LETTER Communicated by Marcus Meister

Asymmetric Dynamics in Optimal Variance Adaptation

Michael DeWeese
Sloan Center, Salk Institute, La Jolla, CA 92037, U.S.A.

Anthony Zador
Salk Institute MNL/S, La Jolla, CA 92037, U.S.A.

It has long been recognized that sensory systems adapt to their inputs. Here we formulate the problem of optimal variance estimation for a broad class of nonstationary signals. We show that under weak assumptions, the Bayesian optimal causal variance estimate shows asymmetric dynamics: an abrupt increase in variance is more readily detectable than an abrupt decrease. By contrast, optimal adaptation to the mean displays symmetric dynamics when the variance is held fixed. After providing several empirical examples and a simple intuitive argument for our main result, we prove that optimal adaptation is asymmetrical in a broad class of model environments. This observation makes specific and falsifiable predictions about the time course of adaptation in neurons probed with certain stimulus ensembles.

1 Introduction

Many real-world signals of interest to both biological and synthetic systems are characterized by their large dynamic range. This dynamic range poses a challenge to both wetware and hardware, which are often constrained to operate within a much more limited dynamic range. For example, photoreceptors in the retina saturate over 2 orders of magnitude of light intensity, yet the retina can operate over 10 orders of magnitude (Barlow & Mollon, 1982). How can a device operating over only 2 orders of magnitude deal with signals that span 10? The retina exploits the nonstationary statistics that characterize light intensity in the real world; on short time scales, light intensity tends to fluctuate over a much smaller range. That is, the retina adapts to the mean light intensity. Adaptation is a basic strategy used by other sensory modalities as well, including the auditory and somatosensory systems (Barlow & Mollon, 1982).

The mean is the simplest statistical characteristic, but it is not the only one. A well-designed system might also be expected to adapt to the variance of a signal. For example, if the input-output response function of a system were sigmoidal, the mean might be used to fix the midpoint of the sigmoid, while the variance might determine the slope at the midpoint. Adaptation to

variance (contrast) is well established in the visual system (Shapley, Enroth-Cugell, Bonds, & Kirby, 1972; Shapley & Enroth-Cugell, 1984; Enroth-Cugell & Robson, 1966; Shapley & Victor, 1978, 1979; Kelly, 1961; deLange, 1958; Giaschi, Douglas, Marlin, & Cynader, 1993; Bonds, 1991; Shapley, 1997).

If a system is to adapt to a signal's nonstationary statistics, it must first estimate those statistics. Optimal estimation of a nonstationary mean is a well-understood problem (Papoulis, 1991). Optimal estimation of a nonstationary variance has received much less attention.

Here we consider the problem of estimating a nonstationary variance. First we define a broad class of processes generated by hidden Markov models that have a well-defined nonstationary variance. We then derive the Bayesian optimal causal[1] estimator of the instantaneous variance. Finally, we show that for many processes within this class, the dynamics of the optimal estimator show an asymmetry: an abrupt increase in variance is more readily detectable than an abrupt decrease. This asymmetry offers a falsifiable test of the hypothesis that sensory systems adapt optimally to nonstationary variance.

2 General Framework

Our goal in this section is to define a class of discrete-time processes s_i with an instantaneous time-varying variance σ_i^2, and then derive the Bayesian optimal causal estimate of the standard deviation at time t_i given its a priori statistics and a realization[2] $s_{j\le i}$ up to t_i. In order to isolate the features of optimal adaptation to variance, we will consider time series with fixed means, and whose third and higher moments are completely determined by the time-varying second moment. In addition, we will study optimal estimation for cases where the mean and variance are covarying. Under both of these conditions, knowing the current value of the variance is tantamount to knowing the whole distribution.

2.1 Markov Generating Process. We begin by writing a discrete-time description of a hidden Markov model with one internal state variable, σ, and one observable output, s:

$$\sigma_i = F_1[\sigma_{i-1}, z_i, u_i, \ldots], \tag{2.1}$$

$$s_i = F_2[\sigma_i, y_i], \tag{2.2}$$

[1] We call our estimator causal since its estimate for the standard deviation at (discrete) time t_i does not depend on its input at any later times t_j, $j > i$.

[2] We use $s_{j<i}$ to indicate all past observations ($\{s_1, s_2, \ldots, s_{i-1}\}$) up to but not including s_i, and $s_{j\le i}$ to indicate all past and present observations ($\{s_1, s_2, \ldots, s_i\}$) up to and including s_i.

where subscripts index time steps and u_i, y_i, z_i, and all other variables appearing after σ_{i-1} in the argument of F_1 are independent and identically distributed (i.i.d.) random variables drawn from their a priori distributions $P(u), P(y), P(z)$, and so forth.[3] The signal s_i will be the input to our estimator. This formulation includes a broad class of models, and it can be generalized to describe arbitrary nonstationary processes by adding more hidden variables. We will consider particular choices of F_2 and $P(y_i)$ in which σ_i^2 can be interpreted as the "instantaneous variance" of the signal, s_i. For example, we will sometimes define $s_i = \sigma_i \times y_i$, where $P(y)$ is a gaussian with unit variance, so that the variance of s_i is given by σ_i^2.

Our task is to use all observations $s_1, s_2, \ldots, s_i$ up to the present time t_i to estimate the current value of σ_i. This would be a trivial problem if s_i were a deterministic function of σ, since then $\sigma = F^{-1}(s)$. We therefore focus on the nontrivial case, where u, y, z, and so forth are stochastic.

2.2 Optimal Bayesian Estimator. We now derive the optimal Bayesian estimator of the standard deviation. Because the generating process (see equation 2.1) is Markovian, at every point in time our knowledge about σ_i is completely summarized by the probability distribution, $P(\sigma_i|s_{j\leq i})$, of σ_i given all previously observed data $s_{j\leq i}$. From this conditional distribution, we can compute specific estimators such as the mean, the mode (most likely value), or any other estimator with whatever properties we choose, but we emphasize that the fundamental object for any estimation task is the distribution $P(\sigma_i|s_{j\leq i})$ itself.

First, we take advantage of the Markovian nature of σ to write an expression for the distribution for the current value of σ given all data up to the last time step as a functional of $P(\sigma_{i-1}|s_{j<i})$:

$$P(\sigma_i|s_{j<i}) = \int d\sigma_{i-1} P(\sigma_i|\sigma_{i-1}) P(\sigma_{i-1}|s_{j<i}), \tag{2.3}$$

where $P(\sigma_i|\sigma_{i-1})$ is the distribution of the current value of σ given its previous value. Next we use Bayes's rule to combine $P(\sigma_i|s_{j<i})$ with a new observation s_i to form $P(\sigma_i|s_{j\leq i})$,

$$\begin{aligned} P(\sigma_i|s_{j\leq i}) &= \frac{P(s_i|\sigma_i, s_{j<i}) P(\sigma_i|s_{j<i})}{P(s_i|s_{j<i})} \\ &= \frac{1}{\Omega} P(s_i|\sigma_i) P(\sigma_i|s_{j<i}), \end{aligned} \tag{2.4}$$

where we have explicitly rewritten the distribution in the denominator as a

[3] We use the convention that all probability distributions are written as $P(\ldots)$; the argument of P determines which probability distribution we mean.

normalization constant, which can be obtained by integrating the numerator:

$$\Omega = \int d\sigma_i P(s_i|\sigma_i) P(\sigma_i|s_{j<i}). \tag{2.5}$$

We also made the substitution $P(s_i|\sigma_i, s_{j<i}) \to P(s_i|\sigma_i)$, which follows from equation 2.2. Combining equations 2.3 and 2.4 yields the recursive formula we seek:

$$P(\sigma_i|s_{j\le i}) = \frac{1}{\Omega} P(s_i|\sigma_i) \int d\sigma_{i-1} P(\sigma_i|\sigma_{i-1}) P(\sigma_{i-1}|s_{j<i}). \tag{2.6}$$

Equation 2.6 is an expression for the distribution of the standard deviation σ_i. It is this distribution that is propagated for further estimates. However, to assess the dynamics of the estimator, it is convenient to compute some scalar function of the distribution, such as its mean or most likely value. In most of the examples that follow, we will use the mean, *but our results do not depend on that choice.*

Since the Bayesian estimation procedure described is central to all that follows, we summarize the procedure:

1. Use the distribution for the standard deviation at t_{i-1} given all data up to that point, $P(\sigma_{i-1}|s_{j<i})$, and the prior distribution for the dynamics of the standard deviation $P(\sigma_i|\sigma_{i-1})$ to compute $P(\sigma_i|s_{j<i})$. Note that $P(\sigma_i|s_{j<i})$ is the distribution for the standard deviation at t_i conditional on all data up to but *not* including the most recent time step t_i.
2. Use Bayes's rule to combine the observed new data point s_i with $P(\sigma_i|s_{j<i})$ to obtain the conditional distribution for the standard deviation at the next time step, $P(\sigma_i|s_{j\le i})$.
3. Compute some scalar from this distribution—for example, its mean or most likely value—to assess dynamics.

