

A Neural Model of High-Acuity Vision in the Presence of Fixational Eye Movements

Alexander G. Anderson¹
and Bruno A. Olshausen²

¹Physics Department, ²School of Optometry,
& ²Helen Wills Neuroscience Institute
University of California, Berkeley
Email: {aga, baolshausen}@berkeley.edu

Kavitha Ratnam
and Austin Roorda

Vision Science Program, School of Optometry
University of California, Berkeley
Email: {kavitha, aroorda}@berkeley.edu

Abstract—Experiments by Ratnam *et al.*[1] demonstrate the benefit of drift eye movements for the discrimination of a diffraction-limited tumbling E sized near the sampling limit of the cone photoreceptor array. Subjects perform better at discriminating the orientation of the E when its projection moves on the retina with the same motion statistics as drift eye movements, but not necessarily correlated to the true eye motion. In order to better understand the neural circuitry that underlies these psychophysical results, we propose a computational model based on a Bayesian ideal observer that attempts to estimate the spatial pattern on the retina given simulated RGC spikes. Our Bayesian model both corroborates the psychophysical measurements and suggests a neural mechanism. We extend previous work by Burak *et al.*[2] by creating a novel, online approximation to the expectation-maximization algorithm that generalizes to the case of continuous eye movements and sparse pattern priors. From this emerges a neural model containing two populations of cells which we hypothesize to exist in primary visual cortex: one that encodes the spatial pattern using a sparse code and another that tracks the eye position and is used to dynamically route information coming from LGN afferents feeding into the pattern cells.

I. INTRODUCTION

Our brains take in sensory data and infer perceptually relevant features of the world. In contrast to our perception, the raw sensory data we receive is typically incomplete, unstable, and noisy. In the case of high acuity vision in humans, there are two important sources of noise and instability: limited channel capacity of responses of retinal ganglion cells and drift eye movements that cause the retinal projection of an object in the world to jitter. Humans with normal 20-20 vision are able to resolve visual features that differ by just a few photoreceptors (e.g., an E versus a F). While we perform this discrimination, the letters drift across the retina over distances much larger than their own sizes. While the brain can, in principle, estimate motion using proprioceptive or efference copy signals, a number of lines of evidence suggest that this is not the case [3] [4] [5]. Thus, in order to properly integrate incoming spikes from the retina, the cortex must estimate the eye's trajectory during drift using the incoming spikes.

Burak *et al.*[2] proposed a model that took an important first step in attempting to solve this problem from a first-principles approach. They defined a probabilistic model and then used a

form of approximate Bayesian inference to factorize form and motion from the incoming retinal spike train. Here we relax a number of their assumptions: 1) We allow for continuous eye movements; 2) We reconstruct continuous gray-valued patterns instead of binary patterns; and 3) We infer a sparse representation of the pattern rather than pixels. We created a novel, online approximation to the EM algorithm in order to handle this more general situation. For comparison, previous papers used a variational mean-field approach [2][6].

Our main hypothesis is that there exists a population of neurons in the primary visual cortex whose retinal receptive fields are dynamically shifted so as to remain fixed in object-centered coordinates. While this question has been investigated previously[7], the results are inconclusive due to conflicting results from different labs [8]. An extension of a recent experiment by McFarland *et al.*[9] provides a means to test this hypothesis in a new way. In particular, they use V1 recordings (with a known stimulus) to fit a spiking neuron model to predict the location of the eye during drift. Their model assumes that the V1 neurons corresponding to the fovea are fixed in retinotopic coordinates. A comparison of the actual eye motion and the inferred eye motion can reveal the extent to which the representation of the pattern in V1 is stabilized in object-centered coordinates. Relatedly, we also predict that there is a collection of neurons that track the eye position since the most recent microsaccade. Snodderly *et al.*[10] find that some V1 cells have varying activation in response to drift and microsaccades (eg. tuned to one or the other, or a combination). The investigation of such cells merits further attention.

While our work shares a superficial resemblance to Rucci *et al.*[11], we use high-contrast stimuli with frequency content above the Nyquist sampling frequency of the retinal cone lattice (50 cycles/deg), whereas Rucci *et al.*[11] use low contrast Gabors at 11 cycles/deg. Furthermore, our model demonstrates a mechanism by which pattern information that has been transformed from the spatial domain to the time domain, via fixational eye movements, may be decoded in cortex.

