nature neuroscience

Decorrelation and efficient coding by retinal ganglion cells

Xaq Pitkow[1] & Markus Meister[2]

An influential theory of visual processing asserts that retinal center-surround receptive fields remove spatial correlations in the visual world, producing ganglion cell spike trains that are less redundant than the corresponding image pixels. For bright, high-contrast images, this decorrelation would enhance coding efficiency in optic nerve fibers of limited capacity. We tested the central prediction of the theory and found that the spike trains of retinal ganglion cells were indeed decorrelated compared with the visual input. However, most of the decorrelation was accomplished not by the receptive fields, but by nonlinear processing in the retina. We found that a steep response threshold enhanced efficient coding by noisy spike trains and that the effect of this nonlinearity was near optimal in both salamander and macaque retina. These results offer an explanation for the sparseness of retinal spike trains and highlight the importance of treating the full nonlinear character of neural codes.

The optic nerve limits how much visual information the eye can transmit to the brain. Early researchers postulated that the retina is designed to use that limited information capacity efficiently, reducing the redundancy in natural scenes by discarding information that the brain has already received from another source in space or time[1,2]. Subsequently, this idea was formalized mathematically[3–8]; images from the natural world have strong, uninformative correlations between the signals carried by different pixels[9]. An efficient encoder could suppress these by spatially filtering the image and thus optimize information transmission. Based on a model of the retina with several simplifying assumptions, one can compute the optimal spatial filter, which resembles the familiar center-surround receptive fields of retinal ganglion cells (RGCs)[5,10]. By computing the difference between the intensity at a point and the average intensity at nearby points, this filter indeed removes spatial correlations in the retinal image, up to some limit determined by photoreceptor noise. This idealized retina model correctly predicts the spatial sensitivity of human vision[6] and several other psychophysical laws[8].

Despite the decorrelation theory's successful predictions, there has been no experimental test of whether neural activity is in fact decorrelated at the putative bottleneck of the optic nerve. One study confirmed that neural firing in the cat's lateral geniculate nucleus is decorrelated in time[11], but there was no test of correlations across space. Another reported both spatial and temporal decorrelation by second-order fly visual neurons[7]. However, the stimuli in this study were still images scanned over the retina, confounding the spatial and temporal contributions to visual processing. A third study found that RGCs oversample visual space, resulting in substantial redundancy[12], but this oversampling may exist either with or without decorrelation relative to the stimulus. Thus, one is still left with these basic questions: does retinal processing indeed decorrelate signals at different spatial locations? If so, does this decorrelation improve coding efficiency?

We inspected spatial and temporal decorrelation in the retina by recording from a population of RGCs while presenting a stimulus with the spatio-temporal correlation structure of natural scenes[9]. We then compared the correlations among RGC spike trains to the correlations between corresponding image locations. To understand how the decorrelation occurs, we separately analyzed the contributions from center-surround receptive fields, noise and sparsifying nonlinearities in the retinal network. We conclude that the dominant effect comes not from the receptive field, but from the nonlinear stimulus-response relationship. These nonlinearities exhibited high response thresholds that led to sparse firing rates. We found that these attributes permitted neurons to transmit information with nearly optimal efficiency.

RESULTS

Our goal was to test whether retinal circuits remove the spatio-temporal correlations present in natural scenes and, if so, to explain whether this helps encode the stimulus efficiently. We measured correlation as a function of distance and time lag in both the visual input and the RGC output. We recorded spike trains from many ganglion cells in the isolated salamander retina under two visual stimuli: naturalistic, which consisted of pseudo-random Gaussian flicker with long-range spatio-temporal correlations such as those of natural scenes (**Fig. 1a** and **Supplementary Fig. 1**), and white noise, which consisted of a flicker stimulus without correlations (**Fig. 1b**). The stimuli were bright in the photopic regime, where the efficient coding theory predicts that decorrelation is the optimal strategy[4,7].

RGCs decorrelate the visual input

The typical ganglion cell responded to such displays with precisely timed bursts of spikes separated by complete silence (**Fig. 1c**).

[1]Department of Brain and Cognitive Sciences, University of Rochester, Rochester, New York, USA. [2]Molecular and Cellular Biology, Center for Brain Science, Harvard University, Cambridge, Massachusetts, USA. Correspondence should be addressed to M.M. (meister@fas.harvard.edu).

Received 20 September 2011; accepted 13 February 2012; published online 11 March 2012; doi:10.1038/nn.3064

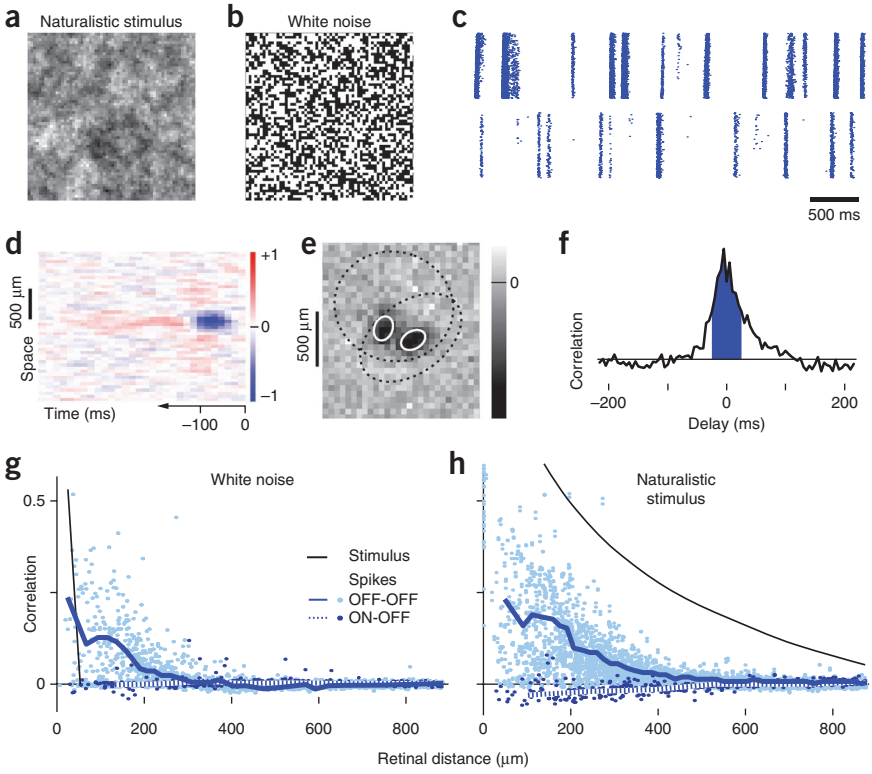

Figure 1 Decorrelation of naturalistic stimuli. (**a**,**b**) Sample frames of naturalistic and white noise stimuli, projected onto a 3.4-mm square on the retina. (**c**) Responses of two RGCs to a short segment of the naturalistic stimulus, displayed as rasters of spikes on 250 identical repeats. (**d**) A sample spatio-temporal receptive field for an OFF ganglion cell, measured as the spike-triggered average stimulus and integrated over one spatial dimension for ease of display. Note the spatial center-surround antagonism (red regions above and below blue) and the biphasic time course (red region left of blue). (**e**) Spatial receptive fields of two OFF cells, including 1-s.d. outlines of the receptive field centers (solid) and surrounds (dotted). (**f**) Cross-correlation function between two ganglion cell spike trains, indicating the frequency of spike pairs as a function of their delay. The shaded area encompasses most of the central peak and indicates the range of delays used to compute the quoted correlation coefficients. (**g**,**h**) Correlation coefficient between the responses of two ganglion cells as a function of their distance under a white noise (**g**) or naturalistic (**h**) stimulus. Each pair of cells contributes a point; lines represent median correlation for pairs at similar distance. Comparisons are restricted within a cell type (solid lines) or across cell types (dashed lines). For reference, the correlation between stimulus pixels is shown (thin lines).

We measured each neuron's spatio-temporal receptive field (**Fig. 1d**,**e**) using the standard reverse-correlation method[13] and then computed the correlation function between the spike trains of any two neurons (**Fig. 1f**). These correlation functions generally showed a central peak ~50 ms wide; this was also the characteristic timescale for variations in the firing rate (**Fig. 1c**). We therefore focused our analysis on the correlations of spike counts in 50-ms time windows (see Online Methods).

We plotted the firing correlation for every pair of ganglion cells against the retinal distance between their receptive field centers (**Fig. 1g**,**h**). This can be compared with the spatial correlations in the stimulus. During white-noise stimulation, the correlation in the firing of ganglion cells greatly exceeded the stimulus correlation out to ~300 µm (**Fig. 1g**). This is because the receptive field centers of nearby RGCs overlap (**Fig. 1e**), and they therefore receive correlated input from their shared photoreceptors. In contrast, under the naturalistic stimulus, neural responses were markedly less correlated than the stimulus pixels (**Fig. 1h**). The ganglion cells exhibited correlations only to ~400 µm distance, whereas the stimulus correlations extended at least twice as far. These observations held for distinct cell types[14,15] that were analyzed separately (data not shown). Thus, the retina decorrelates stimuli with natural image statistics while introducing excess correlation under the unnatural white-noise ensemble. This much is consistent with the classical efficient coding theory.

Decorrelation is primarily achieved by retinal nonlinearities

However, the theory also specifies a decorrelation mechanism, namely RGC receptive fields with antagonistic center and surround regions[2–4,7]. Owing to this antagonism, a RGC fires less to stimuli with low spatial frequency, which drive center and surround equally, and more to those with high spatial frequency[16]. But the low-frequency patterns are precisely those that synchronize nearby neurons. Consequently, a center-surround receptive field should reduce spatial correlations in the retinal output.

To test this, we measured how much decorrelation could be attributed to receptive field filtering. We convolved each spatio-temporal receptive field (**Fig. 1d**) with the naturalistic visual stimulus and analyzed the remaining correlations (**Fig. 2**). Filtering by the receptive field center alone extended the range of correlations, but the addition of the antagonistic surround reduced them below the correlations in the stimulus (**Fig. 2a**), as predicted by the theory, especially at distances beyond one center diameter, ~300 µm (**Fig. 2a**). However, unlike the theoretical prediction, this decorrelation was far from complete. Under the high-contrast stimuli that we used, the optimal linear filters should reduce the correlations to nearly zero[4] for distances greater than the center diameter. Instead, the experimentally measured receptive fields left substantial correlations out to distances twice as great (**Fig. 2a**,**b**), falling far short of the theoretical prediction.

In comparison, the actual decorrelation achieved by the retina was very efficient. The measured correlations between ganglion cell spike trains were suppressed by a factor of ~3 even inside the receptive field center, and by more than tenfold outside (**Fig. 2a**,**b**). Clearly, something other than receptive field filtering is responsible.

Each ganglion cell fired in short stimulus-locked episodes, with some trial-to-trial variation (**Fig. 3**). The correlation between two such spike trains depends on the similarity of their firing events, and therefore on timing, sparseness, and trial-to-trial fluctuations or noise in each firing event (**Fig. 3d**).

The timing of a ganglion cell's firing events is largely determined by its spatio-temporal receptive field, as confirmed by comparing the peaks in the filtered stimulus to those in the firing rate (**Fig. 3b**,**c**). However, measured firing events were narrower than the positive excursions of the linear model (**Fig. 3b**,**c**), presumably resulting from the many documented nonlinearities in the retina's response, including synaptic rectification, depression, gain control, spiking threshold and refractoriness[17–20]. This sparsification means that firing events overlap in time much less than expected from receptive field processing.

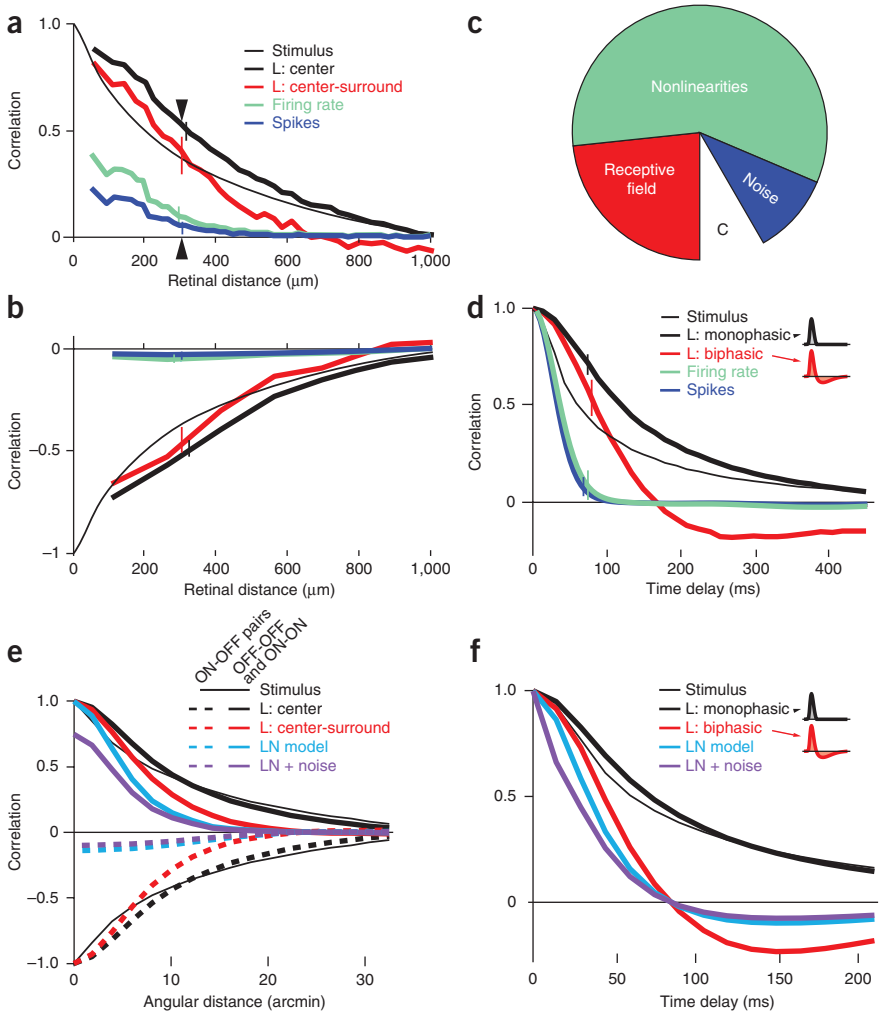

Figure 2 Nonlinearity accounts for much of decorrelation. (**a**,**b**) Spatial correlation functions for neurons and models under naturalistic stimulation. Cells with the same polarity preference (OFF-OFF or ON-ON pairs) have positive correlations (**a**) and those with opposite polarity preferences (OFF-ON pairs) have negative correlations (**b**). Curves are presented as in **Figure 1h** for the stimulus, trial-averaged firing rates, spike trains and linear models. The stimulus correlations are shown with opposite sign for ease of comparison in **b**. Results from many cell pairs are summarized by the median correlation for pairs at similar retinal distance; error bars indicate the central quartiles. L: center and L: center-surround designate linear models using receptive fields including the center component only or both center and surround. (**c**) The origins of decorrelation in different response components. The full circle represents the median correlation present in the stimulus after filtering by the receptive field centers at a retinal distance of 300 μm (arrowheads in **a**). The empty wedge (C) is the much smaller remaining correlation between the ganglion cell spike trains. The red wedge represents the decorrelation caused by lateral inhibition from receptive field surrounds. The difference between the linear response and the observed firing rate is a result of nonlinear processing and is responsible for over half the decorrelation implemented by the retina (green wedge). The trial-to-trial variation contributes an additional small amount of decorrelation (blue wedge). (**d**) Decorrelation in the time domain. Autocorrelation functions of salamander ganglion cell responses and linear models are plotted as a function of delay during naturalistic stimulation. The linear filter's first lobe, ~100 ms wide (inset, black), introduced excess correlation beyond that in the stimulus. The antagonistic second lobe (inset, red) counteracted those, but overcompensated, introducing anticorrelations at long delays. The observed correlations in the firing rate were much smaller still. (**e**,**f**) Spatial (**e**) and temporal (**f**) correlations in macaque RGCs, displayed as in **a**, **b** and **d**. Macaque RGC responses were approximated by an LN model[13,23], using published spatio-temporal receptive field parameters[36] (equations (4–6)) and sigmoidal nonlinearities[23] (equation (10)). The output noise was modeled as sub-Poisson variation (equation (11)) with parameters derived from published spike trains[21] (see Online Methods). The stimulus was scaled in space and time to compensate for the different scales of primate and salamander receptive fields. L, receptive field filter only; LN, including the nonlinearity; LN + noise, including the noise.

Indeed, correlations of the trial-averaged firing rates (**Fig. 3b**) lay far below those of the stimulus and those predicted from receptive field filtering alone (**Fig. 2a**). The effect is especially notable for neurons of opposite response polarity, where the retinal nonlinearities effectively abolish the pairwise correlations (**Fig. 2b**). Finally, we determined that the trial-to-trial fluctuations in different neurons were largely independent under the present stimulus conditions (**Supplementary Fig. 2**). This noise further decorrelates the ganglion cell output (**Fig. 2a**,**b**). Such noise-induced effects are detrimental to efficient coding, but downstream circuits in the brain cannot distinguish decorrelation by noise from that achieved by other means.

One can now compare how much the different aspects of retinal processing contribute to decorrelating the retinal output. For instance, at a distance of 300 μm, the natural stimulus contained strong correlations, but retinal processing subsequently reduced them by a total of 92% (**Fig. 2a**). Of this, the receptive field surround contributed ~25%, the sparsifying nonlinearities contributed ~60% and noise was responsible for ~15% (**Fig. 2c**). Thus, nonlinear processing in retinal circuits is by far the largest contributor to decorrelation at the retinal output, whereas the much-touted center-surround receptive field makes only a minor contribution.

These observations applied to temporal correlations as well. Filtering the stimulus through the receptive field produced a mild reduction in the autocorrelation at short time delays, but also introduced strong anticorrelations at long delays (**Fig. 2d**). This was a result of the biphasic time course of the receptive field (**Fig. 1d**), analogous to the spatial antagonism between center and surround. In comparison, both the trial-averaged firing rate and noisy spike trains showed almost complete decorrelation, down to delays <100 ms (**Fig. 2d**). Again, one concludes that the filtering by receptive fields reduces stimulus correlations only marginally, whereas the sparsifying nonlinearities account for the bulk of temporal decorrelation in the retina.

To assess the generality of these results, we asked whether they also extend to primate retinas and, thus, to our own visual processing. Published spike trains show that macaque RGCs similarly produce sparse bursts separated by silence[21,22], an indication that substantial nonlinear processing occurs in the macaque retina. By analyzing the

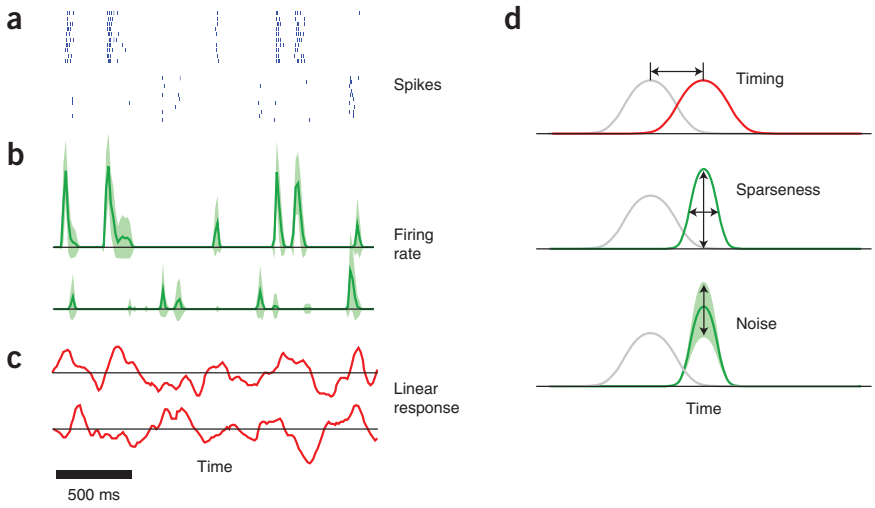

Figure 3 Sparseness in retinal responses. (**a**) Spike rasters for two salamander ganglion cells over ten repetitions of a naturalistic stimulus. Firing events are brief, separated by long silences, and have some trial-to-trial variability. (**b**) Mean firing rates for the same neurons, with shading that indicates the s.d. about the mean in time bins of 50 ms. (**c**) The linear response generated from convolving the stimulus with the spatiotemporal receptive fields of those two cells. This linear model generally captures the times of firing events, but differs markedly in sparseness. (**d**) Depiction of three factors contributing to decorrelation between two caricatured neural responses: event timing, sparseness and noise.

shapes of ganglion cell receptive fields[23,24], we confirmed that the center-surround filter explains only part of the decorrelation at the retinal output (**Fig. 2e**,**f**). On the other hand, the sparsifying nonlinearities in the response[23] make a substantial contribution. Again, they strongly decorrelate responses of opposite polarity (**Fig. 2e**) and they suppress negative temporal correlations at long delays (**Fig. 2f**).

Decorrelation, sparseness and efficient coding in LNP models

To build intuition for these effects and to prepare for further analysis, we considered a simple, tractable model of neural signaling that has enjoyed some popularity in the study of retina, visual cortex and other sensory modalities[25]. In the so-called LNP model, the visual stimulus is first convolved with a linear receptive field (L), producing a time-varying input signal. That signal is passed through an instantaneous nonlinearity (N), typically of sigmoid shape, producing a time-dependent firing rate from which the spike train is generated by a Poisson process (P). The LNP model offers perhaps the simplest instance in which one can analyze the contributions of receptive field, nonlinearity and noise to visual coding.

Consider two such neurons that process a Gaussian-distributed stimulus with different receptive field filters (**Fig. 4**). The outputs of the two filters will be jointly Gaussian variables with a statistical dependency that is fully characterized by the correlation coefficient. Passage through the subsequent nonlinearity always reduces the correlation of the two signals (**Fig. 4b–d**), regardless of the shape of the nonlinearity[26]. For a monotonic sigmoid nonlinearity, a higher threshold produces greater decorrelation (**Fig. 4c**,**d**). An increase in threshold also lowers the mean firing rate, accounting for earlier observations that correlations decrease when firing rates are low[27]. Note that a nonlinearity with a high threshold has qualitatively different effects from one with low threshold: although it suppresses positive correlation coefficients to a certain extent, it almost completely eliminates negative correlations (**Fig. 4c**,**d**). This is because two signals of opposite sign cannot cross threshold at the same time. These effects are very robust under different shapes of the nonlinearity (**Fig. 4c**), and likely explain why the observed anti-correlations between ON and OFF cells are so strongly suppressed by retinal nonlinearities (**Fig. 2b**). Finally, the effect of output noise is simply to reduce the correlation coefficient by a further factor (**Fig. 4e**). In sum, the basic relationships that we found for actual retinal spike trains can be understood in the context of a simple model of nonlinear stochastic processing.

The classical theory of retinal decorrelation attributed that phenomenon to filtering by center-surround receptive fields and explained its purpose as serving the efficient transmission of visual information through the optic nerve. Given that most of the observed decorrelation is instead furnished by the nonlinear response function of the retina, one wonders whether this version of decorrelation is equally beneficial for efficient coding. We explored this in the context of the LNP model and compared the resulting predictions with the measured spike trains.

In the LNP model, the nonlinearity decorrelates if it has a high threshold (**Fig. 4d**), ensuring that each neuron spends much of the time silent except for sharp and sparse firing events. This sparseness is prominent in the ganglion cell responses (**Fig. 1c**) and has been observed across species[11,21,22,28,29]. This seems to be counterproductive for efficient information transmission. Why don't ganglion cells modulate their firing rate continuously to encode different stimulus values? Suppose a neuron must transmit an input signal that changes every time interval Δt by producing spikes during each interval according to a Poisson process with some firing rate. What mapping from input to firing rate maximizes the information rate in the spike train?

To explore this, we compared different monotonic sigmoid nonlinearities, as are often observed in fitting the LNP model to visual neurons[23,30]. These can be described by three parameters: threshold, gain and peak rate (**Fig. 4b**). We took the filtered stimulus to have a normal distribution; this is guaranteed by construction for our Gaussian naturalistic stimulus and by the central limit theorem for the white noise stimulus because the receptive fields extend over many stimulus values in space and time. Next, we compute the mutual information between stimulus and spike train for any shape of the nonlinearity. The information can be increased arbitrarily by simply raising all of the firing rates, so we fixed the mean firing rate to a realistic value for RGCs. That constraint leaves only two free shape parameters for the nonlinearity, for example, the threshold and the gain.