3 Examples

In this section we consider two types of model environment: one in which the variance of the input makes abrupt changes, and another in which it changes smoothly with time. In both cases, we will find that the optimal estimator for the variance responds more quickly to increases than decreases.

3.1 Memoryless Variance Modulation: A Simple Example. To illustrate the estimation procedure, we consider a very simple dynamical environment in which the standard deviation, σ, is drawn independently at each time step,

$$\sigma_i = |z_i|, \tag{3.1}$$

$$s_i = \sigma_i y_i. \tag{3.2}$$

Here both z_i and y_i are gaussian i.i.d. variables with zero mean and unit variance, so that for each time step $P(s_i|\sigma_i) = \exp(-s_i^2/2\sigma_i^2)/\sqrt{2\pi\sigma_i^2}$, $-\infty < s_i < \infty$, and $P(\sigma_i) = 2 \times \exp(-\sigma_i^2/2)/\sqrt{2\pi}$, $0 < \sigma_i < \infty$. On average, then, the signal, s, is distributed according to

$$P(s_i) = \int_0^\infty d\sigma_i P(s_i|\sigma_i)P(\sigma_i) = \int_0^\infty d\sigma_i \frac{e^{-s_i^2/2\sigma_i^2-\sigma_i^2/2}}{\pi\sigma_i} = \frac{K_0(|s_i|)}{\pi}, \qquad (3.3)$$

where K_0 is the modified Bessel function of the second kind. Note that by adding a constant to the right-hand side of equation 3.2, we could give s_i a nonzero mean value, but this would change the optimal estimation strategy only in a superficial way, so we will absorb any nonzero mean into our definition for s.

We use Bayes's theorem to compute the conditional distribution for σ_i given the observation s_i,

$$\begin{aligned} P(\sigma_i|s_{j\le i}) = P(\sigma_i|s_i) &= \frac{P(s_i|\sigma_i)P(\sigma_i)}{P(s_i)} \\ &= \frac{e^{-s_i^2/2\sigma_i^2}e^{-\sigma_i^2/2}}{\sigma_i K_0(|s_i|)}. \end{aligned} \qquad (3.4)$$

This is just equation 2.6 for the special case where σ_i is drawn afresh at each step independent of $\sigma_{j<i}$.

Different scalar functions of this distribution can now be compared. The mean of this distribution is given by

$$\bar{\sigma} \equiv \int_0^\infty d\sigma_i \sigma_i P(\sigma_i|s_i) = \sqrt{\pi/2}\frac{e^{-|s_i|}}{K_0(|s_i|)}. \qquad (3.5)$$

For comparison, the maximum likelihood estimator for σ is obtained by solving $dP(\sigma_i|s_i)/d\sigma_i = 0$,

$$\sigma_{m.l.} = \sqrt{-1/2 + \sqrt{1/4 + s_i^2}}. \qquad (3.6)$$

Despite the rather different forms of these expressions, these two estimators agree well for $s \gtrsim 1$ where the distribution is not too asymmetrical about the peak. (However, as $s \to 0$, $\sigma_{m.l.} \to 0$, while $\bar{\sigma}$ remains finite.) The corresponding estimators involving the variance give essentially the same result: $\bar{\sigma} \approx \sqrt{\bar{\sigma^2}}$ and $\sigma_{m.l.} \approx \sqrt{(\sigma^2)_{m.l.}}$. Thus, our results reflect the statistics of the input, but are not sensitive to the exact form of our estimator, as we will discover in the upcoming sections.

In this simple example, we were able to find closed-form analytic expressions for a variety of estimates of σ_i at each time step. In the next few

sections, we consider dynamics that require the optimal estimator to make use of the history of the process. For these cases, we will obtain analytical results for only the first time step. Fortunately, it will be possible to compute the time course of our optimal estimator $\bar{\sigma}$ (and $\sigma_{m.l.}$) numerically for any prior distribution on σ, and any distribution for the signal being used to probe the estimator.

3.2 Variance Switching.

3.2.1 Two Values. We will now derive the optimal estimator for a simple nonstationary environment. Consider a world where the variance switches between a high-value σ^2_{high} and a low-value σ^2_{low}. We will draw the switching times from a homogeneous Poisson process so that they are totally uncorrelated with each other; knowing exactly when switches have occurred in the past gives no information about when they will occur in the future. We will then construct the optimal causal estimator, $\bar{\sigma}_i$, for the time-dependent standard deviation of this process, which minimizes the root mean squared (rms) error. Note that our estimator has access only to the signal, s, not to the underlying switching times of σ, so even though the switching times of the true σ are uncorrelated, our estimator must incorporate the entire past time course of s to do the optimal job. Even in this impoverished environment, we will find that the optimal estimator will behave in a subtle way: it will respond to an increase in the variance more quickly than a decrease.

The process we are interested in can be written as

$$\sigma_i = z_i\sigma_{i-1} + (1 - z_i)(\sigma_{high} + \sigma_{low} - \sigma_{i-1}), \tag{3.7}$$

$$s_i = \sigma_i y_i. \tag{3.8}$$

Here z_i is a binary variable that assumes values 0 with probability x and 1 with probability $1-x$, and y_i is gaussian with zero mean and unit variance. The probability of switching per time step is thus x. The standard deviation σ_i of the signal s_i now has a memory, with a correlation time $\tau_c \sim \frac{1}{2x}$:

$$\langle \sigma_i \sigma_{i+t} \rangle_{P(s_j)} = \frac{(\sigma_{high} + \sigma_{low})^2}{4} + \frac{(\sigma_{high} - \sigma_{low})^2}{4} e^{-2xt}, \tag{3.9}$$

for times long compared to the time step. Sample realizations of σ and s are shown in Figure 1.

The optimal estimator for this simple problem can make use of the fact that σ takes on only two values by expressing the entire distribution $P(\sigma_i|s_{j\le i})$ as a single parameter: $P_i^{low} \equiv P(\sigma_i = \sigma_{low}|s_{j\le i})$. $1 - P_i^{low}$ is then the probability that $\sigma_i = \sigma_{high}$. Since x is the a priori probability that a switch occurred in the last time step,

$$P(\sigma_i = \sigma_{low}|s_{j<i}) = P_{i-1}^{low}(1 - x) + (1 - P_{i-1}^{low})x \tag{3.10}$$