The rest of this paper is organized as follows. First, we discuss our mathematical formulation of the problem as ap-

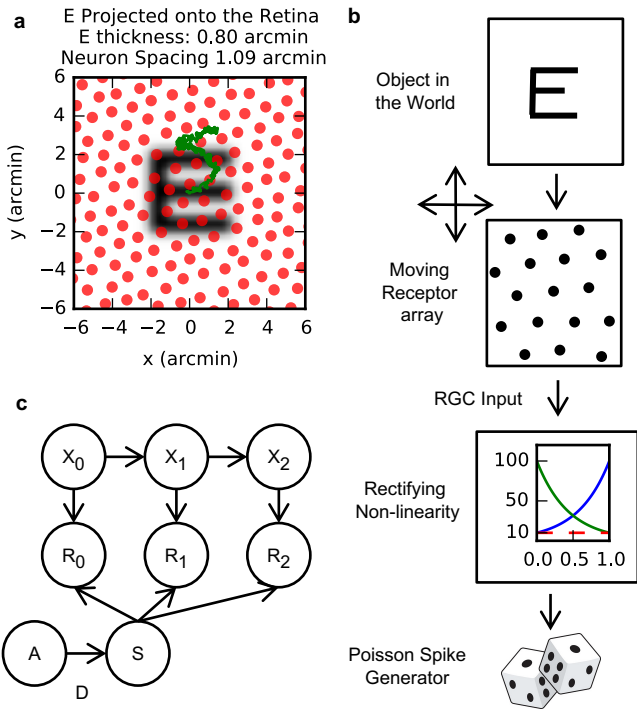


Fig. 1. Top: An upright E projected onto a simulated cone sampling lattice. The width of the E is 0.8 arcmin. The cones are placed in a hexagonal lattice with spacing 1.09 arcmin. The lattice is randomly translated and rotated, and each cone is jittered slightly about its position. The green curve shows 500 ms of a sample eye trajectory collected in association with [1]. Bottom: Spike Generation Model: We start with an object in the world. Next, we project that object onto a photoreceptor array that moves according to fixational eye movements. The dot product between the retinal projection and the cone receptive field gives a number that is passed through a non-linearity to get a firing rate. Finally, that firing rate is converted to spikes using a Poisson spike generator.

proximating a Bayesian ideal observer that seeks to infer a sparse code of the image of an object in the world given RGC spikes. Second, we demonstrate that a sparse pattern prior improves inference. Third, we demonstrate in simulation that an aliased and diffraction-limited E (as in[1]) that is moving produces a richer, decodable signal as compared to a stationary pattern landing on the retina.

II. MATHEMATICAL METHODS

A. Model Notation

For the rest of the paper, we will use the following subscripts consistently: t : time step, b : batch index, i : pixel index, j : RGC index, k : sparse component index.

- 1) S is the spatial pattern to be inferred, in a pixel representation. S_i denotes a particular pixel of the pattern. We constrain S_i to be between 0 and 1. X_i^S denotes the center of pixel i . The pixels are placed in a grid with spacing ds . The simulated projected image of the pattern $I(x)$ is smoothed using Gaussian interpolation, with $\sigma^S = 0.5 * ds$. Then $I(x) = \sum_i S_i N(x; \mu = X_i^S, \sigma = \sigma^S)$ where N denotes a Gaussian.

- 2) A is the vector of sparse coefficients that generate S through a sparse coding dictionary, D . A_k denotes the k th sparse coefficient.
- 3) D is a sparse coding dictionary where D_k is the k th dictionary element.
- 4) X_t^R (sometimes abbreviated as X_t) denotes the amount that the retina has moved since the start of the simulation.
- 5) D_C is the diffusion coefficient of the eye movements, $\lambda_0 = 10$ Hz, $\lambda_1 = 100$ Hz are the baseline and maximum firing rates of the neurons.
- 6) $R_{t,j}$ denotes number of spikes of RGC j in the time window $[t, t + \Delta t]$. Δt is the timestep (usually taken to be 1 ms). We use R as an abbreviation for $R_{t,j}$ for all t and j .
- 7) The j th RGC has a gaussian receptive field $N(x; \mu = X_j^E, \sigma = \sigma_j^E)$. Each RGC can either be an ON cell or OFF cell. We construct a jittered, hexagonal cone lattice with spacing de . Each cone is connected to one ON RGC and one OFF RGC as is the case in the human fovea. $\sigma^E = 0.203 \cdot de$ following [12].

