Adaptive coding for dynamic sensory inference

Wiktor Młynarski[1] & Ann M. Hermundstad[2]

1) Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
2) Janelia Research Campus, Howard Hughes Medical Institute

mlynar@mit.edu, hermundstada@janelia.hhmi.org

Abstract

Behavior relies on the ability of sensory systems to infer properties of the environment from incoming stimuli. The accuracy of inference depends on the fidelity with which behaviorally-relevant properties of stimuli are encoded in neural responses. High-fidelity encodings can be metabolically costly, but low-fidelity encodings can cause errors in inference. Here, we derive adaptive sensory encoding schemes that dynamically navigate this tradeoff between coding cost and inference error. We find that optimal encodings increase the fidelity of the neural response following a change in the stimulus distribution, whereas stimuli originating from a known distribution can be encoded with low fidelity at little cost. We predict dynamical signatures of an encoding scheme designed to preserve only those stimulus details that are relevant for inference, and we demonstrate how known phenomena, such as burst coding and the perception of metamers, can be understood as hallmarks of adaptive encoding for accurate inference.

INTRODUCTION

Biological systems must make inferences about the environment in order to successfully plan and accomplish goals. Inference is the process of estimating behaviorally-relevant properties of the environment from low-level sensory signals registered by neurons in the early sensory periphery [1]. Many perceptual tasks, such as color perception [2], visual speed estimation [3], or sound localization [4, 5], can be understood as probabilistic inference. All of these tasks rely on the estimation of features (such as the speed of an object) that are not explicitly represented by low-level sensory stimuli (such as light signals incident on photoreceptors).

To accurately perform inference, the nervous system can construct an internal model that relates incoming sensory stimuli to behaviorally-relevant properties of the environment [1, 6, 7, 8, 9]. As the environment changes, this internal model must be continually updated with new stimuli [10, 11, 12, 13, 14], and therefore the accuracy of this internal model depends on the fidelity with which incoming stimuli are encoded in neural responses.

The process of encoding sensory stimuli, however, is metabolically expensive [15, 16, 17, 18, 19, 20], and a large body of evidence suggests that sensory systems have evolved to reduce the energetic costs of stimulus coding [15, 21, 22]. These findings provide empirical support for the efficient coding hypothesis [23], which postulates that sensory systems minimize metabolic cost while maximizing the amount of information that is encoded about a stimulus [24, 25, 26].

The goal of maximizing stimulus information does not reflect the fact that different stimuli can have different utility to a system for making inferences about the environment [27, 28, 29, 30]. The relative utility of a stimulus is determined by the potential impact that it can have on the system's belief about the state of the environment; stimuli that sway this belief carry high utility, and they should be encoded with high fidelity in order to minimize error in inference. Stimuli that do not change this belief are not relevant, and they can be encoded with low fidelity without negatively impacting the inference process.

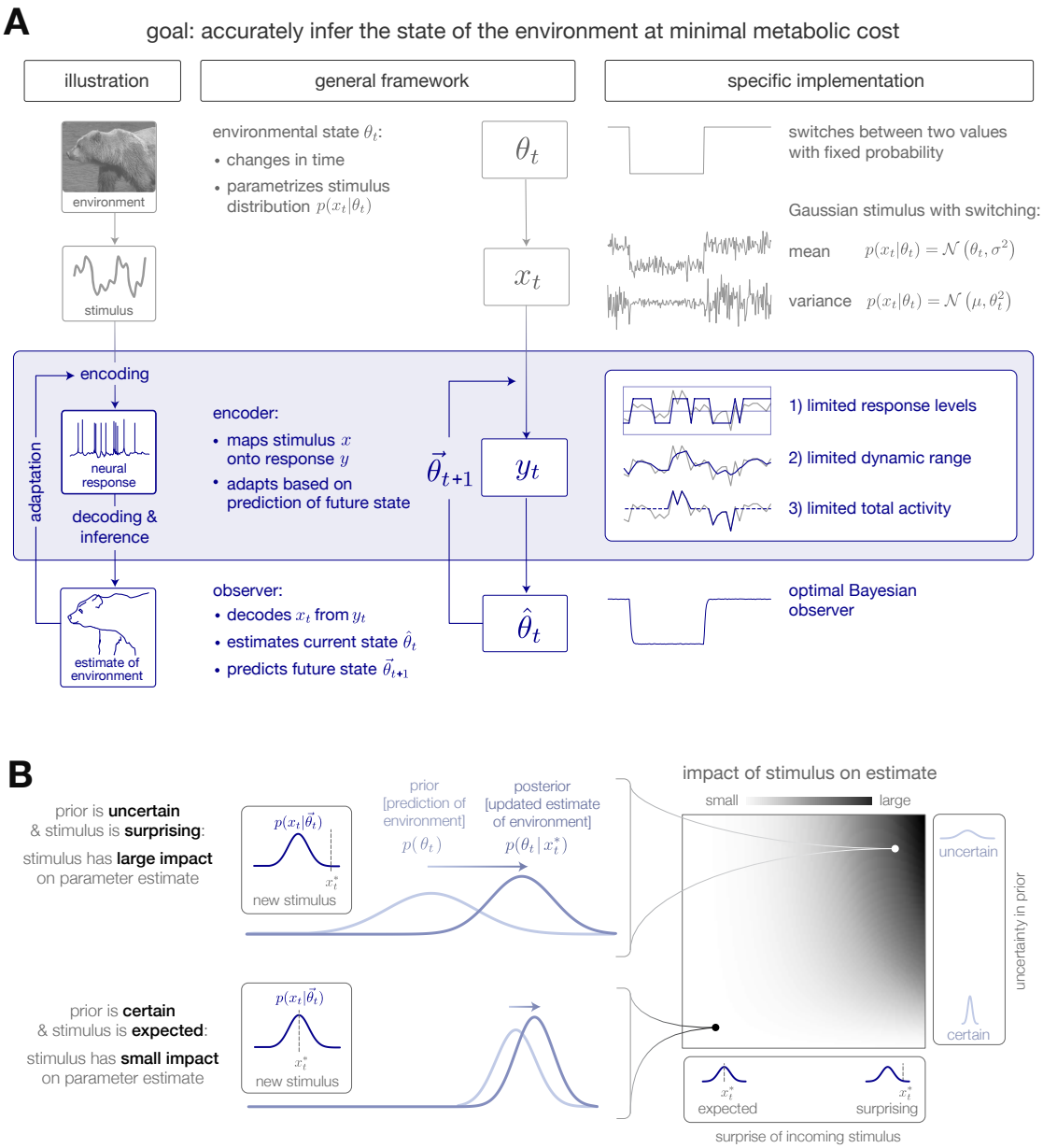


Figure 1: **A framework for dynamically balancing coding cost and inference error. (A)** We develop a framework (left, middle columns) in which a sensory system infers the state of a dynamic environment at minimal metabolic cost. The state of the environment is signaled by sensory stimuli that are encoded in neural responses. To infer this state, the system must decode stimuli from neural responses and use them to update an internal model of the environment. This internal model can then be used to adapt the encoding at earlier stages. We consider a specific implementation (right column) in which an environmental state θ_t switches between two values with fixed probability. This state parameterizes the mean or variance of a Gaussian stimulus distribution, $p(x_t|\theta_t)$. Stimuli x_t are drawn from this distribution and encoded in neural responses y_t. We consider three encoding schemes that reflect limitations on neural response levels, dynamic range, and total activity. Each encoding scheme reduces the fidelity of the neural response and introduces error into any downstream computations. An optimal Bayesian observer uses the neural response to construct an estimate $\hat{\theta}_t$ and a prediction $\vec{\theta}_{t+1}$ of the environmental state. This prediction is fed upstream and used to adapt the encoder. **(B)** Incoming stimuli can have varying impact on the observer's estimate of the environmental state depending on the relationship between the observer's uncertainty and the surprise of the stimulus (heat map; right). We use the example of Bayesian mean estimation [31] to demonstrate that when the observer is uncertain (wide prior $p(\theta_t)$) and the stimulus x_t^* is surprising (x_t^* falls on the edge of the distribution $p(x_t|\vec{\theta}_t)$), the stimulus has a large impact on the observer's estimate. This causes a large shift in the posterior $p(\theta_t|x_t^*)$ (upper left; schematic). In contrast, when the observer is certain and the stimulus is expected, the stimulus has a small impact on the observer's estimate (lower left; schematic).

When the distribution of stimuli changes in time, as in any natural environment, both the belief about the environment [11] and the relative impact of different stimuli on this belief also change in time. Any system that must perform accurate inference with minimal energy must therefore *dynamically* balance the cost of encoding stimuli with the error that this encoding can introduce in the inference process. While studies have separately shown that sensory neurons dynamically adapt to changing stimulus distributions in manners that reflect either optimal encoding [32] or inference [10], the interplay between these two objectives is not understood.

In this work, we develop a general framework for relating low-level sensory encoding schemes to the higher-level processing that ultimately supports behavior. We use this framework to explore the dynamic interplay between encoding, which serves to represent the stimulus with minimal metabolic cost, and inference, which serves to estimate behaviorally-relevant properties of the stimulus with minimal error. To this end, we consider a system whose goal it is to infer the state of a non-stationary environment from a stream of incoming stimuli that are encoded in neural responses. Here, the environmental state represents the underlying cause of the stimulus; it is this underlying cause, and not the stimulus itself, that we consider to be of behavioral relevance. To illustrate the implications of this framework, we consider three neurally-plausible encoding schemes that perform different transformations—discretization, temporal filtering, and stimulus selection—on incoming stimuli. Each scheme reflects a different limitation on neural response properties, and consequently each represents a different strategy for reducing the metabolic cost of encoding.

When optimized for maintaining an accurate estimate of the environmental state, we find that all three encoding schemes can significantly reduce metabolic costs while minimally affecting the inference process. They do so by adjusting the fidelity of the encoding based on the surprise of incoming stimuli and the observer's uncertainty about the state of the environment. When the environment is stable, the system maintains low metabolic costs by encoding incoming stimuli with low fidelity. Following a change in the environment, the system expends more energy to increase the fidelity of the encoding, thereby enabling the system to quickly adapt its model of the environment based on the new distribution of stimuli. This behavior differs significantly from encoding schemes that are designed to accurately reconstruct all details of the stimulus, regardless of the relevance of these details for performing inference. In such cases, the resulting neural responses are more metabolically costly, and the inference process exhibits qualitatively different inaccuracies.

Together, these results predict dynamical signatures that would differentiate an encoding strategy designed to support inference from one designed to reconstruct the stimulus itself. These dynamical signatures also provide a new interpretation of experimentally-observed phenomena such as burst coding, firing-rate adaptation, and perceptual metamers. We argue that these phenomena could arise as a consequence of a dynamic tradeoff between coding cost and inference error.

Framework

Sensory systems must use internal representations of external stimuli to build and update models of the environment. As an illustrative example, consider the task of avoiding a predator (Fig. 1A, left column). The predator is signaled by sensory stimuli, such as patterns of light intensity or chemical odorants, that change over time. To avoid a predator, an organism must first determine whether a predator is present, and if so, which direction the predator is moving, and how fast. This inference process requires that incoming stimuli first be encoded in the spiking activity of sensory neurons. This activity must then be transmitted to downstream neurons that infer the position and speed of the predator.

Not all stimuli will be equally useful for this task, and the relative utility of different stimuli could change over time. When first trying to determine whether a predator is present, it might be crucial to encode stimulus details that could discriminate fur from grass. Once a predator has been detected, however, the details of the predator's fur are not relevant for determining its position and speed. If encoding stimuli is metabolically costly, energy should be devoted to encoding those details of the stimulus that are most useful for inferring the quantity at hand.

We formalize this scenario within a general Bayesian framework consisting of an encoder and an observer (Fig. 1A, middle column). We then use a simplified model to explore the interplay between encoding and inference (Fig. 1A, right column). The notion of metabolic cost is implicit within this framework. First, neurons have a finite representational capacity, such as a finite number of response patterns or a finite dynamic range, that is controlled by their physical response properties. This finite capacity places a limitation on the maximum fidelity with which stimuli can be encoded in neural responses. Second, the use of this finite capacity is metabolically costly; it requires more energy to

encode a stimulus with high fidelity than with low fidelity. We consider three neurally-plausible encoding schemes that reflect these limitations (blue box in Fig. 1A). In one scheme, the encoder is constrained in the total number of distinct responses it can produce at a given time, and uses a discrete set of neural response levels to represent a stimulus. In second scheme, the encoder is constrained in dynamic range and temporal acuity, and filters incoming stimuli in time. Finally, we consider an encoder that is constrained in the total amount of activity that can be used to encode a stimulus, and must therefore selectively encode certain stimuli and not others.

An efficient encoding scheme should devote energy at times when incoming stimuli will have a large impact on the observer's belief about the state of the environment. The relative impact of different stimuli is modulated by the observer's uncertainty about this state, and by the surprise of incoming stimuli (Fig. 1B). When the observer is certain about the state of the environment or when the stimulus is consistent with the observer's belief, the new stimulus has little influence over the observer's belief. Conversely, when the observer is uncertain or when the new observation is surprising, the stimulus has a large impact. Given limited metabolic resources, incoming sensory stimuli should be encoded with high fidelity when the observer is uncertain about the state of the environment, or when incoming stimuli are unexpected given the observer's belief. In what follows, we substantiate this intuition by adapting each encoding scheme over time in a manner that optimally balances coding cost and inference error.

Model environments

To make our considerations concrete, we model an optimal Bayesian observer in a two-state environment (Fig. 1A, right column). Despite its simplicity, this model has been used to study the dynamics of inference in neural and perceptual systems, and can generate a range of complex behaviors [11, 33, 12, 34, 35]. Within this model, the environment is described by a time-varying state variable θ_t that switches randomly between a "low" and a "high" state at a small but fixed hazard rate h (we take $h = 0.01$). This state parametrizes the instantaneous distribution of sensory stimuli, $p(x_t|\theta_t)$. Here, we take θ_t to specify either the mean or the standard deviation of a Gaussian stimulus distribution, and we refer to these as "mean-switching" and "variance-switching" environments, respectively. At each point in time, a single stimulus sample x_t is drawn randomly from this distribution.

In standard Bayesian inference, the observer would have direct access to the incoming stimulus x_t. In our framework, the stimulus is mapped onto a neural response y_t via an adaptive encoding step (blue box in Fig. 1A). Before updating its belief about the environment, the observer must first decode the stimulus from the neural response. This decoded stimulus can then be combined with past evidence and knowledge of environment dynamics to update the observer's belief about the current state of the environment. This belief is summarized by the posterior distribution $p(\theta_t|y_{\tau\leq t})$, which for a two-state environment is given by the probability P_t^L that the environment is in the low state at time t. A prediction about the future state of the environment, which can be computed in an analogous manner, is then fed back upstream and used to adapt the encoder. Taken together, the process can be summarized by the following steps (as detailed in Box 1): encode stimulus, decode stimulus, estimate current state, predict future state, adapt encoder.