At very high thresholds, the information transmission is poor (**Fig. 4f**). In this regime, the neuron reports only the rare threshold crossings, firing a burst of spikes each time to match the mean firing rate. Notably, transmission also drops at low thresholds. In this condition, the neuron fires in many of the time bins, and the spike counts must therefore be low to satisfy the average rate constraint. In a Poisson process, however, low spike counts are associated with high relative variability. Thus, the choice of threshold involves a trade-off between rarely using reliable symbols, such as high spike counts, or frequently using unreliable symbols, such as low spike counts. The optimum is found at an intermediate threshold value.

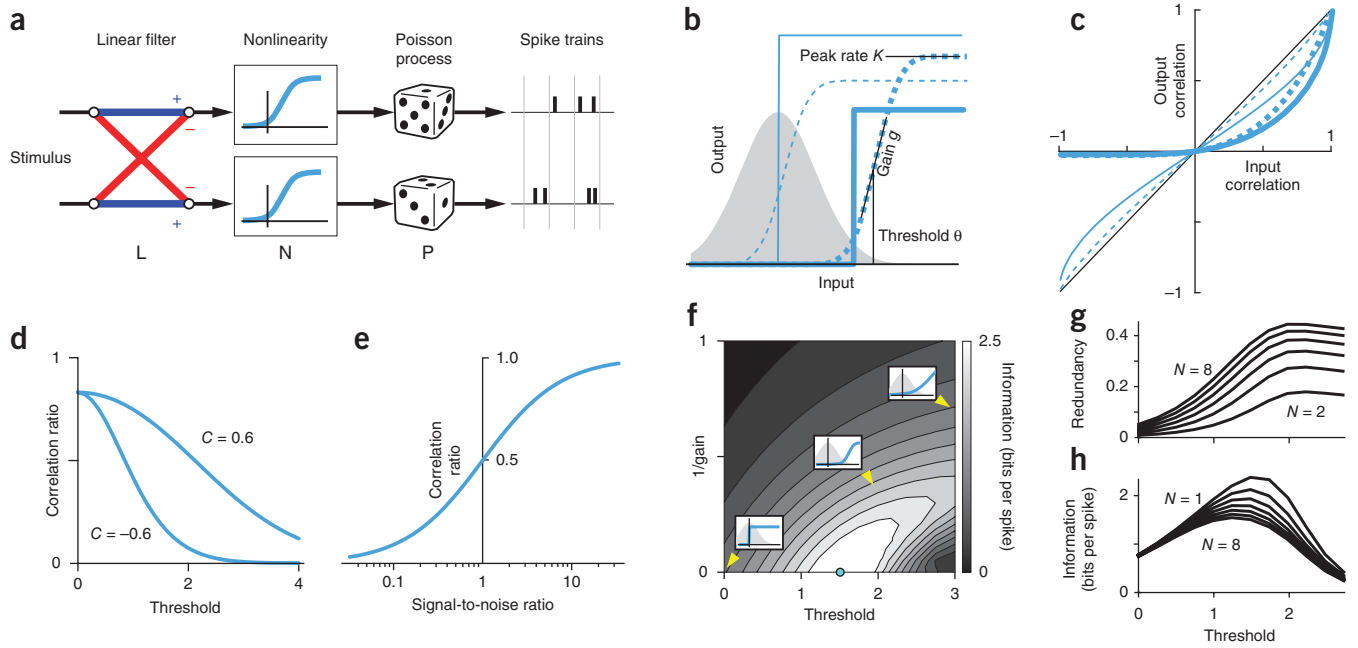

Figure 4 Decorrelation and efficient coding in the LNP model. (**a**) Schematic of two visual neurons that each respond according to the LNP model. For each cell (top and bottom), the stimulus is processed by a linear filter that includes lateral inhibition in space. This signal is passed through a sigmoid nonlinearity and the result modulates the rate of a Poisson process that generates spikes; the spike counts in discrete time windows are the response variable. (**b**) Four sample nonlinearities with sigmoid shape and high or low gain (solid or dashed lines), high or low threshold (thick or thin lines), and various peak rates. The shaded curve indicates the probability distribution of the filtered stimulus signal at the input to the nonlinearity. (**c**) The effects of such a nonlinear transform on the correlations between two jointly Gaussian variables. Note that the output correlation is always less than that of the input. A low threshold (thin lines) affects the correlation only weakly, but at high threshold (thick lines) the output correlation is greatly reduced, especially for negative values. The precise shape of the nonlinearity (dashed versus solid) is less important and the peak rate has no effect. (**d**) The ratio of output correlation to input correlation decreases with increasing threshold, shown here for the sigmoid nonlinearity applied to two variables with input correlation $C = \pm 0.6$. (**e**) When the two outputs are affected by independent additive noise, this reduces the output correlation by a factor determined by the signal-to-noise ratio (equation (14)). (**f**) Influence of the nonlinearity on information transmission. In the framework of the LNP model, the threshold and gain of the sigmoid nonlinearity determine how much information about the stimulus is transmitted by the spikes (grayscale and contour lines). The average firing rate was fixed at 1.1 Hz (the median over the salamander ganglion cells). Threshold and 1/gain are measured in s.d. of the input signal distribution. Insets illustrate nonlinearities (solid lines) at different thresholds and gains relative to the input distribution (shaded area). (**g**,**h**) When multiple neurons receive correlated inputs, raising the threshold makes their outputs more redundant (**g**) even as the total information increases (**h**) and correlation decreases (**d**). All neurons had pairwise correlation coefficients of 0.9, equal thresholds, optimal (infinite) gain and a fixed mean firing rate of 1.1 Hz. The optimal threshold varies only weakly with population size ($N = 1,\ldots,8$).

The optimal gain of the model neuron was infinite (**Fig. 4f**), with complete silence for stimuli below threshold and maximal firing rate for those above. This result runs counter to the conventional view of neural coding as a graded modulation of the firing rate, although related predictions have been in the theory literature for some time[31,32]. For the parameters that characterize a typical salamander ganglion cell from our experiments (coding window Δt=50 ms, average firing rate = 1.1 Hz), the optimal neuron should remain silent 94.5% of the time and fire at 20 Hz the remaining 5.5% of the time. Thus, efficient coding theory predicts that, under the present constraints on firing rate and dynamics, a neuron should indeed fire sparsely, with brief firing events being separated by periods of silence.

Sparse firing enhances coding efficiency of ganglion cells

How close do empirically observed firing rates come to optimal performance? We made the approximation that the dominant source of noise in ganglion cell responses arises at the output, after all of the retina's nonlinear processing has occurred, for example, during spike generation. In that case, the information transmission rate about the stimulus only depends on the probability distribution of the ganglion cell's firing rate, and not on how it is generated (equation (18)). Inspecting that distribution (**Fig. 5a**) reveals that, in most time bins, the measured rate was exactly zero, followed by a long tail in the distribution out to high values. These distributions are fit well by a three-parameter expression (equation (19)). How efficient are these distributions of the firing rate for information transfer?

For comparison, we identified the firing rate distribution with the same mean that used spikes most efficiently. Because real ganglion cell spike trains do not conform exactly to Poisson statistics[17,21,33], we used an empirically fit noise model (equation (11), **Supplementary Fig. 2**). The optimum firing rate distribution, as for the LNP model considered above, was a binary distribution that uses just two firing rates (**Fig. 5b**). But there was a corridor of high efficiency leading to that point, and almost all of the measured rate distributions lay in that domain. Indeed, when we computed the information transmission directly from the spike trains (Online Methods), the median RGC had a coding efficiency of 73% compared with the theoretical optimum (**Fig. 5c**).

Again, we found that these results extend to responses from primate RGCs. Although their mean firing rates were higher, the correlation time of the response, and thus the effective bin width, for spike train signaling was shorter, on the order of Δt = 10 ms[21]. We analyzed

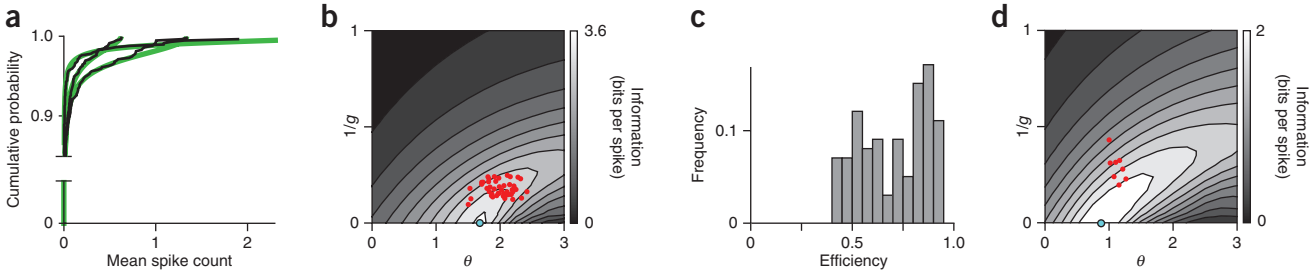

Figure 5 Efficiency of stimulus coding by RGCs. (**a**) Cumulative distribution of the spike count in 50-ms time bins, averaged over multiple repeats of the stimulus. Data (thin lines) for three sample ganglion cells and their fit with a model (thick lines) parametrized by θ, g and K (equation (20)). (**b**) The information transmitted by model firing rate distributions with a fixed mean firing rate of 1.1 Hz, whose shape is parametrized by θ and g. Noise was assumed to be sub-Poisson as observed empirically (equation (11), **Supplementary Fig. 3**). The blue dot indicates the globally maximal rate of information transmission at this mean rate. Red dots indicate the parameters of the rate distribution measured from salamander ganglion cells. These cells have widely varying mean firing rates. The contour plot of information transmission varies slightly with mean rate, but is shown here for illustration purposes only at one typical mean rate. (**c**) Histogram of information efficiencies over the population of salamander RGCs. For each cell, the information rate is calculated directly from the empirical spike counts. To calculate efficiency, we compared this information rate to the maximal information rate possible for the measured mean firing rate (Online Methods). (**d**) Information transmission estimated for macaque RGCs, displayed as in **b**. Red dots are parameters describing the firing rate distribution obtained from published spike rasters in response to white noise stimulation[21]. The contour plot shows the information transmission for different firing rate distributions while fixing the mean rate and time window to typical values, namely 30 Hz and 10 ms, respectively.

published distributions of the firing rate and the trial-to-trial noise[21] and computed information transmission rate as described above. The sparse responses of macaque neurons allowed a transmission rate close to the optimum, with a median efficiency of 81% (**Fig. 5d**). In summary, we found that a treatment of efficient coding theory that incorporates nonlinear transforms and noisy spike trains can explain the paradoxical nature of high-threshold nonlinearities and sparse responses in retinal processing.

DISCUSSION

Our findings extend the application of efficient coding theory in the retina to considerably more realistic conditions. The classic approach treated the early visual system as a linear filter, with graded output signals, Gaussian noise and an average power constraint[4,7], none of which describes the real retina. We allowed for nonlinear processing, a spiking output with stochastic noise and a constraint on the overall firing rate, as might be dictated by metabolic cost[34]. These extensions deliver new insights into the nature of retinal processing.

Two forms of redundancy reduction

We viewed the prominent decorrelation of signals in the retinal output as deriving primarily from two very different mechanisms (**Figs. 2–4**). The first is a linear spatio-temporal filter that implements lateral inhibition in space and biphasic responses in time. This conforms to the classic notion that the retina seeks to reduce redundancy between parallel channels in space and in a channel across time[2], although this reduction is incomplete (**Fig. 2**).

The second, more substantial contribution derives from nonlinear processing in each individual channel (**Figs. 2** and **3**), which efficiently matches visual signals to the available coding symbols (**Figs. 4** and **5**). This second stage reduces the coding redundancy in each output channel resulting from inefficient symbol use. These observations apply for ganglion cells of multiple types in different species, such as salamander and macaque (**Figs. 2** and **5**), suggesting that our extension of the efficient coding framework has some general utility.

Validity of the assumptions

Although the model of retinal processing that we used is considerably more realistic than that described in the classical linear decorrelation theory, it is worth inspecting the remaining approximations. For our first claim, that the receptive field filters contribute only a fraction of the decorrelation, we used a standard method to measure receptive fields, namely reverse correlation of the response to white noise stimuli[13]. Although receptive fields can adapt to the pattern of stimulation[35], we found that surrounds estimated under naturalistic stimulation narrowed only slightly and did not decorrelate any more than those obtained with white noise (data not shown).

Our second claim, that high-threshold nonlinearities enhance efficient coding, assumes that photoreceptor noise is negligible. This is the regime in which the classical theory predicts decorrelation of the retinal output. We also used this assumption to estimate the information rates in spike trains. The high light levels that we used in the experiments were designed to favor low photoreceptor noise. Any remaining input noise would be shared by ganglion cells with overlapping receptive fields, but we found noise correlations to be very small (**Supplementary Fig. 2**). This suggests that most of the noise in the RGC responses arises close to the output, rather than in shared presynaptic sources.

Our analysis of information transmission follows a classic approach[31] and requires choice of a coding window Δt, the timescale on which RGCs can completely change their firing rates to different values. We adopted $\Delta t = 50$ ms for salamander ganglion cells on the basis of the observed width of firing events (**Fig. 3a**,**b**) and their autocorrelation function (**Fig. 2f**). We varied Δt in the analysis and found that the general conclusions were insensitive to small changes in this parameter; sparse firing provides the most efficient code as long as the mean spike count remains considerably less than one spike per time bin.

Our models of signal and noise in the coding window do not specify a particular mechanism of spike generation. However, it is worth noting that the experimentally observed distributions of the firing rate (**Fig. 5a**) and the noise (**Supplementary Fig. 3**) are readily reproduced by mechanistic models such as a leaky integrate-and-fire neuron with Gaussian subthreshold noise (data not shown).

Incomplete decorrelation by receptive fields

We found that the spatial receptive fields of ganglion cells failed to decorrelate retinal signals as completely as would be expected from

the classical theory (**Fig. 2a**). Basically, the antagonistic surround of the receptive field is weaker than predicted. With the high luminance and high contrast that we used, the theory predicts that the integrated strength of the surround should precisely cancel the center (equation (2.4) of ref. 10). The receptive fields that we observed have much weaker surrounds, and thus decorrelate less. This is also evident in preceding work. In macaque RGCs, the surround amounts to only ~50% of the center (see Fig. 10 of ref. 36). Note that, although the original studies always wrote about "retinal filters" for spatial decorrelation, their tests of the theory used comparisons to human psychophysics, and therefore included post-retina stages of decorrelation (see Figs. 1 and 4 of ref. 4).

A plausible explanation of why the linear receptive fields fail to decorrelate much is that they don't entirely reflect what these RGCs compute. Many of these neurons are selective for quite specific visual features, such as motion in a particular direction[37], differential motion[38] or local edges[39]. This selectivity arises from diverse nonlinearities[40] and is poorly represented in the spatio-temporal receptive field. Thus, even neurons with strongly overlapping receptive fields may nonetheless never fire together. This recalls another feature of retinal organization that (so far) cannot be explained by efficient coding principles: the profusion of different ganglion cell types that each appear to compute a different visual message[41].

Nonlinearity and sparseness

Regardless of what a RGC computes, it must communicate the result downstream via noisy spike trains. To optimize information transmission using such a spiking process with a low mean activity, we found that RGCs should be silent most of the time and fire at a high rate only rarely. This expectation holds over all of the experimental conditions that we analyzed, for all of the salamander ganglion cells and for all but one of the macaque neurons. The actual measured nonlinearities were not quite infinitely sharp, but matched the expected threshold closely (**Fig. 5**).

The theory behind this was developed already some time ago. It was discovered by numerical methods that a Poisson process transmits maximal information using a discrete set of firing rates—only two if the maximal rate is strongly limited[31]. The result was later proved analytically in studies of fiber-optic communication[32]. Nevertheless, these facts are poorly appreciated among neuroscientists, even though Poisson models are used ubiquitously. Most of us (ourselves included) assumed that neurons should modulate their firing rate continuously to benefit from all possible rates. This intuition was formalized in an influential study[42] that derived a smooth sigmoid as the optimal shape of the response function. But that treatment was for a continuous output signal, such as membrane potential, and a constant additive noise level. The fact that the spike train is a point process with output-dependent noise ultimately leads to the counter-intuitive step-shaped nonlinearity. This behavior has been derived under a constraint on the maximal firing rate[32]. We found that discrete firing rate distributions also arise when the constraint applies instead to the mean rate (**Fig. 4f**).

The sparse responses of RGCs under naturalistic stimulation can be seen as maximizing coding efficiency in single spike trains in the optic nerve bottleneck. In the cortex, sparse coding has been interpreted differently, as a useful strategy for learning and processing spike patterns[43] or to extract large signals from background noise[44]. These arguments are plausible for highly overcomplete representations, where, unlike in the retina, the number of neurons greatly exceeds the stimulus dimensionality. Still, one might imagine that, even in the cortex, the driving force for sparseness is really communicating efficiently with Poisson spike trains[45].

Decorrelation and efficient coding

Correlation is often considered to be a proxy for information-theoretic redundancy, with the implication that decorrelation somehow improves efficiency. Certainly high correlation does imply strong statistical dependence, but weak correlation need not imply weak dependence; correlation is a second-order measure and fully reflects the redundancy between two signals only if they are normally distributed. For highly non-Gaussian signals, such as neural spike trains and natural images, correlation may be only weakly related to redundancy. For example, the nonlinearity of the LNP model markedly decreases the correlation between neural responses (**Fig. 4d**) while actually increasing their statistical dependency (**Fig. 4g**,**h**). Correlation and efficiency also have a complex relationship. For instance, if two signals are affected by independent noise, this decorrelates them without improving coding efficiency. Nonlinearities invariably decorrelate two Gaussian signals, but may not improve coding efficiency. Nonetheless, many studies of neural signaling simply measure correlation and leave the impression that decorrelation alone is evidence of improved efficiency[46–50].

These examples illustrate that all decorrelations are not created equal. Although many neural circuits perform some decorrelation of their inputs, one must distinguish the various forms of this phenomenon, as they are implemented by very different mechanisms and have different roles for the neural code.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/natureneuroscience/.

Note: Supplementary information is available on the Nature Neuroscience website.

ACKNOWLEDGMENTS

We thank the members of the Meister laboratory, M. Berry, T. Toyozumi and J.-P. Nadal for helpful advice. This work was funded by grants from the US National Institutes of Health to M.M.

AUTHOR CONTRIBUTIONS

X.P. and M.M. designed the study. X.P. performed all of the experiments, analysis and modeling. X.P. and M.M. wrote the article.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/natureneuroscience/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Attneave, F. Some informational aspects of visual perception. *Psychol. Rev.* **61**, 183–193 (1954).
2. Barlow, H.B. Possible principles underlying the transformation of sensory messages. in *Sensory Communication* (ed. Rosenblith, W.A.) 217–234 (MIT Press, Cambridge, MA, 1961).
3. Srinivasan, M.V., Laughlin, S.B. & Dubs, A. Predictive coding: a fresh view of inhibition in the retina. *Proc. R. Soc. Lond. B Biol. Sci.* **216**, 427–459 (1982).
4. Atick, J.J. & Redlich, A.N. What does the retina know about natural scenes? *Neural Comput.* **4**, 196–210 (1992).
5. Atick, J.J. & Redlich, A.N. Convergent algorithm for sensory receptive field development. *Neural Comput.* **5**, 45–60 (1993).
6. Atick, J.J. & Redlich, A.N. Could information theory provide an ecological theory of sensory processing? *Network* **3**, 213–251 (1992).
7. van Hateren, J.H. Real and optimal neural images in early vision. *Nature* **360**, 68–70 (1992).
8. van Hateren, J.H. Spatiotemporal contrast sensitivity of early vision. *Vision Res.* **33**, 257–267 (1993).
9. Field, D.J. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A* **4**, 2379–2394 (1987).
10. Atick, J.J. & Redlich, A.N. Toward a theory of early visual processing. *Neural Comput.* **2**, 308–320 (1990).
11. Dan, Y., Atick, J.J. & Reid, R.C. Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *J. Neurosci.* **16**, 3351–3362 (1996).

12. Puchalla, J.L., Schneidman, E., Harris, R.A. & Berry, M.J. Redundancy in the population code of the retina. *Neuron* **46**, 493–504 (2005).
13. Chichilnisky, E.J. A simple white noise analysis of neuronal light responses. *Network* **12**, 199–213 (2001).
14. Warland, D.K., Reinagel, P. & Meister, M. Decoding visual information from a population of retinal ganglion cells. *J. Neurophysiol.* **78**, 2336–2350 (1997).
15. Segev, R., Puchalla, J. & Berry, M.J. Functional organization of ganglion cells in the salamander retina. *J. Neurophysiol.* **95**, 2277–2292 (2006).
16. Enroth–Cugell, C. & Robson, J.G. Functional characteristics and diversity of cat retinal ganglion cells. Basic characteristics and quantitative description. *Invest. Ophthalmol. Vis. Sci.* **25**, 250–267 (1984).
17. Berry, M.J. & Meister, M. Refractoriness and neural precision. *J. Neurosci.* **18**, 2200–2211 (1998).
18. Burrone, J. & Lagnado, L. Synaptic depression and the kinetics of exocytosis in retinal bipolar cells. *J. Neurosci.* **20**, 568–578 (2000).
19. Demb, J.B., Zaghloul, K., Haarsma, L. & Sterling, P. Bipolar cells contribute to nonlinear spatial summation in the brisk-transient (Y) ganglion cell in mammalian retina. *J. Neurosci.* **21**, 7447–7454 (2001).
20. Field, G.D. & Rieke, F. Nonlinear signal transfer from mouse rods to bipolar cells and implications for visual sensitivity. *Neuron* **34**, 773–785 (2002).
21. Uzzell, V.J. & Chichilnisky, E.J. Precision of spike trains in primate retinal ganglion cells. *J. Neurophysiol.* **92**, 780–789 (2004).
22. Pillow, J.W. *et al*. Spatio-temporal correlations and visual signaling in a complete neuronal population. *Nature* **454**, 995–999 (2008).
23. Chichilnisky, E.J. & Kalmar, R.S. Functional asymmetries in ON and OFF ganglion cells of primate retina. *J. Neurosci.* **22**, 2737–2747 (2002).
24. Croner, L.J., Purpura, K. & Kaplan, E. Response variability in retinal ganglion cells of primates. *Proc. Natl. Acad. Sci. USA* **90**, 8128–8130 (1993).
25. Schwartz, O., Pillow, J.W., Rust, N.C. & Simoncelli, E.P. Spike-triggered neural characterization. *J. Vis.* **6**, 484–507 (2006).
26. Lancaster, H.O. Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika* **44**, 289–292 (1957).
27. de la Rocha, J., Doiron, B., Shea–Brown, E., Josic, K. & Reyes, A. Correlation between neural spike trains increases with firing rate. *Nature* **448**, 802–806 (2007).
28. Berry, M.J., Warland, D.K. & Meister, M. The structure and precision of retinal spike trains. *Proc. Natl. Acad. Sci. USA* **94**, 5411–5416 (1997).
29. Reinagel, P. How do visual neurons respond in the real world? *Curr. Opin. Neurobiol.* **11**, 437–442 (2001).
30. Baccus, S.A. & Meister, M. Fast and slow contrast adaptation in retinal circuitry. *Neuron* **36**, 909–919 (2002).
31. Stein, R.B. The information capacity of nerve cells using a frequency code. *Biophys. J.* **7**, 797–826 (1967).
32. Shamai, S. Capacity of a pulse amplitude modulated direct detection photon channel. *IEE Proc. Commun. Speech Vis.* **137**, 424–430 (1990).
33. Keat, J., Reinagel, P., Reid, R.C. & Meister, M. Predicting every spike: a model for the responses of visual neurons. *Neuron* **30**, 803–817 (2001).
34. Balasubramanian, V. & Berry, M.J. A test of metabolically efficient coding in the retina. *Network* **13**, 531–552 (2002).
35. Hosoya, T., Baccus, S.A. & Meister, M. Dynamic predictive coding by the retina. *Nature* **436**, 71–77 (2005).
36. Croner, L.J. & Kaplan, E. Receptive fields of P and M ganglion cells across the primate retina. *Vision Res.* **35**, 7–24 (1995).
37. Barlow, H.B. & Levick, W.R. The mechanism of directionally selective units in rabbit's retina. *J. Physiol. (Lond.)* **178**, 477–504 (1965).
38. Ölveczky, B.P., Baccus, S.A. & Meister, M. Segregation of object and background motion in the retina. *Nature* **423**, 401–408 (2003).
39. Levick, W.R. Receptive fields and trigger features of ganglion cells in the visual streak of the rabbits retina. *J. Physiol. (Lond.)* **188**, 285–307 (1967).
40. Gollisch, T. & Meister, M. Eye smarter than scientists believed: neural computations in circuits of the retina. *Neuron* **65**, 150–164 (2010).
41. Dacey, D.M. Origins of perception: retinal ganglion cell diversity and the creation of parallel visual pathways. in *The Cognitive Neurosciences* (ed. Gazzaniga, M.S.) 281–301 (MIT Press, Cambridge, Massachusetts, 2004).
42. Laughlin, S.B. A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch. C* **36c**, 910–912 (1981).
43. Olshausen, B.A. & Field, D.J. Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.* **14**, 481–487 (2004).
44. Ringach, D.L. & Malone, B.J. The operating point of the cortex: neurons as large deviation detectors. *J. Neurosci.* **27**, 7673–7683 (2007).
45. van Vreeswijk, C.A. Whence sparseness? *Adv. Neural Inf. Process. Syst.* **13**, 189–195 (2001).
46. Vinje, W.E. & Gallant, J.L. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* **287**, 1273–1276 (2000).
47. Wang, X.J., Liu, Y., Sanchez–Vives, M.V. & McCormick, D.A. Adaptation and temporal decorrelation by single neurons in the primary visual cortex. *J. Neurophysiol.* **89**, 3279–3293 (2003).
48. Rucci, M. & Casile, A. Fixational instability and natural image statistics: implications for early visual representations. *Network* **16**, 121–138 (2005).
49. Cleland, T.A. Early transformations in odor representation. *Trends Neurosci.* **33**, 130–139 (2010).
50. Wiechert, M.T., Judkewitz, B., Riecke, H. & Friedrich, R.W. Mechanisms of pattern decorrelation by recurrent neuronal circuits. *Nat. Neurosci.* **13**, 1003–1010 (2010).

