigma_\xi)$. Since x_t evolves linearly with the product of A, we can express $E[x_{t+1}] = A^t x_0$ and $E[y_{t+1}] = \sum_{i=0}^{t} D_\beta^i C_\beta A^{t-i} x_0$. If we consider the input fixed, our system responds linearly to its input via C_β and D_β. Our predictive information bottleneck will optimize over these filters to maximally extract predictive bits from $\overleftarrow{X}$.

Noting then that the expected value of $x_0 = A^{t-1-1}x_t$, we can write $\overleftarrow{Y} = H(\beta)\overleftarrow{X}$ where

$$H(\beta) = \frac{1}{k} \begin{bmatrix} \sum_{i=0}^{t-1} D_\beta^i C_\beta A^{t-i} \\ \vdots \\ \sum_{i=0}^{t-k} D_\beta^i C_\beta A^{t-i} \end{bmatrix} \begin{bmatrix} (A^{t-1})^{-1} & \dots & (A^{t-k})^{-1} \end{bmatrix}.$$

ANALYTICAL DERIVATION

Since $\overleftarrow{Y}$ is simply a linear combination of Gaussians, it must be Gaussian distributed as well. Using this observation, the definition of mutual information in terms of differential entropy, the formula for the entropy of a Gaussian random variable, $h(X) = \frac{1}{2}\log(2\pi e)^d |\Sigma_X|$, and dropping irrelevant constants, the solution to our predictive information bottleneck problem is

$$\begin{aligned} T^*(\beta) &= \arg\min_{H(\beta)} I(\overleftarrow{X};\overleftarrow{Y}) - \beta I(\overleftarrow{Y};\overrightarrow{X}) \\ &= \arg\min_{H(\beta)} (1-\beta) h(\overleftarrow{Y}) - h(\overleftarrow{Y}|\overleftarrow{X}) + \beta h(\overleftarrow{Y}|\overrightarrow{X}) \\ &= \arg\min_{H(\beta)} (1-\beta)\log|H(\beta)\Sigma_{\overleftarrow{X}}H(\beta)^T + \Sigma_\xi| - \log|\Sigma_\xi| \\ &\qquad + \beta\log|H(\beta)\Sigma_{\overleftarrow{X}|\overrightarrow{X}}H(\beta)^T + \Sigma_\xi|, \end{aligned}$$

since the Schur complement gives us

$$\begin{aligned} \Sigma_{\overleftarrow{Y}|\overleftarrow{X}} &= \Sigma_{\overleftarrow{Y}} - \Sigma_{\overleftarrow{Y}\overleftarrow{X}}\Sigma_{\overleftarrow{X}}^{-1}\Sigma_{\overleftarrow{X}\overleftarrow{Y}} \\ &= H(\beta)\Sigma_{\overleftarrow{X}}H(\beta)^T + \Sigma_\xi - H(\beta)\Sigma_{\overleftarrow{X}}\Sigma_{\overleftarrow{X}}^{-1}\Sigma_{\overleftarrow{X}}H(\beta)^T \\ &= \Sigma_\xi \end{aligned}$$

and

$$\begin{aligned} \Sigma_{\overleftarrow{Y}|\overrightarrow{X}} &= \Sigma_{\overleftarrow{Y}} - \Sigma_{\overleftarrow{Y}\overrightarrow{X}}\Sigma_{\overrightarrow{X}}^{-1}\Sigma_{\overrightarrow{X}\overleftarrow{Y}} \\ &= H(\beta)\Sigma_{\overleftarrow{X}}H(\beta)^T + \Sigma_\xi - H(\beta)\Sigma_{\overleftarrow{X}\overrightarrow{X}}\Sigma_{\overrightarrow{X}}^{-1}\Sigma_{\overrightarrow{X}\overleftarrow{X}}H(\beta)^T \\ &= H(\beta)\Sigma_{\overleftarrow{X}|\overrightarrow{X}}H(\beta)^T + \Sigma_\xi. \end{aligned}$$

By Lemma A1 from [11] we can set $\Sigma_\xi = I$ without loss of generality, and so our solution becomes

$$\begin{aligned} T^*(\beta) = \arg\min_{H(\beta)} &(1-\beta)\log|H(\beta)\Sigma_{\overleftarrow{X}}H(\beta)^T + I| \\ &+ \beta\log|H(\beta)\Sigma_{\overleftarrow{X}|\overrightarrow{X}}H(\beta)^T + I|. \end{aligned}$$