In order to optimize and assess the dynamics of the system, we use the point value $\hat{\theta}_t$ as an estimate of the current state of the environment. The optimal point estimate is computed by averaging the posterior and is guaranteed to minimize the mean squared error between the estimated state $\hat{\theta}_t$ and the true state θ_t in the absence of any encoding [36]. Note that although the environmental states are discrete, this optimal estimate is continuous (Methods). In an analogous manner, we use the point value $\vec{\theta}_{t+1}$ as a prediction of the future state of the environment. For small hazard rates (as considered here), the prediction $\vec{\theta}_{t+1}$ can be approximated by the current estimate $\hat{\theta}_t$ (Methods).

Uncertainty and surprise in model environments

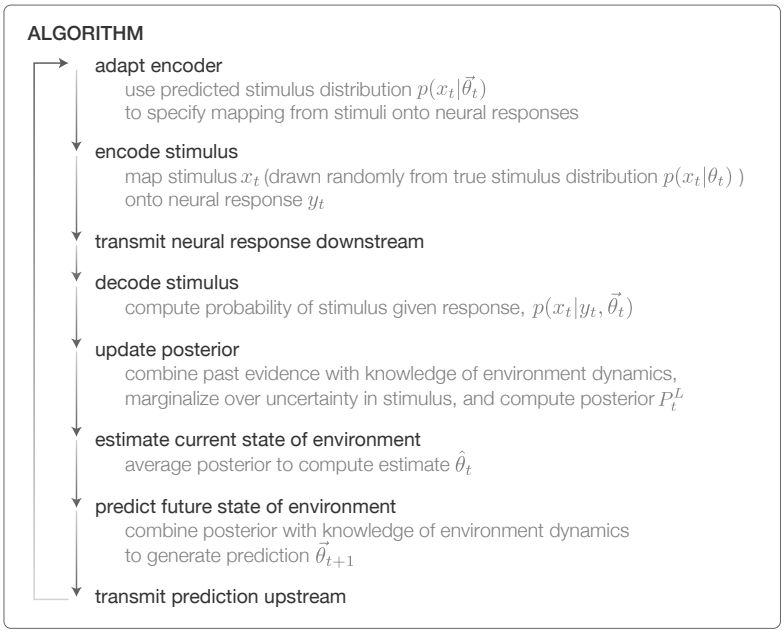
We expect efficient encoding schemes to operate on uncertainty and surprise, each of which relates to the properties of the model environments considered here. In a two-state environment, the observer's uncertainty is captured by the entropy of the posterior, $H(P_t^L)$. When the posterior is either near zero or near one, the observer is certain that the environment is in the high or low state, respectively. When the posterior is near 0.5, the observer is highly uncertain; it believes that the environment is equally likely to be in either the high state or the low state.

The degree to which incoming stimuli are surprising to the observer depends on two factors: the average surprise of the stimuli themselves, which is a property intrinsic to the stimulus distribution, and

the alignment of the observer's belief with this stimulus distribution.

The average surprise of stimuli drawn from a known distribution is measured by the entropy of that distribution. The entropy of a Gaussian distribution, as considered here, is fully specified by its variance. High-variance stimuli will therefore be more surprising to an observer, on average, than low-variance stimuli.

An observer will additionally be surprised if its internal prediction is misaligned with the true state of the environment, and observed stimuli do not match its expectations. For a Gaussian stimulus distribution, this alignment is captured by the relationship between the true mean and variance, μ and σ^2, and the predicted mean and variance, $\vec{\mu}$ and $\vec{\sigma}^2$. When the variance of the true stimulus distribution is fixed but the mean is changing in time, the average surprise of incoming stimuli depends only on the squared difference between the true and predicted mean, $(\mu - \vec{\mu})^2$. The average surprise will thus increase by the same amount in response to increases versus decreases in the mean. As a result, any encoding strategy that relies on stimulus surprise should be symmetric to increases and decreases in the mean.



Box 1: **Algorithm for performing Bayesian inference with adaptively-encoded stimuli.**

Conversely, if the mean of the stimulus distribution is fixed but the variance is changing in time, the average surprise depends on the ratio of the true and predicted variances, $\sigma^2/\vec{\sigma}^2$. Stimuli be will more surprising if the observer's prediction underestimates the variance of the true stimulus distribution. As a result, an encoding strategy that relies on stimulus surprise will be asymmetric to increases and decreases in variance.

RESULTS

We consider three parametric encoding schemes that can be manipulated to adjust the fidelity with which stimuli are mapped onto neural responses. For each scheme, we impose a global constraint that controls the maximum fidelity of the encoding. We then adapt the instantaneous fidelity of the encoding, subject to the global constraint. All three encoding schemes reduce the entropy of the neural response relative to the entropy of the raw stimulus; we use the entropy of the response as a measure of the metabolic cost of encoding [17, 37].

We adapt each encoding strategy in time to optimally balance coding cost and inference error. Because information is lost in the process of mapping stimuli onto neural responses, the estimate $\hat{\theta}_{y,t}$ constructed from the neural response will typically be less accurate than the optimal estimate $\hat{\theta}_{x,t}$ constructed directly from the stimulus. The difference between these estimates, $(\hat{\theta}_{x,t} - \hat{\theta}_{y,t})^2$, is a measure of the *error in inference* induced by mapping a stimulus x_t onto a neural response y_t. This function assigns high error to those mappings that impact the estimate $\hat{\theta}_t$, regardless of how much error is induced between x_t and

y_t. In what follows, we we will use $\hat{\theta}_t$ and $\vec{\theta}_t$ to denote the estimates and predictions constructed from the neural response, unless otherwise noted.

At each timestep, the parameters of the encoding are chosen to minimize the error in inference when averaged over the predicted distribution of stimuli, $p(x_t|\vec{\theta}_t)$. We compare this minimization to one in which the goal is to reconstruct the stimulus itself; in this case, the *error in reconstruction* is given by $(x_t - y_t)^2$. In both cases, the encoding is optimized based on the internal prediction of the environmental state. As a result, the entropy of the neural response (and thus the metabolic cost of the encoding) will depend on how closely this prediction aligns with the true state of the environment.

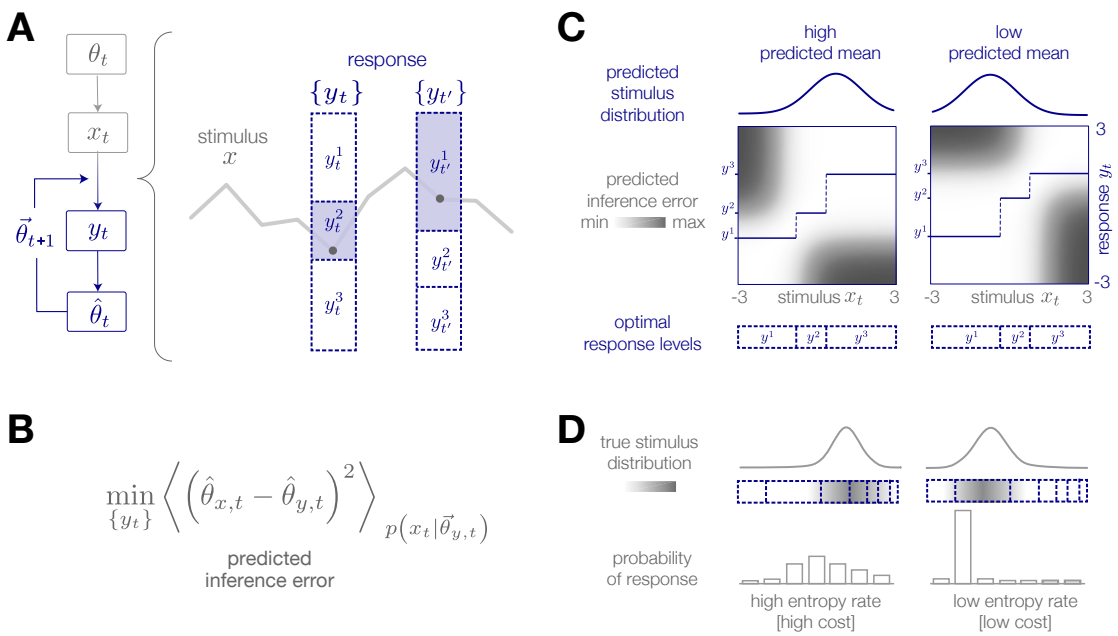


Figure 2: **Encoding via discrete response levels. (A)** (Schematic) At each timestep, an incoming stimulus x_t (gray dot) is mapped onto a discrete neural response level y_t^i (solid blue rectangle) chosen from a set $\{y_t^i\}$ (dotted rectangles). The encoder can dynamically adapt the response levels based on an internal prediction of the stimulus distribution. **(B)** The optimal set of response levels is found at each timestep by minimizing the average predicted error in inference, subject to a constraint on the number of levels. **(C)** The predicted inference error induced by mapping a stimulus x_t onto a neural response y_t varies as a function of the observer's belief P_t^L about the state of the environment (shown for $P_t^L = 0.12$, left; $P_t^L = 0.88$, right). There are many possible ways of mapping stimuli onto neural responses while maintaining low inference error (white regions on both panels). The optimal response levels $\{y^1, y^2, y^3\}$ (solid line and dotted rectangles) are chosen to minimize this error when averaged over the predicted stimulus distribution. **(D)** (Schematic) The probability that each neural response level will be used to encode incoming stimuli is determined by partitioning the true stimulus distribution into the optimal response levels. The resulting probability distribution determines whether the entropy rate of the encoding is high (probability is distributed across many levels; left) or low (probability is concentrated on few levels; right).

To illustrate the dynamic relationship between encoding and inference, we use a "probe" environment that switches between two states at fixed intervals of $1/h$ timesteps. This specific instantiation is not unlikely given the observer's model of the environment [11], and allows us to illustrate average behaviors over many cycles of the environment.

Limited neural response levels

In many systems, neural responses are discrete. Neurons use precise sequences of spikes [38] or discrete firing rate levels [26, 39] to represent continuous stimuli. This inherent discreteness imposes a fundamental limitation on the number of distinct neural responses that can be used to represent a continuous stimulus space. Many studies have argued that sensory neurons make efficient use of limited response

levels by appropriately tuning these levels to match the steady-state distribution of incoming stimuli (e.g. [26, 40]).

Here, we consider an encoder that adaptively maps an incoming stimulus x_t onto a discrete set of neural response levels $\{y_t^i\}$ (Fig. 2A). Because there are many more stimuli than levels, each level must be used to represent multiple stimuli. The number of levels reflects a global constraint on representational capacity; fewer levels indicates a stronger constraint and results in a lower-fidelity encoding.

The encoder can adapt this mapping by expanding, contracting, and shifting the response levels to devote higher fidelity to different regions of the stimulus space. We consider an optimal strategy in which the response levels are chosen to minimize the predicted inference error, subject to a constraint on the number of levels (Fig. 2B, Methods; note that when the variance is switching over time, we define these levels with respect to the absolute deviation from the mean). The predicted inference error induced by encoding a stimulus x_t in a response y_t changes over time as a function of the observer's prediction of the environmental state (Fig. 2C). For example, when the observer predicts that the mean of the stimulus distribution is high (Fig. 2C, left), encodings that alter small stimulus values will cause the highest error, while encodings that alter large stimulus values will have minimal impact (and vice-versa when the observer predicts that the mean is low). Because some stimuli have very little impact on the estimate at a given time, they can be mapped onto the same neural response level without inducing error in the estimate. The optimal response levels are chosen to minimize this error when averaged over the predicted distribution of stimuli. The relative width of each level is a measure of the resolution devoted to different regions of the stimulus space; narrower levels devote higher resolution (and thus higher fidelity) to the corresponding regions of the stimulus space.

While the optimal set of response levels is determined by the observer's prediction of the stimulus distribution, the output of these response levels is determined by the true stimulus distribution. An encoding that devotes higher resolution to stimuli that are likely to occur in the environment will result in a higher entropy rate (and thus higher cost), because many different response levels will be used with relatively high frequency (Fig. 2D, left). In contrast, if an encoding scheme devotes high resolution to surprising stimuli, very few response levels will be used, and the resulting entropy rates and costs will be low (Fig. 2D, right).

We use simple mean- and variance-switching environments (Fig. 3A) to illustrate the behavior of an encoder that is optimized for inference (Fig. 3B) versus stimulus reconstruction (Fig. 3C). These qualitatively different encoding strategies impact both the cost of encoding and the error in inference through feedback from the observer (Fig. 3D-F).

When designed for accurate inference, we find that the optimal encoder devotes its resolution to stimuli that would be likely if the environment were to switch states (Fig. 3B). In a mean-switching environment (left column of Fig. 3), the stimuli that are likely under a switch are also those stimuli that are surprising given the current estimate. When the environment is stable and the observer's prediction is accurate, resolution is devoted to surprising stimuli, and likely stimuli are mapped onto the same response level (Fig. 3B, left). The resulting entropy rate is low (blue curve in Fig. 3D, left). When the environment changes abruptly, stimuli that were surprising under the observer's model become highly likely under the new stimulus distribution, and are given more resolution by the encoder. This results in a large spike in entropy rate, which enables the observer to quickly adapt its estimate to the change (blue curve in Fig. 3E, left). As the observer's estimate begins to align with the new environmental state, resolution once again shifts toward stimuli that are surprising, and the entropy rate drops.

In the case of a variance-switching environment (right column of Fig. 3), stimuli that are likely under a switch are not always surprising given the current estimate. In particular, when variance is high, stimuli that are likely under a switch fall within the bulk of the distribution (rather than within the tails, as in mean estimation), and thus have relatively high likelihood. Because of this, entropy rates remain high during the high-variance state (blue curve in Fig. 3D, right). In the low-variance state, stimuli that would be likely under a switch are again those stimuli that are surprising given the current estimate, and entropy rates remain low.