These quantities are related through the probabilistic graphical model shown in Figure 2. In order to specify a graphical model, we need to specify the probability of each node given the parents of that node. We systematically describe them below:

B. Generation Model

1) *Spiking Model*: We model the spiking of the neurons using a generalized linear model (GLM).

$$\log p(R_{t,j} | S, X_t) = R_{t,j} \log[\lambda_j(S, X_t) dt] - \lambda_j(S, X_t) dt \quad (1)$$

$$\lambda_j(S, X_t) = \exp(\log \lambda_0 + \log(\lambda_1/\lambda_0) \cdot c'_{j,t}) \quad (2)$$

$$c'_{j,t} = c_{j,t} \text{ if } j \in \text{ON} \text{ or } 1 - c_{j,t} \text{ if } j \in \text{OFF} \quad (3)$$

$$c_{j,t} = g * \sum_i S_i T(x_t^R)_{i,j} \quad (4)$$

$$T(x^R)_{i,j} = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{\|x_i^S - x_j^E - x^R\|^2}{2\sigma^2}\right] \quad (5)$$

$$\sigma^2 = (\sigma^S)^2 + (\sigma^E)^2 \quad (6)$$

g is a gain factor that sets the maximum size of $c_{j,t}$ to be 1. In practice, $\lambda_1 \Delta t$ is much less than 1. The equation for $T(x)$ results from taking the dot product of the Gaussian interpolated image and the Gaussian Receptive field of each cone.

2) *Spike Train Generation Process*: In order to generate a spike train for our decoder, we feed a spatial pattern and an eye motion path into our spiking model above. For the eye path, we either generate a diffusive random walk with a diffusion constant D_C^{gen} or we use actual eye motion trajectories collected from an AOSLO [13]. For the objects in the world, we use an E or an MNIST digit (future work will consider natural scenes as well).

C. Priors for the Decoding Algorithm

There are many possible pairs of eye motion paths and spatial patterns that are consistent with the incoming retinal

spikes. In order to deal with this ambiguity, we impose priors on the eye trajectory and the pattern, as shown in Figure 2. If N are all the nodes of the graphical model, then $p(N) = \prod_i p(N_i | N_{\pi(i)})$ where $\pi(i)$ denote the parents of node i . Then all other quantities that we are interested in are computed by marginalizing the resulting distribution. Our algorithm is based on using the EM algorithm to approximate $\text{argmax}_A p(A|R)$ and then expanding the resulting equations using a Gaussian approximation to the data terms.

1) *Motion Prior*: We model the fixational eye movements as a simple diffusion process with a diffusion constant $D_C \equiv D_C^{infer}$:

$$p(X_0) = \delta(X_0) \quad (7)$$

$$-\log p(X_t | X_{t-1}) = \frac{1}{2(D_C/2)\Delta t} (X_t - X_{t-1})^2 + C \quad (8)$$

Note that X_t is a two dimensional vector, so for the overall vector to have a diffusion constant of $D_C \Delta t$, then each individual component has a diffusion constant of $D_C/2 \Delta t$.

2) *Pattern Prior*: Finally, we must specify a prior on the spatial pattern to be inferred. For this we utilize a sparse coding prior:

$$-\log p(S|A) = \delta(S - DA) \quad (9)$$

$$-\log p(A) = \beta \sum_k |A_k| \quad (10)$$

where $\delta(x)$ is a delta function. As a result of this prior, we estimate a sparse code of the pattern instead of directly estimating pixels. We also add a term in the cost function to force the pixels to be in the range $[0, 1]$: $-\log p(S_i) = \gamma * (\Theta(S_i - 1) + \Theta(-S_i))$ where $\Theta(x)$ is 1 if $x > 0$ and zero otherwise and γ is a parameter. It should also be noted that an independent pixel prior as in Burak *et al.*[2] may be seen as a special case of this system when $\beta = 0$ and D is equal to the identity matrix.