ONLINE METHODS

Recording. Experiments were performed on the isolated retina of the larval tiger salamander, superfused with oxygenated Ringer's solution, following protocols approved by the Institutional Animal Care and Use Committee at Harvard University. Action potentials from RGCs were recorded extracellularly with a multi-electrode array[51]. Neurons were selected for analysis if they maintained steady firing rates throughout the 2-h experiments and their spike waveforms could be sorted unambiguously. 103 cells from 9 retinas satisfied this criterion. Classification into different cell types was achieved by agglomerative maximum-linkage clustering according to the Euclidean distances between temporal receptive fields[52]. Of the recorded cells, 6 were classified as ON cells, 18 as slow OFF cells, and 79 as fast OFF cells. Altogether, this yielded 5,356 response pairs, including comparisons across experiments.

Stimulation. Light was projected from a computer monitor onto the photoreceptor layer. The stimulus was a square grid with fields of size 54 μm^2 covering a total area of 3.4 mm^2. The monitor refresh interval was 15 ms. The mean light level at the retina (7×10^{-3} W m^{-2}) was in the regime of photopic vision[51].

The decorrelation theories assumed that light intensities in visual stimuli are drawn from a correlated multivariate normal distribution exhibiting the spatial power spectrum measured for natural scenes, which varies with spatial frequency $\mathbf{k}$ as $1/|\mathbf{k}|^2$. These assumptions neglect objects, edges and textures, but capture pairwise intensity correlations in the visual world. To address these theories directly, we designed spatiotemporal stimuli that approximated the pairwise correlations in natural stimuli and neglected all higher order structure.

We generated the spatial structure of the stimulus $S(\mathbf{x},t)$ by drawing spatial frequency coefficients $\tilde{S}_0(\mathbf{k},t)$ independently every 15 ms from a Gaussian distribution with variance proportional to $1/|\mathbf{k}|^2$. Temporal correlations were introduced by low-pass filtering the spatial frequency coefficients with an exponential of time constant $\tau = 1/|\mathbf{k}|v$, where v is a constant with units of velocity that determines the scaling between space and time. This constant was set to $v = 10°\ s^{-1}$, corresponding to a typical velocity that elicits neural and behavioral responses in salamanders in visual tasks[53]. The spatial frequency coefficients were given by

$$\tilde{S}(\mathbf{k},t) = Ae^{-t/\tau} \circ \tilde{S}_0(\mathbf{k},t) \tag{1}$$

where $\circ$ represents a temporal convolution and the constant A fixed the overall contrast (the ratio of s.d. of luminance to mean luminance) at 35%. An inverse spatial Fourier transform generated each image frame for display (**Fig. 1a**). The overall spatiotemporal power spectrum at spatial frequency $\mathbf{k}$ and temporal frequency ω is

$$\tilde{P}(\mathbf{k},\omega) \propto \frac{1}{|\mathbf{k}|^2} \cdot \frac{1-e^{-2|\mathbf{k}|v\Delta t}}{1+e^{-2|\mathbf{k}|v\Delta t}-2e^{-|\mathbf{k}|v\Delta t}\cos\omega\Delta t} \tag{2}$$

(**Supplementary Fig. 1**). This spectrum closely approximates that of natural movies[54]; in the limit of low spatial frequency it becomes $|\mathbf{k}|^{-3} f(\omega/|\mathbf{k}|)$ for a function f peaked at zero in the ratio $\omega/|\mathbf{k}|$. In this limit the power varies with temporal frequency approximately as ω^{-2}, as observed[54]. The stimulus has qualitative similarities with natural scenes: Large features are more prominent, and persist longer than small details. To compare the results from macaque and salamander, we used the same stimuli, except that the stimulus checker size was scaled in proportion to the mean ganglion cell receptive field radius.

Correlation. The correlation between two signals, x and y, was quantified by the second-order correlation function

$$C_{xy}(\tau) = \frac{\langle \Delta x(t) \cdot \Delta y(t+\tau) \rangle}{\sqrt{\langle \Delta x^2(t) \rangle \langle \Delta y^2(t) \rangle}} \tag{3}$$

where Δx and Δy represent deviations of x and y from their respective means and $\langle \cdot \rangle$ symbolizes an average over time. To reduce high-frequency noise, we first binned each signal into windows of width $\Delta t = 50$ ms for salamander neurons and 10 ms for primate neurons. This sets the time resolution on which the neural responses are analyzed. These values were chosen because they reflect the timescale on which ganglion cell firing varies: the typical duration of a stimulus-evoked burst of spikes (**Fig. 1c**) and the width of the peak in the temporal receptive field (**Fig. 1d**).

As the shared noise sources were small (**Supplementary Fig. 2**), we focused on stimulus-driven correlations, by presenting the same stimulus twice and computing correlations between the spike trains across the two repeats. The correlation measure (equation (3)) was computed the same way for pairs of stimulus values, trial-averaged firing rates, spike trains or the outputs of various functional models. In graphs of correlation versus spatial distance, we plotted the correlation at zero delay, $C_{xy}(0)$. For visualization, we binned the cell pairs by distance into groups of 100 and plotted the median for each group (**Figs. 1g,h** and **2**). Distances were quantified as the separation between the midpoint of the receptive fields.

Receptive fields. To map the receptive fields, we applied a random checkerboard stimulus[51] with a temporal sampling rate of 22 Hz and (54 μm)^2 black or white checkers. To reduce noise in the receptive field estimate, we fitted each neuron's spatiotemporal receptive field with a direct product of a spatial receptive field and a temporal kernel

$$F(\mathbf{x},t) \approx X(\mathbf{x})T(t) \tag{4}$$

using singular value decomposition. Each neuron's position was assigned as the midpoint of a two-dimensional Gaussian fit to its spatial receptive field $X(\mathbf{x})$ (**Fig. 1e**).

For modeling primate receptive fields, we parametrized $X(\mathbf{x})$ and $T(t)$ as

$$X(\mathbf{x}) = \frac{1}{2\pi\sigma_c^2} e^{-|\mathbf{x}-\mu_c|^2/2\sigma_c^2} - a\frac{1}{2\pi\sigma_s^2} e^{-|\mathbf{x}-\mu_s|^2/2\sigma_s^2} \tag{5}$$

$$T(t) = (t/\tau_1)^{n_1} e^{-n_1(t/\tau_1-1)} - b(t/\tau_2)^{n_2} e^{-n_2(t/\tau_2-1)} \tag{6}$$

with spatial parameters drawn from a previous study[55] for the parafovea (5°–10° eccentricity) and temporal parameters drawn from ref. 56.

Given a receptive field $F(\mathbf{x},t)$, we computed the linear prediction $r(t)$ for the neural response by convolution with the stimulus

$$r(t) = \iiint d^2\mathbf{x}\, d\tau\, F(\mathbf{x},\tau) S(\mathbf{x},t-\tau) \tag{7}$$

Nonlinearities. In the LNP model, the linear prediction $r(t)$ is transformed into a firing rate $\rho(t)$ by an instantaneous nonlinearity $N(\cdot)$,

$$\rho(t) = N(r(t)) \tag{8}$$

and then into a spike count n by drawing from a Poisson distribution with that rate

$$P(n|\rho) = \frac{e^{-\rho\Delta t}(\rho\Delta t)^n}{n!} \tag{9}$$

where Δt is the time bin. We parametrized the nonlinearity as a sigmoid using the logistic function

$$N(r) = K/\left(1+e^{-g(r-\theta)}\right) \tag{10}$$

with peak firing rate K, gain g and threshold θ.

If the linear input $r(t)$ follows a normal distribution, one can constrain the mean firing rate of the model neuron to a value μ by setting the peak rate to

$$K = \mu\sqrt{2\pi} \Big/ \int dr\, e^{-\frac{1}{2}r^2} \left(1+e^{-g(r-\theta)}\right)^{-1}$$

Noise. Large bursts of spikes from ganglion cells are more regular than expected from Poisson statistics[57,58], so the Poisson model generally overestimates the noise. For some computations (**Fig. 5b–d**) we used a noise distribution that was measured empirically. For a given mean spike count ρ at a given time during the trial, the measured spike count distributions $P(n|\rho)$ had a width that stayed constant with ρ after an initial Poisson-like growth (**Supplementary Fig. 3**).

These distributions were well-described as a Gaussian distribution on non-negative integer spike counts

$$P(n|\rho) \propto \exp\left[-\frac{(n-n_0(\rho))^2}{2\sigma^2}\right] \quad n=0,1,2,\ldots \tag{11}$$

where

$$n_0(\rho) = a\log\left(1+e^{\rho\Delta t/a}\right)$$

is the center of the Gaussian and σ is the width of the noise distribution. For each σ, the parameter a was set so the conditional mean of the model noise distribution closely approximated the desired mean ρ. The noise width σ was fit by numerically maximizing the log-likelihood,

$$\sum_{t,i}\log P\left(n(t,i)|\rho(t);\sigma\right)$$

where $n(t,i)$ is the measured spike count in bin t during stimulus repetition i.

Our models assume that noise affects the spiking of each neuron independently, whereas nearby ganglion cells share certain noise sources, especially at low light levels[59]. We found that noise correlations at photopic intensities were very small, <0.01 for 90% of pairwise comparisons (**Supplementary Fig. 2**). This justified the independent noise approximation for the great majority of cells, which simplifies the treatment of optimal coding. Another study reported that response models that account for noise correlations in ganglion cell spike trains can extract additional (~20%) visual information[60].

Decorrelation by nonlinearities and noise. The correlation between two LNP model neurons depends on both the nonlinearities and the noise (**Fig. 4**). Suppose that the inputs x and y to two neurons are both normally distributed with zero mean and unit variance and correlation coefficient c. After transformation by the nonlinear function $N(\bullet)$, the correlation coefficient becomes

$$C_{N(x)N(y)} = \frac{\langle N(x)N(y)\rangle - \mu^2}{\sigma^2} \tag{12}$$

where the nonlinear output has mean $\mu = \langle N(x)\rangle$ and variance $\sigma^2 = \langle N^2(x)\rangle - \mu^2$, and where $\langle\ \rangle$ is an expectation over the input distribution

$$P(x,y) = \frac{1}{2\pi\sqrt{1-c^2}}\exp\left(-\frac{1}{2}\frac{x^2-2xyc+y^2}{1-c^2}\right) \tag{13}$$

We computed these expectation values by numerical integration (**Fig. 4b–d**).

Response noise increases the variance without altering the covariance, lowering the correlation. For two conditionally independent signals x and y with (time dependent) trial averages of $\bar{x}$ and $\bar{y}$, the noise is $\delta x = x - \bar{x}$ and $\delta y = y - \bar{y}$. The correlation between the noisy signals x and y is then

$$\begin{aligned} C_{xy} &= \frac{\langle\bar{x}\bar{y}\rangle - \langle\bar{x}\rangle\langle\bar{y}\rangle}{\sqrt{\langle\bar{x}^2\rangle - \langle\bar{x}\rangle^2 + \langle\delta x^2\rangle}\sqrt{\langle\bar{y}^2\rangle - \langle\bar{y}\rangle^2 + \langle\delta y^2\rangle}} \\ &= C_{\bar{x}\bar{y}}\frac{1}{\sqrt{\left(1+1/SNR_x\right)\left(1+1/SNR_y\right)}} \end{aligned} \tag{14}$$

where $C_{\bar{x}\bar{y}}$ is the correlation of the trial-averaged responses and $SNR_x = \left(\langle\bar{x}^2\rangle - \langle\bar{x}\rangle^2\right)\Big/\langle\delta x^2\rangle$ is the ratio of signal variance to noise variance (**Fig. 4e**).

Information and efficiency for a single neuron. To analyze the role of nonlinearity in efficient coding, we computed the mutual information between the stimulus and the ganglion cell spike count in single windows of width Δt. This approximation neglects correlations between spike counts in different bins and spike timing within a bin. The mutual information between stimulus s and the spike count n is

$$I(n;s) = H(n) - H(n|s) \tag{15}$$

where the unconditional entropy $H(n)$ is

$$H(n) = -\sum_{n=0}^{\infty} p(n)\log p(n) \tag{16}$$

and the conditional entropy $H(n|s)$ is

$$H(n|s) = -\int ds\, p(s)\sum_{n=0}^{\infty} p(n|s)\log p(n|s) \tag{17}$$

We calculated the mutual information in two ways: directly from neural responses, and using a response model. For the former, the integrals over all possible stimuli were replaced by temporal averages over the stimulus presentation. For the latter, the integrals over the high-dimensional stimulus ensemble are intractable. However, the model responses depend on the stimulus only through the time-varying firing rate $\rho(t)$. Assuming again that input noise is negligible, this firing rate is a deterministic function of the stimulus. Thus, the conditional entropy given the stimulus equals the conditional entropy given the firing rate, $H(n|s) = H(n|\rho)$, and the mutual information is fully specified by the distribution of firing rates $p(\rho)$, regardless of how those rates arise

$$I(n;s) = H(n) - H(n|s) = H(n) - H(n|\rho) = I(n;\rho) \tag{18}$$

Thus we compute entropies (equations (16–17)) using $p(n) = \int d\rho\, p(n|\rho)\, p(\rho)$ and $p(n|\rho)$ instead of $p(n) = \int ds\, p(n|s)\, p(s)$ and $p(n|s)$.

For the LNP model, the firing rate distribution is produced by the sigmoid nonlinearity $N(r)$ (equation (10)) acting on the Gaussian distributed linear input r. These distributions are parametrized like the logistic function, by the peak rate K, gain g and threshold θ

$$p(\rho) = \frac{K\exp\left[-\frac{1}{2}\left(\frac{1}{g}\log\left(K/\rho-1\right)-\theta\right)^2\right]}{\sqrt{2\pi}g\rho^2\left(K/\rho-1\right)} \tag{19}$$

This family of distributions encompasses a wide range of unimodal and bimodal shapes, including binary rate distributions when $g = \infty$.

To fit each ganglion cell response distribution (**Fig. 5b,d**), we minimized the mean squared difference between the cumulative distribution of the parametric model (equation (19))

$$D(\rho) = \frac{1}{2}\mathrm{erfc}\left(\frac{\log\left(K/\rho-1\right)-g\theta}{\sqrt{2}g}\right) \tag{20}$$

and the cumulative distribution of the measured firing rates. The median parameters over the recorded salamander cells were $K = 48$ Hz, $g = 5.8$ and $\theta = 2.0$. For primate neurons, fits were derived from published spike rasters[58], with median parameters $K = 72$ Hz, $g = 2.8$ and $\theta = 0.95$.

We numerically calculated the mutual information for the response model (**Figs. 4f** and **5b,d**) by substituting the rate distribution (equation (19)) and noise model (equation (11)) into equations (15–17).

Given a neuron's mean firing rate μ, we determined the firing rate distribution that optimizes information transmission by numerically maximizing mutual information (equation (18)) over the parameters g and θ in equation (19), setting

$$K = \mu\sqrt{2\pi}\Big/\int dr\, e^{-\frac{1}{2}r^2}\left(1+e^{-g(r-\theta)}\right)^{-1} \tag{21}$$

to preserve the mean firing rate. Finally, we computed coding efficiency (**Fig. 5c**) by dividing the mutual information for the measured neural responses by this maximal information.

doi:10.1038/nn.3064

Information and redundancy for multiple correlated neurons. To compute the mutual information for a population of N LNP model neurons (**Fig. 4g**,**h**), we allowed the spike counts and firing rates in equations (15–17) to be N-dimensional vectors. We made several simplifications for tractability. First, all models had identical thresholds θ, gains g and peak firing rates K. Next, we assumed the input to the nonlinearities was a multivariate Gaussian with uniform correlation matrix $\mathbf{\Sigma} = c + (1-c)\mathbf{I}$. Finally, we restricted the nonlinearity to have optimal (infinite) gain; each neuron was either silent or fired at a maximal rate K in each time bin.

Given these simplifying assumptions, we can calculate the mutual information for the population. The unconditional probabilities of the vector of binary firing rates are

$$p(\boldsymbol{\rho}) = \int_{O(\boldsymbol{\rho})} d^N\mathbf{r}\, \frac{1}{(2\pi)^{N/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{r}-\theta)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{r}-\theta)\right) \tag{22}$$

with integration over orthants

$$O(\boldsymbol{\rho}) = \bigcap\nolimits_i \left\{\mathrm{sgn}\left(\rho_i - \tfrac{K}{2}\right) r_i > 0\right\}$$

We computed these integrals numerically, exploiting the model's permutation symmetry to reduce the number of integrals.

The model neurons are silent and have zero noise entropy if $\rho_i = 0$, and emit spikes with probability

$$p(n_i|\rho_i = K) = \begin{cases} q & n_i = 0 \\ 1-q & n_i > 0 \end{cases} \tag{23}$$

and noise entropy

$$h(q) = -q\log q - (1-q)\log(1-q) \tag{24}$$

otherwise, with $q = \exp(-K\Delta t)$ for Poisson noise. The conditional entropy (equation (17)) is the average noise entropy across all firing rate patterns

$$H(\mathbf{n}|\boldsymbol{\rho}) = \sum_{\boldsymbol{\rho}} p(\boldsymbol{\rho}) h(q) \sum_i \frac{\rho_i}{K} \tag{25}$$

The unconditional entropy (equation (16)) is computed from the marginal spike count probability over spikes and silences,

$$p(\mathbf{n}) = \sum\nolimits_{\boldsymbol{\rho}} p(\boldsymbol{\rho}) \prod\nolimits_i p(n_i|\rho_i)$$

Redundancy (**Fig. 4g**) measures the difference between the total information conveyed by each neuron considered independently and the information all neurons convey together, compared to the information that could be conveyed if all neurons were independent,

$$R = \left(\sum\nolimits_i I(\rho_i;s) - I(\boldsymbol{\rho};s)\right) \Big/ \sum\nolimits_i I(\rho_i;s)$$

51. Meister, M., Pine, J. & Baylor, D.A. Multi–neuronal signals from the retina: acquisition and analysis. *J. Neurosci. Methods* **51**, 95–106 (1994).
52. Segev, R., Puchalla, J. & Berry, M.J. Functional organization of ganglion cells in the salamander retina. *J. Neurophysiol.* **95**, 2277–2292 (2006).
53. Himstedt, W. Prey selection in salamanders. in *Analysis of Visual Behavior* (eds. Ingale, D.J., Goodale, M.A. & Mansfield, R.J.W.) 47–66 (MIT Press, Cambridge, Massachusetts, 1982).
54. Dong, D.W. & Atick, J.J. Statistics of natural time-varying images. *Network* **6**, 345–358 (1995).
55. Croner, L.J. & Kaplan, E. Receptive fields of P and M ganglion cells across the primate retina. *Vision Res.* **35**, 7–24 (1995).
56. Chichilnisky, E.J. & Kalmar, R.S. Functional asymmetries in ON and OFF ganglion cells of primate retina. *J. Neurosci.* **22**, 2737–2747 (2002).
57. Berry, M.J. & Meister, M. Refractoriness and neural precision. *J. Neurosci.* **18**, 2200–2211 (1998).
58. Uzzell, V.J. & Chichilnisky, E.J. Precision of spike trains in primate retinal ganglion cells. *J. Neurophysiol.* **92**, 780–789 (2004).
59. Schneidman, E., Bialek, W. & Berry, M.J. Synergy, redundancy and independence in population codes. *J. Neurosci.* **23**, 11539–11553 (2003).
60. Pillow, J.W. *et al.* Spatio–temporal correlations and visual signaling in a complete neuronal population. *Nature* **454**, 995–999 (2008).

Learning unbelievable marginal probabilities

Xaq Pitkow
Department of Brain and Cognitive Science
University of Rochester
Rochester, NY 14607
xaq@post.harvard.edu

Yashar Ahmadian
Center for Theoretical Neuroscience
Columbia University
New York, NY 10032
ya2005@columbia.edu

Ken D. Miller
Center for Theoretical Neuroscience
Columbia University
New York, NY 10032
ken@neurotheory.columbia.edu

Abstract

Loopy belief propagation performs approximate inference on graphical models with loops. One might hope to compensate for the approximation by adjusting model parameters. Learning algorithms for this purpose have been explored previously, and the claim has been made that every set of locally consistent marginals can arise from belief propagation run on a graphical model. On the contrary, here we show that many probability distributions have marginals that cannot be reached by belief propagation using any set of model parameters or any learning algorithm. We call such marginals 'unbelievable.' This problem occurs whenever the Hessian of the Bethe free energy is not positive-definite at the target marginals. All learning algorithms for belief propagation necessarily fail in these cases, producing beliefs or sets of beliefs that may even be worse than the pre-learning approximation. We then show that averaging inaccurate beliefs, each obtained from belief propagation using model parameters perturbed about some learned mean values, can achieve the unbelievable marginals.

1 Introduction

Calculating marginal probabilities for a graphical model generally requires summing over exponentially many states, and is NP-hard in general [1]. A variety of approximate methods have been used to circumvent this problem. One popular technique is belief propagation (BP), in particular the sum-product rule, which is a message-passing algorithm for performing inference on a graphical model [2]. Though exact and efficient on trees, it is merely an approximation when applied to graphical models with loops.

A natural question is whether one can compensate for the shortcomings of the approximation by setting the model parameters appropriately. In this paper, we prove that some sets of marginals simply cannot be achieved by belief propagation. For these cases we provide a new algorithm that can achieve much better results by using an ensemble of parameters rather than a single instance.

We are given a set of variables $\boldsymbol{x}$ with a given probability distribution $P(\boldsymbol{x})$ of some data. We would like to construct a model that reproduces certain of its marginal probabilities, in particular those over individual variables $p_i(x_i) = \sum_{\boldsymbol{x}\setminus x_i} P(\boldsymbol{x})$ for nodes $i \in V$, and those over some relevant clusters of variables, $p_\alpha(\boldsymbol{x}_\alpha) = \sum_{\boldsymbol{x}\setminus\boldsymbol{x}_\alpha} P(\boldsymbol{x})$ for $\alpha = \{i_1, \ldots, i_{d_\alpha}\}$. We will write the collection of all these marginals as a vector $\boldsymbol{p}$.