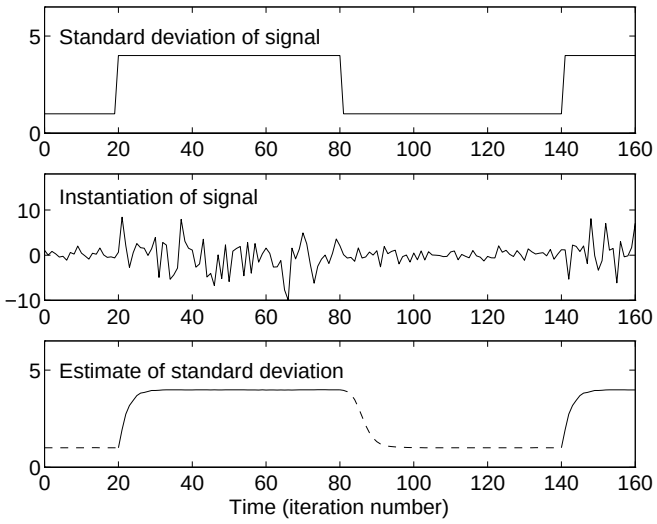

Figure 1: Dynamics of the optimal estimator for two-state variance switching. (Top) The standard deviation during one cycle of the "probe" signal. (Center) A specific instantiation of this signal over the same period of time, generated according to equation 3.7 from the time-varying standard deviation shown at the top; this is the signal s available to the estimator. (Bottom) The optimal causal estimate (obtained from s) of the standard deviation, averaged over 1000 cycles of the signal. For this example, the prior assumes that the signal, s, is gaussian with a standard deviation that randomly switches with probability .001 per time step between $\sigma_{low} = 1$ and $\sigma_{high} = 4$, the same two values used to generate the "probe" (top curve). Notice that the estimate responds more quickly to an increase in the signal's standard deviation (solid portion of curve) than it does after a decrease (dashed curve). This difference in adaptation rates is more apparent in Figure 2.

and $P(\sigma_i = \sigma_{high}|s_{j<i}) = 1 - P(\sigma_i = \sigma_{low}|s_{j<i})$. We now apply Bayes's theorem (see equation 2.4),

$$\begin{aligned} P_i^{low} &= \frac{1}{\Omega} P(\sigma_i = \sigma_{low}|s_{j<i}) P(s_i|\sigma_i = \sigma_{low}) \\ &= \frac{1}{\Omega} \frac{(1-x)P_{i-1}^{low} + x(1-P_{i-1}^{low})}{\sqrt{2\pi(\sigma_{low})^2}} e^{-s_i^2/2(\sigma_{low})^2}, \end{aligned} \tag{3.11}$$

where the normalization is

$$\Omega = \frac{(1-x)P_{i-1}^{low} + x(1-P_{i-1}^{low})}{\sqrt{2\pi(\sigma_{low})^2}} e^{-s_i^2/2(\sigma_{low})^2} +$$

$$\frac{xP_{i-1}^{low} + (1-x)(1-P_{i-1}^{low})}{\sqrt{2\pi(\sigma_{high})^2}} e^{-s_i^2/2(\sigma_{high})^2}. \tag{3.12}$$

Thus, at each time step, equations 3.10 and 3.11 combine the new observation s_i with P_{i-1}^{low} to obtain the updated probability P_i^{low}. Note that for $x = 1/2$, the current estimate for the standard deviation, σ_i, given by equation 3.10, depends on only the current value of the signal s_i, as in the previous section.

The prior in this case assumes that the variance of s switches between two known values at randomly chosen times. The variance spends equal time at the high and low values, and switches between these values are instantaneous in both directions, so the underlying dynamics of σ is totally symmetric under the exchange $\sigma_{low} \leftrightarrow \sigma_{high}$. Despite this, the optimal estimator behaves asymmetrically to increases and decreases in variance. This can be shown analytically for the first time step after abrupt switches in the variance (see the appendix).

To study the entire time course of the adaptation, we use equation 3.11 to compute P_i^{low} numerically at every time step while probing the estimator with a square wave in standard deviation. The estimator is optimized for an *ensemble* of different waveforms of σ (here, the random telegraph signal), but to illustrate the behavior of the estimator, we use a *single* such signal (the square wave). In this case, the probe stimulus is not too unlikely in the estimator's prior distribution, which will not always be the case in later sections. We calculate the estimate for the standard deviation, which minimizes the rms error via

$$\bar{\sigma}_i = P_i^{low}\sigma_{low} + \left(1 - P_i^{low}\right)\sigma_{high}. \tag{3.13}$$

Figure 1 shows the trajectory of this estimate averaged over many periods of the standard deviation square wave. Figure 2 redisplays the portions of the curve immediately following switches in the input standard deviation in a way that makes it easier to compare the response times to abrupt increases and decreases in standard deviation.

The observations s are drawn independently at every time step, so successive presentations of the up and down jumps in standard deviation result in new instantiations of s. It is clear from the figure that the optimal estimate for the standard deviation tracks the upward step faster than the downward step. In the next section, we confirm that this is true for a more complex prior.

3.2.2 Many Values. We now consider a prior in which σ jumps from one value to the next at random times, but unlike the previous example, σ will now assume a spectrum of values rather than just two:

$$\sigma_i = z_i\sigma_{i-1} + (1 - z_i)u_i, \tag{3.14}$$

$$s_i = \sigma_i y_i, \tag{3.15}$$

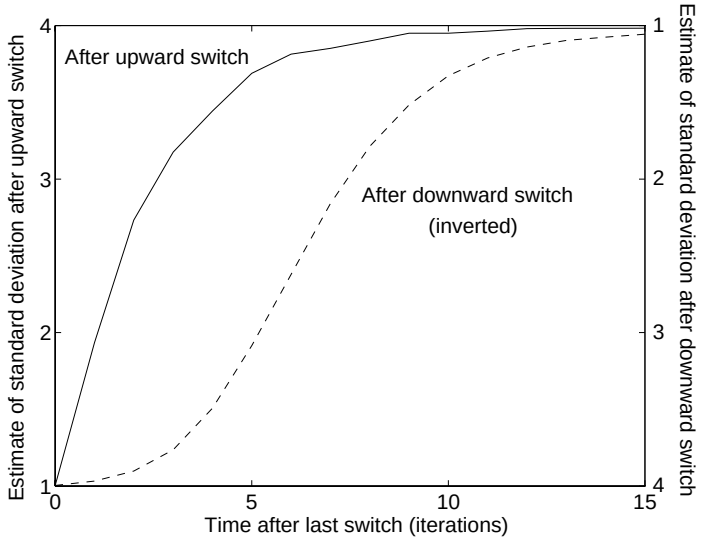

Figure 2: Dynamics of optimal adaptation are faster following an increase than a decrease in variance. Here we have replotted the optimal causal estimate of the standard deviation $\bar{\sigma}$ from the bottom plot of Figure 1 so that the rates of adaptation after increases and decreases can be more easily compared. The solid curve is the average $\bar{\sigma}$ following upward jumps in the signal's standard deviation, as before. The dashed curve is the average $\bar{\sigma}$ following downward jumps, but it has been inverted to make the comparison easier. For both curves, the time of the last switch is set to zero. Clearly, optimal adaptation in this case is faster following abrupt increases in standard deviation than it is after decreases.

where u is uniformly distributed over some finite range of nonnegative values, $[a, b]$, and y and z are gaussian and binary i.i.d., respectively, as in the last section. In other words:

$$P(\sigma_i|\sigma_{i-1}) = xFlat(\sigma_i) + (1-x)\delta(\sigma_i - \sigma_{i-1}), \tag{3.16}$$

where $Flat(\sigma)$ is $1/(b-a)$ whenever $a \leq \sigma \leq b$, and 0 otherwise so that all allowed values for the standard deviation are equally likely. From equation 3.15, s_i is gaussian with variance σ_i^2, so equation 2.6 becomes

$$P(\sigma_i|s_{j\leq i}) = \frac{1}{\Omega}\frac{1}{\sigma_i}\exp\left(-\frac{s_i^2}{2\sigma_i^2}\right)[xFlat(\sigma_i) + (1-x)P(\sigma_{i-1}=\sigma_i|s_{j<i})]. \tag{3.17}$$