Note that our optimization depends on the noncausal covariance matrix $\Sigma_{\overleftarrow{X}|\overrightarrow{X}}$, which the system must learn over time. Taking the derivative of the above minimization and setting it to zero we obtain the following equality,

$$\begin{aligned} \frac{\beta-1}{\beta}(H(\beta)\Sigma_{\overleftarrow{X}|\overrightarrow{X}}H(\beta)^T + I)(H(\beta)\Sigma_{\overleftarrow{X}}H(\beta)^T + I)^{-1}H(\beta) \\ = H(\beta)\Sigma_{\overleftarrow{X}|\overrightarrow{X}}\Sigma_{\overleftarrow{X}}^{-1}. \end{aligned}$$

By diagonalizing $\Sigma_{\overleftarrow{X}|\overrightarrow{X}}\Sigma_{\overleftarrow{X}}^{-1}$, taking the singular value decomposition of $H(\beta)$, and performing a few substitutions, we obtain the optimal solution [11, 12],

$$H^*(\beta) = \begin{bmatrix} \alpha_1 v_1^T \\ \vdots \\ \alpha_k v_k^T \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}$$

where

$$\alpha_i = \sqrt{\frac{\beta(1-\lambda_i)-1}{\lambda_i(v_i^T \Sigma_x v_i)}},$$

and λ_i, v_i are the eigenvalues and eigenvectors, respectively, (in ascending order) of $\Sigma_{\overleftarrow{X}|\overrightarrow{X}}\Sigma_{\overleftarrow{X}}^{-1}$. Intuitively, we can think of this matrix $\Sigma_{\overleftarrow{X}|\overrightarrow{X}}\Sigma_{\overleftarrow{X}}^{-1}$ as killing off the variance in $\overleftarrow{X}$ and then slowly adding back in the bits about the past with the least variance given the future. As β increases, our filter $H^*(\beta)$ loses its **0** rows and keeps more of the higher variance components of $\Sigma_{\overleftarrow{X}|\overrightarrow{X}}\Sigma_{\overleftarrow{X}}^{-1}$.

SIMULATIONS OF ANALYTICAL RESULTS

To visualize these analytical results we drew $A = Q$ from the QR decomposition of random matrices, which effectively renders A a generalized rotation matrix and gives $\overleftarrow{X}$ sinusoidal structure (figure 2). We let each x_t be 10 dimensional, and added progressively more noise to each dimension such that dimension 1 was noiseless and dimension 10 was nearly independent.

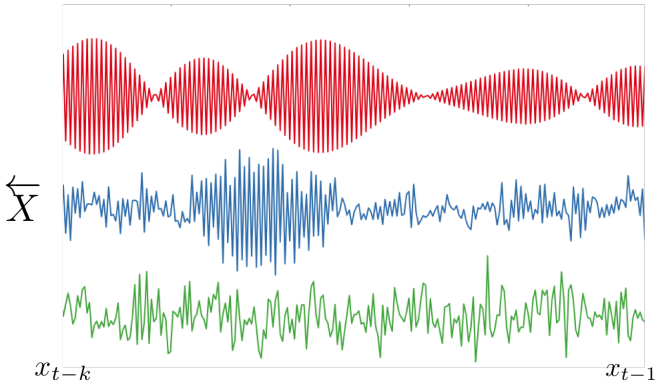

FIG. 2: 3 of $\overleftarrow{X}$'s 10 dimensions. From top to bottom the additive noise increases and the maximum predictive information from that time series decreases.

Over hundreds of iterations we estimated $\Sigma_{\overleftarrow{X}|\overrightarrow{X}}\Sigma_{\overleftarrow{X}}^{-1}$ (figure 3), performed the eigenvalue decomposition of this matrix, and used it to compute $H(\beta)$ and $\overleftarrow{Y} = H(\beta)\overleftarrow{X}$ (figure 4) for an intermediate value of β.

Since $\overleftarrow{X}$ is $mk \times 1$ dimensional, where m is the dimension of a single time point and k is the number of time points, intuitively it should skip over the m dimensions that are very noisy and have little predictive information to offer. Indeed, we can see this periodic nature in the rows of $H(\beta)$ (figure 4).