We assume a model distribution $Q_0(\boldsymbol{x})$ in the exponential family taking the form

$$Q_0(\boldsymbol{x}) = e^{-E(\boldsymbol{x})}/Z \tag{1}$$

with normalization constant $Z = \sum_{\boldsymbol{x}} e^{-E(\boldsymbol{x})}$ and energy function

$$E(\boldsymbol{x}) = -\sum_{\alpha} \boldsymbol{\theta}_\alpha \cdot \boldsymbol{\phi}_\alpha(\boldsymbol{x}_\alpha) \tag{2}$$

Here, α indexes sets of interacting variables (factors in the factor graph [3]), and $\boldsymbol{x}_\alpha$ is a subset of variables whose interaction is characterized by a vector of sufficient statistics $\boldsymbol{\phi}_\alpha(\boldsymbol{x}_\alpha)$ and corresponding natural parameters $\boldsymbol{\theta}_\alpha$. We assume without loss of generality that each $\boldsymbol{\phi}_\alpha(\boldsymbol{x}_\alpha)$ is irreducible, meaning that it cannot be written as a sum of any linearly independent functions that themselves do not depend on any x_i for $i \in \alpha$. We collect all these sufficient statistics and natural parameters in the vectors $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$.

Normally when learning a graphical model, one would fit its parameters so the marginal probabilities match the target. Here, however, we will not use *exact* inference to compute the marginals. Instead we will use *approximate* inference via loopy belief propagation to match the target.

2 Learning in Belief Propagation

2.1 Belief propagation

The sum-product algorithm for belief propagation on a graphical model with energy function (2) uses the following equations [4]:

$$m_{i\to\alpha}(x_i) \propto \prod_{\beta\in N_i\setminus\alpha} m_{\beta\to i}(x_i) \qquad m_{\alpha\to i}(x_i) \propto \sum_{\boldsymbol{x}_\alpha\setminus x_i} e^{\boldsymbol{\theta}_\alpha\cdot\boldsymbol{\phi}_\alpha(\boldsymbol{x}_\alpha)} \prod_{j\in N_\alpha\setminus i} m_{j\to\alpha}(x_j) \tag{3}$$

where N_i and N_α are the neighbors of node i or factor α in the factor graph. Once these messages converge, the single-node and factor beliefs are given by

$$b_i(x_i) \propto \prod_{\alpha\in N_i} m_{\alpha\to i}(x_i) \qquad b_\alpha(\boldsymbol{x}_\alpha) \propto e^{\boldsymbol{\theta}_\alpha\cdot\boldsymbol{\phi}_\alpha(\boldsymbol{x}_\alpha)} \prod_{i\in N_\alpha} m_{i\to\alpha}(x_i) \tag{4}$$

where the beliefs must each be normalized to one. For tree graphs, these beliefs exactly equal the marginals of the graphical model $Q_0(\boldsymbol{x})$. For loopy graphs, the beliefs at fixed points are often good approximations of the marginals. While they are guaranteed to be locally consistent, $\sum_{\boldsymbol{x}_\alpha\setminus x_i} b_\alpha(\boldsymbol{x}_\alpha) = b_i(x_i)$, they are not necessarily globally consistent: There may not exist a single joint distribution $B(\boldsymbol{x})$ of which the beliefs are the marginals [5]. This is why the resultant beliefs are called *pseudomarginals*, rather than simply marginals. We use a vector $\boldsymbol{b}$ to refer to the set of both node and factor beliefs produced by belief propagation.

2.2 Bethe free energy

Despite its limitations, BP is found empirically to work well in many circumstances. Some theoretical justification for loopy belief propagation emerged with proofs that its stable fixed points are local minima of the Bethe free energy [6, 7]. Free energies are important quantities in machine learning because the Kullback-Leibler divergence between the data and model distributions can be expressed in terms of free energies, so models can be optimized by minimizing free energies appropriately.

Given an energy function $E(\boldsymbol{x})$ from (2), the Gibbs free energy of a distribution $Q(\boldsymbol{x})$ is

$$F[Q] = U[Q] - S[Q] \tag{5}$$

where U is the average energy of the distribution

$$U[Q] = \sum_{\boldsymbol{x}} E(\boldsymbol{x})Q(\boldsymbol{x}) = -\sum_{\alpha} \boldsymbol{\theta}_\alpha \cdot \sum_{\boldsymbol{x}_\alpha} \boldsymbol{\phi}_\alpha(\boldsymbol{x}_\alpha) q_\alpha(\boldsymbol{x}_\alpha) \tag{6}$$

which depends on the marginals $q_\alpha(\boldsymbol{x}_\alpha)$ of $Q(\boldsymbol{x})$, and S is the entropy

$$S[Q] = -\sum_{\boldsymbol{x}} Q(\boldsymbol{x}) \log Q(\boldsymbol{x}) \tag{7}$$

Minimizing the Gibbs free energy $F[Q]$ recovers the distribution $Q_0(\boldsymbol{x})$ for the graphical model (1).

The Bethe free energy F^β is an approximation to the Gibbs free energy,

$$F^\beta[Q] = U[Q] - S^\beta[Q] \tag{8}$$

in which the average energy U is exact, but the true entropy S is replaced by an approximation, the Bethe entropy S^β, which is a sum over the factor and node entropies [6]:

$$S^\beta[Q] = \sum_\alpha S_\alpha[q_\alpha] + \sum_i (1 - d_i) S_i[q_i] \tag{9}$$

$$S_\alpha[q_\alpha] = -\sum_{\boldsymbol{x}_\alpha} q_\alpha(\boldsymbol{x}_\alpha) \log q_\alpha(\boldsymbol{x}_\alpha) \qquad S_i[q_i] = -\sum_{x_i} q_i(x_i) \log q_i(x_i) \tag{10}$$

The coefficients $d_i = |N_i|$ are the number of factors neighboring node i, and compensate for the overcounting of single-node marginals due to overlapping factor marginals. For tree-structured graphical models, which factorize as $Q(\boldsymbol{x}) = \prod_\alpha q_\alpha(\boldsymbol{x}_\alpha) \prod_i q_i(x_i)^{1-d_i}$, the Bethe entropy is exact, and hence so is the Bethe free energy. On loopy graphs, the Bethe entropy S^β isn't really even an entropy (*e.g.* it may be negative) because it neglects all statistical dependencies other than those present in the factor marginals. Nonetheless, the Bethe free energy is often close enough to the Gibbs free energy that its minima approximate the true marginals [8]. Since stable fixed points of BP are minima of the Bethe free energy [6, 7], this helped explain why belief propagation is often so successful.

To emphasize that the Bethe free energy directly depends only on the marginals and not the joint distribution, we will write $F^\beta[\boldsymbol{q}]$ where $\boldsymbol{q}$ is a vector of pseudomarginals $q_\alpha(\boldsymbol{x}_\alpha)$ for all α and all $\boldsymbol{x}_\alpha$. Pseudomarginal space is the convex set [5] of all $\boldsymbol{q}$ that satisfy the positivity and local consistency constraints,

$$0 \le q_\alpha(\boldsymbol{x}_\alpha) \le 1 \qquad \sum_{\boldsymbol{x}_\alpha \setminus x_i} q_\alpha(\boldsymbol{x}_\alpha) = q_i(x_i) \qquad \sum_{x_i} q_i(x_i) = 1 \tag{11}$$

2.3 Pseudo-moment matching

We now wish to correct for the deficiencies of belief propagation by identifying the parameters $\boldsymbol{\theta}$ so that BP produces beliefs $\boldsymbol{b}$ matching the true marginals $\boldsymbol{p}$ of the target distribution $P(\boldsymbol{x})$. Since the fixed points of BP are stationary points of F^β [6], one may simply try to find parameters $\boldsymbol{\theta}$ that produce a stationary point in pseudomarginal space at $\boldsymbol{p}$, which is a necessary condition for BP to reach a fixed point there. Simply evaluate the gradient at $\boldsymbol{p}$, set it to zero, and solve for $\boldsymbol{\theta}$.

Note that in principle this gradient could be used to directly minimize the Bethe free energy, but $F^\beta[\boldsymbol{q}]$ is a complicated function of $\boldsymbol{q}$ that usually cannot be minimized analytically [8]. In contrast, here we are using it to solve for the parameters needed to move beliefs to a target location. This is much easier, since the Bethe free energy is linear in $\boldsymbol{\theta}$. This approach to learning parameters has been described as 'pseudo-moment matching' [9, 10, 11].

The $L_{\boldsymbol{q}}$-element vector $\boldsymbol{q}$ is an overcomplete representation of the pseudomarginals because it must obey the local consistency constraints (11). It is convenient to express the pseudomarginals in terms of a minimal set of parameters $\boldsymbol{\eta}$ with the smaller dimensionality $L_{\boldsymbol{\theta}}$ as $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, using an affine transform

$$\boldsymbol{q} = W\boldsymbol{\eta} + \boldsymbol{k} \tag{12}$$

where W is an $L_{\boldsymbol{q}} \times L_{\boldsymbol{\theta}}$ rectangular matrix. One example is the expectation parameters $\boldsymbol{\eta}_\alpha = \sum_{\boldsymbol{x}_\alpha} q_\alpha(\boldsymbol{x}_\alpha) \boldsymbol{\phi}_\alpha(\boldsymbol{x}_\alpha)$ [5], giving the energy simply as $U = -\boldsymbol{\theta} \cdot \boldsymbol{\eta}$. The gradient with respect to those minimal parameters is

$$\frac{\partial F^\beta}{\partial \boldsymbol{\eta}} = \frac{\partial U}{\partial \boldsymbol{\eta}} - \frac{\partial S^\beta}{\partial \boldsymbol{q}} \frac{\partial \boldsymbol{q}}{\partial \boldsymbol{\eta}} = -\boldsymbol{\theta} - \frac{\partial S^\beta}{\partial \boldsymbol{q}} W \tag{13}$$

The Bethe entropy gradient is simplest in the overcomplete representation $\boldsymbol{q}$,

$$\frac{\partial S^\beta}{\partial q_\alpha(\boldsymbol{x}_\alpha)} = -1 - \log q_\alpha(\boldsymbol{x}_\alpha) \qquad \frac{\partial S^\beta}{\partial q_i(x_i)} = (-1 - \log q_i(x_i))(1 - d_i) \tag{14}$$

Setting the gradient (13) to zero, we have a simple linear equation for the parameters $\boldsymbol{\theta}$ that tilt the Bethe free energy surface (Figure 1A) enough to place a stationary point at the desired marginals $\boldsymbol{p}$:

$$\boldsymbol{\theta} = -\left.\frac{\partial S^\beta}{\partial \boldsymbol{q}}\right|_{\boldsymbol{p}} W \tag{15}$$

2.4 Unbelievable marginals

It is well known that BP may converge on fixed points that cannot be realized as marginals of any joint distribution. In this section we show that the converse is also true: There are some distributions whose marginals cannot be realized as beliefs for any set of couplings. In these cases, existing methods for learning often yield poor results, sometimes even worse than performing no learning at all. This is surprising in view of claims to the contrary: [9, 5] state that belief propagation run after pseudo-moment matching can always reach a fixed point that reproduces the target marginals. While BP does technically have such fixed points, they are not always stable and thus may not be reachable by running belief propagation.

Definition 1. *A set of marginals are 'unbelievable' if belief propagation cannot converge to them for any set of parameters.*

For belief propagation to converge to the target — namely, the marginals $\boldsymbol{p}$ — a zero gradient is not sufficient: The Bethe free energy must also be a local minimum [7].[1] This requires a positive-definite Hessian of F^β (the 'Bethe Hessian' H) in the subspace of pseudomarginals that satisfies the local consistency constraints. Since the energy U is linear in the pseudomarginals, the Hessian is given by the second derivative of the Bethe entropy,

$$H = \frac{\partial^2 F^\beta}{\partial \boldsymbol{\eta}^2} = -W^\top \frac{\partial^2 S^\beta}{\partial \boldsymbol{q}^2} W \tag{16}$$

where projection by W constrains the derivatives to the subspace spanned by the minimal parameters $\boldsymbol{\eta}$. If this Hessian is positive definite when evaluated at $\boldsymbol{p}$ then the parameters $\boldsymbol{\theta}$ given by (15) give F^β a minimum at the target $\boldsymbol{p}$. If not, then the target cannot be a stable fixed point of loopy belief propagation. In Section 3, we calculate the Bethe Hessian explicitly for a binary model with pairwise interactions.

Theorem 1. *Unbelievable marginal probabilities exist.*

Proof. Proof by example. The simplest unbelievable example is a binary graphical model with pairwise interactions between four nodes, $\boldsymbol{x} \in \{-1, +1\}^4$, and the energy $E(\boldsymbol{x}) = -J \sum_{(ij)} x_i x_j$. By symmetry and (1), marginals of this target $P(\boldsymbol{x})$ are the same for all nodes and pairs: $p_i(x_i) = \frac{1}{2}$ and $p_{ij}(x_i = x_j) = \rho = (2 + 4/(1 + e^{2J} - e^{4J} + e^{6J}))^{-1}$. Substituting these marginals into the appropriate Bethe Hessian (22) gives a matrix that has a negative eigenvalue for all $\rho > \frac{3}{8}$, or $J > 0.316$. The associated eigenvector $\boldsymbol{u}$ has the same symmetry as the marginals, with single-node components $u_i = \frac{1}{2}(-2 + 7\rho - 8\rho^2 + \sqrt{10 - 28\rho + 81\rho^2 - 112\rho^3 + 64\rho^4})$ and pairwise components $u_{ij} = 1$. Thus the Bethe free energy does not have a minimum at the marginals of these $P(\boldsymbol{x})$. Stable fixed points of BP occur only at local minima of the Bethe free energy [7], and so BP cannot reproduce the marginals $\boldsymbol{p}$ for any parameters. Hence these marginals are unbelievable. □

Not only do unbelievable marginals exist, but they are actually quite common, as we will see in Section 3. Graphical models with multinomial or gaussian variables and at least two loops always have some pseudomarginals for which the Hessian is not positive definite [12]. On the other hand, all marginals with sufficiently small correlations are believable because they are guaranteed to have a positive-definite Bethe Hessian [12]. Stronger conditions have not yet been described.

2.5 Bethe wake-sleep algorithm

When pseudo-moment matching fails to reproduce unbelievable marginals, an alternative is to use a gradient descent procedure for learning, analagous to the wake-sleep algorithm used to train Boltzmann machines [13]. The original rule can be derived as gradient descent of the Kullback-Leibler

[1] Even this is not sufficient, but it is necessary.

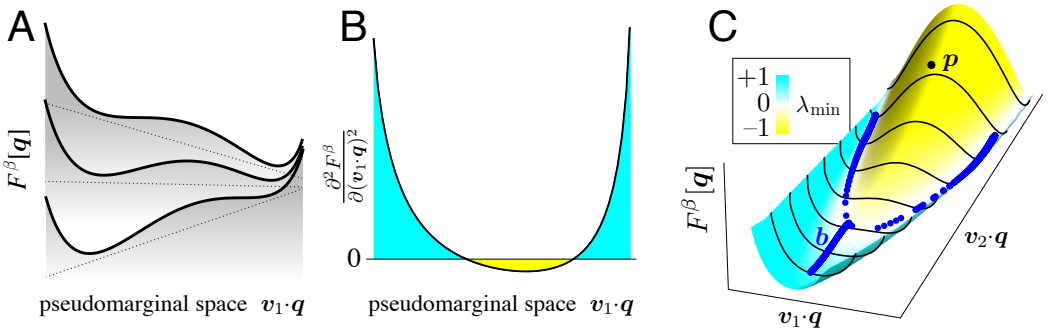

Figure 1: Landscape of Bethe free energy for the binary graphical model with pairwise interactions. (**A**) A slice through the Bethe free energy (solid lines) along one axis $\boldsymbol{v}_1$ of pseudomarginal space, for three different values of parameters $\boldsymbol{\theta}$. The energy U is linear in the pseudomarginals (dotted lines), so varying the parameters only changes the tilt of the free energy. This can add or remove local minima. (**B**) The second derivatives of the free energies in (A) are all identical. Where the second derivative is positive, a local minimum can exist (cyan); where it is negative (yellow), no parameters can produce a local minimum. (**C**) A two-dimensional slice of the Bethe free energy, colored according to the minimum eigenvalue $\lambda_{\min}$ of the Bethe Hessian. During a run of Bethe wake-sleep learning, the beliefs (blue dots) proceed along $\boldsymbol{v}_2$ toward the target marginals $\boldsymbol{p}$. Stable fixed points of BP can exist only in the believable region (cyan), but the target $\boldsymbol{p}$ resides in an unbelievable region (yellow). As learning equilibrates, the fixed points jump between believable regions on either side of the unbelievable zone.

divergence between the target $P(\boldsymbol{x})$ and the graphical model $Q(\boldsymbol{x})$ (1),

$$D_{\mathrm{KL}}[P||Q] = \sum_{\boldsymbol{s}} P(\boldsymbol{x}) \log \frac{P(\boldsymbol{x})}{Q(\boldsymbol{x})} = F[P] - F[Q] \tag{17}$$

where F is the Gibbs free energy (5) using the energy function (2). Here we use a new cost function, the 'Bethe divergence' $D_\beta[\boldsymbol{p}||\boldsymbol{b}]$, by replacing these free energies by Bethe free energies [14] evaluated at the true marginals $\boldsymbol{p}$ and at the beliefs $\boldsymbol{b}$ obtained from BP fixed points,

$$D_\beta[\boldsymbol{p}||\boldsymbol{b}] = F^\beta[\boldsymbol{p}] - F^\beta[\boldsymbol{b}] \tag{18}$$

We use gradient descent to optimize this cost, with gradient

$$\frac{dD_\beta}{d\boldsymbol{\theta}} = \frac{\partial D_\beta}{\partial \boldsymbol{\theta}} + \frac{\partial D_\beta}{\partial \boldsymbol{b}} \frac{\partial \boldsymbol{b}}{\partial \boldsymbol{\theta}} \tag{19}$$

The data's free energy does not depend on the beliefs, so $\partial F^\beta[\boldsymbol{p}]/\partial \boldsymbol{b} = 0$, and fixed points of belief propagation are stationary points of the Bethe free energy, so $\partial F^\beta[\boldsymbol{b}]/\partial \boldsymbol{b} = 0$. Consequently $\partial D_\beta/\partial \boldsymbol{b} = 0$. Furthermore, the entropy terms of the free energies do not depend explicitly on $\boldsymbol{\theta}$, so

$$\frac{dD_\beta}{d\boldsymbol{\theta}} = \frac{\partial U(\boldsymbol{p})}{\partial \boldsymbol{\theta}} - \frac{\partial U(\boldsymbol{b})}{\partial \boldsymbol{\theta}} = -\boldsymbol{\eta}(\boldsymbol{p}) + \boldsymbol{\eta}(\boldsymbol{b}) \tag{20}$$

where $\boldsymbol{\eta}(\boldsymbol{q}) = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x})$ are the expectations of the sufficient statistics $\boldsymbol{\phi}(\boldsymbol{x})$ under the pseudomarginals $\boldsymbol{q}$. This gradient forms the basis of a simple learning algorithm. At each step in learning, belief propagation is run, obtaining beliefs $\boldsymbol{b}$ for the current parameters $\boldsymbol{\theta}$. The parameters are then changed in the opposite direction of the gradient,

$$\Delta\boldsymbol{\theta} = -\epsilon \frac{dD_\beta}{d\boldsymbol{\theta}} = \epsilon(\boldsymbol{\eta}(\boldsymbol{p}) - \boldsymbol{\eta}(\boldsymbol{b})) \tag{21}$$

where ϵ is a learning rate. This generally increases the Bethe free energy for the beliefs while decreasing that of the data, hopefully allowing BP to draw closer to the data marginals. We call this learning rule the Bethe wake-sleep algorithm.

Within this algorithm, there is still the freedom of how to choose initial messages for BP at each learning iteration. The result depends on these initial conditions because BP can have several stable

fixed points. One might re-initialize the messages to a fixed starting point for each run of BP, choose random initial messages for each run, or restart the messages where they stopped on the previous learning step. In our experiments we use the first approach, initializing to constant messages at the beginning of each BP run.

The Bethe wake-sleep learning rule sometimes places a minimum of F^β at the true data distribution, such that belief propagation can give the true marginals as one of its (possibly multiple) fixed points. However, for the reasons provided above, this cannot occur where the Bethe Hessian is not positive definite.

2.6 Ensemble belief propagation

When the Bethe wake-sleep algorithm attempts to learn unbelievable marginals, the parameters and beliefs do not reach a fixed point but instead continue to vary over time (Figure 2A,B). Still, if learning reaches equilibrium, then the temporal average of beliefs is equal to the unbelievable marginals.

Theorem 2. *If the Bethe wake-sleep algorithm reaches equilibrium, then unbelievable marginals are matched by the belief propagation fixed points averaged over the equilibrium ensemble of parameters.*

Proof. At equilibrium, the time average of the parameter changes is zero by definition, $\langle \Delta\boldsymbol{\theta} \rangle_t = 0$. Substitution of the Bethe wake-sleep equation, $\Delta\boldsymbol{\theta} = \epsilon(\boldsymbol{\eta}(\boldsymbol{p}) - \boldsymbol{\eta}(\boldsymbol{b}(t)))$ (20), directly implies that $\langle \boldsymbol{\eta}(\boldsymbol{b}(t)) \rangle_t = \boldsymbol{\eta}(\boldsymbol{p})$. The deterministic mapping (12) from the minimal representation to the pseudomarginals gives $\langle \boldsymbol{b}(t) \rangle_t = \boldsymbol{p}$. □

After learning has equilibrated, fixed points of belief propagation occur with just the right frequency so that they can be averaged together to reproduce the target distribution exactly (Figure 2C). Note that none of the individual fixed points may be close to the true marginals. We call this inference algorithm *ensemble* belief propagation (eBP).

Ensemble BP produces perfect marginals by exploiting a constant, small amplitude learning, and thus assumes that the correct marginals are perpetually available. Yet it also works well when learning is turned off, if parameters are drawn randomly from a gaussian distribution with mean and covariance matched to the equilibrium distribution, $\boldsymbol{\theta} \sim \mathcal{N}(\bar{\boldsymbol{\theta}}, \Sigma_{\boldsymbol{\theta}})$. In the simulations below (Figures 2C–D, 3B–C), $\Sigma_{\boldsymbol{\theta}}$ was always low-rank, and only one or two principle components were needed for good performance. The gaussian ensemble is not quite as accurate as continued learning (Figure 3B,C), but the performance is still markedly better than any of the available fixed points.

If the target is not within a convex hull of believable pseudomarginals, then learning cannot reach equilibrium: Eventually BP gets as close as it can but there remains a consistent difference $\boldsymbol{\eta}(\boldsymbol{p}) - \boldsymbol{\eta}(\boldsymbol{b})$, so $\boldsymbol{\theta}$ must increase without bound. Though possible in principle, we did not observe this effect in any of our experiments. There may also be no equilibrium if belief propagation at each learning iteration fails to converge.

3 Experiments

The experiments in this section concentrate on the Ising model: N binary variables, $\boldsymbol{s} \in \{-1, +1\}^N$, with factors comprising individual variables x_i and pairs x_i, x_j. The energy function is $E(\boldsymbol{x}) = -\sum_i h_i x_i - \sum_{(ij)} J_{ij} x_i x_j$. Then the sufficient statistics are the various first and second moments, x_i and $x_i x_j$, and the natural parameters are h_i, J_{ij}. We use this model both for the target distributions and the model.