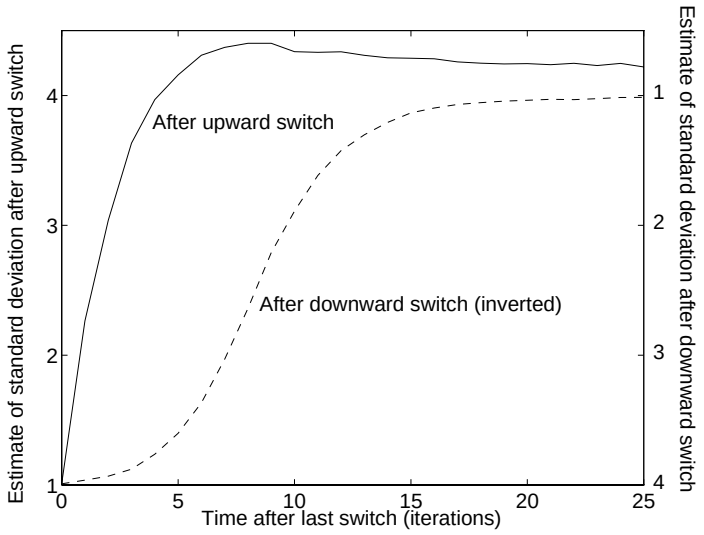

Figure 3: Dynamics of optimal adaptation for a uniform, "jumping" prior probed with a step. As in the plot in Figure 2, the solid curve shows the dynamics of average $\bar{\sigma}$ following an upward jump in standard deviation from 1 to 4, while the dashed curve (inverted) shows the behavior following a downward jump from 4 to 1. Here, the estimator is optimized for a gaussian distributed signal with a standard deviation that jumps with probability per time step (x) of .001 to a new value drawn at random from all values between 0.1 and 8. Note that following an upward jump, the initial response is more rapid, the asymptotic behavior is slower, and the estimate overshoots the correct value.

The estimate for the standard deviation that minimizes the rms error is given by the mean of this distribution:

$$\bar{\sigma}_i \equiv \int d\sigma_i \sigma_i P(\sigma_i | s_{j \le i}). \tag{3.18}$$

Figure 3 illustrates the dynamics of $\bar{\sigma}$ for $x = 1/1000$ and $\sigma_{high}/\sigma_{low} = 4$. Once again, for times immediately following a jump in variance, the initial upward adaptation is faster than downward. For longer times, however, the upward adaptation asymptotes more slowly. This is related to the fact that the number of independent examples of the signal needed to get an estimate of either the mean or the standard deviation to some fixed level of accuracy grows quadratically with standard deviation.

Curiously, the upward adaptation tends to overshoot before asymptoting to its final value, provided that the larger of the two standard deviation values we are probing with is not too close to the upper cutoff (8 in this

case). There is a simple explanation. After receiving an unexpectedly large value for s_i, the estimator infers that the variance must have jumped to a higher value. Initially, it weights all standard deviation values between s_i and the upper cutoff about equally, since its prior on the standard deviation is flat. After measuring the signal for many time steps, the estimator homes in on the true value.

Neither of the two curves in Figure 3 is well fit by a single exponential, nor are they obviously related by some other simply parameterizable family of curves. For this reason, the effective rate of adaptation for either curve is a function of the delay since the last jump in the probe variance, which makes quantitative comparisons between different parameter settings difficult. Qualitatively, the differences between the curves are more pronounced as $\sigma_{high}/\sigma_{low}$ is increased or x is decreased. In the other extreme, if we set $\sigma_{low} = \sigma_{high}$, then no adaptation is required. In the limit of long delays, the behavior is predictable from simple statistics. Intermediate delays give mixed results, which depend on the details of the prior. In the limit of short delays, we observe universal behavior: the asymmetrical dynamics, which we have emphasized.

If we minimize the error of our estimate of the variance rather than the standard deviation, we arrive at essentially the same results, whether we use a flat a priori distribution for the standard deviation or the variance. In addition, the dynamics of the mode, $\sigma_{m.l.}$, of $P(\sigma_i|s_{j\le i})$ are essentially the same as for the mean and display the same asymmetry.

So far we have probed our estimator with a signal that is not too unlikely in its expected input ensemble. This will not be the case in the next section.

3.3 Smoothly Changing Variance. We now consider a prior in which the variance changes smoothly with time according to diffusive dynamics (or a random walk in the discrete version shown here) with reflecting boundaries at a and b:

$$\sigma_i = \begin{cases} \sigma_{i-1} + 2D\Delta t z_i, & \text{if } a \le \sigma_{i-1} + 2D\Delta t z_i \le b \\ 2b - \sigma_{i-1} - 2D\Delta t z_i, & \text{if } \sigma_{i-1} + 2D\Delta t z_i > b, \\ 2a - \sigma_{i-1} - 2D\Delta t z_i, & \text{if } \sigma_{i-1} + 2D\Delta t z_i < a, \end{cases} \tag{3.19}$$

$$s_i = \sigma_i y_i, \tag{3.20}$$

where $b > a > 0$, Δt is the duration of each time step, D is the (one dimensional) diffusion constant, and y and z are both gaussian distributed i.i.d. processes with unit variance and zero mean. Far from the boundaries, this implies that

$$P(\sigma_i|\sigma_{i-1}) = \frac{1}{\sqrt{4\pi D\Delta t}} \exp\left(-\frac{(\sigma_i - \sigma_{i-1})^2}{4D\Delta t}\right) \tag{3.21}$$

and

$$P(s_i|\sigma_i) = \frac{e^{-s_i^2/2\sigma_i^2}}{\sqrt{2\pi\sigma_i^2}}. \tag{3.22}$$

D is defined as the inverse of the correlation time constant, $D = 1/\tau_c$. Like the example in the previous section, the underlying dynamics of σ are symmetric with respect to increases and decreases in standard deviation. By imposing reflecting boundary conditions, we ensure that the time-averaged distribution for the standard deviation is flat, so that any asymmetry in adaptation time is not due to the relative likelihood of big and small standard deviation values.

Combining equations 3.21 and 3.22 with equation 2.6 as before, we obtain an expression for updating the conditional distribution for the standard deviation after receiving s_i:

$$P(\sigma_i|s_{j\le i}) = \frac{1}{\Omega}\frac{e^{-s_i^2/2\sigma_i^2}}{\sqrt{2\pi\sigma_i^2}}\int_a^b d\sigma_{i-1}\frac{1}{\sqrt{4\pi D\Delta t}} \times \exp\left(-\frac{(\sigma_i-\sigma_{i-1})^2}{4D\Delta t}\right)P(\sigma_{i-1}|s_{j<i}), \tag{3.23}$$

where[4] we have again introduced a and b, the lower and upper cutoffs for σ. We will again use the mean of this distribution as our estimate for the current standard deviation, $\bar{\sigma}_i$ (see equation 3.18).

For the first time step after an abrupt change in variance, we can derive a compact expression for the rate of adaptation following an upward jump in variance $\sigma_{low} \to \sigma_{high}$ (see the appendix):

$$\text{rate}_{up} \equiv \hat{\sigma}_i^{up} - \sigma_{low} = D\Delta t\frac{2}{\sigma_{low}}\left[\left(\frac{\sigma_{high}}{\sigma_{low}}\right)^2 - 1\right]. \tag{3.24}$$

Dividing this by the corresponding expression for a downward jump yields:

$$\frac{\text{rate}_{up}}{\text{rate}_{down}} \equiv \frac{\hat{\sigma}_i^{up} - \sigma_{low}}{\hat{\sigma}_i^{down} - \sigma_{high}} = \left(\frac{\sigma_{high}}{\sigma_{low}}\right)^3. \tag{3.25}$$

This is certainly asymmetric, with the upward adaptation faster than the downward adaptation for any pair of variances.

[4] For brevity, we dropped some terms on the right-hand side of equation 3.23 which we included in our algorithm to enforce our reflecting boundary conditions. Since we always placed our boundaries far from the extreme values of our probe stimulus compared to $D\Delta t$, our results were independent of the boundary conditions.