D. Algorithm for Inferring Pattern and Motion from Spikes

In extending previous models of vision in the presence of fixational eye movements, there are a number of key themes driving this work. First, we want the algorithm to be causal (eg. you cannot use information from the future to infer your current state). Second, we want the algorithm to be an online algorithm. That is to say that the algorithm has a finite memory buffer that it updates using observations. A Kalman filter is a good example of an algorithm of satisfying these two requirements. Third, we want the algorithm to work with a pattern representation that does not necessarily consist of pixels, but where each neuron could have a structured (e.g., oriented) receptive field as in V1. Fourth, we want the algorithm to be implementable in a neural circuit.

Here, we describe an algorithm with such properties. For emphasis, we note that the approach in [2] does not generalize to this situation and we developed a novel approach to handle this more general setting. The algorithm requires storing three variables in memory:

- 1) $q_t(X_t)$ is the algorithm's current estimate of the position of the eye at time t . Concretely, we write $q_t(X_t) = \sum_b W_{t,b} \cdot \delta(X_t, X_{t,b})$ where $X_{t,b}$ is a collection of positions, and $W_{t,b}$ are the corresponding weights.
- 2) \hat{A}_t , is a vector of size N_{sp} (the number of sparse coefficients) that represents the algorithm's estimate of the underlying spatial pattern, represented as a sparse code, after looking at spikes in the time interval $[0, t]$.
- 3) \hat{H}_t is a matrix of size N_{sp} by N_{sp} that represents the inverse of the covariance associated with our estimate of A_t after looking at spikes in the time interval $[0, t]$.

The algorithm consists of the following steps:

1. Initialization: set $\hat{A}_0 = 0$ and $H_0 = 0$.
2. Update q :

$$q_{t+1}(X_{t+1}) \sim p(R_{t+1} | X_{t+1}, S = D\hat{A}^t) \sum_{X_t} p(X_{t+1} | X_t) q_t(X_t) \quad (11)$$

We use sequential importance sampling with resampling in order to evaluate this equation as in [14].

3. Update the estimate of the sparse coefficients:

$$\hat{A}^{t+1} = \text{argmin}_A [E_g(A) + E_r^{t+1}(A) + E_p(A)] \quad (12)$$

$$E_g(A) = \frac{1}{2} (A - \hat{A}^t)^T H_t (A - \hat{A}^t) \quad (13)$$

$$-E_r^{t+1}(A) = \langle \log p(R_{t+1} | X_{t+1}, S = DA) \rangle_{q_{t+1}(X_{t+1})} \quad (14)$$

$$-E_p(A) = [\log p(A)] - (A - \hat{A}^t) \frac{\partial \log p(A)}{\partial A} \Big|_{A=\hat{A}^t} \quad (15)$$

$$+\gamma * \sum_i \Theta(S_i - 1) + \Theta(-S_i) \quad (16)$$

where $\Theta(x) = x$ for $x \geq 0$, and zero otherwise. The minimization is executed using the FISTA algorithm [15]. A neural interpretation emerges when writing out the FISTA equations for this minimization (similar to the locally competitive algorithm of Rozell *et al.*[16]).

4. Update the value for the Hessian:

$$\hat{H}^{t+1} = \exp\left(-\frac{\Delta t}{\tau}\right) \hat{H}^t + \frac{\partial^2}{\partial A^2} E_r^{t+1}(A) \Big|_{A=\hat{A}^{t+1}} \quad (17)$$

where τ is a time constant for forgetting the Hessian.

It can clearly be seen that this gives us an online algorithm as we take the previous state, $(q, \hat{A}, \hat{H})_t$, combine it with new data R_{t+1} , and calculate the new state $(q, \hat{A}, \hat{H})_{t+1}$.