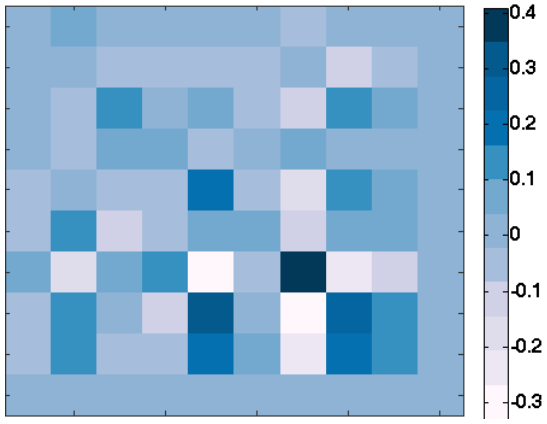

FIG. 3: A visualization of $\Sigma_{\overleftarrow{X}|\overrightarrow{X}}\Sigma_{\overleftarrow{X}}^{-1}$ for our example problem with progressively noisier dimensions running from left to right and top to bottom. Since $\sigma_{x_{t-i}|x_{t+j}}$ is large when x_{t-i}'s variance is unexplained by x_{t+j}, the dimensions in the top left corner represent the subspace of $\overleftarrow{X}$ correlated most with the future.

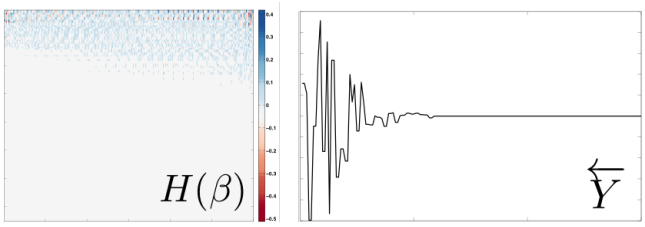

FIG. 4: On the left are the entries of $H(\beta)$ corresponding to how the system optimally combines past time points to preserve predictive information for an intermediate value of β. The right time series is the optimal response given β, where left time points are less delayed. The drop to baseline indicates that responding longer does not increase $\beta I(\overleftarrow{Y};\overrightarrow{X})$ enough to offset $I(\overleftarrow{X};\overleftarrow{Y})$.

LEARNING PREDICTIVE FILTERS FOR NATURALISTIC IMAGES

Curious about what the predictive filter would look like for natural image processing, we generated a naturalistic 1-d stimulus from a database of natural images. After picking a pixel location randomly, we would traverse the image via a random walk for five seconds before switching to a new image (figure 5). Although this stimulus is only locally Gaussian, the optimal predictive filter adapted to the timescale of the "saccades" - no matter how high β was tuned, $\overleftarrow{Y}$ never allocated bits beyond 5 seconds in the past (figure 6).

PREDICTIVE FILTERS IN THE RETINA

The retina is one system where it would be highly advantageous to selectively extract predictable infor-

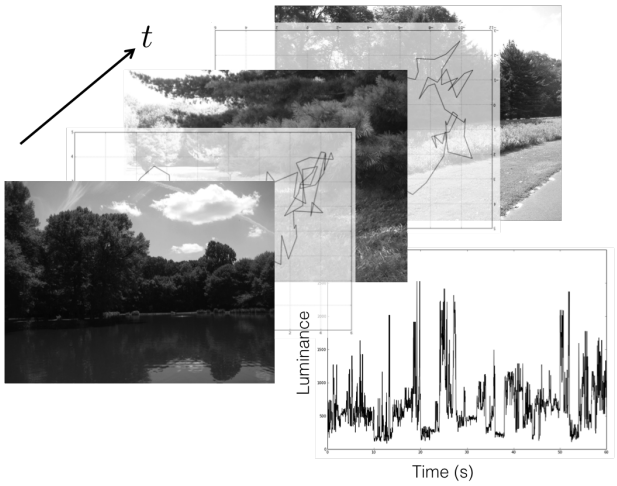

FIG. 5: Our stimulus of natural images. We switch images every 5 seconds, mimicking saccades, and perform a random walk over the image mimicking fixational eye movements.

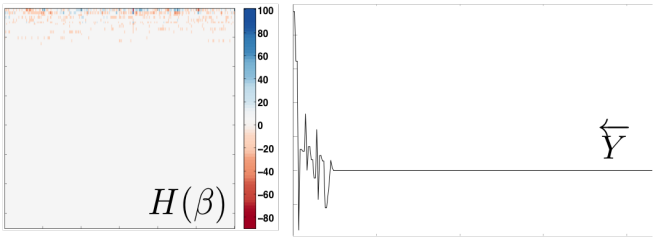

FIG. 6: The optimal filter $H(\beta)$ and system output $\overleftarrow{Y}$ for the natural image stimulus.

mation while minimizing the total number of bits transmitted. The retina comprises three primary layers of cells - photoreceptors, bipolar cells, and retinal ganglion cells. Since photoreceptors and bipolar cells roughly serve to linearly filter the raw rhodopsin activity while retinal ganglion cells have a threshold that generates the first action potentials in the visual pathway, researchers typically model the retina with linear-nonlinear models. One particularly well-known fact of these models is that the filters measured via reverse correlation are strongly biphasic (figure 6) [4, 13].