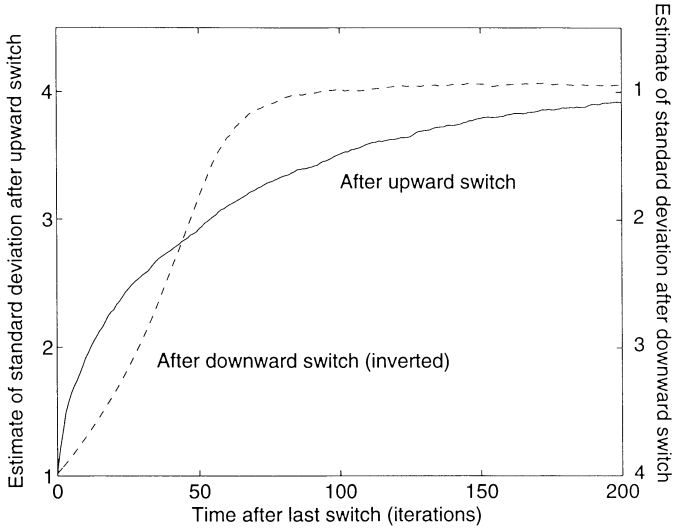

Figure 4: Dynamics of optimal adaptation for a diffusive prior probed with a step. As in Figures 2 and 3, the solid curve shows the dynamics of the average $\bar{\sigma}$ following an upward jump in standard deviation from 1 to 4, while the dashed curve (inverted) shows the behavior following a downward jump from 4 to 1. In this case, the estimator is optimized for a gaussian distributed signal with a standard deviation that diffuses (i.e., takes a random walk) between 0.05 and 5.0 with reflecting walls at the boundaries. The correlation time (the inverse of the one-dimensional diffusion constant) of the diffusion process is 300 time steps. Note that following an upward jump, the initial response is more rapid and the asymptotic behavior is slower.

For the complete dynamics, we must compute the solution numerically. To do this, we discretize the probability distribution $P(\sigma_i|s_{j\leq i})$ between the reflecting boundaries at .05 to 4.95 in bins of width .025. The results for $D\Delta t = 1/300$ are plotted in Figure 4.

For the sake of comparison between the different models, we can relate the correlation time of this diffusing prior to that of the switching prior of the last section by estimating the expected time for the standard deviation to diffuse from σ_{low} to σ_{high}. For $\sigma_{high} = 4$, $\sigma_{low} = 1$ and $D\Delta t = \Delta t/\tau_c = 1/300$, the average time for the rms displacement of the standard deviation to reach $\sigma_{high} - \sigma_{low} = 3$ is

$$3 = \sqrt{2t/\tau} \Rightarrow t = 4.5\tau \sim 1400 \text{ time steps}, \tag{3.26}$$

so the correlation time for these parameter settings roughly corresponds to that of the switching prior with the probability of switching per time step (x) near 1/1400.

As in the previous examples, adaptation immediately following an upward jump is faster than it is after a downward jump. As before, the upward adaptation is slower to asymptote to its final value due to the larger value of the final variance. Once again, we find that neither of the two curves is well fit by a single exponential, which makes quantitative comparisons between different parameter settings complicated for intermediate to long times after jumps in standard deviation. Qualitatively, the differences between the two curves become more pronounced with increasing $\sigma_{high}/\sigma_{low}$ and decreasing $D\Delta t$. Unlike the previous example, the diffusion prior produces monotonic behavior with no overshoot, so the curves cross at some intermediate time during the adaptation. Another difference is that in this example, the curves are independent of the boundary conditions, whereas the dynamics of the last example were strongly cutoff dependent. Clearly, qualitatively universal behavior is present immediately after abrupt changes in the variance, while the dynamics are prior dependent at later times.

As in the previous section, we repeated the analysis of this section for a diffusing variance rather than standard deviation and found no significant difference in our results whether we minimized the error of the standard deviation or the variance. Again, the mode, $\sigma_{m.l.}$, of $P(\sigma_i|s_{j\leq i})$ displayed the same asymmetrical dynamics as did the mean of that distribution.

To summarize our results so far, we computed the optimal estimator $\bar{\sigma}$ for three processes with different dynamics for their respective time-varying variances. At any single moment in time, these processes were all gaussian with zero mean. In each case, when dynamics were probed with a square wave of standard deviation, the adaptation to an upward jump was faster than for a downward jump. We now consider nongaussian signal distributions and environments where the mean and variance fluctuate together.

3.4 Other Signal Distributions.

3.4.1 Nongaussian Signal Distributions and Nonflat Priors on the Standard Deviation. In all of the previous examples, we considered nonstationary signals that were gaussian distributed at every moment in time. We repeated each of these examples for exponentially distributed signals, and the results were qualitatively the same as for gaussian signals in every case. In fact, the asymmetry in the dynamics persisted even if we biased the prior distribution on the standard deviation to favor smaller values. Specifically, we sometimes used a flat prior for $1/\sigma$ rather than for σ so that $P(\sigma) = d\sigma/\sigma^2$.

As we explain in section 4, the adaptation dynamics are largely determined by the occurrence of outliers in the signal distribution immediately after abrupt changes in the probe stimulus. For that reason, the dynamics are roughly independent of the exact shape of the distribution everywhere

except in the tails. In the appendix, we prove that under weak assumptions, the dynamics are asymmetric immediately following jumps in the probe stimulus for the binary switching standard deviation model, provided that the tails fall off like $\exp(-as_i^n)$ for all positive a and n. This is a rich set of functions for the shape of the tails for which the distribution has well-defined moments at all orders. The proof is valid for a distribution of nearly any shape, so long as it does not get arbitrarily small anywhere except in the tails. For the diffusing standard deviation model with $P(s_i|\sigma_i) \propto \exp(-as_i^n)$, we can derive a compact expression for the ratio of the rate of adaptation immediately following an upward jump in variance $\sigma_{low} \rightarrow \sigma_{high}$ versus the corresponding downward jump (see the appendix),

$$\frac{\text{rate}_{up}}{\text{rate}_{down}} \equiv \frac{\hat{\sigma}_i^{up} - \sigma_{low}}{\hat{\sigma}_i^{down} - \sigma_{high}} = \left(\frac{\sigma_{high}}{\sigma_{low}}\right)^{n+1}, \tag{3.27}$$

which is clearly asymmetric for all positive a and n, though the exact ratio depends on n.

3.4.2 Simultaneously Adapting to the Mean and Variance. In each of the examples above, the optimal estimate of the standard deviation $\bar{\sigma}$ is a function of the absolute value of the signal s since $P(-s_i|\sigma_i) = P(s_i|\sigma_i)$ in each case. Therefore, our results would be the same for a different signal s', which is positive definite:[5]

$$P(s_i'|\sigma_i) = \begin{cases} 2P(s_i|\sigma_i) & \text{if } s_i' > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3.28}$$

A gaussian, nonnegative signal distributed according to

$$P(s_i'|\sigma_i) = \sqrt{\frac{2}{\pi\sigma_i^2}} \exp\left(-\frac{(s_i')^2}{2\sigma_i^2}\right) \tag{3.29}$$

has an "instantaneous" mean, μ_i', and standard deviation, σ_i', given by

$$\mu_i' = \sigma_i\sqrt{\frac{2}{\pi}}, \tag{3.30}$$

$$\sigma_i' = \sigma_i\sqrt{1 - \frac{2}{\pi}}. \tag{3.31}$$

For an exponentially distributed nonnegative signal,

$$P(s_i'|\sigma_i) = \frac{1}{\sigma_i} \exp\left(-\frac{s_i'}{\sigma_i}\right), \tag{3.32}$$

[5] Here we assume that the probability of s_i being exactly zero is of measure zero.

the "instantaneous" mean and standard deviation are the same:

$$\mu_i' = \sigma_i' = \sigma_i. \tag{3.33}$$

For all the models we have studied, the dynamics of optimal adaptation look the same whether the mean is constant or tightly coupled to the fluctuating standard deviation.

4 Discussion

4.1 Main Result. A system whose inputs span a large dynamic range can use adaptation to exploit their nonstationary statistics. Adaptation requires an estimate of the present statistics of the signal based on its recent past. Optimal estimation of the mean of a nonstationary signal is a well-studied problem. Here we have considered optimal estimation of the second-order statistics.