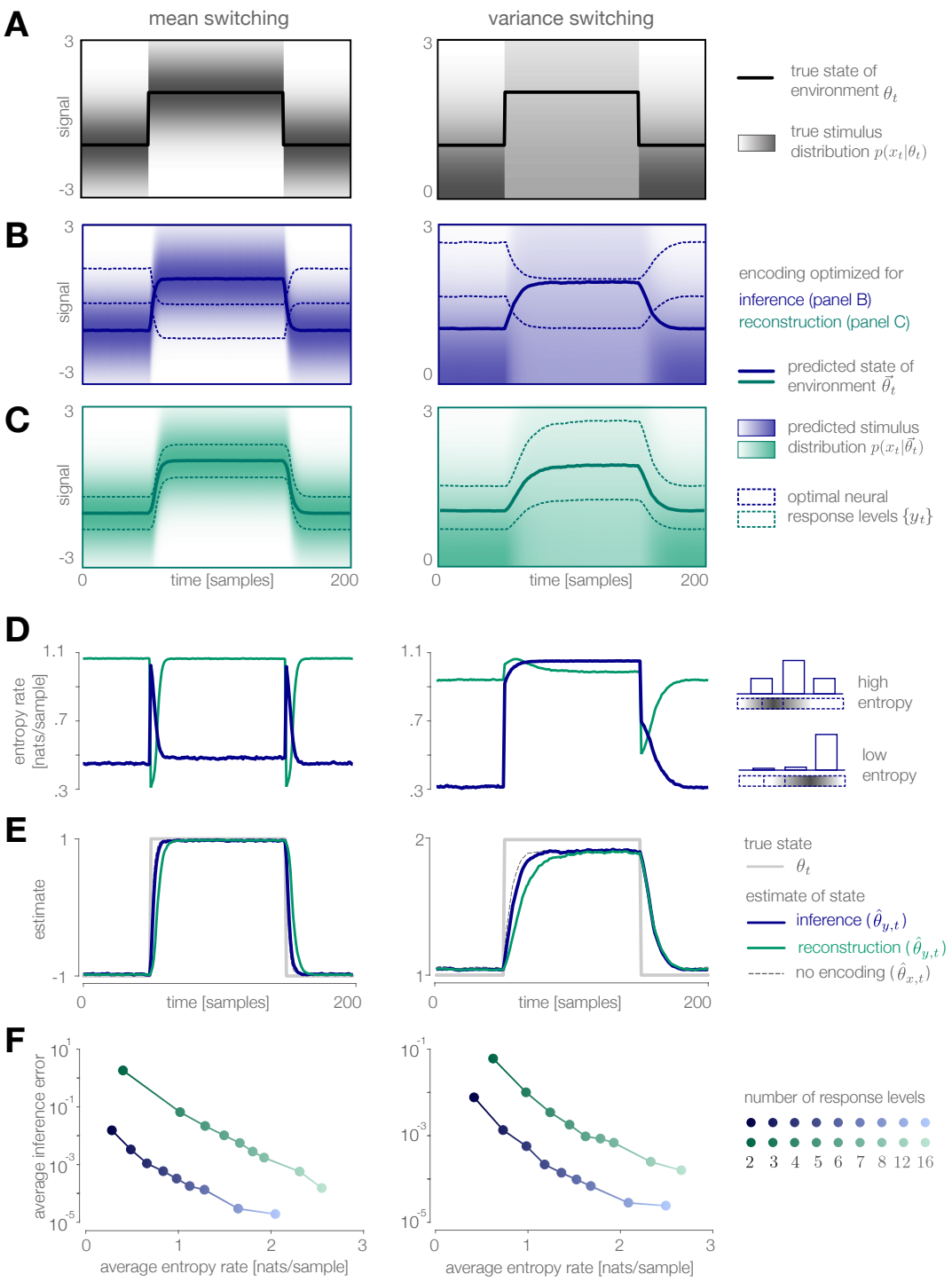


Figure 3: **Dynamic inference with optimally-adapted response levels. (A)** We consider a probe environment in which a state θ_t (solid line) switches between two values at fixed time intervals. This state parametrizes the mean (left) or the variance (right) of a Gaussian stimulus distribution (heatmap). **(B, C)** Optimal response levels (dotted lines) are chosen to minimize error in inference (blue) or stimulus reconstruction (green) based on the predicted stimulus distribution $p(x_t|\vec{\theta}_t)$ (heatmap). Results are shown for three response levels. All probability distributions in panels A-C are scaled to the same range, $[0, .4]$. **(B)** Response levels optimized for inference devote higher resolution (narrower levels) to stimuli that would be likely to occur if the environment were to switch states. **(C)** Response levels optimized for stimulus reconstruction devote higher resolution to stimuli that are likely to occur given the current prediction.

Figure 3: **Continued from previous page. (D)** The entropy rate of the encoding is found by partitioning the true stimulus distribution (heatmap in panel A) based on the optimal response levels (dotted lines in panels B-C). Abrupt changes in the environment induce large changes in entropy rate that are symmetric for mean estimation (left) but asymmetric for variance estimation (right). Apparent differences in the baseline entropy rate for low- versus high-mean states arise from numerical instabilities. **(E)** Encoding induces error into the estimate $\hat{\theta}_t$. Errors are larger if the encoding is optimized for stimulus reconstruction than for inference. The error induced by upward and downward switches is symmetric for mean estimation (left) but asymmetric for variance estimation (right). In the latter case, errors are larger when inferring upward switches in variance. **(F)** Increasing the number of response levels decreases the average inference error but increases the cost of encoding. Across all numbers of response levels, an encoding optimized for inference (blue) achieves lower error at lower cost than an encoding optimized for stimulus reconstruction (green). All results in panels A-C and E are averaged over 500 cycles of the probe environment. Results in panel D were computed using the average response levels shown in panels B-C. Results in panel F were determined by computing time-averages of the results in panels D-E.

When designed for accurate reconstruction of the stimulus, we find that the optimal encoder devotes its resolution to stimuli that are likely given the current environmental state (Fig. 3C). When the environment is stable and the observer's prediction is accurate, the resulting entropy rate is high, regardless of the environment. When the mean switches upward or downward, stimuli that were highly likely under the observer's model become less likely under the new stimulus distribution, and are given less resolution by the encoder. This results in a local drop in the entropy rate (green curve in Fig. 3D, left), which slows the observer's detection of changes in the environment (green curve in Fig. 3E, left). The same is true when the variance switches downward. When the variance switches upwards, stimuli that were surprising under the observer's model become likely under the new stimulus distribution, and stimuli that were likely remain likely. As a result, the entropy rate increases (green curve in Fig. 3D, right).

The observed differences between discrete encoders optimized for inference (Fig. 3B) versus stimulus reconstruction (Fig. 3C) imply that optimal stimulus discretization depends strongly on the objective of the encoding. These results agree with a recent study [39], which also found differences between discrete stimulus encodings optimized for information maximization versus stimulus reconstruction.

Whether optimizing for inference or stimulus reconstruction, the entropy rate, and thus the coding cost, changes dynamically over time in a manner that is tightly coupled with the inference error. The average inference error can be reduced by increasing the number of response levels, but this induces a higher average coding cost (Fig. 3F). Regardless of the number of levels, a strategy optimized for inference achieves lower inference error at lower coding cost than a strategy optimized for stimulus reconstruction.

To illustrate how this encoding scheme could translate into neural responses, we consider a simple mapping from the discrete response levels of the encoder onto binary spike patterns. As an example, we assign the patterns [00], [01], [10], and [11] to an encoder with four discrete response levels. The pattern with the lowest spike count, [00], is assigned to the response level that is predicted to be used with highest probability based on the optimal partitioning shown in Fig. 3B (and vice-versa for the pattern with the highest spike count). This procedure, known as entropy coding, guarantees a minimal code length that uses the fewest number of spikes [41]. Although the predicted stimulus distribution determines how spike patterns are assigned to response levels, it is the true stimulus distribution that determines the frequency with which each pattern is used (Fig. 4A).

When the environment is changing, incoming stimuli are encoded by a diverse set of spike patterns (Fig. 4B). This leads to a transient increase in firing rate that closely resembles the entropy rate of the optimal encoder in Fig. 3D (left). As shown in Fig. 3E (left), this transient increase enables the observer to adapt its prediction to the new stimulus distribution. At times when the environment is stationary and the observer maintains an accurate prediction, stimuli that are consistent with this prediction are mapped onto the same spike pattern ([00]), and the firing rate is low. Once encoded in the same pattern, the responses to physically different stimuli can no longer be distinguished from one another. This gives rise to the counterintuitive result that stimuli become *less distinguishable* to the observer as its model of the environment becomes *more accurate* (Fig. 4C-D). This behavior arises as a consequence of an encoding strategy that is optimized for inference; the converse result arises from a strategy optimized for stimulus reconstruction. This behavior is closely tied to the idea of metamers: stimuli that differ in their physical characteristics but elicit the same percept [42, 43]. Our results suggest that the likelihood of a metamer should change over time as a function of the alignment between the observer's internal prediction and the true state of the environment.

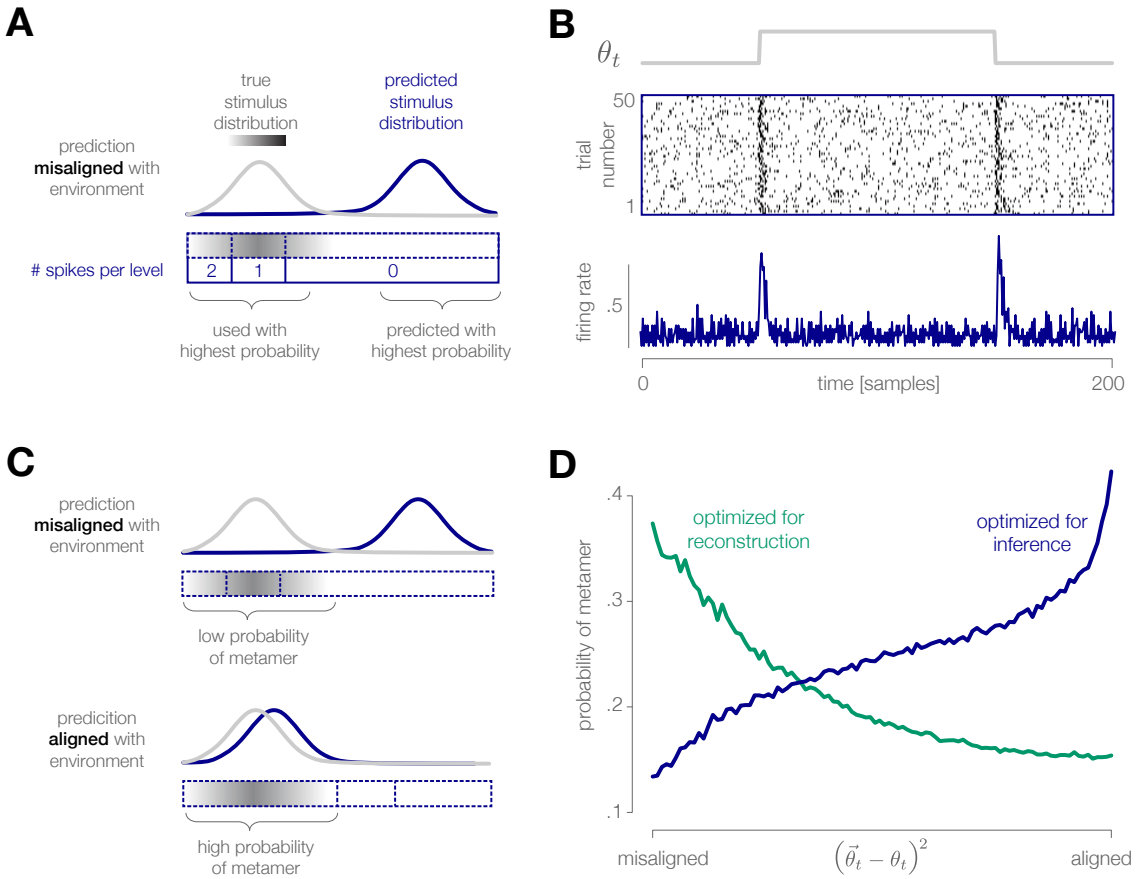


Figure 4: **Response levels optimized for inference give rise to bursts and metamers. (A)** (Schematic) Low spike patterns are assigned to response levels with the highest probability given the predicted stimulus distribution (and vice-versa). When this prediction is misaligned with the true stimulus distribution, high spike patterns are used with higher probability. **(B)** A spike raster (upper row) is simulated by drawing patterns of spikes according to the set of response levels optimized for inference (simulated for an encoder with four response levels operating in a mean-switching environment; see text). Bursts of spiking activity follow changes in the stimulus distribution. The resulting firing rates (lower row; averaged over 400 cycles of the probe environment) closely resemble the entropy rates shown in the left panel of Fig. 3D. A similar match between firing rates and entropy rates was observed across environments (mean- and variance-switching) and across optimization schemes (response levels optimized for inference and reconstruction). **(C)** (Schematic) If the observer's prediction is misaligned with the true state of the environment, the stimuli that are most likely to occur in the environment are mapped onto different neural responses. The probability of a metamer—defined by the set of stimuli that give rise to the same neural response—is low. Conversely, when the prediction is closely aligned with the environment, the probability of a metamer is high. **(D)** When an encoding is optimized for inference, the probability of a metamer increases as the observer's prediction aligns with the environment, because an increasing fraction of stimuli are mapped onto the same neural response. The opposite trend is observed when the encoding is optimized for stimulus reconstruction (illustrated for an encoder with eight response levels operating in a mean-switching environment).

Limited gain and temporal acuity

Neural responses have limited gain and temporal acuity, a feature that is often captured by linear filters. For example, neural receptive fields are often characterized as linear temporal filters, sometimes followed by a nonlinearity [44, 38]. The properties of these filters are known to dynamically adapt to changing stimulus statistics (e.g. [45, 46]), and numerous theoretical studies have suggested that such filters are adapted to maximize the amount of information that is encoded about the stimulus [24, 47, 48].

Here, we consider an encoder that implements a very simple temporal filter:

$$y_t = \alpha_t x_t + (1 - \alpha_t) x_{t-1} \tag{1}$$

where $\alpha_t \in [0.5, 1]$ is a coefficient that specifies the shape of the filter and controls the instantaneous fidelity of the encoding. When $\alpha_t = 0.5$, the encoder computes the average of current and previous stimuli by combining them with equal weighting, and the fidelity is minimal. When $\alpha_t = 1$, the encoder transmits the current stimulus with perfect fidelity (i.e., $y_t = x_t$).

The encoder can adapt α_t in order to manipulate the instantaneous fidelity of the encoding (Fig. 5A). We again consider an optimal strategy in which the value of α_t is chosen to minimize the predicted inference error, subject to a constraint on the predicted entropy rate of the encoding (Fig. 5B, Methods). The entropy constraint favors a local reduction in coding fidelity through the manipulation of gain and temporal correlation. The fixed multiplier β reflects a global constraint on representational capacity; larger values of β correspond to stronger constraints and reduce the maximum fidelity of the encoding.

The amount of error induced by averaging consecutive stimuli depends not only on the strength of the averaging (value of α_t), but also on the observer's certainty about the stimulus distribution (value of P_t^L) (Fig. 5C). When the observer is uncertain, strong averaging induces high error. When the observer is certain, the error is small, regardless of how incoming stimuli are combined. In this limit, even strong averaging of incoming stimuli exerts little influence on the observer's estimate, and the observer need not rely strongly on sensory data. The same is not true of the predicted entropy rate of the encoding. This entropy rate depends on the predicted entropy of the stimulus distribution itself; the more variable the stimulus distribution, the higher its entropy, and the costlier it is to encode. Thus, the predicted entropy rate of the encoding depends on the strength of averaging and on the observer's prediction of stimulus variance; if the variance is fixed, the entropy rate depends only on the strength of averaging (Fig. 5C, upper right panel). These two terms—entropy and error—are combined to find the optimal value of α_t given the observer's prediction about the state of the environment (Fig. 5D).