III. COMPARISON OF DIFFERENT PATTERN PRIORS

One key assumption of previous work [2] is that the decoding model uses an independent pixel prior in the decoding of the spatial pattern. This results in pattern cells that essentially have single pixel receptive fields. Here we demonstrate that not only are we able to infer the pattern in a sparse coding basis, but using the sparse prior improves inference (Fig. 3). We train a positive-only sparse code with 81 dictionary elements on the MNIST dataset downsampled by a factor of two (which has dimension $14^2 = 196$). As another baseline, we do the same training with the sparsity penalty equal to zero (non-sparse prior). Finally, we compare this to an independent pixel prior.

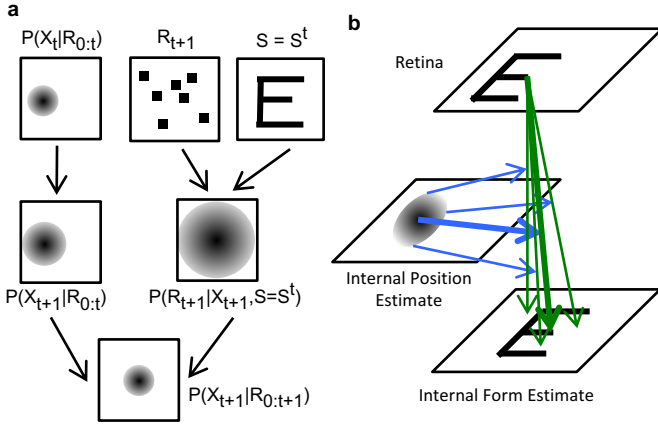


Fig. 2. Top: Graphical model underlying our model. The spikes (R) are observed and the sparse coefficients (A) and eye position (X) must be simultaneously inferred. Our solution alternates between two steps. Middle: In the first step, we fix the estimate of the pattern and update the estimate of eye position (shown as a probability cloud). When we predict the position at the next point in time, the uncertainty of the position estimate grows ($P(X_{t+1}|R_{0:t})$). By comparing the current estimate of the pattern with the incoming spikes and probabilistically combining that information with the prediction, we get an updated position estimate, $P(X_{t+1}|R_{0:t+1})$. Bottom: In the second step, we use the internal position estimate to dynamically route incoming spikes to the correct part of the internal form (pattern) estimate.

We see that performance is the best when we infer the pattern using a sparse prior.

IV. MOTION BENEFIT FOR HIGH ACUITY VISION

Ratnam *et al.*[1] show that fixational drift eye movements are beneficial for the perception of high acuity targets. One hypothesis to explain these results is that a moving stimulus generates a more informative retinal signal than a stationary stimulus, due to gaps and inhomogeneities in the cone sampling lattice. We support this hypothesis with our model (see Fig. 4). Summary of results:

- 1) A diffraction-limited tumbling letter E defined by line stroke widths of 0.8 arcmin projected on a cone lattice with 1.09 arcmin spacing is better recovered by our algorithm when the object is moving.
- 2) A simple pattern prior using a dictionary consisting of 2x2 pixel blocks is sufficient for generating a recognizable E.
- 3) At this size, the decoder frequently makes errors in absolute position of the stimulus but preserves relative spatial relationships within the pattern.

V. CONCLUSION

This work uses mathematical modeling and psychophysical experiments based on diffraction-limited stimuli to investigate the neural circuitry underlying human high-acuity vision. In addition to suggesting a neural mechanism for our ability to have 20-20 vision in the presence of eye movements, our model reproduces psychophysical measurements.

It should be noted that while the model recovers an explicit stabilized representation of the object, it is also possible