We have shown that optimal adaptation to variance leads to asymmetrical dynamics. In particular, the optimal estimate for the variance tracks an abrupt increase in variance more closely than an abrupt decrease in environments, where the true dynamics of the variance is symmetric. This is true whether the mean is fixed or allowed to vary with the standard deviation. It is easy to show that this is not a feature of optimal adaptation to a time-varying mean in environments where the variance is fixed and the mean fluctuates in a symmetric fashion.

4.2 Intuition Behind the Main Result. Our basic result can be readily understood with the following intuitive argument. Consider the generic case for the instantaneous signal distribution, where $P(s_i|\sigma_i)$ has a single maximum that roughly coincides with its mean, and parameterize the distribution so that changing σ_i merely rescales s_i while leaving the mean unchanged (e.g., $P(s_i|\sigma_i)$ could be gaussian with zero mean and a standard deviation of σ_i). When we probe with an upward jump in variance, the estimator expects to see an s that is not much larger than the old standard deviation, but instead measures a value that is on average equal to the new, larger standard deviation. After receiving this unlikely outlier, the estimator immediately infers that the standard deviation has increased. On the other hand, if the standard deviation jumps down, then the estimator will most likely receive an s that is much smaller than the old standard deviation, but this is near the peak of the distribution of s for any standard deviation, so it will wait for more data before lowering its estimate for the current standard deviation.

This argument is based on the occurrence of outliers in the likelihood $P(s_i|\sigma_i)$ of the observed signal s_i given the current estimate for the standard deviation σ_i. If many independent presentations of the signal can be observed in the average time it takes for the standard deviation to fluc-

tuate between the high and low values of the probe stimulus, we should expect outliers in the likelihood distribution to dominate the behavior of our estimator immediately following jumps in the true standard deviation. This separation between the time scales of signal detection and changes in the environment is what makes the problem of adaptation interesting. If the environment completely changes between observations, then the optimal estimation strategy does not require dynamic adaptation, as we saw in section 3.1.

As long as the prior distribution for the dynamics of the standard deviation is not too asymmetrical, our intuition about the likelihood should hold for the full Bayesian treatment. This is true no matter what form the likelihood takes—it can be asymmetric, multimodal, finite for positive values only, or something else—provided that it approaches zero only out in the tail(s). In the appendix, we prove that optimal adaptation is indeed faster after abrupt increases in the standard deviation for a large class of model environments, just as we saw for the previous examples.

4.3 Connection with Experiment. It is common to identify the firing rate of a cell in response to a stimulus of given intensity with its state of adaptation. Our formulation of the variance estimation problem does not specify what aspect of the output of any given cell type will reflect its state of adaptation, so we are not restricting ourselves to cells whose firing rates encode the mean or variance of a sensory signal. Implicit in our framework is the idea that adaptation is a useful property that is not due solely to fatigue of cellular mechanisms or saturation of their inputs. It is well documented that adaptation occurs even when saturation is not present (Shapley & Victor, 1978).

We have stated our results in terms of ratios of adaptation times rather than absolute times, partly because we found a very general behavior for this ratio. In order to convert discrete time into experimental units, one would need to measure the "integration time" of the biological system being studied. In other words, observations of the signal at consecutive time steps in our framework represent effectively independent measurements, so our discrete time steps reflect both the filtering of the sensory system and the effective noise at the input. If the system is visual, then the spatial statistics of scenes from the creature's environment can be convolved into the temporal statistics of the signal impinging on local regions of the retina through saccades and body motion.

In qualitative agreement with our findings, recent experiments by Smirnakis, Berry, Warland, Bialek, and Meister (1997) have shown that ganglion cells in the salamander and rabbit retinas adapt faster to abrupt increases than decreases in variance. In these experiments, the activity of individual ganglion cells was monitored while the retina was stimulated with light whose contrast switched between a high and low value while the mean was fixed, as in Figure 1 (top). Between the jumps, in contrast, the trial-averaged

firing rate followed an exponential path, the time constant of which was always faster after upward jumps in contrast.

In section 3.4.2, we showed that optimal adaptation is asymmetrical in several environments in which the mean fluctuates with the standard deviation. In natural scenes, the mean and standard deviation of light intensity are probably correlated. Naively, one might imagine that the mean and standard deviation are proportional since they have the same units. This would be the case in a room with spatially fixed light sources all controlled by a single dimmer switch and walls that have a reflectance that does not depend on the intensity of the light. Natural environments could in principle be much more complex than this.[6] Unfortunately, the true relationship between the mean and standard deviation of natural scenes is not yet fully understood, but it is reasonable to assume that the standard deviation will be roughly proportional to the mean light level in many environments. For example, for one data set, the local standard deviation is roughly three times the mean with a correlation coefficient of about 1/2 (Ruderman, personal communication).

It is well known that sensory cells (photoreceptors and ganglion cells in the retina) adapt more quickly to sudden increases than decreases in the mean light intensity (Barlow & Mollon, 1982). This is usually attributed to limitations of the biological machinery, but our results indicate that this behavior could reflect the optimal strategy for the environment they evolved in. These ideas give parameter-free predictions once the relevant biological constraints and input statistics are known.

Appendix: A Proof and Some Exact Results

A.1 Proof of Asymmetric Dynamics for Binary Variance Switching. We will now prove that under weak assumptions, optimal adaptation is faster immediately following upward jumps in the standard deviation σ of its input signal s than it is for downward jumps for a system that expects the standard deviation to switch between two values, σ_{low} and σ_{high}, at random times (as in section 3.2.1). For clarity, we will prove our result for a specific signal distribution and then state a more general result, which can be proved following the same steps.

[6] As a toy example of an environment where the mean and standard deviation are not proportional, consider a bird's-eye view of a plowed field with parallel furrows running from north to south. Early in the day, when the sun is low on the horizon, the mean light level is low, and the standard deviation is comparable to the mean due to the shadows cast into the furrows. At noon, when the sun is high in the sky, there are few shadows since the sun can illuminate the bottoms and shallow sides of the furrows, so the entire scene is equally illuminated, resulting in a standard deviation much smaller than the (high) mean light intensity.

We make the following assumptions:

1. The standard deviation switches between a low value, σ_{low}, and a high value, $\sigma_{high} > 2\sigma_{low}$, at (discrete) times drawn from a homogeneous Poisson process; x is the probability of switching per time step.
2. The probability distribution of the (nonnegative) signal for fixed standard deviation is bounded everywhere except in the tail, $y \leq P(s_i|\sigma_i) \leq z$ for $0 \leq s_i \leq \Lambda\sigma_i$, and falls off exponentially in the tail, $P(s_i|\sigma_i) \propto \exp(-a(s_i/\sigma_i))$ for $s_i > \Lambda\sigma_i$. Note that $P(s_i|\sigma_i)$ need not be monotonically decreasing for $0 \leq s_i \leq \Lambda\sigma_i$.
3. The estimator makes many observations of the signal in the average time required for the standard deviation to drift between σ_{low} and σ_{high}: $x \ll 1$. All other parameters are of order unity so that $x \ll \{1, y/z, y\sigma_{low}, \sigma_{low}/\sigma_{high}, \Lambda\}$.