When designed to support accurate inference, we find that the optimal encoder devotes higher fidelity at times when the observer is uncertain, and at times when the predicted stimulus variance is high. In the latter case, this relates directly to stimulus surprise; stimuli that are drawn from a high-variance distribution are more surprising on average. In a mean-switching environment, the stimulus variance is fixed (Fig. 6A, left), and thus the fidelity depends only on the observer's uncertainty. At times when the environment is stable and the observer's prediction is accurate, stimuli can be strongly averaged without adversely impacting the estimate (Fig. 6B, D, left). When the environment switches abruptly, the uncertainty in the observer's prediction grows. As a result, the encoder transiently increases the fidelity of the encoding by reducing the amount of averaging. This transient increase enables the observer to quickly adapt its estimate (Fig. 6D, left). The entropy constraint impacts the overall fidelity of the encoding; the stronger the constraint, the more strongly stimuli are averaged, and the lower the fidelity. The accuracy of the inference, however, is highly robust to the strength of the entropy constraint; even when incoming stimuli are strongly averaged ($\alpha_t = 0.5$), the encoder transmits the mean of two consecutive samples, which is precisely the statistic that the observer is trying to estimate.

In a variance-switching environment, both the observer's uncertainty and the predicted stimulus variance change in time (Fig. 6A, right). More variable stimuli are encoded with higher fidelity and weaker averaging than less variable stimuli. When the variance changes abruptly, the observer's uncertainty grows, leading to an additional increase in fidelity (Fig. 6B, right). Both the encoder and the observer are slower to respond to changes in variance than to changes in mean, and the accuracy of the inference is more sensitive to the strength of the entropy constraint (Fig. 6D, right).

When designed to accurately reconstruct the stimulus, the fidelity of the optimal encoder depends only on the predicted stimulus variance, and not on the observer's uncertainty. In a mean-switching environment, the variance is fixed (Fig. 6A), and thus the fidelity is flat across time. In a variance-switching environment, the fidelity increases with the predicted variance of incoming stimuli, not because variable stimuli are more surprising, but rather because they are larger in magnitude and can lead to higher errors in reconstruction (Fig. 6C). As the strength of the entropy constraint increases, the encoder devotes proportionally higher fidelity to high-variance stimuli because they have a greater impact on the reconstruction error.

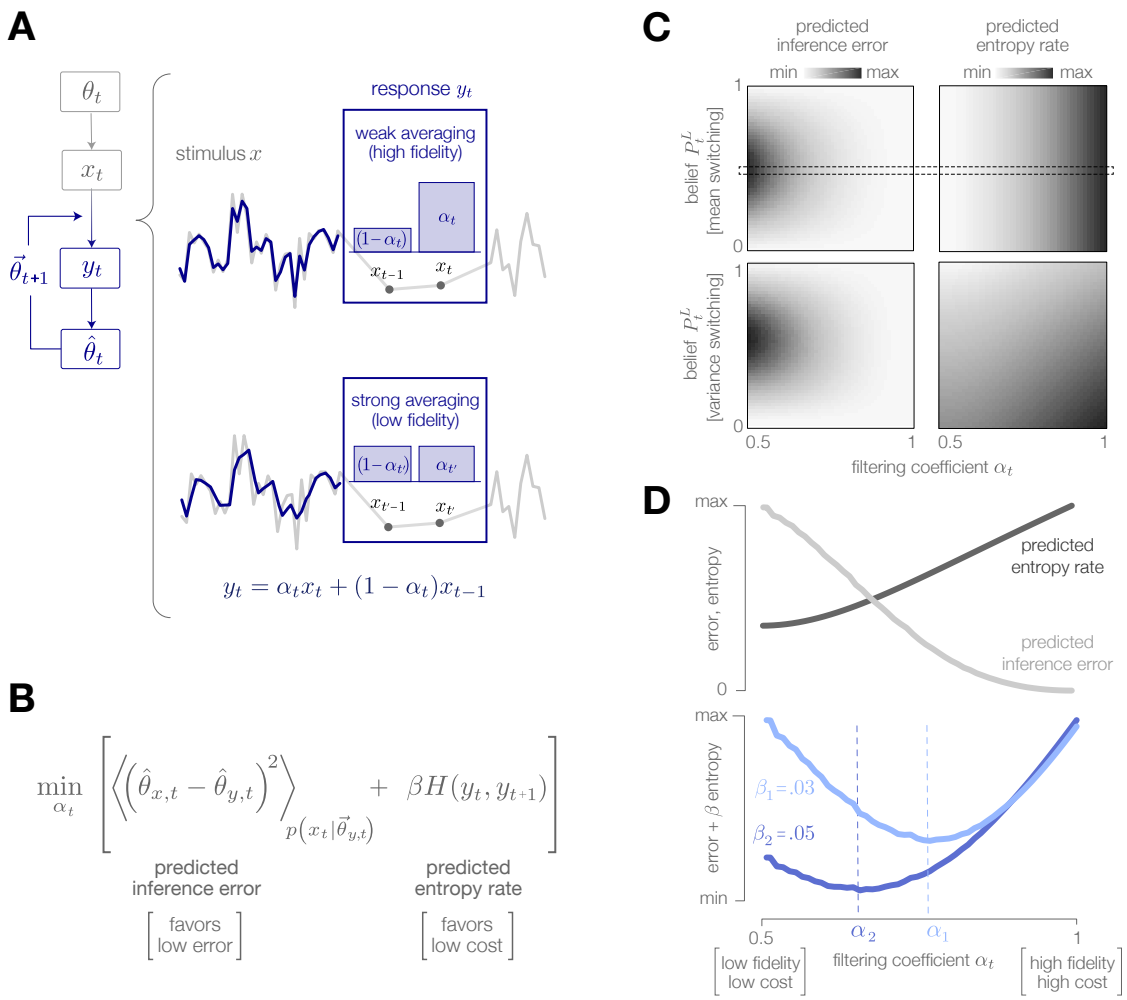


Figure 5: **Encoding via temporal filtering.** **(A)** (Schematic) At each timestep, incoming stimuli are combined via a linear filter with a coefficient α_t. When $\alpha_t = 1$, the encoder transmits individual stimuli with perfect fidelity. When α_t is close to 0.5, the encoder transmits the average of two consecutive stimulus samples. The encoder can dynamically adapt the filter coefficient based on the predicted stimulus distribution. **(B)** The optimal filter coefficient is chosen to minimize the average predicted inference error, subject to a constraint on the predicted entropy rate of the encoding. The multiplier β determines the relative strength of the constraint. **(C)** The average predicted inference error (left column) depends on the filter coefficient α_t and on the observer's belief P_t^L about the state of the environment. The predicted entropy rate of the encoding (right column) depends on the observer's belief only as it relates to the prediction of stimulus variance; when the variance is fixed (upper panel), the entropy shows no dependence on P_t^L. **(D)** The optimal filter coefficient is found by balancing error (gray) and entropy (black) given a prediction of the environmental state (shown for $P_t^L = 0.5$ in a mean-switching environment, corresponding to the horizontal cross section in panel C). The optimal balance is determined by the strength β of the entropy constraint; a stronger constraint (darker blue curve) produces a smaller optimal value of α_t.

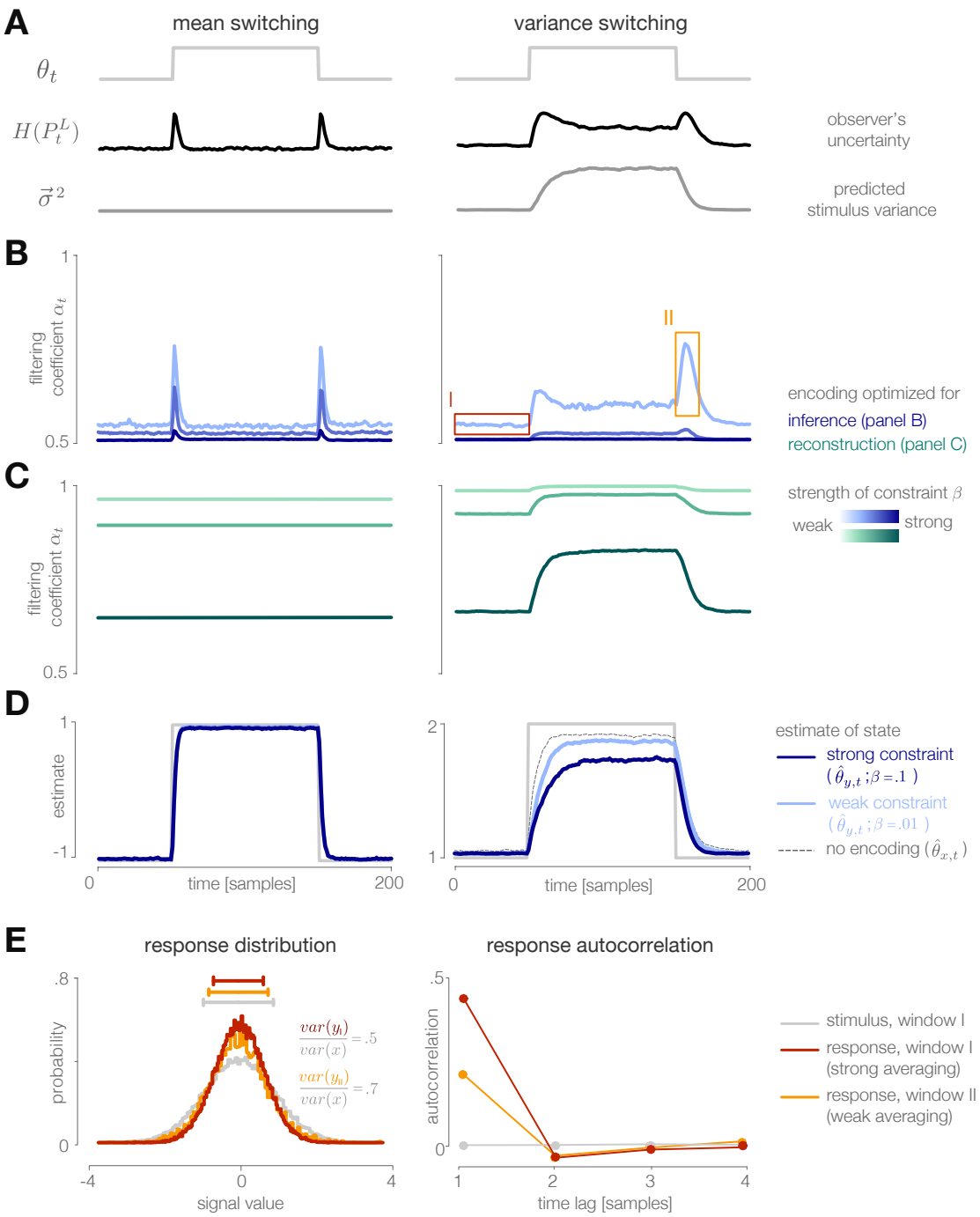


Figure 6: **Dynamic inference with optimally-adapted temporal filters.** **(A)** The observer's uncertainty (computed as the binary entropy in the posterior, $H(P_t^L)$) is largest when the environment is changing. The predicted stimulus variance (a proxy for both the predicted magnitude of the stimulus distribution, and the predicted surprise of incoming stimuli) is constant in a mean-switching environment (left) but variable in a variance-switching environment (right) (computed using a filter coefficient optimized for inference with a weak entropy constraint, corresponding to the lightest blue curves in panel B). **(B, C)** Optimal values of the filter coefficient α_t are chosen to minimize error in inference (blue) or stimulus reconstruction (green), subject to a constraint on predicted entropy. Darker colors indicate stronger constraints. **(B)** Filters optimized for inference devote high fidelity at times when the observer is uncertain and stimuli are predicted to be surprising. Shown for $\beta = 0.02, 0.1, 1$ (left) and $\beta = 0.01, 0.1, 1$ (right). **(C)** Filters optimized for reconstruction devote fidelity at times when the magnitude of the stimulus is predicted to be high. Shown for $\beta = 0.01, 0.1, 1$. **(D)** Filtering induces error into the estimate $\hat{\theta}_t$. Strong filtering has minimal impact on mean estimation (left), but induces large errors in the estimation of high variances (right).

Figure 6: **Continued from previous page. (E)** Distributions of filtered neural responses (red/orange, extracted from timepoints within the temporal windows indicated in panel A) differ from the distribution of incoming stimuli (gray). At times when the filtering coefficient is small (red; measured from $t = 1$ to $t = 50$ within window I), the distribution of neural responses has a lower variance (left panel) and higher autocorrelation (right panel) than the distribution of stimuli. When the filtering coefficient is large (orange; measured from $t = 153$ to $t = 160$ within window II), the distributions of responses and stimuli have more similar variance and autocorrelation. (Note that windows I and II are not shown to scale). All results in panels A-D are averaged over 800 cycles of the probe environment.

These results make specific experimental predictions about dynamical signatures of adaptation in neural filters. When the environment is stable (window I in Fig. 6B), we find that the gain of the optimal encoder is low. The resulting distribution of neural responses has much lower variance than the incoming distribution of stimuli, and is more correlated in time (red curves in Fig. 6E). After a change in the environment (window II in Fig. 6B), the distribution of neural responses is more similar to the distribution of stimuli, and is less correlated in time (orange curves in Fig. 6E). Together, these results suggest that when an inferential system is adapted to the stimulus distribution, gain should be low, and the temporal extent of the filter should be large. Changes in the state of the environment should cause the optimal filter to increase its gain and decrease its extent of temporal correlation.

Limited neural activity

Sensory neurons show sparse activation during natural stimulation [49, 50, 51], an observation that is often interpreted as a signature of coding cost minimization [52, 37]. In particular, early and intermediate sensory neurons may act as gating filters, selectively encoding only highly-informative features of the stimulus [53, 54]. Such a selection strategy reduces the number of spikes transmitted downstream.

Here, we consider an encoder that selectively transmits only those stimuli that are surprising and are therefore likely to change the observer's belief about the state of the environment. The average predicted surprise of incoming stimuli is given by the entropy $H[p(x_t|\vec{\theta}_t)]$ of the predicted stimulus distribution. When the observer's prediction is inaccurate, the measured surprise will differ from the predicted surprise by an amount equal to the cross-entropy between the predicted and true stimulus distributions (which is larger than the predicted surprise; see Methods for details). This difference can be used to selectively encode stimuli at times when the observer's estimate is inaccurate.