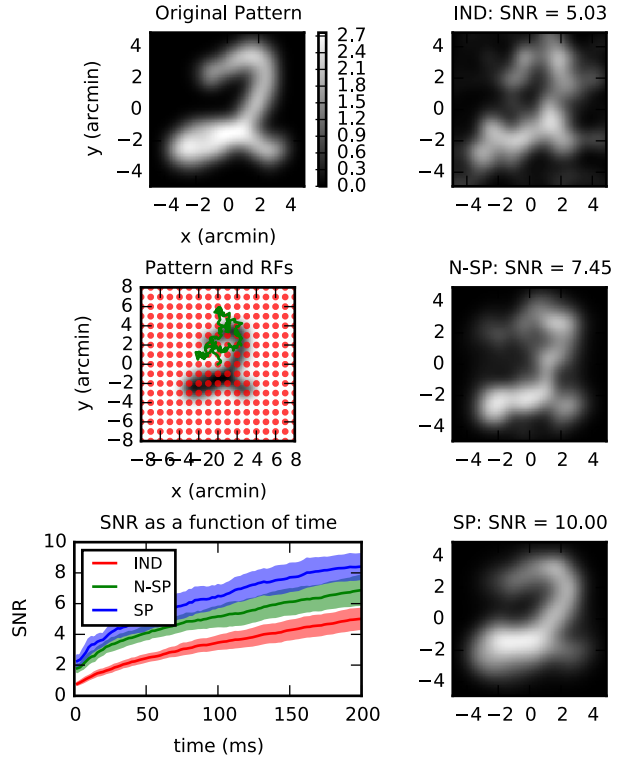


Fig. 3. Top: Results of the decoding process. The top left shows the object projected onto a cone mosaic. The top right shows an exponential moving average of the spikes of the ON and OFF cells. The bottom left shows the inferred pattern after 200 ms of spikes. The bottom right shows the true eye path (green) and the estimated eye path (blue) plus or minus one standard deviation. The path was generated with a diffusion constant of 100 arcmin²/sec. Cone spacing is 1 arcmin. The pattern is defined on a 14 × 14 array with each pixel spaced apart by 0.7 arcmin. The pattern is inferred using a sparse coding basis trained on MNIST. Bottom: As a function of time, decoding with a sparse coding prior works better than the independent pixel prior and the non-sparse prior ($p = 0.005$). The shaded region reflects plus or minus half a standard deviation over different trials.

that these computations could be done in a non-stabilized representation that still integrates information optimally. In particular, we could have a population of cells, \bar{A}_t , that represent a sparse code of the pattern translated by the current eye position. While this would simplify part of the model, we would need to update \bar{A}_{t+1} from \bar{A}_t and $X_{t+1} - X_t$. This would require the circuit to know how to compute a translation in an arbitrary direction in the current encoding of the pattern (e.g. if T_X is the translation operator in pixel space, then the circuit would need to implement $T'_X \approx D^{-1}T_X D$ which is a translation operator in the sparse code space). In our preliminary experiments with such model, we found it difficult to model the translation operator. More theoretical work on a translation operator that acts on a sparse code of an pattern could enable such a model.

ACKNOWLEDGMENT

The authors would like to thank members of the Redwood Center for Theoretical Neuroscience (David Zipser, Dylan Paiton, Eric Weiss, Jesse Livezey, Brian Cheung, and Vasha