Consider the probability that the optimal estimator will correctly detect an upward jump from σ_{low} to σ_{high}. More precisely, we want the probability that $\bar{\sigma}_i = \sigma_{high}$ one time step after our estimator was "sure" that the standard deviation was at a low value, σ_{low}, given that the true standard deviation is currently at the high value, σ_{high}:

$$\begin{aligned} P_{up} &\equiv \int ds_i P(s_i|\sigma_i = \sigma_{high}) P(\sigma_i = \sigma_{high}|s_i, \sigma_{i-1} = \sigma_{low}) \\ &= \int ds_i P(s_i|\sigma_i = \sigma_{high}) \frac{P(s_i|\sigma_i = \sigma_{high}) P(\sigma_i = \sigma_{high}|\sigma_{i-1} = \sigma_{low})}{P(s_i|\sigma_{i-1} = \sigma_{low})} \\ &= \int ds_i \frac{P(s_i|\sigma_i = \sigma_{high})}{1 + \frac{1-x}{x} \frac{P(s_i|\sigma_i = \sigma_{low})}{P(s_i|\sigma_i = \sigma_{high})}}, \end{aligned} \tag{A.1}$$

where we have used Bayes's theorem between the first and second lines, and reintroduced x, the probability of switching per time step. Our claim is that this is greater than the probability that the estimator will correctly detect a downward jump in the reverse situation:

$$P_{down} = \int ds_i \frac{P(s_i|\sigma_i = \sigma_{low})}{1 + \frac{1-x}{x} \frac{P(s_i|\sigma_i = \sigma_{high})}{P(s_i|\sigma_i = \sigma_{low})}}. \tag{A.2}$$

Under the conditions described above, $P(s_i|\sigma_i = \sigma_{high})/P(s_i|\sigma_i = \sigma_{low})$ is never less than y/z for all allowed values of s_i, so we can always set x sufficiently small in equation A.2 to guarantee that P_{down} is proportional to x:

$$\begin{aligned} P_{down} &= x \int_0^{\infty} ds_i \frac{P^2(s_i|\sigma_i = \sigma_{low})}{P(s_i|\sigma_i = \sigma_{high})} + \mathcal{O}\left[x^2\right] \\ &= x \int_0^{\Lambda\sigma_{high}} ds_i \frac{P^2(s_i|\sigma_i = \sigma_{low})}{P(s_i|\sigma_i = \sigma_{high})} \; + \end{aligned}$$

$$x\int_{\Lambda\sigma_{high}}^{\infty} ds_i \frac{a\sigma_{high}}{\sigma_{low}^2}\exp\left[-a\left(\frac{2}{\sigma_{low}}-\frac{1}{\sigma_{high}}\right)s_i\right]+\mathcal{O}[x^2]$$
$$\leq x\left[\left(\frac{z}{y}\right)^2+\frac{\sigma_{high}^2}{2\sigma_{high}\sigma_{low}-\sigma_{low}^2}\right]+\mathcal{O}[x^2].$$

We have made use of the fact that $0<\Lambda\sigma_{high}\leq 1/y$ in the last line.

Clearly, P_{down} is proportional to our small parameter x as long as $x \ll \sigma_{low}/\sigma_{high}$ since z/y is of order one. However, $P(s_i|\sigma_i=\sigma_{high})/P(s_i|\sigma_i=\sigma_{low})$ grows arbitrarily large as we move further out in the tail, so P_{up} can remain finite as $x\to 0$. If we restrict our attention to $s_i>\Lambda\sigma_{high}$ so we are in the tail of the distribution for either value of the standard deviation, the denominator of the integrand in the last line of equation A.1 is much greater than one only for

$$s_i \ll s' \equiv \frac{\ln\left[\frac{\sigma_{high}}{\sigma_{low}}\frac{1}{x}\right]}{a\left(1/\sigma_{low}-1/\sigma_{high}\right)}. \tag{A.3}$$

If $s'\geq\sigma_{high}\Lambda$, then we can exploit the fact that the integrand in equation A.1 is nowhere negative to rigorously bound P_{up} from below by isolating the contribution to the integral from the region $s_i>s'$:

$$\begin{aligned}P_{up} &\geq \int_{s'}^{\infty} ds_i \frac{P(s_i|\sigma_i=\sigma_{high})}{1+\frac{1-x}{x}\frac{P(s_i|\sigma_i=\sigma_{low})}{P(s_i|\sigma_i=\sigma_{high})}}\\ &\geq N\int_{s'}^{\infty} ds_i \frac{\exp\left[-as_i/\sigma_{high}\right]}{2}\\ &\geq \frac{y\sigma_{high}}{2}\frac{e^{a\Lambda}}{a}\left[\frac{\sigma_{low}}{\sigma_{high}}x\right]^{\left(\frac{1}{\sigma_{high}/\sigma_{low}-1}\right)},\end{aligned} \tag{A.4}$$

where we have replaced the normalization factor N with its smallest possible value $ye^{a\Lambda}$ for fixed a and Λ. By choosing a probe stimulus with $\sigma_{high}>2\sigma_{low}$, we can make the first term in the final line of equation A.4 sublinear in x, so that P_{up} is larger than P_{down} for x sufficiently small.

On the other hand, if s' is less than $\sigma_{high}\Lambda$, then we can safely integrate over the entire tail region of $P(s_i|\sigma_{high})$ to get a different bound:

$$\begin{aligned}P_{up} &\geq \int_{\sigma_{high}\Lambda}^{\infty} ds_i \frac{P(s_i|\sigma_i=\sigma_{high})}{1+\frac{1-x}{x}\frac{P(s_i|\sigma_i=\sigma_{low})}{P(s_i|\sigma_i=\sigma_{high})}}\\ &\geq N\int_{\Lambda\sigma_{high}}^{\infty} ds_i \frac{\exp\left[-as_i/\sigma_{high}\right]}{2}\\ &\geq \frac{y\sigma_{high}}{2a},\end{aligned} \tag{A.5}$$

which is greater than P_{down} for all a such that

$$a < \frac{y\sigma_{high}}{2x}\left[\left(\frac{z}{y}\right)^2 + \frac{\sigma_{high}^2}{2\sigma_{high}\sigma_{low} - \sigma_{low}^2}\right]^{-1}. \tag{A.6}$$

If a is too large to satisfy this inequality, then we focus on the region in the tail of $P(s_i|\sigma_{low})$ but not of $P(s_i|\sigma_{high})$:

$$\begin{aligned} P_{up} &\geq \int_{\Lambda\sigma_{low}}^{\Lambda\sigma_{high}} ds_i \frac{P(s_i|\sigma_i = \sigma_{high})}{1 + \frac{1-x}{x}\frac{P(s_i|\sigma_i=\sigma_{low})}{P(s_i|\sigma_i=\sigma_{high})}} \\ &\geq \int_{\Lambda\sigma_{low}}^{\Lambda\sigma_{high}} ds_i \frac{y}{1 + \frac{1}{x}\frac{z\exp(a\Lambda - as_i/\sigma_{low})}{y}} \\ &\geq \frac{\sigma_{low}y}{a}\ln\left[\frac{xy\exp[a\Lambda(\sigma_{high}/\sigma_{low}-1)]+z}{xy+z}\right] \\ &\geq \frac{\sigma_{low}y}{a}\left(a\Lambda(\sigma_{high}/\sigma_{low}-1) + \ln\left[\frac{xy}{xy+z}\right]\right) \\ &\geq y\Lambda(\sigma_{high}-\sigma_{low}) - \mathcal{O}[x\ln(x)], \end{aligned} \tag{A.7}$$

where we have assumed that a is at least of order $1/x$ in the last line, which is the one case not covered by the two previous bounds on P_{up}. Again, Λ, $y\sigma_{low}$, and $y\sigma_{high}$ are all much greater than x, so this bound is greater than our upper bound for P_{down}.

We have shown that the optimal estimator adapts more quickly to abrupt increases than decreases in the standard deviation of its input for the conditions enumerated above. By following the same steps, one can prove the same result by relaxing condition 2 to include all signal distributions with either one or two tails that decrease monotonically with $|s_i|$; each tail must take the same form for any value of the standard deviation, so that changing the standard deviation amounts to rescaling s and renormalizing:

$$\begin{aligned} P(s|\sigma_{high}) = \frac{\sigma_{low}}{\sigma_{high}} P\left(\frac{\sigma_{low}}{\sigma_{high}}s \,\middle|\, \sigma_{low}\right) \\ \text{for } \left(\frac{s_i}{\sigma_i} < -\Lambda_L\right)\bigcup\left(\frac{s_i}{\sigma_i} > \Lambda_R\right). \end{aligned} \tag{A.8}$$

The tails must fall off faster than a power law to ensure that the variance and all higher moments are well defined. We can also permit a region about zero that is less than y and even zero so long as the form of $P(s_i|\sigma_i)$ in this region remains unchanged up to a factor of order one (i.e., s does *not* scale with σ in this region), as would be the case for a detector whose sensitivity falls off for small signal strengths placed in an environment where low signal values always occur with some finite probability. In other words,

$y < P(s_i|\sigma_i) < z$ for $-\sigma_i\Lambda_L \le s_i \le -\sigma_{high}\lambda_L$ and $\sigma_{high}\lambda_R \le s_i \le \sigma_i\Lambda_R$; but $P(s_i|\sigma_i = \sigma_{high}) = P(s_i|\sigma_i = \sigma_{low})$ for $-\sigma_{high}\lambda_L \le s_i \le \sigma_i\lambda_R$. In each case one has to check that the ratio $\sigma_{high}/\sigma_{low}$ is sufficiently large to complete the proof.