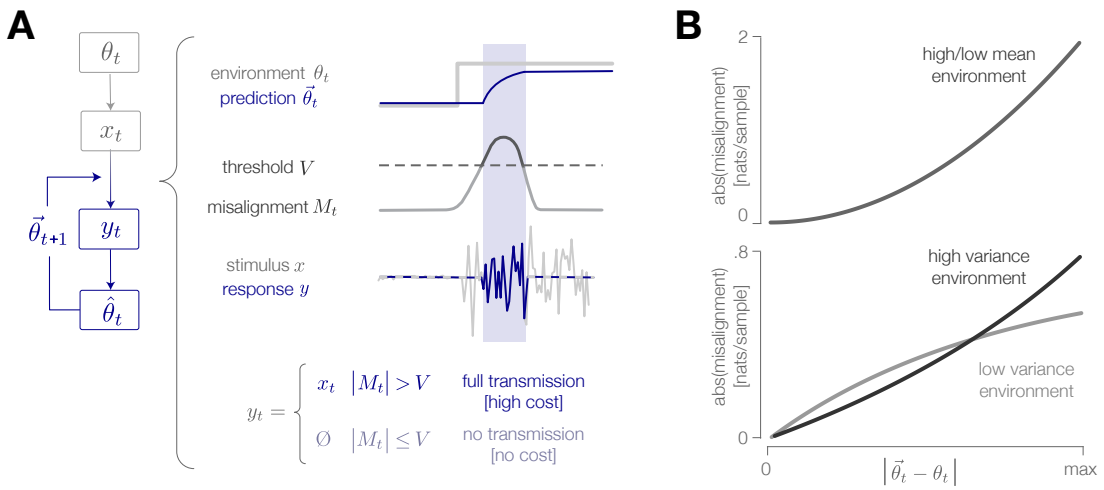


Figure 7: **Encoding via dynamic stimulus selection. (A)** (Schematic) At each timestep, the encoder computes the misalignment between the predicted and measured surprise of incoming stimuli. If the misalignment exceeds a threshold, the stimulus is encoded with perfect fidelity; otherwise, the stimulus is not encoded. **(B)** The misalignment signal (computed here analytically; see Methods) depends on the relationship between the predicted and true state of the environment. When the mean is changing over time (upper panel), the misalignment depends only on the absolute difference between the true and predicted mean. When the variance is changing over time (lower panel), the misalignment also depends on the true variance of the environment.

In practice, however, the encoder does not have access to the entropy of the true stimulus distribution. Instead, it must measure surprise directly from incoming stimulus samples. The measured surprise of each incoming stimulus sample is given by its negative log probability, $-\log p(x_t|\vec{\theta}_t)$. We consider an encoder that computes a running average of the measured surprise, and compares this running average to the predicted surprise. In this way, the encoder can heuristically assess whether a change in the stimulus distribution had occurred by computing the "misalignment" M_t between the predicted and measured stimulus distributions:

$$M_t = -\frac{1}{T}\sum_{\tau=0}^{T} \log p(x_{t-\tau}|\vec{\theta}_t) - H(x_t|\vec{\theta}_t) \tag{2}$$

The misalignment is computed over a time window T, which ensures that the observer's prediction does not gradually drift from the true value in cases where surprising stimuli are not indicative of a change in the underlying stimulus distribution (in what follows, all results are generated with $T = 10$).

The misalignment signal is both non-stationary and non-Gaussian. Optimizing an encoding scheme based on this signal would require deriving the corresponding optimal observer model, which is difficult to compute in the general case. We instead propose a heuristic (albeit sub-optimal) solution, in which the encoder selectively encodes the current stimulus with perfect fidelity ($y_t = x_t$) when recent stimuli are sufficiently surprising and the magnitude of the misalignment signal exceeds a threshold V (Fig. 7A). When the magnitude of the misalignment signal falls below the threshold, stimuli are not encoded ($y_t = \emptyset$). At these times, the observer does receive any information about incoming stimuli, and instead marginalizes over its internal prediction to update its estimate (Methods). The value of the threshold reflects a constraint on overall activity; higher thresholds result in stronger criteria for stimulus selection, which decreases the maximum fidelity of the encoding.

Because the misalignment signal is directly related to the surprise of incoming stimuli, it is symmetric to upward and downward switches in the mean of the stimulus distribution, but it is asymmetric to switches in variance and has a larger magnitude in the high-variance state (Fig. 7B depicts analytical values of the misalignment signal).

When the mean of the stimulus distribution changes in time, very few stimuli need to be encoded, on average, to maintain an accurate estimate of the environmental state (Fig. 8A-B, left). When the environment is stable and the observer's prediction is accurate, the misalignment signal is small, and very few stimuli are transmitted to the observer. When the mean changes abruptly, the misalignment signal begins to increase because the observer's prediction is no longer aligned with the distribution of incoming stimuli. Once the misalignment signal exceeds the threshold, incoming stimuli are transmitted and used to adapt the observer's prediction. Because it requires several stimulus samples to increase the misalignment signal above threshold, there is a delay between the switch in the environment and the burst of transmitted stimuli. This delay, which is proportional to the size of the threshold, slows the observer's detection of the change (Fig. 8C, left).

When the variance changes in time, the average surprise of incoming stimuli also changes in time. When the variance is low and the observer's prediction is accurate, the misalignment signal is small, and very few stimuli are transmitted (Fig. 8A-B, right). When the variance abruptly increases, the misalignment signal grows quickly, both because the observer's prediction is no longer accurate, and because the average surprise of the incoming stimulus distribution increases. A large proportion of stimuli are transmitted, and the observer quickly adapts to the change. If the threshold is sufficiently high, however, the observer's prediction never fully aligns with the true state. When the variance abruptly decreases, the incoming stimulus distribution is less surprising on average, and therefore a greater number of stimulus samples is needed before the misalignment signal exceeds threshold. As a result, observer is slower to detect decreases in variance than increases (Fig. 8C, right).

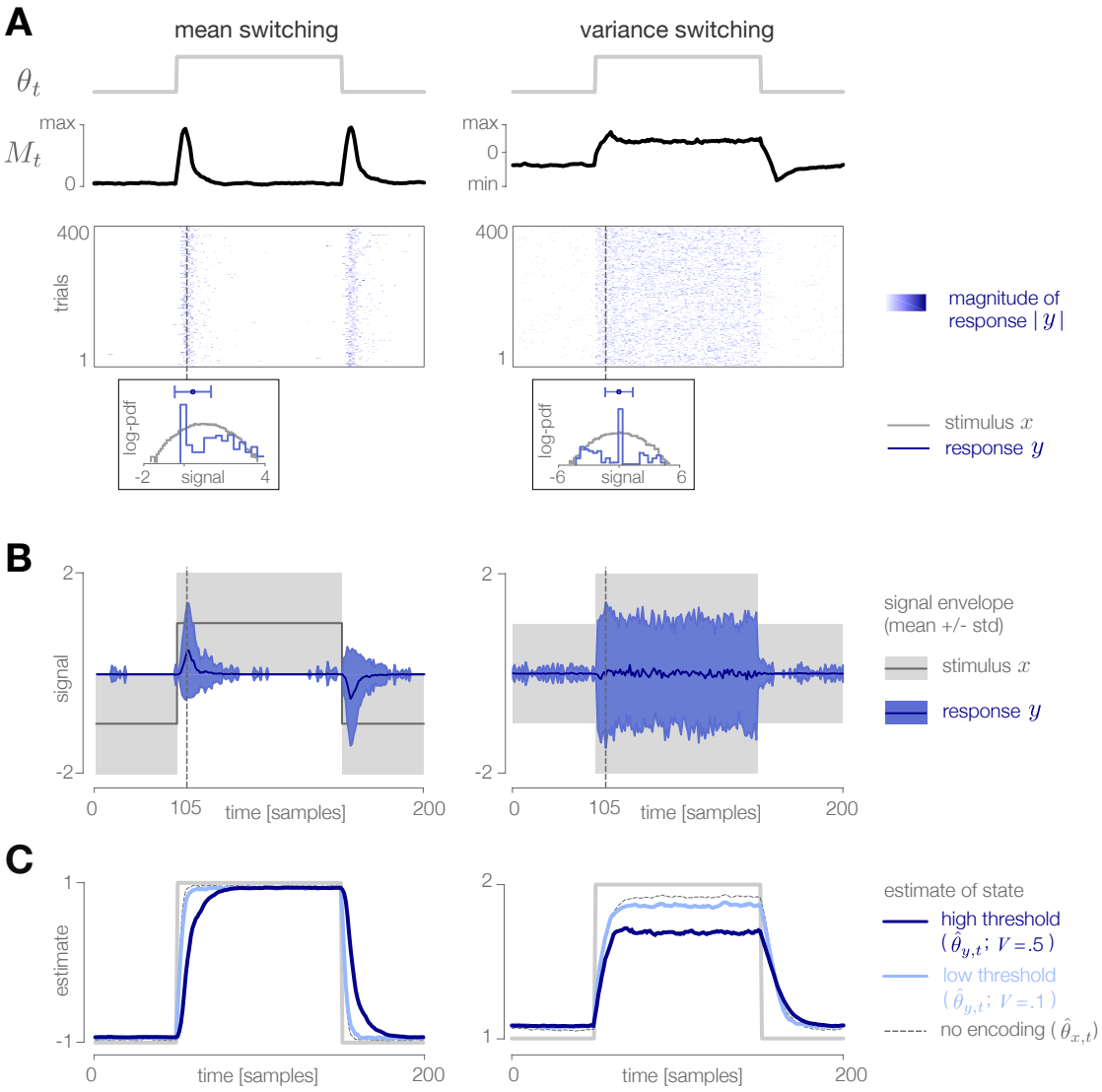


Figure 8: **Dynamic inference with stimulus selection.** **(A)** When the environment is changing, the observer's prediction is misaligned with the state of the environment. When this misalignment is large, stimuli are transmitted in full ($y_t = x_t$; blue). When this misalignment falls below a threshold (shown here for $V = 0.5$), stimuli are not transmitted at all ($y_t = 0$; white). The colorbar indicates the magnitude of the response y_t, scaled relative to the magnitude of the largest response in each panel. Insets: At times when the environment is changing (shown for $t = 105$), the distribution of responses (blue) is sparser the distribution of stimuli (gray), due to the large proportion of stimuli that are not encoded (indicated by the large peak in probability mass at 0). **(B)** The distribution of encoded stimuli changes over time, as can be seen by comparing the envelope of the stimulus distribution (gray) with the envelope of the neural responses (blue). Left: When the mean of the stimulus distribution changes abruptly, a large proportion of stimuli are encoded, and the mean of the neural response (blue line) approaches the mean of the stimulus distribution (black line). At times when the mean is stable, very few stimuli are encoded, and the mean of the neural response drops to zero. Right: When the variance is low, very few stimuli are encoded. When the variance increases, the average surprise of incoming stimuli increases, and a large proportion of stimuli are encoded. The envelope of the neural response expands and approaches the envelope of the stimulus distribution. **(C)** Higher thresholds slow the observer's detection of changes in the mean (left), and cause the observer to underestimate high variances (right). Threshold values are scaled relative to the maximum analytical value of the misalignment signal in the mean- and variance-switching environment (shown in the upper and lower panels of Fig. 7B, respectively). Results in panels B and C are averaged over 800 cycles of the probe environment.

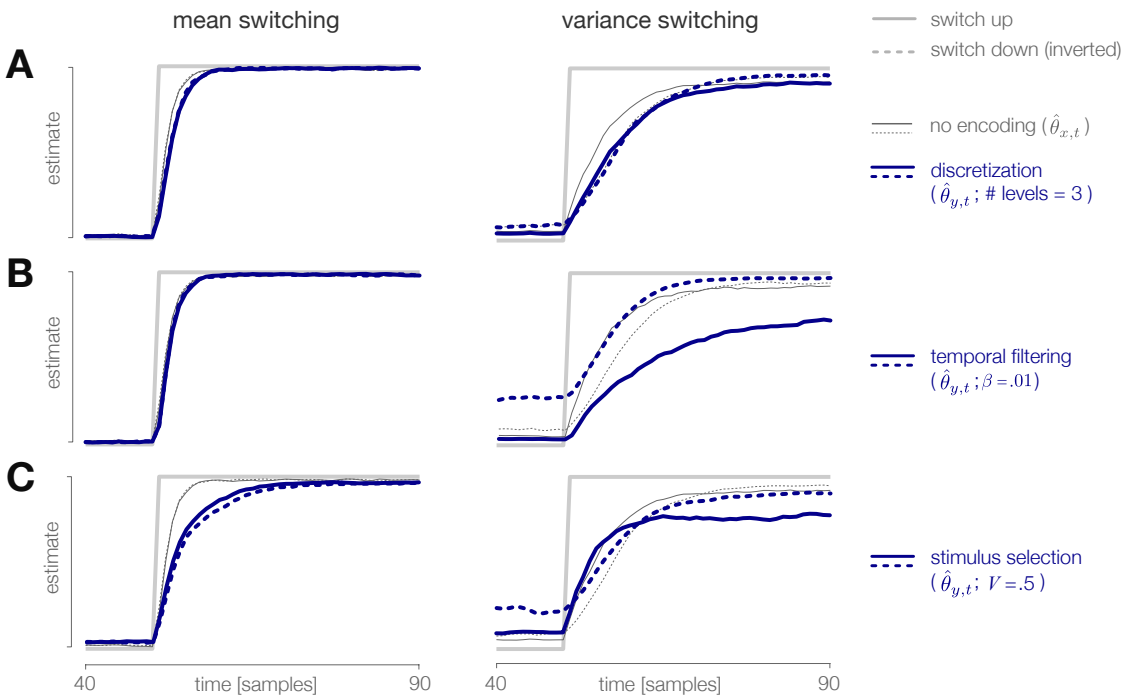


Figure 9: **Different encoding schemes induce different patterns error in inference.** Encoding schemes optimized for inference introduce inaccuracies in the estimates $\hat{\theta}_{y,t}$ relative to estimates $\hat{\theta}_{x,t}$ produce by the ideal Bayesian observer in the absence of an encoding. These inaccuracies are shown here for **(A)** discretization, **(B)** filtering, and **(C)** stimulus selection. For ease of comparison, downward switches are inverted vertically and overlaid on upward switches. In all cases, estimates of the mean are symmetric to upward and downward switches, while estimates of variance show asymmetries. **(A)** Encoding via discretization has minimal impact on the speed and accuracy of mean estimation (left), but slows the detection of upward switches in variance (right). **(B)** Encoding via filtering has minimal impact on the speed and accuracy of mean estimation (left). It significantly slows the detection of upward switches in variance, and causes the system to underestimate high variances (right). **(C)** Encoding via stimulus selection slows the detection of switches in the mean, but has minimal impact on the final accuracy of the inference (left). When variance in changing, this encoding causes the system to underestimate high variances (right).

Dynamical signatures of adaptive coding schemes

The encoding schemes considered here were each designed to reduce the fidelity with which stimuli are encoded in neural responses. This, in turn, reduced the accuracy of the observer's estimate with respect to the ideal Bayesian observer in the absence of an encoding. Interestingly, however, we found that the dynamical signatures of these inaccuracies could differentiate each encoding scheme (Fig. 9).

In all three encoding schemes, we found that the accuracy of inference was symmetric to upward and downward switches of the mean. Encoding via stimulus selection had the largest impact on inference accuracy; several stimulus samples were required before the observer began updating its estimate, and the estimate was slower to converge to the true mean of the stimulus distribution (Fig. 9C, left). Encoding via filtering and discretization had much weaker effects on accuracy (Fig. 9A-B, left).