For several decades the dominant explanation of these biphasic filters is that they serve to reduce redundancy in visual information that is highly correlated [14]. However, could we derive these filters from first principles? In particular, if the retina is indeed trying to only keep predictive information, the solution to our optimization problem should be these biphasic filters.

To directly compare our solution with measured convolutional filters, we modify our system slightly to be

$$\overleftarrow{X} = \begin{bmatrix} x_t \\ x_{t-1} \\ \vdots \\ x_{t-k-n+2} \end{bmatrix}, \quad \overrightarrow{X} = \begin{bmatrix} x_{t+1} \\ \vdots \\ x_{t+k} \end{bmatrix}, \quad \text{and} \quad \overleftarrow{Y} = \begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-k+1} \end{bmatrix},$$

where

$$\begin{aligned} y_t &= \sum_{i=0}^{n-1} W_{-i,\beta}\, x_{t-i} + \xi \\ &= W_\beta \overleftarrow{X}_{t:t-n+1} + \xi \end{aligned}$$

and $W_\beta = [W_{0,\beta}, W_{-1,\beta}, \ldots, W_{-n+1,\beta}]$. We specifically let each $W_{-i,\beta}$ be scalar, such that we can write $\overleftarrow{Y} = T(\beta)\overleftarrow{X}$ where

$$T(\beta) = \begin{bmatrix} W_\beta & & \mathbf{0} & \\ \mathbf{0} & W_\beta & \mathbf{0} & \\ & & \ddots & \\ & \mathbf{0} & & W_\beta \end{bmatrix},$$

is Toeplitz and $\overleftarrow{Y}$ is simply the convolution $W_\beta * \overleftarrow{X}$.

Although we've already reduced the dimensionality from k^2qm to n where $n << k$, we further reduce the dimensionality of finding W_β by parameterizing the filter $H(\beta)$ as the product of sinusoids (with parameters frequency and phase) and a Gaussian filter (with variable width).

Stimulus

Temporal correlations in natural images are typically described as having a k/f power spectrum, where k is some positive constant. We generated such a dominantly low-frequency stimulus by transforming white noise in the frequency domain and then performing an inverse Fourier transform to return it to the temporal domain (figure 7). After convolving the stimulus with a given filter, we numerically compute the average $I(\overleftarrow{X};\overleftarrow{Y}) - \beta I(\overleftarrow{Y};\overrightarrow{X})$ over tens of iterations for 10^4 frames of stimuli each.

We minimized this estimate of $I(\overleftarrow{X};\overleftarrow{Y}) - \beta I(\overleftarrow{Y};\overrightarrow{X})$ as a function of W_β using both the Broyden-Fletcher-Goldfarb-Shanno algorithm, which has good performance even when the error landscape is not smooth, and simulated annealing. While we initially anticipated that simulated annealing would better find the global minimum since the information bottleneck value as a function of the filter's parameters was non-convex, simulated annealing would consistently cool at non-optimal values. The figures here are all results from the BFGS runs.

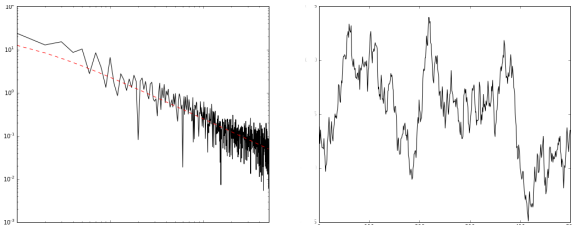

FIG. 7: The $1/f$ power spectrum (left) and time series (right) of our 1d naturalistic stimuli. The power spectrum on the left is a log-log plot with the theoretical value in red and the simulated value in black.

Results

To some degree, this optimization routine is "doomed to succeed" - since optimality depends on the exact value of β we choose, there is a large space of optimal filters. Nonetheless, the choice of β determines how highly a system values compression versus prediction. In our optimization, we found that high values of β resulted in non-biological filters (inset, figure 8), whereas low values of β yielded optimal filters that were nearly identical to biphasic filters measured in real retinas (figure 8).