We parameterize pseudomarginals as $\{q_i^+, q_{ij}^{++}\}$ where $q_i^+ = q_i(x_i = +1)$ and $q_{ij}^{++} = q_{ij}(x_i = x_j = +1)$ [8]. The remaining probabilities are linear functions of these values. Positivity constraints and local consistency constraints then appear as $0 \le q_i^+ \le 1$ and $\max(0, q_i^+ + q_j^+ - 1) \le q_{ij}^{++} \le \min(q_i^+, q_j^+)$. If all the interactions are finite, then the inequality constraints are not active [15]. In

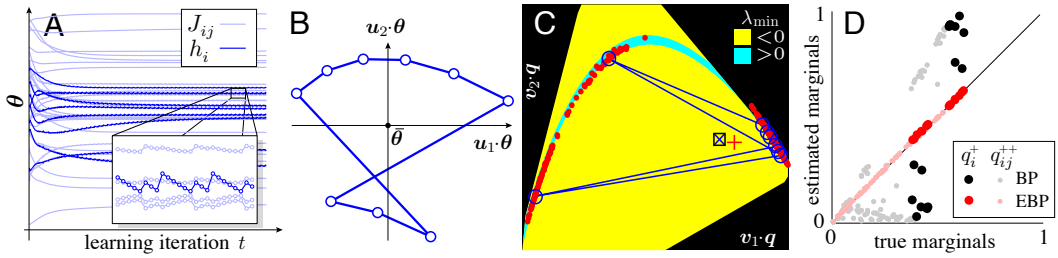

Figure 2: Averaging over variable couplings can produce marginals otherwise unreachable by belief propagation. (**A**) As learning proceeds, the Bethe wake-sleep algorithm causes parameters $\boldsymbol{\theta}$ to converge on a discrete limit cycle when attempting to learn unbelievable marginals. (**B**) The same limit cycle, projected onto their first two principal components $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ of $\boldsymbol{\theta}$ during the cycle. (**C**) The corresponding beliefs $\boldsymbol{b}$ during the limit cycle (blue circles), projected onto the first two principal components $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ of the trajectory through pseudomarginal space. Believable regions of pseudomarginal space are colored with cyan and the unbelievable regions with yellow, and inconsistent pseudomarginals are black. Over the limit cycle, the average beliefs $\bar{\boldsymbol{b}}$ (blue $\times$) are precisely equal to the target marginals $\boldsymbol{p}$ (black $\square$). The average $\bar{\boldsymbol{b}}$ (red $+$) over many fixed points of BP (red dots) generated from randomly perturbed parameters $\bar{\boldsymbol{\theta}} + \delta\boldsymbol{\theta}$ still produces a better approximation of the target marginals than any of the individual believable fixed points. (**D**) Even the best amongst several BP fixed points cannot match unbelievable marginals (black and grey). Ensemble BP leads to much improved performance (red and pink).

this parameterization, the elements of the Bethe Hessian (16) are

$$-\frac{\partial^2 S^\beta}{\partial q_i^+ \partial q_j^+} = \delta_{i,j}(1-d_i)\left[(q_i^+)^{-1} + (1-q_i^+)^{-1}\right] + \delta_{j\in N_i}\left[(1-q_i^+-q_j^++q_{ij}^{++})^{-1}\right] \tag{22a}$$
$$+ \delta_{i,j}\sum_{k\in N_i}\left[(q_i^+-q_{ik}^{++})^{-1} + (1-q_i^+-q_k^++q_{ik}^{++})^{-1}\right]$$

$$-\frac{\partial^2 S^\beta}{\partial q_i^+ \partial q_{jk}^{++}} = -\delta_{i,j}\left[(q_i^+-q_{ik}^{++})^{-1} + (1-q_i^+-q_k^++q_{ik}^{++})^{-1}\right] \tag{22b}$$
$$-\delta_{i,k}\left[(q_i^+-q_{ij}^{++})^{-1} + (1-q_i^+-q_j^++q_{ij}^{++})^{-1}\right]$$

$$-\frac{\partial^2 S^\beta}{\partial q_{ij}^{++} \partial q_{k\ell}^{++}} = \delta_{ij,k\ell}\left[(q_{ij}^{++})^{-1} + (q_i^+-q_{ij}^{++})^{-1} + (q_j^+-q_{ij}^{++})^{-1} + (1-q_i^+-q_j^++q_{ij}^{++})^{-1}\right] \tag{22c}$$

Figure 3A shows the fraction of marginals that are unbelievable for 8-node, fully-connected Ising models with random coupling parameters $h_i \sim \mathcal{N}(0, \frac{1}{3})$ and $J_{ij} \sim \mathcal{N}(0, \sigma_J)$. For $\sigma_J \gtrsim \frac{1}{4}$, most marginals cannot be reproduced by belief propagation with any parameters, because the Bethe Hessian (22) has a negative eigenvalue.

We generated 500 Ising model targets using $\sigma_J = \frac{1}{3}$, selected the unbelievable ones, and evaluated the performance of BP and ensemble BP for various methods of choosing parameters $\boldsymbol{\theta}$. Each run of BP used exponential temporal message damping of 5 time steps [16], $\boldsymbol{m}^{t+1} = a\boldsymbol{m}^t + (1-a)\boldsymbol{m}_{\text{undamped}}$ with $a = e^{-1/5}$. Fixed points were declared when messages changed by less than 10^{-9} on a single time step. We evaluated BP performance for the actual parameters that generated the target (1), pseudomoment matching (15), and at best-matching beliefs obtained at any time during Bethe wake-sleep learning. We also measured eBP performance for two parameter ensembles: the last 100 iterations of Bethe wake-sleep learning, and parameters sampled from a gaussian $\mathcal{N}(\bar{\boldsymbol{\theta}}, \Sigma_{\boldsymbol{\theta}})$ with the same mean and covariance as that ensemble.

Belief propagation gave a poor approximation of the target marginals, as expected for a model with many strong loops. Even with learning, BP could never get the correct marginals, which was guaranteed by selection of unbelievable targets. Yet ensemble belief propagation gave excellent results. Using the exact parameter ensemble gave orders of magnitude improvement, limited by the

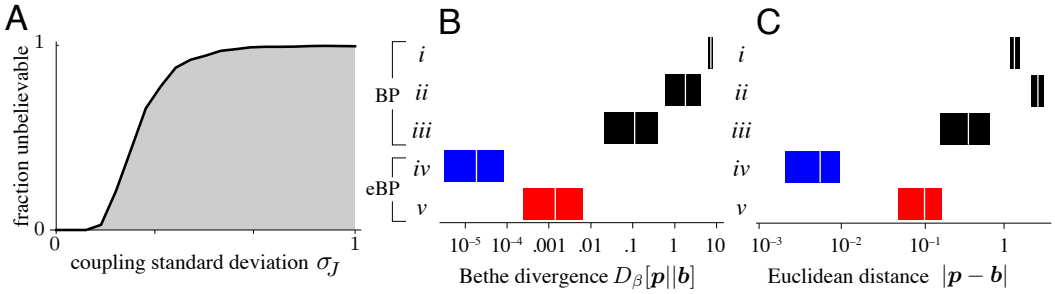

Figure 3: Performance in learning unbelievable marginals. (**A**) Fraction of marginals that are unbelievable. Marginals were generated from fully connected, 8-node binary models with random biases and pairwise couplings, $h_i \sim \mathcal{N}(0, \frac{1}{3})$ and $J_{ij} \sim \mathcal{N}(0, \sigma_J)$. (**B**,**C**) Performance of five models on 370 unbelievable random target marginals (Section 3), measured with Bethe divergence $D_\beta[\boldsymbol{p}||\boldsymbol{b}]$ (B) and Euclidean distance $|\boldsymbol{p}-\boldsymbol{b}|$ (C). Target were generated as in (A) with $\sigma_J = \frac{1}{3}$, and selected for unbelievability. Bars represent central quartiles, and white line indicates the median. The five models are: (*i*) BP on the graphical model that generated the target distribution, (*ii*) BP after parameters are set by pseudomoment matching, (*iii*) the beliefs with the best performance encountered during Bethe wake-sleep learning, (*iv*) eBP using exact parameters from the last 100 iterations of learning, and (*v*) eBP with gaussian-distributed parameters with the same first- and second-order statistics as *iv*.

number of beliefs being averaged. The gaussian parameter ensemble also did much better than even the best results of BP.

4 Discussion

Other studies have also made use of the Bethe Hessian to draw conclusions about belief propagation. For instance, the Hessian reveals that the Ising model's paramagnetic state becomes unstable in BP for large enough couplings [17]. For another example, when the Hessian is positive definite throughout pseudomarginal space, then the Bethe free energy is convex and thus BP has a unique fixed point [18]. Yet the stronger interpretation appears to be underappreciated: When the Hessian is not positive definite for some pseudomarginals, then BP can never have a fixed point there, for any parameters.

One might hope that by adjusting the parameters of belief propagation in some systematic way, $\boldsymbol{\theta} \to \boldsymbol{\theta}_{\mathrm{BP}}$, one could fix the approximation and so perform exact inference. In this paper we proved that this is a futile hope, because belief propagation simply can never converge to certain marginals. However, we also provided an algorithm that does work: Ensemble belief propagation uses BP on several different parameters with different fixed points and averages the results. This approach preserves the locality and scalability which make BP so popular, but corrects for some of its defects at the cost of running the algorithm a few times. Additionally, it raises the possibility that a systematic compensation for the flaws of BP might exist, but only as a mapping from individual parameters to an ensemble of parameters $\boldsymbol{\theta} \to \{\boldsymbol{\theta}_{\mathrm{eBP}}\}$ that could be used in eBP.

An especially clear application of eBP is to discriminative models like Conditional Random Fields [19]. These models are trained so that known inputs produce known inferences, and then generalize to draw novel inferences from novel inputs. When belief propagation is used during learning, then the model will fail even on known training examples if they happen to be unbelievable. Overall performance will suffer. Ensemble BP can remedy those training failures and thus allow better performance and more reliable generalization.

This paper addressed learning in fully-observed models only, where marginals for all variables were available during training. Yet unbelievable marginals exist for models with hidden variables as well. Ensemble BP should work as in the fully-observed case, but training will require inference over the hidden variables during both wake and sleep phases.

One important inference engine is the brain. When inference is hard, neural computations may resort to approximations, perhaps including belief propagation [20, 21, 22, 23, 24]. It would be undesirable for neural circuits to have big blind spots, *i.e.* reasonable inferences it cannot draw, yet that is precisely what occurs in BP. By averaging over models with eBP, this blind spot can be eliminated. In the brain, synaptic weights fluctuate due to a variety of mechanisms. Perhaps such fluctuations allow averaging over models and thereby reach conclusions unattainable by a deterministic mechanism.

Acknowledgments

The authors thank Greg Wayne for helpful conversations.

References

[1] Cooper G (1990) The computational complexity of probabilistic inference using bayesian belief networks. Artificial intelligence 42: 393–405.

[2] Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann Publishers, San Mateo CA.

[3] Kschischang F, Frey B, Loeliger H (2001) Factor graphs and the sum-product algorithm. IEEE Transactions on Information Theory 47: 498–519.

[4] Bishop C (2006) Pattern recognition and machine learning. Springer New York.

[5] Wainwright M, Jordan M (2008) Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning 1: 1–305.

[6] Yedidia JS, Freeman WT, Weiss Y (2000) Generalized belief propagation. In: IN NIPS 13. MIT Press, pp. 689–695.

[7] Heskes T (2003) Stable fixed points of loopy belief propagation are minima of the Bethe free energy. Advances in Neural Information Processing Systems 15: 343–350.

[8] Welling M, Teh Y (2001) Belief optimization for binary networks: A stable alternative to loopy belief propagation. In: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., pp. 554–561.

[9] Wainwright MJ, Jaakkola TS, Willsky AS (2003) Tree-reweighted belief propagation algorithms and approximate ML estimation by pseudo-moment matching. In: AISTATS.

[10] Welling M, Teh Y (2003) Approximate inference in Boltzmann machines. Artificial Intelligence 143: 19–50.

[11] Parise S, Welling M (2005) Learning in markov random fields: An empirical study. In: Joint Statistical Meeting. volume 4.

[12] Watanabe Y, Fukumizu K (2011) Loopy belief propagation, Bethe free energy and graph zeta function. arXiv cs.AI.

[13] Hinton G, Sejnowski T (1983) Analyzing cooperative computation. Proceedings of the Fifth Annual Cognitive Science Society, Rochester NY .

[14] Welling M, Sutton C (2005) Learning in markov random fields with contrastive free energies. In: Cowell RG, Ghahramani Z, editors, In Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics. Society for Artificial Intelligence and Statistics, pp. 397-404.

[15] Yedidia J, Freeman W, Weiss Y (2005) Constructing free-energy approximations and generalized belief propagation algorithms. IEEE Transactions on Information Theory 51: 2282–2312.

[16] Mooij J, Kappen H (2005) On the properties of the Bethe approximation and loopy belief propagation on binary networks. Journal of Statistical Mechanics: Theory and Experiment : P11012.

[17] Mooij J, Kappen H (2005) Validity estimates for loopy belief propagation on binary real-world networks. In: Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, pp. 945–952.

[18] Heskes T (2004) On the uniqueness of loopy belief propagation fixed points. Neural Computation 16: 2379–2413.

[19] Lafferty J, McCallum A, Pereira F (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning : 282–289.

[20] Litvak S, Ullman S (2009) Cortical circuitry implementing graphical models. Neural Computation 21: 3010–3056.

[21] Steimer A, Maass W, Douglas R (2009) Belief propagation in networks of spiking neurons. Neural Computation 21: 2502–2523.

[22] Ott T, Stoop R (2007) The neurodynamics of belief propagation on binary markov random fields. In: Schölkopf B, Platt J, Hoffman T, editors, Advances in Neural Information Processing Systems 19, Cambridge, MA: MIT Press. pp. 1057–1064.

[23] Shon A, Rao R (2005) Implementing belief propagation in neural circuits. Neurocomputing 65–66: 393–399.

[24] George D, Hawkins J (2009) Towards a mathematical theory of cortical micro-circuits. PLoS computational biology 5: 1–26.

A Neural Computation for Visual Acuity in the Presence of Eye Movements

Xaq Pitkow[1¤], Haim Sompolinsky[2,3], Markus Meister[3,4*]

1 Program in Biophysics, Harvard University, Cambridge, Massachusetts, United States of America, **2** Racah Institute of Physics and Center for Neural Computation, Hebrew University, Jerusalem, Israel, **3** Center for Brain Science, Harvard University, Cambridge, Massachusetts, United States of America, **4** Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts, United States of America

Humans can distinguish visual stimuli that differ by features the size of only a few photoreceptors. This is possible despite the incessant image motion due to fixational eye movements, which can be many times larger than the features to be distinguished. To perform well, the brain must identify the retinal firing patterns induced by the stimulus while discounting similar patterns caused by spontaneous retinal activity. This is a challenge since the trajectory of the eye movements, and consequently, the stimulus position, are unknown. We derive a decision rule for using retinal spike trains to discriminate between two stimuli, given that their retinal image moves with an unknown random walk trajectory. This algorithm dynamically estimates the probability of the stimulus at different retinal locations, and uses this to modulate the influence of retinal spikes acquired later. Applied to a simple orientation-discrimination task, the algorithm performance is consistent with human acuity, whereas naive strategies that neglect eye movements perform much worse. We then show how a simple, biologically plausible neural network could implement this algorithm using a local, activity-dependent gain and lateral interactions approximately matched to the statistics of eye movements. Finally, we discuss evidence that such a network could be operating in the primary visual cortex.

Citation: Pitkow X, Sompolinsky H, Meister M (2007) A neural computation for visual acuity in the presence of eye movements. PLoS Biol 5(12): e331. doi:10.1371/journal.pbio.0050331

Introduction

People with normal visual acuity are able to resolve visual features that subtend a single arc minute of visual angle. For the letters "F" and "P" on a Snellen eye chart, this corresponds to a difference of just a few photoreceptors (Figure 1). As we try to resolve these tiny features, fixational eye movements jitter them across the retina over distances substantially greater than the features themselves (Figure 1). How can we have such fine acuity when our eyes are moving so much?

If the brain knew the complex eye movement trajectory, then it could realign the retinal responses before processing them further. However, central visual circuits probably do not have access to the eye movement trajectory at a sufficiently fine scale. Fixational eye movements arise from imperfect compensation for head and body movements [1,2] and motor noise [3], so it is unlikely that the visual system has a reliable estimate of the resultant image motion. Although there are both efference copies of eye movement signals and proprioceptive feedback, they have a limited accuracy of several degrees [4,5], which is inadequate for tracking the much smaller movements during fixation. Thus, any estimate the brain makes about fine fixational eye movements is probably driven by visual input alone [6,7].

Unfortunately, visual processing in the retina introduces noise, leaving the brain with uncertainty both about the stimulus shape itself and about the precise trajectory the stimulus traces on the retina. The retina's output neurons—the retinal ganglion cells—are not perfectly reliable in their response to stimulation, and even without stimulation, they fire action potentials at a substantial rate. For brief, small stimuli on a featureless background, the total stimulated retinal response may consist of just a few tens of spikes. The brain must distinguish these spikes from the many hundreds of spontaneous spikes that reflect only noise. The usual remedy would be to accumulate many spikes over time until the signal emerges from the noise; but this is difficult because the fixational eye movements scatter the desired responses across space.

Thus we recognize a challenge for visual acuity in the presence of eye movements: To identify the stimulus, the brain needs to know the precise stimulus trajectory; yet to track the stimulus trajectory, the brain needs to identify which neural spikes are stimulated and which are only noise.

Presented with this challenge, what strategy could the brain use to achieve the visual acuity that humans exhibit? We will show that naive decodings of retinal spike trains that neglect the eye movements perform poorly at discriminating fine visual features. We derive a significantly better strategy that exploits the fact that eye movements are continuous to estimate the stimulus position on the retina and give greater weight to retinal spikes originating near this position. Surprisingly, we found that this strategy is attainable by a

Academic Editor: David Burr, Istituto di Neurofisiologia, Italy

Received March 27, 2007; **Accepted** November 9, 2007; **Published** December 27, 2007

Abbreviations: V1, primary visual cortex

* To whom correspondence should be addressed. E-mail: meister@fas.harvard.edu

¤ Current address: Center for Theoretical Neuroscience, Columbia University, New York, New York, United States of America

Author Summary

Like a camera, the eye projects an image of the world onto our retina. But unlike a camera, the eye continues to execute small, random movements, even when we fix our gaze. Consequently, the projected image jitters over the retina. In a camera, such jitter leads to a blurred image on the film. Interestingly, our visual acuity is many times sharper than expected from the motion blur. Apparently, the brain uses an active process to track the image through its jittering motion across the retina. Here, we propose an algorithm for how this can be accomplished. The algorithm uses realistic spike responses of optic nerve fibers to reconstruct the visual image, and requires no knowledge of the eye movement trajectory. Its performance can account for human visual acuity. Furthermore, we show that this algorithm could be implemented biologically by the neural circuits of primary visual cortex.

simple neural network whose properties are consistent with functional and anatomical features of primary visual cortex.

Results

Psychophysics

For concreteness, we choose a simple task to analyze: An observer is asked to discriminate between two tiny oriented bars that span 1 or 2 arcmin of visual angle. In the retina's fovea, this stimulus affects just a few cone photoreceptors, each collecting light from a region about 0.5 arcmin in diameter. Each cone drives approximately one On-type and one Off-type ganglion cell, and conversely, each ganglion cell receives its input from just one cone [8]. This means that at any given instant, the brain receives information about the stimulus from spiking in a small cluster of retinal ganglion cells, but the identity of those cells changes continually as the stimulus jitters across the retina. We tested human subjects on this discrimination task and found that despite these challenges, many human subjects can actually perform well above chance (Figure 2, see also [9,10]).

It is plausible that the finest human acuity might be limited primarily by the information available in the retina rather than by later constraints or losses. For example, our ability to detect dim lights in absolute darkness is ultimately limited by photon shot noise at the rod photoreceptor. In bright light—the condition considered here—noise introduced by retinal processing greatly exceeds photon shot noise [11–13]. Correspondingly, human thresholds on fine acuity tasks are worse by a factor of ten than expected from ideal processing of photon counts [9,10]. Instead, human performance on simple visual tasks is more compatible with the limitations from noisy retinal ganglion cell spikes [14,15]. If acuity is in fact limited by the retinal spike trains, then the brain must make efficient use of these spikes to extract the relevant information.

Markov Decoder Model

We now present a strategy for accumulating information about position and orientation of the small stimulus bar on the retina. This strategy decodes the observed spike trains from retinal ganglion cells using prior knowledge about the statistics of those spikes and the statistics of eye movements. The output of the decoder is a moment-to-moment estimate of the bar's orientation.

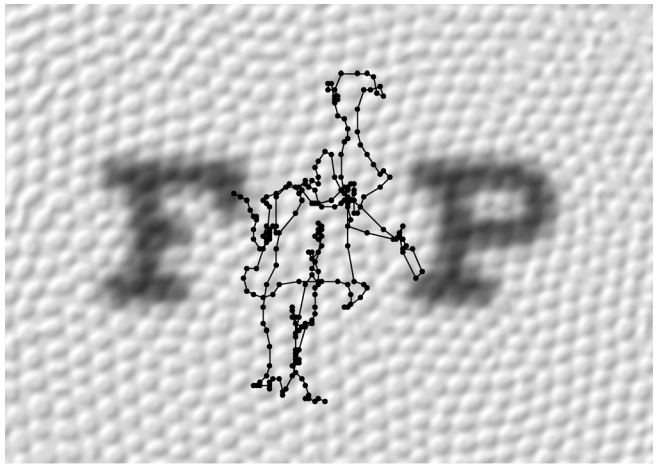

Figure 1. The Neighboring Letters "F" and "P" on the 20/20 Line of the Snellen Eye Chart, Blurred by a Gaussian of Diameter 0.5 arcmin and Projected onto an Image of the Foveal Cone Mosaic (Photoreceptor Image Modified from [92])

The 1-arcmin features that distinguish the letters extend over only a few photoreceptors. Also shown is a sample fixational eye movement trajectory for a standing subject (courtesy of [25]), sampled every 2 ms for a duration of 500 ms and then smoothed with a 4-ms boxcar filter.
doi:10.1371/journal.pbio.0050331.g001

The decoder assumes a model of retinal ganglion cell spike generation, shown in Figure 3A, which includes random eye movements, optical blur, spatial receptive fields, temporal filtering, rectification, and probabilistic spiking. Each stimulus is a small, dark, oriented rectangle that jitters across the retina. The eye's optics introduce a spatial blur, implemented by a Gaussian filter with a 0.5 arcmin diameter. We assume this image is sensed by photoreceptors arranged on a square lattice, each activating one Off-type ganglion cell. We neglect the On-type cells because they will generate only a weak response to the small, dark stimulus (see Discussion). For the same reason, we neglect the broad, but shallow, surrounds of Off-cells, which are usually approximately 50 times weaker than the receptive field center [16]. Furthermore, we first assume for simplicity that ganglion cells report on the instantaneous light intensity in their receptive field center; later, we will consider implications of including a temporal filter like that in Figure 3D. Under these assumptions, when a stimulus with orientation S is at position $\mathbf{x}$, a model retinal ganglion cell at position $\mathbf{y}$ fires action potentials with Poisson statistics at the instantaneous time-dependent rate $r_S(\mathbf{y}-\mathbf{x})$ depicted in Figure 3B, ranging from a peak value $r_{\max}$ at positions near the stimulus to the background firing rate r_0 at large distances. In bright conditions, retinal ganglion cells respond to a contrast of 100% (black on white) with a spike rate of $r_{\max} \sim 100$ Hz [17]. Far from the stimulus, we assume neurons fire spontaneously with rates on the order of $r_0 \sim 10$ Hz [18,19].

In weighting the retinal responses properly, the decoder takes into account the statistics of the trajectories that are traced by the fixed stimulus on the moving retina. Fixational eye movements are classified into three types of motion: microsaccades, drift, and tremor [20]. Microsaccades are not thought to play a role in fine visual tasks [21–23], though they

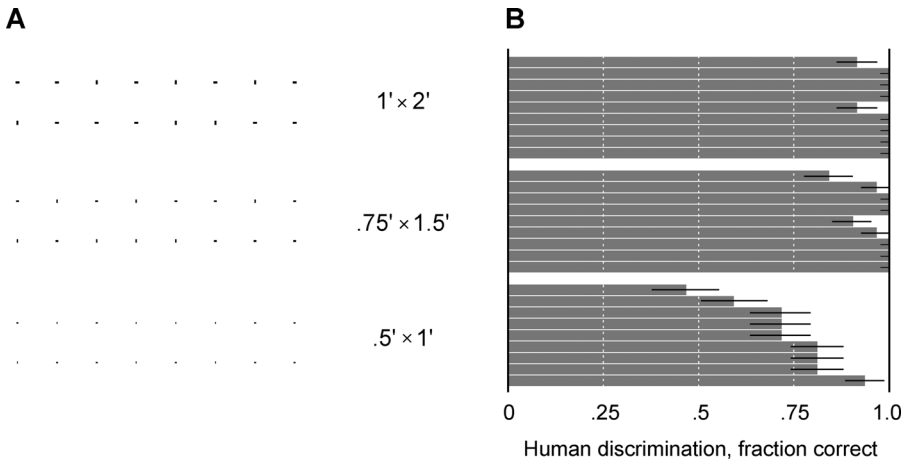

Figure 2. The Discrimination Task

(A) Tiny horizontal and vertical stimuli, sized to subtend 0.5 × 1, 0.75 × 1.5, and 1 × 2 arcmin^2 when viewed at a distance of 88 cm.
(B) Performance of nine human participants on this task, measured by the fraction of correct guesses out of 32 trials. Error bars represent the 68% confidence interval.
doi:10.1371/journal.pbio.0050331.g002

may contribute to peripheral vision [24]. Tremor has very low amplitude, less than a photoreceptor diameter. We therefore concentrate on the drift component, which has the properties of a random walk [3], with modest deviations on short and long timescales [25]. For simplicity, we assume that the fixational eye movements are described by a spatially discrete random walk across the photoreceptor lattice with an effective diffusion constant of $D \sim 100$ arcmin^2/s (see Materials and Methods).