In contrast, the accuracy of inference was asymmetric to upward and downward switches in variance. Such asymmetries were previously observed in the ideal Bayesian observer model, where the observer was shown to respond more slowly to decreases in variance than increases [11]. Here, we found that each encoding scheme impacted this asymmetry in a qualitatively different way. We found that encoding via discretization removed this asymmetry; the observer was equally fast to detect upward and downward switches (Fig. 9A, right). Encoding via temporal filtering reversed the asymmetry; the observer was much slower to detect *increases* in variance than decreases, and underestimated the value of the high-variance state (Fig. 9B, right). Finally, encoding via stimulus selection preserved the asymmetry, but caused the observer to underestimate the high-variance state (Fig. 9C, right).

Together, these observations suggest that an efficient system could prioritize some encoding schemes over others when the goal is to quickly infer specific changes in the environmental state (such as increases versus decreases in variance), rather than to accurately infer all possible changes.

DISCUSSION

Organisms rely on incoming sensory stimuli to build and update internal models of their environment. When the environment is changing over time, both the details of the incoming stimuli and their relative utility for behavior can also change. Here, we analyzed encoding schemes that could exploit these changes in order to reduce metabolic costs while maintaining accurate inferences about the environment. To balance coding cost and inference error, we showed that efficient systems should increase energy use at times when the environment is changing, and reduce energy use when the environment is stable.

We demonstrated this behavior using three neurally-plausible encoding schemes. Each scheme reflected a different limitation on neural response properties, and each implemented a dynamic mapping from stimuli onto neural responses. In two cases, the encoder was adapted over time to minimize inference error based on an internal prediction of the stimulus distribution. In the third case, the encoder implemented dynamic stimulus selection in a manner that supported inference without explicit optimization. At times when the stimulus distribution changed abruptly, we found that all three schemes exhibited a transient increase in the fidelity with which stimuli were encoded in neural responses. This enabled the system to quickly adapt its estimate of the environment based on the new distribution of incoming stimuli. Once the estimate was aligned with the new state of the environment, the fidelity of the encoding decreased in order to reduce metabolic costs. Depending on the specific encoding scheme, changes in fidelity were marked by changes in the diversity of neural response patterns (Fig. 3D), the temporal correlations in neural activity (Fig. 6E), or the total amount of neural activity (Fig. 8A).

Stimulus surprise and observer uncertainty determine optimal coding strategies during sensory inference

The strategies employed by each encoder reflected both the uncertainty of the observer and the surprise of incoming stimuli. When response levels were limiting, the optimal encoder devoted higher fidelity to stimuli that were predicted to occur if the environment switched states; in most cases, these stimuli were also those that were predicted to be the most surprising. When gain and temporal resolution were limiting, the optimal encoder devoted higher fidelity at times when the observer was uncertain and incoming stimuli were predicted to be surprising. Neither of these two encoders was explicitly designed to encode based on surprise and uncertainty; rather, this result fell out of an optimization that balanced coding cost and inference error. We further demonstrated that an encoder explicitly designed to transmit stimuli based on surprise could perform accurate inference with very limited numbers of sensory stimuli. Together, these results show that stimulus surprise and observer uncertainty capture the changing impact that stimuli can have on the estimation process. By devoting fidelity to those stimuli with the greatest impact, a system can significantly reduce coding costs while maintaining accurate inference.

This behavior changed dramatically when encoding schemes were optimized for stimulus reconstruction, rather than inference. When response levels were limiting, the optimal encoder devoted higher fidelity to stimuli that were predicted to be probable. Abrupt changes in the environment were accompanied by a decrease in the entropy rate of the neural response, an effect that has been observed experimentally [32]. When gain and temporal resolution were limiting, higher fidelity was devoted to stimuli that were predicted to be larger in magnitude. In both cases, coding costs were higher on average than when the same scheme was optimized for inference.

Bursts signal salient changes in the environment

To maintain low metabolic costs, we found that optimal encoders adapt their encoding strategies in response to the changing utility of incoming stimuli. Such changes were signaled by elevated periods of response variability, temporal decorrelation, or total activity. Burst-like changes in each of these properties served to increase the fidelity of the neural response, and enabled the system to quickly respond to informative changes in the stimulus distribution. In the nervous system, bursts of high-frequency activity are thought to be important for reliably conveying salient changes in an organism's surroundings [55, 56, 57]. For example, in the lateral geniculate nucleus, bursts of spikes are predominantly triggered by structured features of natural stimuli [57], and in the lateral line lobe of the weakly electric fish,

neurons burst in response to electric field distortions similar to those elicited by prey [56]. A close relationship between bursts of spikes and the amplitude of behavioral responses was also observed in the cricket auditory system [55, 58]. Here, we propose that a broad range of bursting phenonema could be indicative of an active data selection process controlled by a top-down prediction of an incoming stimulus distribution, and could thus serve as an efficient strategy for encoding changes in the environment.

Metamers emerge from resource-constrained encoding of stationary stimulus distributions

A central idea of this work is that stimuli that are not useful for a statistical estimation task need not be encoded. This was most notably observed during periods in which an observer maintained an accurate prediction of a stationary stimulus distribution. Here, different stimuli could be encoded by the same neural response without impacting the accuracy of the observer's prediction. This process ultimately renders stimuli indistinguishable, because information that would differentiate stimuli is lost during encoding. Stimuli that are physically different but perceptually indistinguishable, referred to as metamers, have been shown in some cases to arise as a consequence of spatial and temporal averaging of stimulus statistics [42, 43]. In order to perform accurate inference, we showed that strong averaging is an optimal strategy for encoding stimuli from a known stationary distribution, which in turn predicts that different stimuli should become less discriminable over time. Such a phenomenon has been recently observed in human perception of auditory textures that are generated from stationary sound sources such as flowing water, humming wind, or large groups of animals [43]. Human listeners are readily capable of distinguishing short excerpts of sounds generated by such sources. Surprisingly, however, when asked to tell apart long excerpts of auditory textures, performance sharply decreases. We propose that this behavior reflects adaptive encoding for accurate perceptual inference.

Efficient use of metabolic resources yields diverse signatures of suboptimal inference

We used an ideal Bayesian observer to illustrate the dynamic relationship between encoding and inference. Ideal observer models have been widely used to establish fundamental limits of performance on different sensory tasks [29, 59, 3]. The Bayesian framework in particular has been used to identify signatures of optimal performance on statistical estimation tasks [60, 36], and a growing body of work suggests that neural systems explicitly perform Bayesian computations [14, 7, 61, 8]. In line with recent studies [62, 63], we examined the impact of limited metabolic resources on such probabilistic neural computations.

While numerous studies have identified signatures of near-optimal performance in both neural coding (e.g. [10]) and perception (e.g. [30, 3]), the ideal observer framework can also be used to identify deviations from optimality. Such deviations have been ascribed to noise [59] and suboptimal neural decoding [64]. Here, we propose that statistical inference can deviate from optimality as a consequence of efficient, resource-constrained stimulus coding. We observed deviations from optimality in both the speed and accuracy of inference, and we found that these deviations could be used to differentiate the underlying scheme that was used to encode incoming stimuli. It might therefore be possible to infer underlying constraints on neural coding by analyzing patterns of suboptimal inference.

Connections to existing work

Efficient coding of task-relevant information has been studied before, primarily within the framework of the Information Bottleneck (IB) method [27]. The IB framework provides a general theoretical approach for extracting task-relevant information from sensory stimuli, and it has been successfully applied to the study of neural coding in the retina [28] and in the auditory cortex [65]. However, the optimal IB solutions tend to be intractable for complex stimulus distributions [66, 67, 68], and as such are unlikely to be instantiated by biological systems [68]. A recent study proposed a Bayesian Efficient Coding (BEC) framework, which postulates that a metabolically-constrained sensory system could minimize an arbitrary error function that could, as in IB, be chosen to reflect task-relevant information [69]. However, neither framework (IB nor BEC) explicitly addresses the issue of sensory coding in non-stationary environments, where the relevance of different stimuli can change in time. Here, we addressed the issues of both tractability and non-stationarity in the context of perceptual inference.

Our approach also bears conceptual similarities to the predictive coding framework proposed by Rao and Ballard [48], in which low-level sensory neurons support accurate stimulus reconstruction by

encoding the residual error between an incoming stimulus and a top-down prediction of the stimulus. Our encoding schemes similarly use top-down predictions to encode useful deviations in the stimulus distribution. Importantly, however, the goal here was not to reconstruct the stimulus itself, but rather to infer the underlying properties of a changing stimulus distribution. To this end, we considered encoding schemes that could use top-down predictions to adaptively adjust their strategies over time based on the predicted utility of different stimuli for supporting inference.

This work synthesizes different theoretical frameworks in an effort to clarify their mutual relationship. In this broad sense, our approach aligns with recent studies that aim to unify frameworks such as efficient coding and Bayesian inference [69], as well as concepts such as efficient, sparse, and predictive coding [70].

Outlook

Efficient coding and probabilistic inference are two prominent frameworks in theoretical neuroscience that address the separate questions of how stimuli can be encoded at minimal cost, and how stimuli can be used to support accurate inferences. In this work, we bridged these two frameworks within a dynamic setting. We examined optimal strategies for encoding sensory stimuli while minimizing the error that such encoding induces in the inference process, and we contrasted these with strategies designed to optimally reconstruct the stimulus itself. These two goals could correspond to different regimes of the same sensory system [17], and future work will explore strategies for balancing these regimes depending on task requirements. In order to test the implications of this work for physiology and behavior, it will be important to generalize this framework to more naturalistic stimuli, noisy encodings, and richer inference tasks. At present, our results identify broad signatures of a dynamical balance between metabolic costs and task demands that could provide a unifying explanation of phenomena in both neural and perceptual systems.

Acknowledgements

We thank John Briguglio, Vivek Jayaraman, Yarden Katz, Emily Mackevicius, and Josh McDermott for useful discussions and feedback on the manuscript. WM was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. AMH was supported by the Howard Hughes Medical Institute.

Competing Interests

The authors declare no competing interests.

METHODS

A Optimal Bayesian inference with adaptively-encoded stimuli

We describe a class of discrete-time environmental stimuli x_t whose statistics are completely characterized by a single time-varying environmental state variable θ_t.

We then consider the scenario in which these stimuli are encoded in neural responses, and it is these neural responses that must be used to construct the posterior probability over environmental states. In what follows, we derive the optimal Bayesian observer for computing this posterior given the history of neural responses.

In a full Bayesian setting, the observer should construct an estimate of the stimulus distribution, $p(x_t)$, by marginalizing over its uncertainty in the estimate of the environmental state θ_t (i.e., by computing $p(x_t) = \int d\theta_t p(x_t|\theta_t)p(\theta_t)$). For simplicity, we avoid this marginalization by assuming that the observer's belief is well-summarized by the average of the posterior, which is captured by the point value $\hat{\theta}_t = \int d\theta_t\, \theta_t p(\theta_t)$ for estimation, and $\vec{\theta}_{t+1} = \int d\theta_{t+1}\, \theta_{t+1} p(\theta_{t+1})$ for prediction. The average of the posterior is an optimal scalar estimate that minimizes the mean-squared-error between the estimated and true states of the environment, and is known to provide a good description of both neural [11] and perceptual [12] dynamics. The observer then uses these point values to condition its prediction of the stimulus distribution, $p(x_t|\vec{\theta}_t)$. Conditioning on a point estimate guarantees that the observer's prediction of the environment belongs to the same family of distributions as the true environment. This is not guaranteed to be the case when marginalizing over uncertainty in θ_t. For example, if the posterior assigns non-zero probability mass to two different mean values of a unimodal stimulus distribution, the predicted stimulus distribution could be bimodal, even if the true stimulus distribution is always unimodal. We verified numerically that the key results of this work are not affected by approximating the full marginalization with point estimates.

When the timescale of the environment dynamics is sufficiently slow, the point prediction $\vec{\theta}_{t+1}$ can be approximated by the point estimate $\hat{\theta}_t$. In the two-state environments considered here, the probability that the environment remains in the low state from time t to time $t+1$ is equal to $P^L_{t+1} = P^L_t(1-h) + (1-P^L_t)h$, where h is the hazard rate [11]. For the small hazard rate used here ($h = 0.01$), $P^L_{t+1} = 0.99P^L_t + 0.01(1-P^L_t)$, and the estimate $\hat{\theta}_t$ is therefore a very close approximation of the prediction $\vec{\theta}_{t+1}$. All results presented in the main text were computed using this approximation (i.e., $\vec{\theta}_{t+1} \approx \hat{\theta}_t$). With this approximation, the optimal Bayesian observer computes the approximate posterior distribution $p(\theta_t|y_{\tau\leq t}, \hat{\theta}_{\tau<t})$, conditioned on the history of neural responses $y_{\tau\leq t}$ and the history of point estimates $\hat{\theta}_{\tau<t}$. In the remainder of the Methods, we will formulate all derivations and computations in terms of the history of past estimates (up to and including time $t-1$), with the understanding that these estimates can be used as approximate predictions of the current state at time t.

With these simplifications, the general steps of the inference process can be broken down as follows:

1. **Encoder:** maps incoming stimuli $x_{\tau\leq t}$ onto a neural response y_t by sampling from the "encoding distribution" $p(y_t|x_{\tau\leq t}, \hat{\theta}_{\tau<t})$
2. **Decoder:** uses Bayes' rule to estimate the likelihood of a stimulus x_t given the neural response y_t, as given by the "decoding distribution" $p(x_t|y_t, \hat{\theta}_{\tau<t})$
3. **Observer:** uses the neural response y_t to update the posterior $p(\theta_t|y_{\tau\leq t}, \hat{\theta}_{\tau<t})$. This can be broken down into the following steps, in which the observer:
   (a) Combines the previous posterior $p(\theta_{t-1}|y_{\tau<t}, \hat{\theta}_{\tau<t-1})$ with knowledge of environment dynamics $p(\theta_t|\theta_{t-1})$ to compute the probability of θ_t given all past data, $p(\theta_t|y_{\tau<t}, \hat{\theta}_{\tau<t-1})$
   (b) Uses Bayes' rule to incorporate a new stimulus x_t and form $p(\theta_t|x_t, y_{\tau<t}, \hat{\theta}_{\tau<t-1})$
   (c) Marginalizes over the uncertainty in x_t using the decoding distribution $p(x_t|y_t, \hat{\theta}_{\tau<t})$, thereby obtaining the updated posterior $p(\theta_t|y_{\tau\leq t}, \hat{\theta}_{\tau<t})$ (which can be averaged to compute the point

estimate $\hat{\theta}_t$)

(d) Combines the updated posterior with knowledge of environment dynamics $p(\theta_{t+1}|\theta_t)$ to generate a predicted posterior $p(\theta_{t+1}|y_{\tau\leq t}, \hat{\theta}_{\tau<t})$ (which can be averaged to compute the point prediction $\vec{\theta}_{t+1}$)

4. **Feedback loop:** sends the prediction back upstream to update the encoder.