Not only is this result consistent with the hypothesis

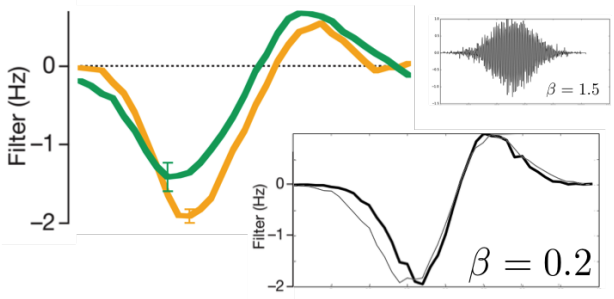

FIG. 8: On the left are real filters measured in tiger salamander [4], while the bottom right are the optimal filters for $\beta = 0.2$ (bottom) and $\beta = 1.5$ (top) obtained after two different initializations of the BFGS minimization of $I(\overleftarrow{X};\overleftarrow{Y}) - \beta I(\overleftarrow{Y};\overrightarrow{X})$.

that the retina is selectively extracting predictable information, it also suggests that the retina is in a regime where the transmission of information is highly costly, and must throw away almost all information except that small amount which is highly indicative of the future.

CONCLUSIONS

The predictive information bottleneck provides a model-free way of quantifying the ability of a system to extract predictable information from any arbitrary input - whether it be photons hitting the eye or a data matrix. Under certain assumptions, we now know that the problem of extracting the most predictive information from the past boils down to learning the structure of $\Sigma_{\overleftarrow{X}|\overrightarrow{X}}\Sigma^{-1}_{\overleftarrow{X}}$ over time. Lastly, the similarity of our optimally predictive filters and measured retinal filters suggests that biological organisms may already be exploiting this design principle.

I would like to thank Stephen Baccus, Surya Ganguli, Jascha Sohl-Dickstein, and Niru Maheswaranathan for their guidance and suggestions. I would also like to thank the TAs and Professor Ng for their support.

* Electronic address: lmcintosh@stanford.edu

[1] Jeremy E Niven and Simon B Laughlin. Energy limitation as a selective pressure on the evolution of sensory systems. *Journal of Experimental Biology*, 211(11):1792–1804, 2008.

[2] Tim Gollisch and Markus Meister. Eye smarter than scientists believed: neural computations in circuits of the retina. *Neuron*, 65(2):150–164, 2010.

[3] Stephen A Baccus, Bence P Ölveczky, Mihai Manu, and Markus Meister. A retinal circuit that computes object motion. *The Journal of Neuroscience*, 28(27):6807–6817, 2008.

[4] Toshihiko Hosoya, Stephen A Baccus, and Markus Meister. Dynamic predictive coding by the retina. *Nature*, 436(7047):71–77, 2005.

[5] Mandyam V Srinivasan, Simon B Laughlin, and Andreas Dubs. Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 216(1205):427–459, 1982.

[6] Joseph J Atick. Could information theory provide an ecological theory of sensory processing? *Network: Computation in neural systems*, 3(2):213–251, 1992.

[7] Simon B Laughlin. A simple coding procedure enhances a neuronÂ's information capacity. *Z. Naturforsch*, 36(910-912):51, 1981.

[8] Vijay Balasubramanian and Michael J Berry. A test of metabolically efficient coding in the retina. *Network: Computation in Neural Systems*, 13(4):531–552, 2002.

[9] Horace B Barlow. Possible principles underlying the transformation of sensory messages. *Sensory communication*, pages 217–234, 1961.

[10] Stephanie E Palmer, Olivier Marre, II Berry, J Michael, and William Bialek. Predictive information in a sensory population. *arXiv preprint arXiv:1307.0225*, 2013.

[11] Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for gaussian variables. In *Journal of Machine Learning Research*, pages 165–188, 2005.

[12] Felix Creutzig, Amir Globerson, and Naftali Tishby. Past-future information bottleneck in dynamical systems. *Physical Review E*, 79(4):041925, 2009.

[13] Stephen A Baccus and Markus Meister. Fast and slow contrast adaptation in retinal circuitry. *Neuron*, 36(5):909–919, 2002.

[14] Joseph J Atick and A Norman Redlich. Towards a theory of early visual processing. *Neural Computation*, 2(3):308–320, 1990.