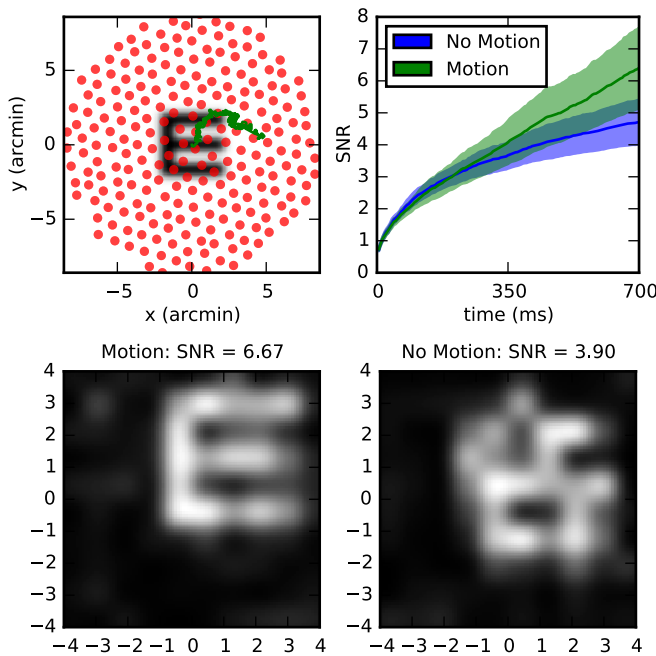


Fig. 4. Top: SNR of the reconstruction of the E as a function of time, averaged over 40 trials (width is plus-minus half a standard deviation). Red shows the case of a moving retina using actual eye movements[1] Blue shows the case of no motion. (The difference at 700 ms has $p = 0.00002$ using the Kolmogorov-Smirnov 2 sided statistic on 2 samples). A simple pattern prior is used to reconstruct the E (as opposed to a prior that forces the inferred pattern to be one of the four orientations of the E) in which the dictionary consists of blocks of 2×2 pixels, with no sparsity imposed. The entire pattern is defined on a 20×20 array. The size of the E and the lattice are the same as in Fig. 1. Bottom: Typical results of these simulations (the organization of the figure is the same as Fig. 3).

Dutell) and the Roorda lab for useful feedback. This material is based upon work supported by the National Science Foundation under Grant No. DGE 1106400 (AGA) and IIS-1111765 (AGA, BAO). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The experimental work was supported by NIH grant EY023591 (AR), Foundation Fighting Blindness (AR), and the Irvin M. Borish/Essilor Ezell Fellowship (KR).

REFERENCES

- [1] K. Ratnam, W. Harmening, and A. Roorda, "Fixational eye movements improve visual performance at the sampling limit," *Journal of Vision*, vol. 15, no. 12, 2015.
- [2] Y. Burak, U. Rokni, M. Meister, and H. Sompolinsky, "Bayesian model of dynamic image stabilization in the visual system," *Proceedings of the National Academy of Sciences*, vol. 107, no. 45, pp. 19 525–19 530, 2010.
- [3] B. L. Guthrie, J. D. Porter, D. L. Sparks *et al.*, "Corollary discharge provides accurate eye position information to the oculomotor system," *Science*, vol. 221, no. 4616, pp. 1193–1195, 1983.
- [4] I. Donaldson, "The functions of the proprioceptors of the eye muscles," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 355, no. 1404, pp. 1685–1754, 2000.

- [5] I. Murakami and P. Cavanagh, "Visual jitter: evidence for visual-motion-based compensation of retinal slip due to small eye movements," *Vision research*, vol. 41, no. 2, pp. 173–186, 2001.
- [6] X. Pitkow, H. Sompolinsky, and M. Meister, "A neural computation for visual acuity in the presence of eye movements," *PLoS biology*, vol. 5, no. 12, p. e331, 2007.
- [7] B. Motter and G. Poggio, "Dynamic stabilization of receptive fields of cortical neurons (vi) during fixation of gaze in the macaque," *Experimental brain research*, vol. 83, no. 1, pp. 37–43, 1990.
- [8] M. Gur and D. M. Snodderly, "Visual receptive fields of neurons in primary visual cortex (v1) move in space with the eye movements of fixation," *Vision research*, vol. 37, no. 3, pp. 257–265, 1997.
- [9] J. M. McFarland, A. G. Bondy, B. G. Cumming, and D. A. Butts, "High-resolution eye tracking using v1 neuron activity," *Nature communications*, vol. 5, 2014.
- [10] D. M. Snodderly, I. Kagan, and M. Gur, "Selective activation of visual cortex neurons by fixational eye movements: implications for neural coding," *Visual neuroscience*, vol. 18, no. 2, pp. 259–277, 2001.
- [11] M. Rucci, R. Iovin, M. Poletti, and F. Santini, "Miniature eye movements enhance fine spatial detail," *Nature*, vol. 447, no. 7146, pp. 852–855, 2007.
- [12] D. I. Macleod, D. R. Williams, and W. Makous, "A visual nonlinearity fed by single cones," *Vision research*, vol. 32, no. 2, pp. 347–363, 1992.
- [13] A. Roorda and J. L. Duncan, "Adaptive optics ophthalmoscopy," *Annual review of vision science*, vol. 1, p. 19, 2015.
- [14] A. Doucet and A. M. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later," *Handbook of Nonlinear Filtering*, vol. 12, pp. 656–704, 2009.
- [15] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.
- [16] C. J. Rozell, D. H. Johnson, R. G. Baraniuk, and B. A. Olshausen, "Sparse coding via thresholding and local competition in neural circuits," *Neural computation*, vol. 20, no. 10, pp. 2526–2563, 2008.