In what remains of Section A, we derive the general equations for the full inference process in the presence of both encoding and decoding. In Section B, we derive the specific forms of the inference equations in a simplified, two-state environment. We first focus on the general equations of the observer model (Section B.2). We then describe the forms of the encoding and decoding distributions implemented by the three different encoding schemes considered in this paper, and detail how the parameters of each encoder can be optimized based on the observer's prediction of the environmental state (Sections B.3-B.6). In Section C, we describe the numerical approximations used to simulate the results presented in the main paper.

A.1 Environment dynamics

We consider a non-stationary environment with Markovian dynamics. The dynamics of the environmental state variable θ_t are then specified by the distribution $p(\theta_t|\theta_{t-1})$. At each time t, the value of θ_t specifies the distribution of stimuli $p(x_t|\theta_t)$.

A.2 Encoder

We consider an encoder that maps incoming stimuli $x_{\tau\leq t}$ onto a neural response y_t. We assume that the encoder has access to the history of estimates $\hat{\theta}_{\tau<t}$ (fed back from a downstream observer) to optimally encode incoming stimuli via the "encoding distribution", $p(y_t|x_{\tau\leq t}, \hat{\theta}_{\tau<t})$.

A.3 Decoder

Because the observer does not have direct access to the stimulus, it must first decode the stimulus from the neural response. We assume that the decoder has access to the instantaneous neural response y_t and this history of past estimates $\hat{\theta}_{\tau<t}$. The decoder must use these signals to marginalize over past stimuli $x_{\tau<t}$ and compute the likelihood of the response y_t conditioned on the current stimulus x_t (this likelihood will later be used to update the observer's posterior):

$$p(y_t|x_t, \hat{\theta}_{\tau<t}) = \int dx_{\tau<t} \underbrace{p(y_t|x_t, x_{\tau<t}, \hat{\theta}_{\tau<t})}_{\substack{\text{encoding}\\ \text{distribution}}} p(x_{\tau<t}|\hat{\theta}_{\tau<t}) \tag{3}$$

The decoder must then invert this distribution (using Bayes' rule) to estimate the probability of the stimulus x_t given the response y_t and past estimates $\hat{\theta}_{\tau<t}$:

$$\begin{aligned} p(x_t|y_t, \hat{\theta}_{\tau<t}) &= \frac{p(y_t|x_t, \hat{\theta}_{\tau<t})p(x_t|\hat{\theta}_{\tau<t})}{p(y_t|\hat{\theta}_{\tau<t})} \\ &= \frac{p(y_t|x_t, \hat{\theta}_{\tau<t})p(x_t|\hat{\theta}_{t-1})}{Z(y_t, \hat{\theta}_{\tau<t})} \end{aligned} \tag{4}$$

where we have written the distribution in the denominator as a normalization constant obtained by integrating the numerator:

$$Z(y_t, \hat{\theta}_{\tau<t}) = \int dx_t \, p(y_t|x_t, \hat{\theta}_{\tau<t}) p(x_t|\hat{\theta}_{t-1}) \tag{5}$$

In what follows, we refer to $p(x_t|y_t, \hat{\theta}_{\tau<t})$ (defined in Eq. 4) as the "decoding distribution".

A.4 Observer

The optimal observer should use the decoding distribution to marginalize over its uncertainty about the true value of the stimulus x_t and thereby obtain the posterior probability of θ_t given past stimuli $y_{\tau\leq t}$ and past estimates $\hat{\theta}_{\tau<t}$. To do this, we first write an expression for the probability of θ_t given all data up to (but not including) the current timestep:

$$p(\theta_t|y_{\tau<t}, \hat{\theta}_{\tau<t-1}) = \int d\theta_{t-1} \, p(\theta_t|\theta_{t-1}) \underbrace{p(\theta_{t-1}|y_{\tau<t}, \hat{\theta}_{\tau<t-1})}_{\substack{\text{posterior from} \\ \text{previous timestep}}} \tag{6}$$

where the prior is taken to be the posterior from the last timestep, and the distribution $p(\theta_t|\theta_{t-1})$ governs the dynamics of the environment.

This distribution can then be combined with a new stimulus x_t:

$$\begin{aligned} p(\theta_t|x_t, y_{\tau<t}, \hat{\theta}_{\tau<t-1}) &= \frac{p(x_t|\theta_t, y_{\tau<t}, \hat{\theta}_{\tau<t-1}) p(\theta_t|y_{\tau<t}, \hat{\theta}_{\tau<t-1})}{p(x_t|y_{\tau<t}, \hat{\theta}_{\tau<t-1})} \\ &= \frac{p(x_t|\theta_t) p(\theta_t|y_{\tau<t}, \hat{\theta}_{\tau<t-1})}{\Omega(x_t, y_{\tau<t}, \hat{\theta}_{\tau<t-1})}. \end{aligned} \tag{7}$$

As before, we have written the distribution in the denominator as a normalization constant obtained by integrating the numerator:

$$\Omega(x_t, y_{\tau<t}, \hat{\theta}_{\tau<t-1}) = \int d\theta_t p(x_t|\theta_t) p(\theta_t|y_{\tau<t}, \hat{\theta}_{\tau<t-1}) \tag{8}$$

Finally, we marginalize over the unknown value of the signal x_t using the decoding distribution $p(x_t|y_t, \hat{\theta}_{\tau<t})$ to obtain the updated posterior distribution:

$$p(\theta_t|y_{\tau\leq t}, \hat{\theta}_{\tau<t}) = \int dx_t \, p(\theta_t|x_t, y_{\tau<t}, \hat{\theta}_{\tau<t-1}) p(x_t|y_t, \hat{\theta}_{\tau<t}) \tag{9}$$

To form a prediction about the future state of the environment, the observer should combine its belief $p(\theta_t|y_{\tau\leq t}, \hat{\theta}_{\tau<t})$ about the current state of the environment with the knowledge $p(\theta_{t+1}|\theta_t)$ of the environment dynamics in a manner analogous to Eq. 6.

A.5 Computing point estimates

The posterior can be used to compute a point estimate $\hat{\theta}_t$ and prediction $\vec{\theta}_{t+1}$ of the environmental state:

$$\hat{\theta}_t = \int d\theta_t \theta_t p(\theta_t|y_{\tau\leq t}, \hat{\theta}_{\tau<t}) \equiv \hat{\theta}_{y,t} \tag{10}$$

$$
\begin{aligned}
\vec{\theta}_{t+1} &= \int d\theta_{t+1}\, \theta_{t+1} p(\theta_{t+1}|y_{\tau\leq t}, \hat{\theta}_{\tau<t}) \\
&= \int d\theta_{t+1} \theta_{t+1} \int d\theta_t\, p(\theta_{t+1}|\theta_t) p(\theta_t|y_{\tau\leq t}, \hat{\theta}_{\tau<t})
\end{aligned} \tag{11}
$$

The point estimate given in Eq. 10 is referred to in the main text as "$\hat{\theta}_{y,t}$". We distinguish this from the point estimate "$\hat{\theta}_{x,t}$", which was derived in [11] in the absence of encoding/decoding.

B Model environments

B.1 Environment dynamics

We consider a two-state environment in which the state θ_t can take one of two values, θ^L and θ^H. At each timestep, the environment can switch states with a constant probability h, referred to as the "hazard rate". The hazard rate fully specifies the dynamics of the environment:

$$
\theta_t = z_t \theta_{t-1} + (1 - z_t)(\theta^L + \theta^H - \theta_{t-1}) \tag{12}
$$

where z_t is a binary random variable equal to 1 with probability h and 0 with probability $1 - h$.

We take θ_t to parametrize either the mean μ or the standard deviation σ of a Gaussian stimulus distribution:

$$
p(x_t|\theta_t) = \begin{cases} \mathcal{N}(x_t; \theta_t, \sigma^2), & \text{mean-switching environment } (\theta_t = \mu) \\ \mathcal{N}(x_t; \mu, \theta_t^2), & \text{variance-switching environment } (\theta_t = \sigma) \end{cases} \tag{13}
$$

B.2 Observer

In a two-state environment, the posterior distribution $p(\theta_t|y_{\tau\leq t}, \hat{\theta}_{\tau<t})$ can be summarized by a single value $P_t^L = p(\theta_t = \theta^L|y_{\tau\leq t}, \hat{\theta}_{\tau<t})$, which is the probability that the environment is in the low state at time t.

Given the posterior P_{t-1}^L at the previous timestep, the distribution for θ_t given all past data is given by:

$$
p(\theta_t = \theta^L|y_{\tau<t}, \hat{\theta}_{\tau<t}) = (1 - h)P_{t-1}^L + h(1 - P_{t-1}^L) \tag{14}
$$

where h is the a priori probability that a switch occurred at the current timestep. This distribution can then be combined with a new stimulus x_t:

$$
\begin{aligned}
p(\theta_t = \theta^L|x_t, y_{\tau<t}, \hat{\theta}_{\tau<t}) &= \frac{p(x_t|\theta_t = \theta^L) p(\theta_t = \theta^L|y_{\tau<t}, \hat{\theta}_{\tau<t})}{\Omega(x_t, y_{\tau<t}, \hat{\theta}_{\tau<t})} \\
&= \frac{\mathcal{N}(x_t; \mu_L, \sigma_L^2) \left[(1 - h)P_{t-1}^L + h(1 - P_{t-1}^L)\right]}{\Omega(x_t, y_{\tau<t}, \hat{\theta}_{\tau<t})}
\end{aligned} \tag{15}
$$

The variables (μ_L, σ_L) and (μ_H, σ_H) correspond to mean and standard deviation of the stimulus distribution in the low and high states, respectively, and their values vary depending on the type of the environment (mean-switching versus variance-switching).

To obtain the updated posterior P_t^L, we marginalize over the decoding distribution $p(x_t|y_t, \hat{\theta}_{\tau<t})$:

$$P_t^L = p(\theta_t = \theta^L | y_{\tau \leq t}, \hat{\theta}_{\tau<t}) = \int dx_t \, p(\theta_t = \theta^L | x_t, y_{\tau<t}, \hat{\theta}_{\tau<t}) p(x_t | y_t, \hat{\theta}_{\tau<t})$$
$$= \int dx_t \, \frac{\mathcal{N}(x_t; \mu_L, \sigma_L^2) \left[(1-h) P_{t-1}^L + h(1 - P_{t-1}^L) \right]}{\Omega(x_t, y_{\tau<t}, \hat{\theta}_{\tau<t})} p(x_t | y_t, \hat{\theta}_{\tau<t}) \quad (16)$$

The posterior can be used to construct a new point-estimate $\hat{\theta}_t$ of the environmental state:

$$\hat{\theta}_t = P_t^L \theta^L + (1 - P_t^L) \theta^H, \quad (17)$$

where $1 - P_t^L = P_t^H$ is the probability that the environment is in the high state at time t. Note that although the environmental states are discrete, the optimal Bayesian observer maintains a continuous estimate $\hat{\theta}_t$.

To form a prediction about the future state of the environment, the observer first combines the posterior P_t^L with knowledge of environment dynamics (in a manner analogous to Eq. 14), and then computes a point prediction (in a manner analogous to Eq. 17):

$$P_{t+1}^L = P_t^L (1-h) + (1 - P_t^L) h \quad (18)$$
$$\vec{\theta}_{t+1} = P_{t+1}^L \theta^L + (1 - P_{t+1}^L) \theta^H \quad (19)$$

For small hazard rates (as considered here), the predicted value $\vec{\theta}_{t+1}$ is very close to the current estimate $\hat{\theta}_t$. For simplicity, we approximate the prediction $\vec{\theta}_{t+1}$ by the estimate $\hat{\theta}_t$. This estimate is then fed back upstream and used to update the encoder. In the general case, however, one should compute the full predicted posterior via Eq. 18, and use this full prediction to optimize the encoder.

B.3 Encoder/Decoder

The posterior (given in Eq. 16) is a function of the decoding distribution $p(x_t | y_t, \hat{\theta}_{\tau<t})$, which depends on the encoding distribution $p(y_t | x_{\tau \leq t}, \hat{\theta}_{\tau<t})$ through Eqs. 3-4. In what follows, we derive the encoding and decoding distributions for the three encoding schemes considered in this paper. All three encoding schemes are noiseless; as a result, the encoding distribution $p(y_t | x_{\tau \leq t}, \hat{\theta}_{\tau<t})$ reduces to a delta function in each case. This encoding distribution can then be used to derive the decoding distribution, from which it is straightforward to compute the posterior P_t^L via Eq. 16 (and similarly any estimates and predictions derived from the posterior).

Each encoding scheme considered here was parametrized by one or more encoding parameters. In two of the three encoding schemes, these parameters were chosen to minimize an error function $E(x, y)$, subject to a constraint on the fidelity of the encoding. We defined this error function with respect to inference or stimulus reconstruction:

$$E(x, y) = \begin{cases} \left(\hat{\theta}_x - \hat{\theta}_y \right)^2, & \text{error in inference} \\ (x - y)^2, & \text{error in reconstruction} \end{cases} \quad (20)$$

where $\hat{\theta}_y$ was defined in Eq. 10, and $\hat{\theta}_x$ was derived in [11].

B.4 Limited neural response levels: encoding via discretization

B.4.1 Encoder

Here, we consider a quantization (instantaneous discretization) of the stimulus space that maps the current stimulus x_t onto one of a discrete set of values $\{y_t^i\}$, where $i = 1, 2, ...N$ labels distinct response levels. This mapping is performed deterministically by choosing the response level that minimizes the instantaneous error $E(x_t, \{y_t^i\})$:

$$\begin{aligned} y_t &= \underset{y_t^i}{\text{argmin}}\ E(x_t, \{y_t^i\}) \\ &= y_t^i \end{aligned} \tag{21}$$

We can therefore write the encoding distribution as a delta function:

$$p(y_t | x_{\tau \leq t}, \hat{\theta}_{\tau < t}) = \delta(x_t - y_t^i), \tag{22}$$

where the set of response levels $\{y_t^i\}$ implicitly contains the dependence on $\hat{\theta}_{t-1}$.

B.4.2 Decoder

The decoder must estimate the probability of a stimulus x_t, given that the observed response was y_t^i. In principle, the response y_t^i could have been generated by any stimulus in the range $[y_t^{i,L}, y_t^{i,H}]$, where $y_t^{i,L}$ and $y_t^{i,H}$ are the lower and upper bounds of the bin represented by level y_t^i, respectively.

The decoding distribution can then be written as a truncated Gaussian distribution:

$$p(x_t | y_t, \hat{\theta}_{\tau < t}) = \begin{cases} \frac{1}{Z(y_t^{i,L}, y_t^{i,H}, \hat{\theta}_{t-1})} \mathcal{N}(x_t; \hat{\mu}_{t-1}, \hat{\sigma}_{t-1}^2), & y_t^{i,L} < x_t < y_t^{i,H} \\ 0, & \text{otherwise} \end{cases} \tag{23}$$

where $Z(y_t^{i,L}, y_t^{i,H}, \hat{\theta}_{t-1})$ is a normalization constant. For simplicity, we approximated this truncated Gaussian distribution with a delta function:

$$p(x_t | y_t, \hat{\theta}_{\tau < t}) \approx \delta(x_t - y_t^i) \tag{24}$$

We verified numerically that this approximation did not impact our results.