For a random walk trajectory, the probability of the current position depends only on its most recent previous position. This attribute, in combination with the assumption that retinal responses are memoryless, allows us to write a differential equation for the probability distribution $P(S, \mathbf{x}, t)$ of the stimulus orientation S and current location $\mathbf{x}$, given all the spikes observed before time t:

$$\frac{\partial}{\partial t}P(S,\mathbf{x},t) = \sum_{\mathbf{y}} \lambda_{\mathbf{y}}(t) f_S(\mathbf{y}-\mathbf{x}) P(S,\mathbf{x},t) - r_S^{tot}(\mathbf{x})P(S,\mathbf{x},t) + D\nabla^2 P(S,\mathbf{x},t) \quad (1)$$

(see Protocol S1 for a derivation). In this equation, $\lambda_{\mathbf{y}}(t) = \sum_{t_{\mathbf{y}}} \delta(t - t_{\mathbf{y}})$ stands for the observed spike train of the retinal neuron $\mathbf{y}$ at time t; $f_S(\mathbf{y}-\mathbf{x}) = \ln(r_S(\mathbf{y}-\mathbf{x})/r_0)$ reflects the expected firing-rate profile generated by the stimulus; $r_S^{tot}(\mathbf{x}) = \sum_{\mathbf{y}} r_S(\mathbf{y}-\mathbf{x})$ denotes the total expected firing rate of the retinal ganglion cell array; and ∇^2 represents a discrete version of a second-order spatial derivative operator. On a square lattice, $\nabla^2 P(S,\mathbf{x},t) = \frac{1}{a^2}(\sum_{\Delta\mathbf{x}} P(S,\mathbf{x}+\Delta\mathbf{x},t) - 4P(S,\mathbf{x},t))$, where $\mathbf{x}+\Delta\mathbf{x}$ ranges over the four nearest neighbors of $\mathbf{x}$ on the lattice (Figure 3C), and a is the distance between lattice points.

Equation 1, also known as a Fokker-Planck equation, describes a reaction-diffusion system [26]. There are three sources of changes in the stimulus posterior probabilities $P(S, \mathbf{x}, t)$. The first term,

$$\sum_{\mathbf{y}} \lambda_{\mathbf{y}}(t) f_S(\mathbf{y}-\mathbf{x}) P(S,\mathbf{x},t), \quad (2)$$

implies that each spike of a retinal neuron $\mathbf{y}$ results in a multiplicative update of the stimulus posterior probabilities $P(S, \mathbf{x}, t)$ by a factor $r_S(\mathbf{y}-\mathbf{x})/r_0$ (as shown in Materials and Methods), thus increasing the likelihoods of stimulus positions $\mathbf{x}$ near the firing retinal neuron, where $r_S(\mathbf{y}-\mathbf{x})/r_0$ is large. The second term,

$$-r_S^{tot}(\mathbf{x})P(S,\mathbf{x},t), \quad (3)$$

represents the "negative" evidence accumulating during quiescent periods. In between retinal spikes, $P(S, \mathbf{x}, t)$ decays exponentially with a decay rate that equals the total expected firing rate of the retinal array with the stimulus S at position $\mathbf{x}$. In the present case, in which the total activation of the retina is the same regardless of the orientation and position of the stimulus, we ignore this term since it does not affect the relative values of the posterior distribution for different orientations S or positions $\mathbf{x}$. These first two terms represent the local "reaction" terms. The last term,

$$D\nabla^2 P(S,\mathbf{x},t), \quad (4)$$

is the "diffusion" term; it describes the lateral spread of the posterior probability across the retina during the time between retinal spikes. This spread accounts for the expected stimulus movements due to the fixational eye movements. The rate of spread is given by D, the diffusion constant of the fixational eye movements. The initial condition for solving Equation 1 is specified by $P(S, \mathbf{x}, 0)$, which is the initial probability distribution of the stimulus orientation and position prior to observing any spikes. We will assume that it is uniform over the entire range of positions and orientations. Finally, we note that Equation 1 technically yields the posterior probability only up to an overall normalization factor (see Protocol S1 for a strictly normalized version). This is unimportant for discrimination, since only the relative values of P for different orientations matter. However, in numerical work, one must supplement Equation 1 by a divisive normalization, periodically dividing all components of P by the sum $\sum_{S,\mathbf{x}} P(S,\mathbf{x},t)$ over space and orientation (see Materials and Methods).

This decoder of retinal spike trains can be applied to a

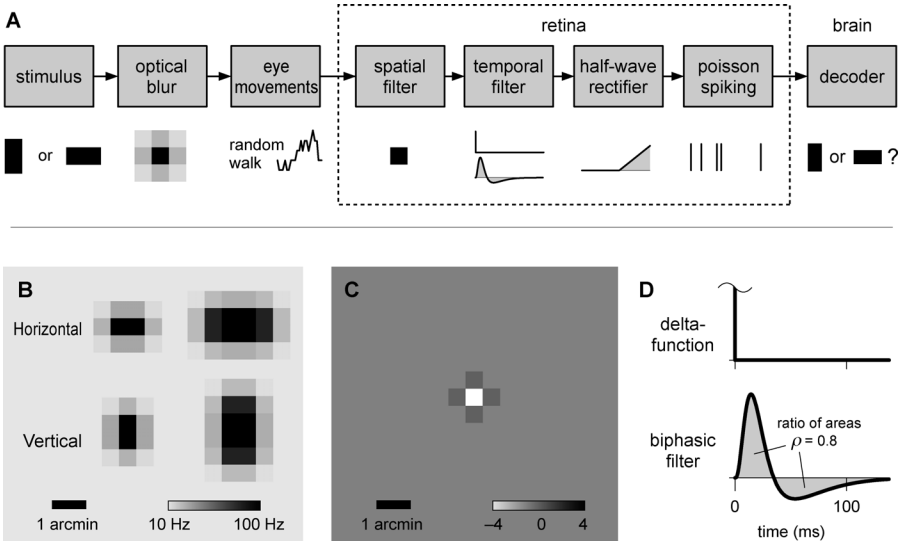

Figure 3. Models of Spike Generation and Decoding

(A) A block diagram of the features in the model visual system; see text for details.
(B) Firing-rate profiles $r_S(\mathbf{y})$ induced by horizontal and vertical stimuli on the model foveal lattice. Left: 0.5 × 1 arcmin^2. Right: 1 × 2 arcmin^2.
(C) A graphical representation of the discrete second-derivative operator used to calculate diffusion rates.
(D) The temporal filters that model retinal ganglion cells use to convert the time-varying light intensity into the instantaneous firing rate.
doi:10.1371/journal.pbio.0050331.g003

variety of tasks. For instance, in a localization task with a stimulus of known orientation, S, the estimate of the stimulus position $\mathbf{x}$ is given by $\mathbf{x}^{\text{estimate}} = \arg\max_{\mathbf{x}} P(S, \mathbf{x}, t)$. In a discrimination task in which only the orientation needs to be determined, a sum over the irrelevant position variable yields $S^{\text{estimate}} = \arg\max_{S} \sum_{\mathbf{x}} P(S, \mathbf{x}, t)$.

The first-order differential equation (Equation 1) implies that the posterior probability can be updated in a way that depends only on the current posterior probability and the current evidence from spikes. This is possible because the assumed process of generating spikes depends only on the current stimulus location. This is an example of what is known as a Markov process, more specifically, a hidden Markov process because the location variable is not observed directly. We will call this decoder of the spike trains the "Markov decoder." It will yield optimal decisions if the Markov assumptions accurately describe the spike generation process.

Visualizing the Markov Decoder Algorithm

We illustrate the performance of the decoder in Figure 4A–4E, using spike trains from a one-dimensional model retina. In the first task (Figure 4A–4C), the stimulus shape is known, so the only uncertainty is its location. The stimulus follows a random walk trajectory, generating the instantaneous firing-rate pattern (Figure 4A), and eliciting extra spikes for neurons along its path while other neurons produce spontaneous spikes at a lower rate (Figure 4B). The Markov decoder collects all the retinal spikes and solves Equation 1 to estimate the posterior probability distribution over positions (see Materials and Methods for numerical details). The result is displayed in Figure 4C.

In the particular trial depicted, the task of localizing the stimulus appears quite difficult, even with only one spatial dimension: In any given time slice, the evidence provided by retinal spikes is rather weak. Nonetheless, the accumulated evidence over time provides a good estimate of the stimulus trajectory. As evidence from the spiking neurons accumulates, the decoder locks onto and tracks the true stimulus location.

In a second task, the decoder must discriminate between two possible stimulus shapes moving on a one-dimensional retina (Figure 4D and 4E). Because one dimension does not allow for horizontal and vertical bars, we take the shape variable S to refer to two stimuli related by reflection (Figure 4D, inset). Again, these probabilities evolve according to the reaction-diffusion dynamics of Equation 1, where incoming spikes lead the probability distributions to track the stimulus, but now there is a competition for probability between two stimulus shapes. The Markov decoder may make errors in position, stimulus identity, or both, depending on the particular spike trains it observed, but on average, it discriminates between the two stimulus shapes with an accuracy well above chance.

Non-Markovian Spike Generation with Temporal Filtering

So far, we have assumed that the retinal ganglion cells report on the instantaneous light intensity, but this is not realistic. Primate photoreceptors react slowly, with integration times on the order of 25 ms [27], yet the eye movements' diffusion constant of 100 arcmin^2/s implies that the stimulus typically moves one photoreceptor diameter in under a millisecond. Therefore, the firing of retinal ganglion cells cannot track the light intensity as it fluctuates on this rapid timescale. More realistically, the ganglion cells respond to the light intensity in their receptive field averaged by a biphasic temporal filter like that shown in Figure 3D [28].

This temporal filtering has an important implication: Since the spiking probability depends on an extended history of stimulus positions, the spikes cannot be interpreted optimally by the Markov decoder. One can generalize Equation 1 to derive the optimal decoder in this situation. The posterior

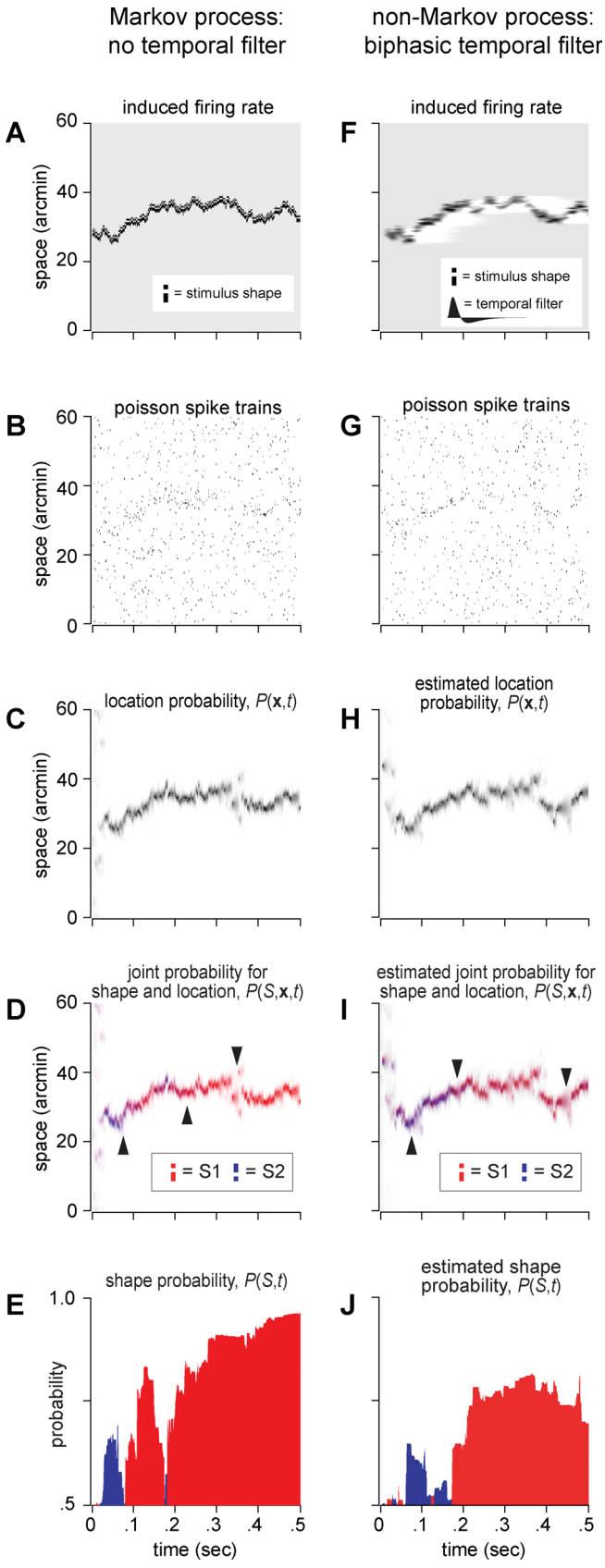

Figure 4. Simulations of the Markov Decoder (Equation 1) for a Small Stimulus Moving on a One-Dimensional Model Retina

(A–E) Spike generation by a Markov process.
(F–J) Spike generation by a non-Markov process that includes the biphasic temporal filter from Figure 3D.
(A and F) Firing rate induced by a stimulus moving on the retina with a random walk diffusion constant of 100 arcmin^2/s. The stimulus shape activates three neurons in the pattern shown in the inset. The background rate is 10 Hz, and the peak stimulated rate is 100 Hz.
(B and G) Poisson retinal spike trains drawn from this instantaneous firing rate. Each row corresponds to a neuron, spaced every 0.5 arcmin.
(C and H) Evolution of the location probability $P(\mathbf{x}, t)$ for a known stimulus shape S (inset in [A]), but an unknown location $\mathbf{x}$, derived from the spike trains shown in the previous panel.
(D and I) Decoder behavior when the stimulus can instead take one of two possible shapes, but the true shape is unknown. The two stimuli each activate three retinal neurons, in mirror-image patterns (inset). The spike trains now induce two spatial distributions of the posterior probability $P(S, \mathbf{x}, t)$, plotted in shades of red and blue.
(E and J) Shape probability $P(S, t) = \sum_{\mathbf{x}} P(S, \mathbf{x}, t)$, colored red for the correct stimulus identity and blue for the incorrect one. In these trials, we see that once the decoder coalesces around the stimulus location, it first attributes a greater probability to the wrong stimulus (leftmost arrow in [D] and [I]) before accumulating enough evidence for the correct stimulus (middle arrow). The decoder can lose track of the stimulus briefly (e.g., at rightmost arrow) but continues to favor the correct stimulus until the end of the trial. Note that (E) reflects the true posterior probabilities, whereas in (J), the Markov decoder can only estimate them because the spike generation process includes temporal filtering that the decoder neglects.
doi:10.1371/journal.pbio.0050331.g004

probability distribution now extends over all possible random walk trajectories within the temporal range of the filter. There are approximately 10^8 such trajectories leading up to each stimulus location, and propagating their probability distribution is numerically unwieldy. It also seems improbable that the brain takes such an approach. These arguments apply strictly to the optimal decoder, but there may exist useful and efficient nonoptimal decoders. In fact, we found that the simple Markov decoder still performs well at the discrimination task, despite the mismatch between the encoding process and the decoder's assumptions.

To explore this, we generated retinal ganglion cell spikes (Figure 4F and 4G) with a model that includes a biphasic temporal filter (Figure 3D). The filtering adds a motion smear to the stimulus, which renders the output spike trains more ambiguous. Despite its ignorance of the temporal filtering, the decoder can still track the stimulus location, with a small delay due to the filter (Figure 4H). Furthermore, the decoder successfully accumulates information about the stimulus shape (Figure 4I and 4J).

Performance of the Markov Decoder

We now evaluate the Markov decoder's performance on the original visual task: to discriminate whether a small jittering bar is oriented horizontally or vertically. Here, we modeled the retina and the decoder using two spatial dimensions and simulated many trials of the discrimination task. For every trial, we selected a random stimulus orientation and trajectory, filtered the instantaneous light intensity with a biphasic temporal filter, rectified the result to calculate the expected firing rates for all retinal neurons over time, and generated Poisson spike trains with these firing rates (Figure 3). We then applied the decoder algorithm to these spike trains by numerically solving Equation 1 and selecting the orientation estimated to be more probable. Performance was quantified as the fraction of trials in which the decoder guessed correctly.

The results of these simulations show that the Markov

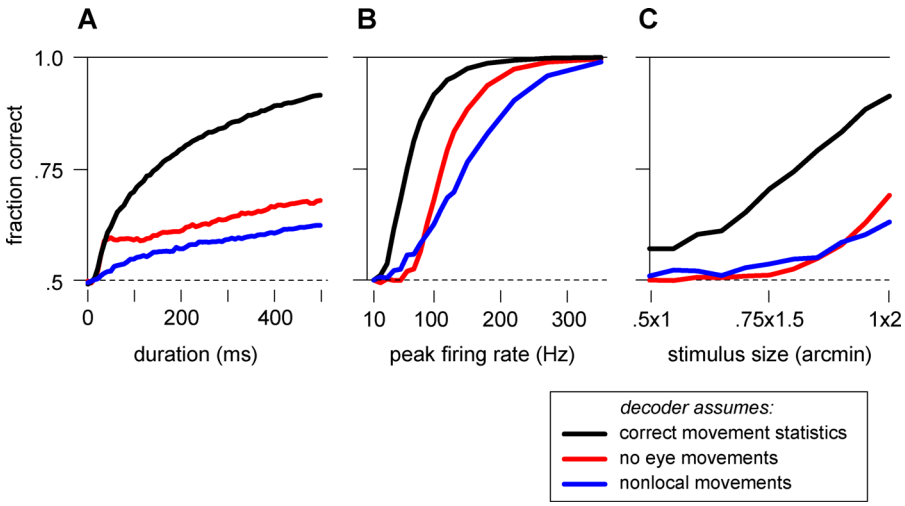

Figure 5. Model Performance on the Horizontal versus Vertical Discrimination Task Shown in Figure 2

Performance is measured by simulating retinal responses, calculating decisions based on those responses, and computing the fraction of correct decisions (see Materials and Methods). When fixational eye movements jitter the stimulus, the Markov decoder is able to perform well on the task by accounting for the eye movement statistics (black curves). Two naive decoders are also applied to this task, one that assumes the stimulus is fixed (red) and one that assumes maximum uncertainty about those movements (blue). Performance is shown as a function of stimulus duration (A), peak stimulated firing rate (B), and stimulus size (C). Where not otherwise specified, the parameters for these simulations are background firing rate of 10 Hz, a peak stimulated rate of 100 Hz, a stimulus of 1×2 arcmin^2, a duration of 500 ms, and a diffusion constant of 100 arcmin^2/s.

doi:10.1371/journal.pbio.0050331.g005

decoder's performance is generally compatible with human performance. The decoder is able to reliably discriminate horizontal from vertical within a few hundred milliseconds (Figure 5A) using spikes generated at biologically realistic rates around 100 Hz (Figure 5B). Like humans, the Markov decoder finds discrimination very challenging with the smallest stimuli, and fairly routine for the largest (compare Figures 2B and 5C).

Importance of Accounting for Fixational Eye Movements

The Markov decoder can be used to evaluate the importance of accounting for fixational eye movements in estimating the stimulus shape or orientation. Specifically, we ask the question: how much better does the Markov decoder perform compared to strategies that ignore the eye movement statistics?

Two naive strategies can be proposed: The first assumes that there are no eye movements. This amounts to using a Markov decoder, but setting its presumed diffusion constant to zero. Another strategy recognizes that the eye moves approximately every 0.6 ms (the average time between random walk steps on the square lattice), but is otherwise ignorant of the eye movement statistics; it conservatively assumes that jumps to all stimulus positions are equally likely.

Naturally, the decoder that uses the correct diffusion statistics works best, but simulations reveal that it outperforms the two naive decoders by a large margin (Figure 5). For very brief stimuli of the same duration as the transient retinal response (~30 ms), the decoder that assumes a fixed stimulus and the decoder that knows the correct movement statistics perform equally well, because temporal filtering does not allow the responses to track the stimulus movements. Yet, under typical viewing conditions, such a duration is too brief for human subjects to discriminate the stimulus shapes. As the decoder integrates information beyond the temporal filter's persistence time, the movements become relevant and the naive algorithm essentially blurs the stimulus even more. The decoder giving equal odds to all locations at all times relies only on the rare coincidences when multiple stimulated neurons spike in tight synchrony. Eventually, this naive decoder can manage to discriminate the stimuli, but it requires a much longer time or many more spikes than the Markov decoder.

Robustness

How robust is the algorithm to imperfections in implementation? The key parameter that incorporates the statistics of the eye movements is the assumed diffusion constant. As shown above, if the decoder assumes that the eye movements are much faster or much slower than they really are, then the performance degrades substantially. However, between these two extremes, there is a broad range of assumed diffusion constants that causes only a few percent of extra mistakes (Figure 6A). In fact, the decoder benefits slightly from assuming a lower diffusion constant, probably due to the apparent stimulus persistence caused by temporal blurring. This demonstrates that it is essential to account for eye movements, but the algorithm proposed here is robust to misestimates of the movement statistics.

Every time the decoder receives a retinal spike, the estimated stimulus probability rises locally by a factor proportional to the expected stimulated firing rate divided by the background rate (Materials and Methods, Equation 10), which reflects the confidence in the new information brought by a retinal spike. Changing this factor in the Markov decoder would be expected to alter its performance. However, we found that performance is remarkably insensitive to this variable over a wide range of values (Figure 6B).

Finally, we may ask whether the decoder performance is sensitive to the assumed stimulus shapes. Each retinal spike increases the estimated stimulus probability at all those locations where a stimulus could potentially have caused that

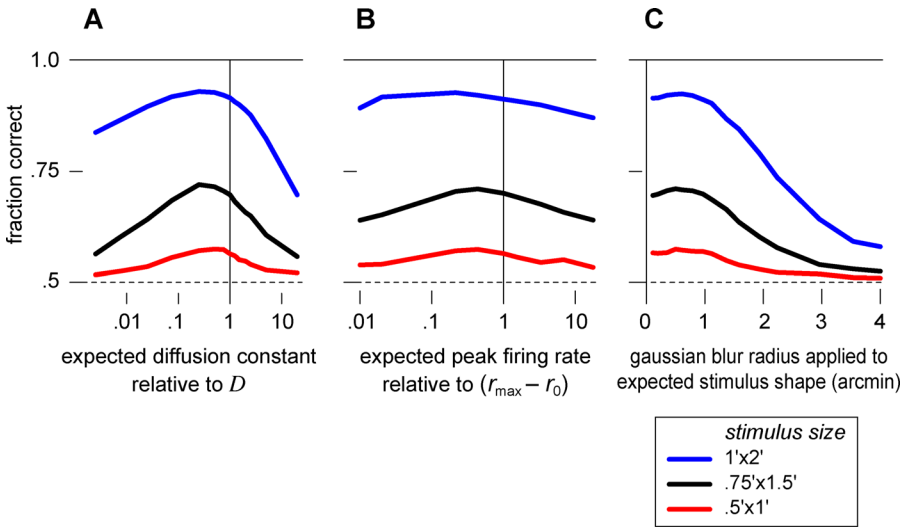

Figure 6. Markov Decoder Robustness to Mismatched Parameters

(A) Discrimination performance when the decoder's estimate for the trajectory statistics is wrong: The stimulus is known to perform a random walk on the retina, but the diffusion constant is misestimated. The performance is optimal for estimated values close to the actual diffusion constant and declines gently on either side.