For example, if the tail with the slowest rate of decrease falls off like $\exp[-a|s_i/\sigma_i|^n]$, then our result holds provided that $\sigma_{high}/\sigma_{low} > (n+1)/n$. This is a very general set of functions that give well-defined moments. As one might expect, the faster the tail(s) fall off, the more inclusive is the range of values of the probe's standard deviation.

A.2 Analytic Expressions for Smoothly Changing Variance. We can derive an exact expression for the rate of adaptation to the standard deviation σ of a stochastic variable s immediately after an abrupt jump in the standard deviation for a system optimized for a diffusing standard deviation. The quantity we want to calculate is the average value of $\bar{\sigma}$ one time step after our estimator was "sure" that the standard deviation was at a low value, σ_{low}, given that the standard deviation is currently at a high value, σ_{high}. We will assume that s is gaussian for our derivation, and then state the general solution for $P(s|\sigma) \propto \exp(-a|s|^n)$ for all positive a and n.

We begin our derivation by replacing $P(\sigma_{i-1}|s_{j<i})$ in equation 3.23 with a dirac delta function, $\delta(\sigma_{i-1} - \sigma_{low})$, and average over the value of the new data point, s_i, in the conditional distribution $P(s_i|\sigma_i = \sigma_{high})$:

$$\begin{aligned}\sigma_i^{up} &\equiv \int d\sigma_i \sigma_i \int ds_i P(s_i|\sigma_i = \sigma_{high}) P(\sigma_i|s_i, \sigma_{i-1} = \sigma_{low}) \\ &= \int ds_i \frac{e^{-s_i^2/2(\sigma_{high})^2}}{\sqrt{2\pi(\sigma_{high})^2}} \frac{\int d\sigma_i \exp\left(-\frac{s_i^2}{2\sigma_i^2} - \frac{(\sigma_i-\sigma_{low})^2}{4D\Delta t}\right)}{\int d\sigma_i \frac{1}{\sigma_i} \exp\left(-\frac{s_i^2}{2\sigma_i^2} - \frac{(\sigma_i-\sigma_{low})^2}{4D\Delta t}\right)}. \end{aligned} \tag{A.9}$$

The amount that this differs from σ_{low} is proportional to the effective rate of adaptation immediately after the upward jump in variance. The effective upward adaptation rate can then be compared to the corresponding difference for a switch from σ_{low} to σ_{high}.

The main trick for solving the integrals in equation A.9 is to make use of the fact that each time step is short compared to the correlation time of the diffusion process, which allows us to expand about $D\Delta t = 0$:

$$\begin{aligned}\frac{1}{\sqrt{4D\Delta t}} \exp\left(-\frac{(\sigma_i - \sigma_{low})^2}{4D\Delta t}\right) \to \delta(\sigma_i - \sigma_{low}) \\ + \frac{2D\Delta t}{2!} \frac{\partial^2 \delta(\sigma_i - \sigma_{low})}{\partial \sigma_i^2}. \end{aligned} \tag{A.10}$$

We do not write this as an equality since it is valid to make this substitution only when the gaussian appears inside an integral over its argument σ_i. To

derive this substitution, consider a gaussian with a small variance v multiplied by a doubly differentiable function f inside an integral. Expanding f in a Taylor series gives:

$$\int dx \frac{e^{-x^2/2v}}{\sqrt{2\pi v}} f(x) \approx f(0) + \frac{v}{2!} \frac{\partial^2 f}{\partial x^2}\bigg|_{x=0}. \tag{A.11}$$

This is exactly what we would have gotten by making the above substitution for the gaussian within the integral and integrating by parts.

With this substitution, we expand both integrals over σ_i to first order in $D\Delta t$, then expand the full s_i integrand to first order and perform the integral to obtain:

$$\sigma_i^{up} = \sigma_{low} + D\Delta t \frac{2}{\sigma_{low}} \left[\left(\frac{\sigma_{high}}{\sigma_{low}} \right)^2 - 1 \right]. \tag{A.12}$$

Finally, we repeat this procedure for a downward jump from σ_{high} to σ_{low} and find a simple form for the ratio of upward-to-downward adaptation rates immediately after a jump in the standard deviation:

$$\frac{\sigma_i^{up} - \sigma_{low}}{\sigma_{high} - \sigma_i^{down}} = \left(\frac{\sigma_{high}}{\sigma_{low}} \right)^3. \tag{A.13}$$

This result is valid for gaussian s; by repeating the calculation for $P(s_i|\sigma_i) = \exp(-a|s_i|^n)$, we can derive a more general expression:

$$\frac{\sigma_i^{up} - \sigma_{low}}{\sigma_{high} - \sigma_i^{down}} = \left(\frac{\sigma_{high}}{\sigma_{low}} \right)^{n+1}, \tag{A.14}$$

which is valid for all positive a and n. Again we see that optimal adaptation is faster following abrupt increases in standard deviation for signal distributions with well-defined moments at all orders.

Acknowledgments

We thank W. Bialek, E. J. Chichilnisky, B. Pearlmutter, and R. Shapley for many useful discussions. We are especially grateful to M. Berry, M. Meister, S. Smirnakis, and D. Warland for sharing their findings with us before publication. This work was supported by the Sloan Foundation for Theoretical Neuroscience.

References

Barlow, H., & Mollon, J. (1982). *The senses* (2nd ed.). Cambridge: Cambridge University Press.

Bonds, A. (1991). Temporal dynamics of contrast gain in single cells of the cat striate cortex. *Visual Neuroscience, 6*, 239–255.

deLange, H. (1958). Research into the dynamic nature of the human fovea—cortex systems with intermittent and modulated light. I. Attenuation characteristics with white and colored light. *J. Opt. Soc. Am., 48*, 777–784.

Enroth-Cugell, C., & Robson, J. (1996). The contrast sensitivity of retinal ganglion cells of the cat. *J. Physiol., 187*, 517–552.

Giaschi, D., Douglas, R., Marlin, S., & Cynader, M. (1993). The time course of direction-selective adaption in simple and complex cells in cat striate cortex. *J. Neurophysiol., 70*, 2024–2034.

Kelly, D. (1961). Visual responses to time-dependent stimuli. I. *J. Opt. Soc. Am., 51*, 422.

Papoulis, A. (1991). *Probability, random variables and stochastic processes* (3rd ed.). New York: McGraw-Hill.

Shapley, R. (1997). Adapting to the changing scene. *Current Biology, 7* (7), 421–423.

Shapley, R., & Enroth-Cugell, C. (1984). Visual adaptation and retinal gain controls. In N. N. Osborne & G. J. Chader (Eds.), *Progress in retinal research* (vol. 3, pp. 263–346). Oxford: Pergamon Press.

Shapley, R., Enroth-Cugell, C., Bonds, A., & Kirby, A. (1972). Gain control in the retina and retinal dynamics. *Nature, 236*, 352–353.

Shapley, R., & Victor, J. (1978). The effect of contrast on the transfer properties of cat retinal ganglion cells. *J. Physiol., 285*, 275–298.

Shapley, R., & Victor, J. (1979). Nonlinear spatial summation and the contrast gain control of cat retinal ganglion cells. *J. Physiol., 290*, 141–161.

Smirnakis, S., Berry, M., Warland, D., Bialek, W., & Meister, M. (1997). Retinal processing adapts to image contrast and spatial scale. *Nature, 385*, 69–73.

Received January 29, 1997; accepted November 14, 1997.