(B) Performance as a function of the expected stimulated firing rate, parameterized as $(r^{est}_{max} - r_0)/(r_{max} - r_0)$.

(C) Performance as a function of the expected stimulus size, obtained by convolving the true stimulus shape with a spatial Gaussian of the specified radius. In each of these plots, parameters are the same as in Figure 5.

doi:10.1371/journal.pbio.0050331.g006

spike. If the expected stimuli differ from the true stimuli, then this probability increases over the wrong set of locations, leading to suboptimal performance. To explore this, we set the decoder's expected stimulus shape to be larger than the true shape by various amounts (Figure 6C). Enlargement up to about 1 arcmin produced no noticeable change in the decoder's performance, but larger discrepancies of about 2 arcmin led to significant decline. This behavior can be understood as follows: a misestimate of the stimulus size effectively leads to excessive smearing of the positional information. This must be compared to the diffusional smearing that occurs as the stimulus moves in the typical time between informative spikes, which amounts to approximately 1 arcmin. Thus the Markov decoder is hardly affected by misestimates in stimulus shape smaller than this amount.

In summary, the Markov decoder is robust to various parameters that encompass its a priori assumptions about the stimulus. If the decoder allows activity to diffuse at an approximately correct rate, and expects shapes not dramatically larger than the true stimuli, then it can achieve good discrimination performance.

Network Implementation

Despite the apparent complexity of the differential equation governing the Markov decoder, its dynamics map directly onto a simple neural network with a structure consistent with many known properties of visual cortex. For clarity, we will first introduce a network that estimates the location probabilities for a given stimulus shape, and then show the extension required for shape discrimination.

Figure 7 depicts a network that implements the Markov decoder algorithm for estimating the location of a stimulus with a known orientation S. The network has three types of neurons: the retinal neurons, a hidden layer of decoder neurons, and an inhibitory neuron. Each neuron in the hidden layer is associated with a spatial location, $\mathbf{x}$, and its activity at time t represents the estimated posterior probability (up to a normalization factor) that the stimulus is present at that location, $P(S, \mathbf{x}, t)$. The feedforward input to each hidden layer neuron $\mathbf{x}$ consists of spikes from retinal locations $\mathbf{y}$, weighted by a spatial receptive field $f_S(\mathbf{y} - \mathbf{x}) = \ln(r_S(\mathbf{y} - \mathbf{x})/r_0)$, which ranges from zero far from the stimulus to a peak of $\ln(r_{max}/r_0)$. The weighted retinal input is then multiplied by a variable gain proportional to the activity of the postsynaptic neuron, $P(S, \mathbf{x}, t)$. This gated retinal input implements the contribution of Equation 2 to the update of the estimated posterior probability. The neurons in the network interact through lateral connections mimicking the diffusion operator (Equation 4 and Figure 3C). Recall that the diffusion operator takes the summed probability of the nearest-neighbors of a given location, $\sum_{\Delta\mathbf{x}} P(S, \mathbf{x} + \Delta\mathbf{x}, t)$, and subtracts $4P(S, \mathbf{x}, t)$ from this in order to conserve probability. In the network, conservation of activity is not required, so the subtraction can be omitted: when the change in P is simply proportional to P, the solution is an exponential decay that scales P uniformly at all locations, leaving the relative values of the activity unaltered. Thus, lateral excitatory connections are sufficient to implement the diffusion term in the network. For the same reason, the network does not need any representation of the local decay term, Equation 3, which also scales all activities equally. Finally, the network includes a global divisive inhibition to maintain network activity at a stable level despite the various excitatory interactions.

To extend this framework to the discrimination task, we need two copies of the network that differ by their orientation tuning (Figure 8). In the "horizontal" network, representing $P(H, \mathbf{x}, t)$, the neurons are tuned to horizontal stimuli, hence their receptive fields are determined by $r_H(\mathbf{y} - \mathbf{x})$ (Figure 3B);

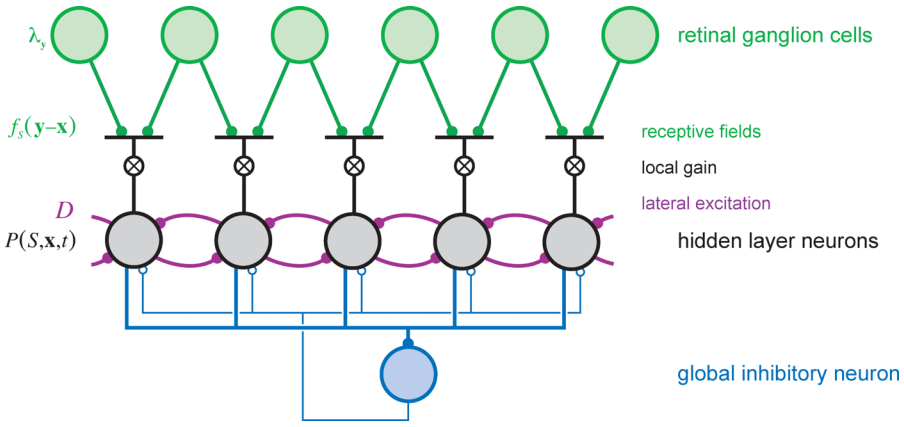

Figure 7. Schematic for a Network Implementation of the Markov Decoder (Equation 1)

Spikes from retinal neurons (green, top layer) are collected by neurons in a hidden layer (black, middle layer) with linear receptive fields $f_S(\mathbf{y}-\mathbf{x})$ and a local gain that is set by activity in the recipient neuron. Global divisive inhibition is driven by the total activity of all neurons in the hidden layer through a pooling neuron (blue, bottom neuron).

doi:10.1371/journal.pbio.0050331.g007

correspondingly, in the "vertical" network, representing $P(V,\mathbf{x},t)$, the receptive fields are related to $r_V(\mathbf{y}-\mathbf{x})$. For the discrimination task, the retinal position is irrelevant; comparing the pooled activity from each subnetwork is sufficient to discriminate between the stimulus orientations. Note that the lateral excitatory connections in this network architecture are orientation specific because fixational eye movements translate the stimulus, but do not appreciably rotate it: orientation, but not position, is preserved. On the other hand, the stabilizing divisive normalization must be global across orientations to ensure a meaningful comparison between the two orientations.

Discussion

Fixational eye movements pose a major challenge for vision since they scatter weak signals about fine stimulus features across the retina. We addressed this challenge mathematically by deriving an algorithm that guesses the orientation of a stimulus, given spiking responses from a model retina and prior knowledge about its function. It accomplishes this by collecting and sorting the scattered feature information in a systematic way, weighting retinal spikes according to an estimated probability that those spikes reflect stimulus features and not noise.

Biological Implementation

As described above, the decoder algorithm has a direct mapping onto an abstract neural network, and we will argue that primary visual cortex (V1) has many properties well suited to instantiate this network with real neurons. Specifically, we take the hidden layer neurons in Figure 7 to be cortical cells that receive inputs from the retina via the thalamus.

For good performance, these neurons should integrate retinal spikes using linear, oriented receptive fields of the same shape and size as the visual stimuli (Figure 8). We showed that the decoder's performance was robust to mismatches between the true stimuli and the expected stimuli (Figure 5B and 5C), so these receptive fields need be only approximately tuned to the stimulus size and strength. Linear oriented receptive fields are a well-established characteristic of cortical simple cells [29]. For stimuli subtending only a few human photoreceptors, we require a receptive field of just 1 or 2 arcmin in size. Receptive fields for cortical neurons dedicated to foveal vision are notoriously difficult to measure, notably due to technical problems associated with fixational eye movements. In macaques, receptive fields have been reported as small as 3 arcmin, slightly larger than the macaque's cone resolution of about 1.7 arcmin [30,31]. Therefore, cortical neurons are likely to exist with receptive fields of the appropriate size. Although equivalent measurements are unavailable for human cortex, our finest acuity may well be mediated by cortical neurons driven by an oriented set of just a few cones.

To account for fixational eye movements, the neural network must be organized retinotopically so that local stimulus movements correspond to local interactions in cortex. This is, of course, a known property of V1 [32,33]. Because fixational eye movements are largely independent in each eye [34], the fine retinal positioning of the stimulus is also independent for the two eyes: Proper accounting for stimulus movement, therefore, requires that lateral excitation should not cross eyes. Ocular dominance columns [35] are thus seen as a necessary feature if cortex is to accommodate fixational eye movements. Eye movements are best handled before the signals from the two eyes are mixed, favoring a locus in the lateral geniculate nucleus (LGN) or in V1 for the proposed network.

Eye movements are expected to simply translate visual features, but not rotate them, and these expectations should be built into circuitry. Activity in the model decoder network diffuses across space through lateral excitatory connections between nearby neurons, but only those with similar orientation preferences. In the early visual system, the required iso-orientation facilitation has been observed psychophysically [36–38], anatomically [39–42], and physiologically [43–45]. Lateral diffusion of activity has also been directly imaged in visual cortex [46].

As the eye drifts, the retina moves rigidly in world coordinates. But since the size of cortical receptive fields increases with distance from the fovea [30,47,48], fixational eye movements do not move stimuli across many receptive

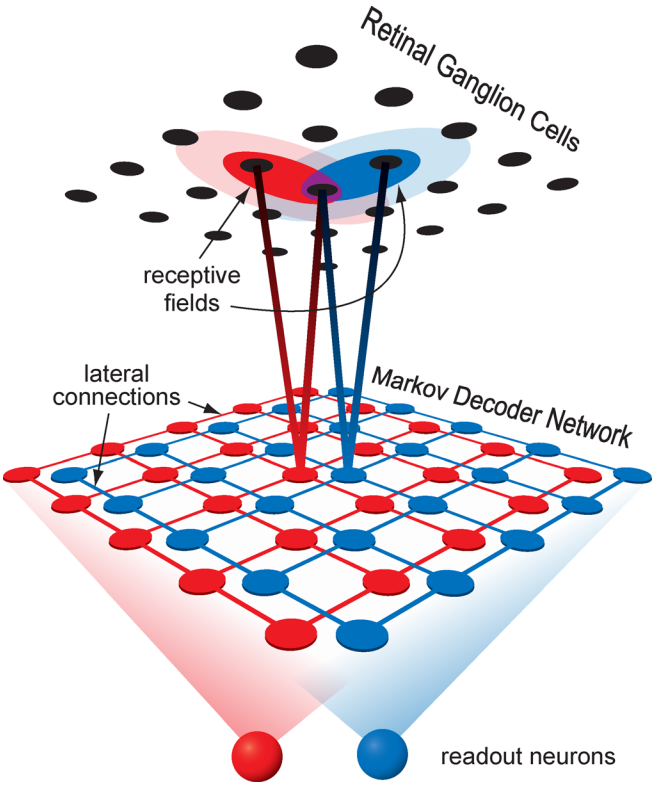

Figure 8. Two Independent but Competing Subnetworks, Each Structured as in Figure 7, Receive Input from the Same Retinal Ganglion Cells, but Use Different Receptive Fields

The total activity in each subnetwork is pooled by two readout neurons. The more active readout neuron indicates the network's estimate of the stimulus orientation.

doi:10.1371/journal.pbio.0050331.g008

fields in the periphery. Accordingly, there is no need to compensate for fixational eye movements in the periphery. We expect, therefore, to see some aspects of the cortical network that are specialized for foveal vision. Consistent with this, more of striate cortex is dedicated to responses from the fovea than can be explained by the density of retinal ganglion cells [49–51], and lateral suppression and facilitation differ between central and peripheral vision [37].

The Markov decoder requires that the lateral facilitatory interactions induce localized changes in the gain for new input spikes. Such multiplicative gain modulations have indeed been observed in the visual cortex [52,53]. A number of neural mechanisms have been invoked to create neural multipliers [54–60]. One potential mechanism involves the postsynaptic NMDA (n-methyl-d-aspartic acid) receptor, a glutamate-gated ion channel with a voltage sensitivity that causes it to open only when the postsynaptic potential is sufficiently large. In the visual cortex, NMDA activation has been shown to produce a multiplicative effect on input gain [61]. Synapses between cortical layers and within layers have different NMDA and AMPA (alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid) receptor distributions, so that lateral inputs may be simply additive, whereas feedforward input may experience a variable gain [62], as required by the Markov decoder architecture.

With an accelerating nonlinearity and excitatory interactions, this network has a positive feedback loop that would cause the activity to quickly diverge. Normalization will maintain stability, but the normalization must be global and orientation independent so that neural activities can be compared on the same scale. Previously described wide-field divisive normalization [63–66] can serve this purpose, although other global homeostatic mechanisms would function as well.

In our forced-choice task, the accumulated evidence for the horizontal and vertical stimuli must be compared. This can be accomplished downstream by a final winner-take-all computation in which the total activity in each subnetwork is pooled and then compared [67]. This type of computation must take place somewhere in the brain any time a decision must be reached, and various biological implementations have been proposed for this operation [68,69].

Whereas the input to the network consists of discrete spikes, the network units themselves represent the stimulus probability, which is a continuous variable. This variable might be most simply encoded by the collective firing rate of a cluster of neurons [70], especially given that the number of cells representing the visual field expands dramatically from the retina to the visual cortex [71]. Alternatively, the computation might well proceed with discrete spikes: model networks of spiking neurons tend to produce similar behavior as rate models with continuous variables, so long as the spikes are not too strongly correlated [72].

In summary, all the key elements of a Markov decoder for short line segments are present in the neural circuitry of primary visual cortex. One essential feature, namely monocular processing, is no longer available beyond V1. We therefore propose that V1 functions as a dynamic network to accumulate information on fine stimulus features in the face of fixational eye movements.

Human Performance versus Model Performance

We presented psychophysical results indicating that human subjects could reliably discriminate between horizontal and vertical stimuli measuring 1×2 arcmin (100% accuracy; Figure 2), but that the task was barely achievable when the stimulus was half that size (70% accuracy). Using biologically reasonable parameters, a Markov decoder of retinal spike trains attains comparable, but slightly weaker, performance (90% and 60%, respectively; Figure 5). What additional information do humans have that might account for this discrepancy? Here, we consider several aspects of realistic visual processing that were ignored by the Markov decoder.

We treated only Off-type retinal ganglion cells, but there are equally many On-type cells in the fovea, and in principle, they could also contribute to discrimination. An On cell is suppressed when a small, dark stimulus on a light background enters its receptive field, and is then excited when the stimulus exits. These responses are unreliable because the reduction in firing rate from the background of 10 Hz is detectable only after 100 ms of silence, and the excitatory response is slow and weak. We explored this further with explicit simulation of both On and Off cells: The decoder performance improved very little (unpublished data), less than required to fully account for human acuity.

Human fixational eye movements are not exactly random walks. Instead, they exhibit some small persistence of velocity on a timescale of 2 ms and antipersistence on a timescale of 100 ms [25,73]. To explore how these details affect the Markov

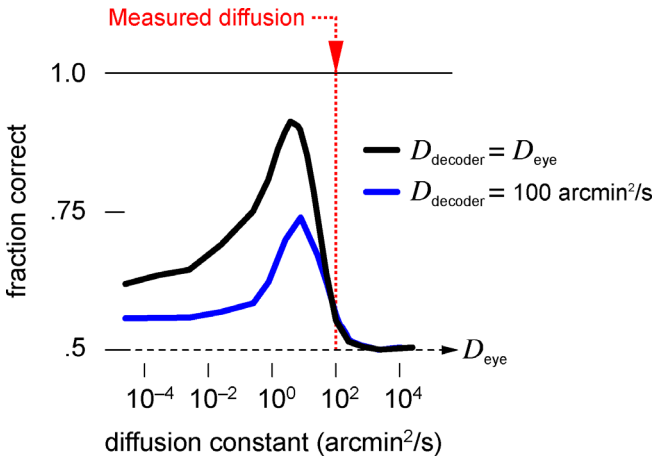

Figure 9. Markov Decoder Discrimination Performance as a Function of Eye Movement Diffusion Constant

The decoder's assumed diffusion constant is either held fixed (blue) or covaried with that of the eye (black). The measured diffusion constant for eye movements is marked in red. These simulations used a biphasic filter with perfectly matched positive and negative lobes, which is the filter that most favors large eye movements. The stimulus measured 0.5 × 1 arcmin^2; otherwise, parameters were as in Figure 5.

doi:10.1371/journal.pbio.0050331.g009

decoder's discrimination performance, we performed additional simulations. Antipersistence at long times can be explained by occasional microsaccades that periodically deflect the eye toward its starting position. Such occasional jerks of the image hardly affected a Markov decoder ignorant of microsaccades (unpublished data): After each stimulus jump, there was only a slight delay until the tails of the diffusing posterior distribution encountered the elevated spike rate at the new stimulus location. The persistence of eye movements at short times is consistent with velocity correlations lasting just a few milliseconds [73]. For a given diffusion constant, a persistent random walk lingers longer at each retinal location than a pure Markov random walk, leading to slightly stronger responses. Correspondingly, simulations showed that the Markov decoder's performance improves modestly with the introduction of a short persistence time (unpublished data).

As discussed above, the Markov decoder is suboptimal because of the temporal blurring of the stimulus before spike generation. The optimal decoder must keep track of all possible histories affecting the current firing rate, rather than only the last stimulus position, and the computational effort rapidly becomes prohibitive. Strategies have been proposed to simplify the decoding of such processes [74,75], but these require complicated learning algorithms and do not lend themselves to straightforward neural implementation. A simpler strategy for improving the Markov decoder might be to first process the retinal spike trains with a temporal filter designed to "undo" temporal integration in the retina. This could plausibly take place in the thalamus [76].

Finally, the real visual system enjoys two additional benefits that were not available to the Markov decoder. The first is global image motion: Our human observers viewed the tiny bar stimuli on a white sheet posted within a laboratory scene. As the eye moves, this peripheral background image moves coherently upon the retina, providing additional global motion cues that the brain could perhaps incorporate to improve perception. Second, our model for retinal responses used the most-random spike pattern for a given firing rate, namely a Poisson process. By contrast, real retinal ganglion cells fire more precisely [11,12] and could thus be more informative, even for a Markov decoder.

Are Fixational Eye Movements Helpful or Harmful?

One commonly held view is that fixational eye movements actually improve vision by preventing the decay of retinal responses that occurs under static stimuli [20]. For example, Rucci and Desbordes have demonstrated that for moderately large, noisy stimuli, orientation discrimination is worse when the image is stabilized on the retina, a result they attribute to a loss of the image motion that would otherwise refresh, and possibly structure, neural activity [77]. In contrast, here we have described these eye movements as a hindrance rather than a help. The transient nature of retinal ganglion cell responses does imply that a fixed stimulus will elicit fewer spikes than a moving stimulus, diminishing the signal that the brain receives. But if the eye movements are too large, then the light intensity is spread thinly over many cells, decreasing each individual response while increasing the positional uncertainty and thus the noise [78]. Between the limits of no eye movement and very large eye movements, an optimum exists. This should occur with eye movements that shift the stimulus to a new set of retinal ganglion cells just as the initial response starts to truncate, and no sooner. For a stimulus area s and transient response duration τ, this occurs when the diffusion constant is $D \sim s/4\tau$. For tiny stimuli ($s = 0.5 \times 1$ arcmin^2) and biphasic temporal kernels with $\tau = 35$ ms, the predicted optimum of $D \sim 3$ arcmin^2/s is more than a full order of magnitude smaller than the naturally occurring eye movements of approximately 100 arcmin^2/s.

To explore this further, we computed the Markov decoder's performance as a function of the eye movement diffusion constant (Figure 9). In one condition, the decoder's assumption about the diffusion constant is held fixed while the eye movement statistics vary; this models a psychophysical experiment in which a viewer's gaze is artificially stabilized. In another condition, the decoder's assumed diffusion constant varies to match the eye movement statistics, approximately optimizing the decoder performance. In both cases, there is an optimum for D near the value predicted above, and the model acuity is dramatically worse than this optimum when the natural diffusion statistics are used. Natural eye movements are therefore substantially larger than optimal for this fine acuity task, implying that they do indeed present a problem for fine visual acuity that the brain must solve.

Predictions

The Markov decoder model yields psychophysical and physiological predictions. We argued that fixational eye movements are unknown to the brain, so using an eye tracker to replace the natural fixational eye movements with exogenous jitter movements, such as eye trajectories recorded from a previous trial, should not affect fine acuity, a prediction supported by recent evidence [79]. We also argued that, for very small stimuli on a featureless background, natural eye movements are larger than optimal: therefore, partially stabilizing the retinal image should improve our finest acuity so long as enough motion remains to avoid

prematurely truncating retinal responses (Figure 9). Although discrimination of larger stimuli does benefit from eye movements [79], there are indications that fine acuity is improved by stabilization [80].

There are two major physiological predictions. First, activity in V1 neurons should locally modulate the gain for feedforward input originating from the retina. Without this modulation, the advantage of using prior expectations is lost. Second, if the neural interactions in V1 are to correctly encode the probabilistic expectations given by random walk eye movement statistics, then the interactions should implement a diffusion operator, which entails that the time delay to reach maximal interaction strength should scale as the square of the interaction distance. This should be observable both directly, as lateral excitatory currents, and indirectly, through the time course of the resulting gain modulation.

The Bayesian Framework

The essential aspect of the Markov decoder we have described is that information of one type attunes the observer to other, related information. In the present context, the decoder expects that responses to oriented line segments are correlated across space and time due to fixational eye movements, and thus these expected responses are enhanced. Other statistical regularities produce expectations as well. For example, strings of line segments often occur together in contours. Correspondingly, collinear iso-orientation facilitation has been hypothesized to subserve contour integration [41,81], and can be viewed as another instance of the principle of enhancing responses to expected signals. More generally, expectations should increase the gain for information that is relevant to the current task, but when that information is irrelevant, then expectations may instead reduce the gain.

The probabilistic processing of information has generated substantial interest as a general framework for neural computation, often designated "Bayesian computation" due to the use of Bayes' rule in calculating probabilities. Human perception has been shown in several conditions to behave according to this rule [82–84]. Experimental evidence also hints that the cortex may be implementing Bayesian inference on a neural level [85]. Modeling studies have suggested how networks of neurons could make these probabilistic inferences [75,86–89]. One study of particular relevance also describes a neural network for approximately Bayesian decoding of arbitrary hidden Markov processes [90].

Although our mathematical formalism is closely related to previous work, we have made several advances in applying the Bayesian paradigm. First, we identified a concrete biological puzzle of considerable practical importance: how can humans see with high acuity when fixational eye movements rapidly jitter the stimulus over a large area? Second, previous Bayesian computations treated neural signals that were poorly constrained by experiment, so the performance of these computations could be characterized only qualitatively. In contrast, retinal signals are well studied, enabling us to make quantitative comparisons between model and human performance. Third, previous studies predominantly described the formal structure of Bayesian computations, whereas we identified a simple and biologically plausible mapping of the probabilistic calculations onto cortical circuitry.

Outlook

The decoder we have described is optimized for discriminating the orientation of line segments, but human acuity extends to more complex tasks, such as telling "F" from "P." Within our formalism, optimal discrimination of arbitrary shapes would require receptive fields tuned to those shapes, whereas the early visual system appears to encode oriented edges, with more complex feature selectivity arising only later in higher brain regions. Therefore, this Markov decoder by itself cannot account for discrimination in complex acuity tasks. However, we propose that it functions as a useful preprocessor that reduces the confounding effects of fixational eye movements before passing signals to subsequent cortical regions for high-level processing.

If the stimulus contains several lines of multiple orientations, the decoder's output will have several peaks that correspond to the individual oriented segments. These peaks will track the stimulus pattern as it is scanned over the retina. This output can then be processed by subsequent networks tuned to more complex patterns. Simulations show that such a pattern detector identifies an arrangement of oriented bars better when it is provided with the output of a Markov decoder than with signals from similar decoders that fail to properly account for eye movements (see figure in Protocol S1). Thus, the Markov decoder elaborates the conventional model of V1 as extracting oriented image elements, and improves over this static model through dynamic processing that partially corrects for eye movements.

In real-world acuity tasks, we do not perceive the incessant motion of the image upon our retinas, but rather perceive a stable image in world coordinates. Nonetheless, our internal representation early in the visual pathway stores visual information in a retinal coordinate system [91]. This moving frame of reference must eventually be superseded before our stable perceptions arise and decisions are reached. The network proposed here could be viewed as creating an intermediate coordinate system: the most current information is represented in retinal coordinates, but the nonlinear operations of the network effectively shift the past retinal coordinates into improved alignment. We may view this neural computation as a step towards invariant world coordinates.

Materials and Methods

Psychophysics. Three groups of small horizontal and vertical stimuli like those in Figure 2 were presented at a distance of 4 m and were scaled to subtend the angles 0.5×1, 0.75×1.5, and 1×2 arcmin^2. Stimuli were printed in black ink on white paper. Ambient lighting generated a luminance of 86 candelas/m^2 for the white background, and 20-fold dimmer for the black stimuli. Room features provided global motion cues, which we did not seek to eliminate. Nine subjects were asked to discriminate between the stimuli while standing, and were not provided with error feedback. Subjects were free to view the stimuli as long as they liked, typically taking a few seconds per stimulus. Performance was reported as the fraction of correct answers out of 32 attempts for each condition. Error bars were given as 68% confidence interval around the mean, assuming a binomial distribution of correct guesses and a uniform prior over the fraction correct. Other experiments with briefly flashed stimuli showed that reliable discrimination was already achieved within 500 ms (unpublished data).

Simulations. We generated model retinal responses for the discrimination task in the following steps: the stimulus orientation S was chosen randomly to be either horizontal or vertical, a random walk trajectory was constructed, and the stimulus light intensity profile was moved along this random walk trajectory; the dynamic

light intensity at each retinal position was filtered by a temporal kernel, then passed through a threshold rectifier to yield the instantaneous firing rate; this rate drove an inhomogeneous Poisson generator to produce the spike train for the retinal neuron at that location. We passed these spikes to the Markov decoder implementing Equation 1, which returned a guess of the stimulus identity. These steps are depicted in Figure 3A and described in detail below.

In both the simulations of retinal spike trains and in the Markov decoder, we modeled the fovea as a square lattice of cone photoreceptors. In the human retina, cones are spaced every 0.5 arcmin, and the receptive fields of retinal ganglion cells each consist of a single cone. Correspondingly, the model ganglion cells had square receptive fields separated by 0.5 arcmin. For numerical work, we simulated a 16×16 arcmin^2 array with wraparound boundary conditions, which was sufficiently large for the relevant values of the diffusion constant and the diffusion time, yet small enough for fast simulations.

The stimulus itself consisted of a rectangle with size z and a 1×2 aspect ratio oriented in either the vertical or horizontal direction. Optical blur was produced by convolving the stimulus with a Gaussian modulation transfer function of diameter $2\sigma = 0.5$ arcmin [10]. The stimulus at location $\mathbf{x}$ induced an instantaneous spatial light absorption profile at retinal positions $\mathbf{y}$ of

$$I_S(\mathbf{y}, \mathbf{x}) = e^{-|\mathbf{y}|^2/2\sigma^2} \circ U_{1,1}(\mathbf{y}) \circ U_{z,2z}(\mathbf{x} - \mathbf{y}), \tag{5}$$

where $\circ$ denotes a convolution operation, and $U_{a,b}(\mathbf{x})$ represents a two-dimensional box profile with dimensions a and b. The resultant stimulus profile is shown in Figure 3B.

We modeled fixational eye movements as a random walk that shifts the stimulus across the retina. The one-sided power spectrum of a one-dimensional random walk is given by $D/\pi^2 f^2$, where f is the temporal frequency. Eizenman et al. [3] reported one-sided power spectra with f^{-2} dependence for the horizontal component of fixational eye movements, from which we inferred a two-dimensional diffusion constant of $D = 100$ arcmin^2/s. Corroborating results come from direct measurements of squared eye displacement as a function of time lag [25]; fitting these data with a straight line of slope $4D$ expected from a random walk yielded diffusion constants of the same magnitude, 100 arcmin^2/s.

We simulated the trajectory of the stimulus as a random walk on a discrete spatial lattice, but continuous in time. After an infinitesimal time interval dt, the probability of stepping to a nearest neighbor location is $dt \cdot D/a^2$, where D is the diffusion constant, and a the distance between lattice points. After many such time steps over a finite interval Δt, the probability that the walker has moved a distance Δx horizontally and Δy vertically can be expressed in series form:

$$P(\Delta x, \Delta y, \Delta t) = F(\Delta x, \Delta t) \cdot F(\Delta y, \Delta t)$$

$$F(\Delta x, \Delta t) = \frac{1}{N} \sum_{j=0}^{N-1} \exp\left(i 2\pi \frac{j \Delta x}{N a}\right) \exp\left[-\frac{2D\Delta t}{a^2}\left(1 - \cos 2\pi \frac{j}{N}\right)\right], \tag{6}$$

where N is the number of points on a side of the square lattice. For speedy simulations, we chose a constant sampling interval $\Delta t = 0.7$ ms and drew independent random walk steps from this distribution; finer temporal sampling produced nearly identical results (unpublished data).

The spatial stimulus profile was moved around the model retina according to the random walk. This produced a temporal sequence of light intensities within each retinal ganglion cell's receptive field, which was then convolved with the parameterized biphasic temporal filter (Figure 3D)

$$h(t) = \frac{t^n}{\tau_1^{n+1}} e^{-t/\tau_1} - \rho \frac{t^n}{\tau_2^{n+1}} e^{-t/\tau_2} \tag{7}$$

to produce a temporally blurred stimulus (Figure 4F). The parameters were chosen as $\tau_1 = 5$ ms, $\tau_2 = 15$ ms, $n = 3$, and $\rho = 0.8$ for all simulations [28] except Figure 9, for which $\rho = 1$ to maximize the performance improvement attributable to eye movements. Finally, this spatiotemporal profile was offset by the background firing rate r_0, half-wave rectified to prevent negative firing rates, and scaled so that the maximum possible firing rate was given by $r_{\max}$. The typical firing-rate parameters we used were $r_0 = 10$ Hz and $r_{\max} = 100$ Hz unless otherwise specified.

The Markov decoder operated on one trial of all ganglion cell spike trains to produce a guess for the stimulus identity, according to the differential equation (Equation 1). This equation can be solved iteratively, moving from spike to spike. When neuron $\mathbf{y}$ produces a spike at time $t_\mathbf{y}$, the diffusion term (Equation 4) is negligible compared to the spiking term (Equation 2), so we have only

$$\frac{\partial}{\partial t} P(S, \mathbf{x}, t) = \delta(t - t_\mathbf{y}) f_S(\mathbf{y} - \mathbf{x}) P(S, \mathbf{x}, t) \tag{8}$$

Dividing both sides by $P(S, \mathbf{x}, t)$ and substituting $f_S(\mathbf{x}) = \ln(r_S(\mathbf{x}/r_0)$, we see that

$$\frac{\partial}{\partial t} \ln P(S, \mathbf{x}, t) = \delta(t - t_\mathbf{y}) \ln \frac{r_S(\mathbf{y} - \mathbf{x})}{r_0}. \tag{9}$$

Integrating the delta function over the spike from time $t_\mathbf{y}^-$ to time $t_\mathbf{y}^+$ we find that the log-probability jumps at spike times by $\ln(r_S(\mathbf{x}/r_0)$, which means that the probability itself is multiplied:

$$P(S, \mathbf{x}, t_\mathbf{y}^+) = \frac{r_S(\mathbf{y} - \mathbf{x})}{r_0} P(S, \mathbf{x}, t_\mathbf{y}^-) \tag{10}$$

In the absence of spikes, only the terms of Equations 3 and 4 contribute to the differential equation (Equation 1), so the probability distribution $P(S, \mathbf{x}, t)$ both decays and diffuses laterally across space. Because the two oriented stimuli both produce the same total spike rate from the retinal array regardless of position, the decay term (Equation 3) does not alter the relative probabilities, and we therefore neglect it. The diffusion term (Equation 4) can be implemented most efficiently in the spatial frequency domain $\tilde{P}(S, \mathbf{k}, t)$, where the diffusion operator $D\tilde{\nabla}^2$ simply multiplies its operand. The solution to

$$\frac{\partial}{\partial t} \tilde{P}(S, \mathbf{k}, t) = D\tilde{\nabla}^2 \tilde{P}(S, \mathbf{k}, t) \tag{11}$$

during a spike-free interval $[t, t + \Delta t]$ is

$$\tilde{P}(S, \mathbf{k}, t + \Delta t) = \exp\left[D\Delta t \tilde{\nabla}^2\right] \tilde{P}(S, \mathbf{k}, t). \tag{12}$$

For computational speed, we sampled the decoder's activity every 0.7 ms. Between samples, the probability distribution was multiplied in the Fourier domain according to Equation 12, and at the sample times, the probabilities were multiplied in the spatial domain following Equation 10: once for each spike that occurred since the last sample time. Thus we were able to execute the ideal observer algorithm by multiplication alternately in the spatial domain and the frequency domain. To ensure stability in the absence of the decay term (Equation 3), at every sampling time, we rescaled the posterior probability by its sum, $\sum_{S,\mathbf{x}} P(S, \mathbf{x}, t)$, recovering a properly normalized probability.

These estimated posterior probabilities can be displayed as a function of space and time, as in Figure 4. Or to reach a decision in the discrimination task, we summed the probabilities over all positions after the specified stimulus duration T to obtain the posterior probability for orientation, $P(S, T)$; the orientation with the greatest probability counted as the decoder's guess. By repeating this process many times (10^4 iterations) and calculating the fraction of correct trials, we quantified the performance for this ideal strategy for various parameter sets, as plotted in Figures 5, 6, and 9.

Supporting Information

Protocol S1. The Derivation of the Markov Decoder Equation

Found at doi:10.1371/journal.pbio.0050331.sd001 (1.2 MB PDF).

Acknowledgments

The authors thank Ralf Engbert and Reinhold Kliegl for their eye movement data, and Daniel Fisher, Maneesh Sahani, and an anonymous referee for helpful conversations and suggestions.

Author contributions. XP, HS, and MM conceived and designed the experiments, analyzed the data, and wrote the paper. XP performed the experiments.

Funding. XP and MM were supported by a National Institutes of Health grant. The work of HS was partially supported by a grant of the US-Israel Binational Science Foundation.

Competing interests. The authors have declared that no competing interests exist.

References

1. Skavenski AA, Hansen RM, Steinman RM, Winterson BJ (1979) Quality of retinal image stabilization during small natural and artificial body rotations in man. Vision Res 19: 675–683.
2. Tomlinson RD (1990) Combined eye-head gaze shifts in the primate. III. Contributions to the accuracy of gaze saccades. J Neurophysiol 64: 1873–1891.
3. Eizenman M, Hallett PE, Frecker RC (1985) Power spectra for ocular drift and tremor. Vision Res 25: 1635–1640.
4. Guthrie BL, Porter JD, Sparks DL (1983) Corollary discharge provides accurate eye position information to the oculomotor system. Science 221: 1193–1195.
5. Donaldson IM (2000) The functions of the proprioceptors of the eye muscles. Philos Trans R Soc Lond B Biol Sci 355: 1685–1754.
6. Murakami I, Cavanagh P (1998) A jitter after-effect reveals motion-based stabilization of vision. Nature 395: 798–801.
7. Murakami I, Cavanagh P (2001) Visual jitter: evidence for visual-motion-based compensation of retinal slip due to small eye movements. Vision Res 41: 173–186.
8. Schein SJ (1988) Anatomy of macaque fovea and spatial densities of neurons in foveal representation. J Comp Neurol 269: 479–505.
9. Geisler WS (1984) Physical limits of acuity and hyperacuity. J Opt Soc Am A 1: 775–782.
10. Geisler WS, Davila KD (1985) Ideal discriminators in spatial vision: two-point stimuli. J Opt Soc Am A 2: 1483–1497.
11. Berry MJ 2nd, Meister M (1998) Refractoriness and neural precision. J Neurosci 18: 2200–2211.
12. Uzzell VJ, Chichilnisky EJ (2004) Precision of spike trains in primate retinal ganglion cells. J Neurophysiol 92: 780–789.
13. Chichilnisky EJ, Rieke F (2005) Detection sensitivity and temporal resolution of visual signals near absolute threshold in the salamander retina. J Neurosci 25: 318–330.
14. Hennig MH, Wörgötter F (2004) Eye micro-movements improve stimulus detection beyond the nyquist limit in the peripheral retina. Adv Neural Inf Process Syst 16: 1475–1482.
15. Wachtler T, Wehrhahn C, Lee BB (1996) A simple model of human foveal ganglion cell responses to hyperacuity stimuli. J Comput Neurosci 3: 73–82.
16. Croner L, Kaplan E (1995) Receptive fields of P and M ganglion cells across the primate retina. Vision Res 35: 7–24.
17. Shapley RM, Victor JD (1978) The effect of contrast on the transfer properties of cat retinal ganglion cells. J Physiol 285: 275–298.
18. Frechette ES, Sher A, Grivich MI, Petrusca D, Litke AM, et al. (2005) Fidelity of the ensemble code for visual motion in primate retina. J Neurophysiol 94: 119–135.
19. Troy JB, Lee BB (1994) Steady discharges of macaque retinal ganglion cells. Vis Neurosci 11: 111–118.
20. Martinez-Conde S, Macknik SL, Hubel DH (2004) The role of fixational eye movements in visual perception. Nat Rev Neurosci 5: 229–240.
21. Winterson BJ, Collewijn H (1976) Microsaccades during finely guided visuomotor tasks. Vision Res 16: 1387–1390.
22. Kowler E, Steinman RM (1979) Miniature saccades: eye movements that do not count. Vision Res 19: 105–108.
23. Bridgeman B, Palca J (1980) The role of microsaccades in high acuity observational tasks. Vision Res 20: 813–817.
24. Martinez-Conde S (2006) Fixational eye movements in normal and pathological vision. Prog Brain Res 154: 151–176.
25. Engbert R, Kliegl R (2004) Microsaccades keep the eyes' balance during fixation. Psychol Sci 15: 431–436.
26. Turing AM (1952) The chemical basis of morphogenesis. Phil Trans Royal Soc Lon B 237: 37–72.
27. Schneeweis DM, Schnapf JL (1999) The photovoltage of macaque cone photoreceptors: adaptation, noise, and kinetics. J Neurosci 19: 1203–1216.
28. Chichilnisky EJ, Kalmar RS (2002) Functional asymmetries in ON and OFF ganglion cells of primate retina. J Neurosci 22: 2737–2747.
29. Hubel DH, Wiesel TN (1959) Receptive fields of single neurones in the cat's striate cortex. J Physiol 148: 574–591.
30. Dow BM, Snyder AZ, Vautin RG, Bauer R (1981) Magnification factor and receptive field size in foveal striate cortex of the monkey. Exp Brain Res 44: 213–228.
31. Snodderly DM, Gur M (1995) Organization of striate cortex of alert, trained monkeys (Macaca fascicularis): ongoing activity, stimulus selectivity, and widths of receptive field activating regions. J Neurophysiol 74: 2100–2125.
32. Talbot SA, Marshall WH (1941) Physiological studies on neural mechanisms of visual localization and discrimination. Amer J Ophthal 24: 1255–1263.
33. Tootell RB, Switkes E, Silverman MS, Hamilton SL (1988) Functional anatomy of macaque striate cortex. II. Retinotopic organization. J Neurosci 8: 1531–1568.
34. Steinman RM, Collewijn H (1980) Binocular retinal image motion during active head rotation. Vision Res 20: 415–429.
35. Hubel DH, Wiesel TN, Stryker MP (1978) Anatomical demonstration of orientation columns in macaque monkey. J Comp Neurol 177: 361–380.
36. Polat U, Sagi D (1993) Lateral interactions between spatial channels: suppression and facilitation revealed by lateral masking experiments. Vision Res 33: 993–999.
37. Xing J, Heeger DJ (2000) Center-surround interactions in foveal and peripheral vision. Vision Res 40: 3065–3072.
38. Adini Y, Sagi D, Tsodyks M (1997) Excitatory-inhibitory network in the visual cortex: psychophysical evidence. Proc Natl Acad Sci U S A 94: 10426–10431.
39. Bosking WH, Zhang Y, Schofield B, Fitzpatrick D (1997) Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. J Neurosci 17: 2112–2127.
40. Gilbert CD, Wiesel TN (1989) Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. J Neurosci 9: 2432–2442.
41. Sincich LC, Blasdel GG (2001) Oriented axon projections in primary visual cortex of the monkey. J Neurosci 21: 4416–4426.
42. Malach R, Amir Y, Harel M, Grinvald A (1993) Relationship between intrinsic connections and functional architecture revealed by optical imaging and in vivo targeted biocytin injections in primate striate cortex. Proc Natl Acad Sci U S A 90: 10469–10473.
43. Ts'o DY, Gilbert CD, Wiesel TN (1986) Relationships between horizontal interactions and functional architecture in cat striate cortex as revealed by cross-correlation analysis. J Neurosci 6: 1160–1170.
44. Kapadia MK, Ito M, Gilbert CD, Westheimer G (1995) Improvement in visual sensitivity by changes in local context: parallel studies in human observers and in V1 of alert monkeys. Neuron 15: 843–856.
45. Polat U, Mizobe K, Pettet MW, Kasamatsu T, Norcia AM (1998) Collinear stimuli regulate visual responses depending on cell's contrast threshold. Nature 391: 580–584.
46. Grinvald A, Lieke EE, Frostig RD, Hildesheim R (1994) Cortical point-spread function and long-range lateral interactions revealed by real-time optical imaging of macaque monkey primary visual cortex. J Neurosci 14: 2545–2568.
47. Hubel DH, Wiesel TN (1974) Uniformity of monkey striate cortex: a parallel relationship between field size, scatter, and magnification factor. J Comp Neurol 158: 295–305.
48. Wilson JR, Sherman SM (1976) Receptive-field characteristics of neurons in cat striate cortex: changes with visual field eccentricity. J Neurophysiol 39: 512–533.
49. Van Essen DC, Newsome WT, Maunsell JH (1984) The visual field representation in striate cortex of the macaque monkey: asymmetries, anisotropies, and individual variability. Vision Res 24: 429–448.
50. Azzopardi P, Cowey A (1993) Preferential representation of the fovea in the primary visual cortex. Nature 361: 719–721.
51. Azzopardi P, Cowey A (1996) The overrepresentation of the fovea and adjacent retina in the striate cortex and dorsal lateral geniculate nucleus of the macaque monkey. Neuroscience 72: 627–639.
52. McAdams CJ, Maunsell JH (1999) Effects of attention on the reliability of individual neurons in monkey visual cortex. Neuron 23: 765–773.
53. Treue S, Martinez Trujillo JC (1999) Feature-based attention influences motion processing gain in macaque visual cortex. Nature 399: 575–579.
54. Koch C, Poggio T (1992) Multiplying with synapses and neurons. In: McKenna T, Davis J, Zornetzer S, editors. Single neuron computation. Boston: Academic Press. pp. 315–345.
55. Mel BW (1993) Synaptic integration in an excitable dendritic tree. J Neurophysiol 70: 1086–1101.
56. Koch C, Segev I (2000) The role of single neurons in information processing. Nat Neurosci 3: 1171–1177.
57. Chance FS, Abbott LF, Reyes AD (2002) Gain modulation from background synaptic input. Neuron 35: 773–782.
58. Murphy BK, Miller KD (2003) Multiplicative gain changes are induced by excitation or inhibition alone. J Neurosci 23: 10040–10051.
59. Mehaffey WH, Doiron B, Maler L, Turner RW (2005) Deterministic multiplicative gain control with active dendrites. J Neurosci 25: 9968–9977.
60. Gabbiani F, Krapp HG, Koch C, Laurent G (2002) Multiplicative computation in a visual neuron sensitive to looming. Nature 420: 320–324.
61. Fox K, Sato H, Daw N (1990) The Effect of varying stimulus intensity on NMDA-receptor activity in cat visual cortex. J Neurophysiol 64: 1413–1428.
62. Rivadulla C, Sharma J, Sur M (2001) Specific roles of NMDA and AMPA receptors in direction-selective and spatial phase-selective responses in visual cortex. J Neurosci 21: 1710–1719.
63. Heeger DJ (1992) Normalization of cell responses in cat striate cortex. Vis Neurosci 9: 181–197.
64. Carandini M, Heeger DJ, Movshon JA (1997) Linearity and normalization in simple cells of the macaque primary visual cortex. J Neurosci 17: 8621–8644.
65. Cavanaugh JR, Bair W, Movshon JA (2002) Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. J Neurophysiol 88: 2530–2546.
66. Webb BS, Dhruv NT, Solomon SG, Tailby C, Lennie P (2005) Early and late mechanisms of surround suppression in striate cortex of macaque. J Neurosci 25: 11666–11675.
67. Lee DK, Itti L, Koch C, Braun J (1999) Attention activates winner-take-all competition among visual filters. Nat Neurosci 2: 375–381.
68. Coultrip R, Granger RH, Lynch G (1992) A cortical model of winner-take-all competition via lateral inhibition. Neural Netw 5: 47–54.
69. Antón PS, Granger RH, Lynch G (1992) Temporal information processing

in synapses, cells, and circuits. In: McKenna T, Davis J, Zornetzer S, editors. Single neuron computation. Boston: Academic Press. pp. 291–313.
70. Shadlen MN, Newsome WT (1998) The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. J Neurosci 18: 3870–3896.
71. Callaway EM (2005) Structure and function of parallel pathways in the primate early visual system. J Physiol 566: 13–19.
72. Shriki O, Hansel D, Sompolinsky H (2003) Rate models for conductance-based cortical neuronal networks. Neural Comput 15: 1809–1841.
73. Mergenthaler K, Engbert R (2007) Modeling the control of fixational eye movements with neurophysiological delays. Phys Rev Lett 98: 138104.
74. Huys QJM, Zemel RS, Natarajan R, Dayan P (2007) Fast population coding. Neural Comput 19: 460–497.
75. Zemel RS, Huys QJM, Natarajan R, Dayan P (2005) Probabilistic computation in spiking populations. In: Saul L, Weiss Y, Bottou L, editors. Advances in neural information processing systems 17. Cambridge (Massachusetts): MIT Press. pp. 1609–1616.
76. Dong DW, Atick JJ (1995) Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus. Network 6: 159–178.
77. Rucci M, Desbordes G (2003) Contributions of fixational eye movements to the discrimination of briefly presented stimuli. J Vis 3: 852–864.
78. Pelli DG (1985) Uncertainty explains many aspects of visual contrast detection and discrimination. J Opt Soc Am A 2: 1508–1532.
79. Rucci M, Iovin R, Poletti M, Santini F (2007) Miniature eye movements enhance fine spatial detail. Nature 447: 851–854.
80. Riggs LA, Ratliff F, Cornsweet JC, Cornsweet TN (1953) The disappearance of steadily fixated visual test objects. J Opt Soc Am 43: 495–501.
81. Sigman M, Cecchi GA, Gilbert CD, Magnasco MO (2001) On a common circle: natural scenes and Gestalt rules. Proc Natl Acad Sci U S A 98: 1935–1940.
82. Brainard DH, Longere P, Delahunt PB, Freeman WT, Kraft JM, et al. (2006) Bayesian model of human color constancy. J Vis 6: 1267–1281.
83. Kording KP, Ku SP, Wolpert DM (2004) Bayesian integration in force estimation. J Neurophysiol 92: 3161–3165.
84. Stocker AA, Simoncelli EP (2006) Noise characteristics and prior expectations in human visual speed perception. Nat Neurosci 9: 578–585.
85. Shadlen MN, Newsome WT (2001) Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. J Neurophysiol 86: 1916–1936.
86. Deneve S, Latham PE, Pouget A (2001) Efficient computation and cue integration with noisy population codes. Nat Neurosci 4: 826–831.
87. Deneve S (2005) Bayesian inference in spiking neurons. Adv Neural Inf Process Syst 17: 353–360.
88. Rao RP (2005) Hierarchical Bayesian inference in networks of spiking neurons. Adv Neural Inf Process Syst 17: 1113–1120.
89. Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. Nat Neurosci 9: 1432–1438.
90. Rao RP (2004) Bayesian computation in recurrent neural circuits. Neural Comput 16: 1–38.
91. Gur M, Snodderly DM (1997) Visual receptive fields of neurons in primary visual cortex (V1) move in space with the eye movements of fixation. Vision Res 37: 257–265.
92. Roorda A, Williams DR (1999) The arrangement of the three cone classes in the living human eye. Nature 397: 520–522.