B.4.3 Determining the optimal response levels

At each point in time, the optimal set of response levels $\{y_t^i\}^*$ was found by minimizing the following equation:

$$\{y_t^i\}^* = \underset{\{y_t^i\}}{\text{argmin}} \left\langle E(x_t, \{y_t^i\}) \right\rangle_{p(x_t | \hat{\theta}_{t-1})} \tag{25}$$

subject to a hard constraint on the number of response levels. When optimizing for mean-switching environments, we defined the error function with respect to the raw stimulus and neural response (i.e.,

$E = E(x, y)$). When optimizing for variance-switching environment, we defined the error function with respect to the absolute value of the stimulus and neural response (i.e., $E = E(|x|, |y|)$). We computed $\langle E(x_t, \{y_t^i\})\rangle$ numerically; see Section C.3.1.

B.5 Limited gain and temporal acuity: encoding via temporal filtering

B.5.1 Encoder

In this encoding scheme, we consider a simple temporal filter parameterized by the coefficient α_t. This filter linearly combines current (x_t) and past (x_{t-1}) stimuli:

$$y_t = \alpha_t x_t + (1 - \alpha_t) x_{t-1} \tag{26}$$

The encoding distribution is then given by:

$$p(y_t | x_{\tau \leq t}, \hat{\theta}_{\tau < t}) = \delta\left(y_t - (\alpha_t x_t + (1 - \alpha_t) x_{t-1})\right), \tag{27}$$

where the filtering coefficient α_t implicitly contains the dependence on $\hat{\theta}_{t-1}$.

B.5.2 Decoder

The encoding is a function of both current and past stimuli. The decoder, however, only has access to the current response y_t. In order to estimate the probability that this response was generated by the stimulus x_t, the decoder must first use the internal estimates $\hat{\theta}_{\tau<t}$ to marginalize over uncertainty in past stimuli $x_{\tau<t}$. This was first outlined in Eq. 3, which reduces here to:

$$\begin{aligned}
p(y_t | x_t, \hat{\theta}_{\tau<t}) &= \int dx_{\tau<t}\, p(y_t | x_t, x_{\tau<t}, \hat{\theta}_{\tau<t}) p(x_{\tau<t} | \hat{\theta}_{\tau<t}) \\
&= \int dx_{t-1}\, p(y_t | x_t, x_{t-1}, \vec{\theta}_t) p(x_{t-1} | \hat{\theta}_{t-1}) \\
&= \int dx_{t-1}\, \delta\left(y_t - (\alpha_t x_t + (1 - \alpha_t) x_{t-1})\right) \mathcal{N}(x_{t-1}; \hat{\mu}_{t-1}, \hat{\sigma}^2_{t-1}) \\
&= \frac{1}{(1 - \alpha_t)} \mathcal{N}\left(\frac{y_t - \alpha_t x_t}{(1 - \alpha_t)}; \hat{\mu}_{t-1}, \hat{\sigma}^2_{t-1}\right) \\
&= \mathcal{N}\left(y_t; \alpha_t x_t + (1 - \alpha_t)\hat{\mu}_{t-1}, (1 - \alpha_t)^2 \hat{\sigma}^2_{t-1}\right)
\end{aligned} \tag{28}$$

The decoder can then use Bayes' rule to invert this distribution and determine the probability of the stimulus x_t given the response y_t:

$$\begin{aligned}
p(x_t | y_t, \hat{\theta}_{\tau<t}) &= \frac{p(y_t | x_t, \hat{\theta}_{\tau<t}) p(x_t | \hat{\theta}_{\tau<t})}{Z(y_t, \hat{\theta}_{\tau<t})} \\
&= \frac{1}{Z(y_t, \hat{\theta}_{\tau<t})} \mathcal{N}\left(y_t; \alpha_t x_t + (1 - \alpha_t)\hat{\mu}_{t-1}, (1 - \alpha_t)^2 \hat{\sigma}^2_{t-1}\right) \mathcal{N}(x_t; \hat{\mu}_{t-1}, \hat{\sigma}^2_{t-1})
\end{aligned} \tag{29}$$

In its current form, this decoding distribution is written as a Gaussian over the variable y_t. Ultimately, the observer must use this decoding distribution to marginalize over uncertainty in x_t. In Appendix I,

we walk through the algebra needed to rewrite this distribution as Gaussian over x_t. The final form of this distribution in given by:

$$p(x_t|y_t, \hat{\theta}_{\tau<t}) = \mathcal{N}\left(x_t; \frac{\alpha_t y_t - (1-\alpha_t)(2\alpha_t - 1)\hat{\mu}_{t-1}}{1 - 2\alpha_t + 2\alpha_t^2}, \left(\frac{(1-\alpha_t)^2}{1 - 2\alpha_t + 2\alpha_t^2}\right)\hat{\sigma}_{t-1}^2\right) \tag{30}$$

B.5.3 Determining the optimal filter coefficient

The optimal filtering coefficient α_t^* was found by minimizing the following equation:

$$\alpha_t^* = \underset{\alpha_t}{\operatorname{argmin}} \left\langle E(x_t, y_t)\right\rangle_{p(x_t|\hat{\theta}_{t-1})} + \beta H(y_t, y_{t+1}|\hat{\theta}_{\tau<t}) \tag{31}$$

The error term, $\langle E(x_t, y_t)\rangle$, was computed numerically; see Section C.3.2. The entropy term, $H(y_t, y_{t+1}|\hat{\theta}_{\tau<t})$, can be computed analytically (see Appendix II for details):

$$\begin{aligned} H(y_t, y_{t+1}|\hat{\theta}_{\tau<t}) &= H(y_{t+1}|y_t, \hat{\theta}_{\tau<t}) + H(y_t|\hat{\theta}_{\tau<t}) \\ &= \frac{1}{2}\log_e\left(4\pi^2 e^2\left(\alpha_t^2\hat{\sigma}_{t-1}^2 + \frac{(1-\alpha_t)^4}{1-2\alpha_t+2\alpha_t^2}\hat{\sigma}_{t-1}^2\right)\left(\alpha_t^2\hat{\sigma}_{t-1}^2 + (1-\alpha_t)^2\hat{\sigma}_{t-1}^2\right)\right) \end{aligned} \tag{32}$$

B.6 Limited neural activity: encoding via dynamic stimulus selection

B.6.1 Encoder

In this encoding scheme, the encoder uses the misalignment signal M_t to determine whether or not to encode and transmit the stimulus x_t. If the magnitude of the misalignment signal exceeds the threshold V, the stimulus is encoded and transmitted. Otherwise, the stimulus is not encoded, and a "null symbol" is transmitted to the observer. For the purposes of computing the encoding and decoding distributions, we use $y_t = 0$ to denote the null symbol (in the main text, we denoted the null symbol by $y_t = \emptyset$).

This encoding is a deterministic mapping of the stimulus x_t onto the response y_t, dependent upon the misalignment signal M_t. The encoding distribution can thus be written in a probabilistic form as a mixture of two delta functions:

$$p(y_t|x_t, \hat{\theta}_{\tau<t}) = \begin{cases} \delta(y_t - x_t), & |M_t| > V \\ \delta(y_t), & |M_t| \leq V \end{cases} \tag{33}$$

where M_t implicitly contains the dependence on $\hat{\theta}_{t-1}$.

B.6.2 Decoder

In this scheme, the form of the decoding distribution depends on whether or not the encoder transmits the stimulus x_t. If the stimulus was encoded and transmitted, there is no uncertainty in its value, and the decoding distribution is a delta function about y_t. If the stimulus was not encoded and the null symbol was instead transmitted, the decoder can only assume that the stimulus came from the estimated stimulus distribution $p(x_t|\hat{\theta}_{t-1})$.

The decoding distribution therefore takes the following form:

$$p(x_t|y_t, \hat{\theta}_{\tau<t}) = \begin{cases} \delta(x_t - y_t), & y_t \neq 0 \\ p(x_t|\hat{\theta}_{t-1}), & y_t = 0 \end{cases} \tag{34}$$

B.6.3 Determining the misalignment signal

In defining this encoding scheme, our aim was to construct a heuristic "misalignment" signal that would alert the encoder to a change in the stimulus distribution. One candidate is a signal that tracks the average surprise of incoming stimuli, given the internal estimate of the environmental state.

The surprise associated with a single stimulus x_t is equal to the negative log-likelihood of the stimulus given the estimate $\hat{\theta}_{t-1}$:

$$S(x_t) = -\log p(x_t|\hat{\theta}_{t-1}) \tag{35}$$

The average surprise of incoming stimuli, obtained by averaging over the true stimulus distribution $p(x_t|\theta_t)$, is equal to cross-entropy between the true and estimated stimulus distributions:

$$H(x_t; \theta_t, \hat{\theta}_{t-1}) = \int dx_t \, S(x_t) p(x_t|\theta_t) \tag{36}$$

$$= H(x_t; \theta_t) + D_{KL}[p(x_t|\theta_t) \| p(x_t|\hat{\theta}_{t-1})], \tag{37}$$

where the second term in Eq. 37 is the Kullback-Leibler divergence of the estimated stimulus distribution from the true stimulus distribution.

The cross-entropy is equal to the entropy of the true stimulus distribution when the observer's estimate is accurate (i.e., when $\hat{\theta}_{t-1} = \theta_t$), and increases with the inaccuracy of the observer's estimate. To construct a signal that deviates from zero (rather than from the entropy of the stimulus distribution) whenever observer's estimate is inaccurate, we subtract the estimated entropy $H(x_t; \hat{\theta}_{t-1})$ from the cross-entropy to define the "misalignment signal":

$$M_t = H(x_t; \theta_t, \hat{\theta}_{t-1}) - H(x_t; \hat{\theta}_{t-1}) \tag{38}$$

$$= H(x_t; \theta_t) + D_{KL}[p(x_t|\theta_t) \| p(x_t|\hat{\theta}_{t-1})] - H(x_t; \hat{\theta}_{t-1}) \tag{39}$$

The magnitude of this signal is large whenever the average surprise of incoming stimuli differs from the estimated surprise, and monotonically increases as a function of the difference between the true and estimated states of the environment. In the case of a Gaussian distribution, the misalignment signal reduces to:

$$M_t = \frac{1}{2}\log(2\pi e \sigma_t^2) + \left(\log \frac{\hat{\sigma}_{t-1}}{\sigma_t} + \frac{\sigma_t^2 + (\mu_t - \hat{\mu}_{t-1})^2}{2\hat{\sigma}_t^2} - \frac{1}{2}\right) - \frac{1}{2}\log(2\pi e \hat{\sigma}_{t-1}^2) \tag{40}$$

where μ_t and σ_t are the mean and standard deviation of the true stimulus distribution, respectively, and $\hat{\mu}_{t-1}$ and $\hat{\sigma}_{t-1}$ are the estimated values of the same parameters. The analytical values of this misalignment signal are plotted in Fig. 7B.

In practice, the encoder does not have access to the parameters of the true stimulus distribution, and must therefore estimate the misalignment signal directly from incoming stimulus samples. This is discussed in more detail in Section C.3.3.

C Numerical Simulations

C.1 Environment parameters

All results were generated using a probe environment in which the state θ_t switched between two fixed values, θ^L and θ^H, every 100 time samples (corresponding to a hazard rate of $h = 0.01$). A single cycle of this probe environment consists of 200 time samples, for which the environment is in the low state ($\theta_t = \theta^L$) for the first 100 time samples and in the high state ($\theta_t = \theta^H$) for the second 100 time samples. In the main text, we averaged results over multiple cycles of the probe environment.

For the mean-switching environment, the state θ_t parametrized the mean of the stimulus distribution and switched between $\mu = \theta^L = -1$ and $\mu = \theta^H = 1$. The standard deviation was fixed to $\sigma = 1$. For the variance-switching environment, θ_t parametrized the standard deviation of the stimulus distribution and switched between $\sigma = \theta^L = 1$ and $\sigma = \theta^H = 2$. The mean was fixed to $\mu = 0$.

C.2 Updating the posterior

On each timestep, a single stimulus x_t was drawn randomly from $p(x_t|\theta_t)$. The stimulus was encoded, decoded, and used to update the posterior P_t^L. Updating the posterior requires marginalizing over the decoding distribution $p(x_t|y_t, \hat{\theta}_{\tau<t})$ (as given by Eq. 9). We approximated this marginalization numerically via Monte-Carlo simulation. At each time step, we generated 200 samples from the decoding distribution specified by each encoding scheme (for reference, the decoding distributions are given in Eqs. 24, 30, and 34). Individual samples were then used to compute separate estimates of the posterior, and the resulting set of estimates was averaged over samples. Results were robust to the number of samples used, provided that this number exceeded 50. In the case of encoding via discretization, we found that results were not sensitive to the inclusion of this marginalization step. We therefore computed all results for the discretization encoding scheme in the absence of marginalization by using the neural response y_t to directly update the posterior. This posterior forms the basis of all estimates $\hat{\theta}_t$ and predictions $\vec{\theta}_{t+1}$.

C.3 Optimizing the encoding

For two of the three encoding schemes (discretization and temporal filtering), the estimate $\hat{\theta}_{t-1}$ was used to optimize a set of encoding parameters (the set of neural response levels $\{y_t^i\}$ in the case of discretization, and the filtering coefficient α_t in the case of temporal filtering). To perform these optimizations, we discretized the posterior P_t^L into 100 values equally spaced between 0 and 1. This resulted in a set of 100 discretized values of the estimate state $\hat{\theta}_{bin}$. We found the optimal encoding parameters for each value of $\hat{\theta}_{bin}$ (described in detail in the following sections); this resulted in 100 sets of optimal response levels (given a fixed number of levels), and 100 values of the filtering coefficient α (given a fixed constraint strength β). On each timestep of the simulation, the true estimate $\hat{\theta}_t$ was mapped onto the closest discretized value $\hat{\theta}_{bin}$. The corresponding encoding parameters were then used to encode the incoming stimulus x_t. Additional details of each optimization procedure are described in the following sections.

C.3.1 Limited neural response levels: encoding via discretization

Response levels were chosen to optimize the following objective function:

$$\{y_t^i\}^* = \underset{\{y_t^i\}}{\operatorname{argmin}} \left\langle E(x_t, \{y_t^i\}) \right\rangle_{p(x_t|\vec{\theta}_t)} \tag{41}$$

The optimal set of response levels $\{y_t^i\}^*$ was found numerically using Lloyd's algorithm [41] (often referred to as K-means clustering). The algorithm takes the following as inputs: a set of points to be clustered $\{x\}$ (corresponding to stimulus samples), a number of quantization levels N (corresponding to the number

of neural response levels), and a distortion measure $d(x, y)$ (corresponding to the error function $E(x, y)$). The goal of the algorithm is to find a quantization (what we referred to as a discretization of the stimulus space) that minimizes the average value of the distortion.

The values of the quantization levels, $y^1, \ldots, y^N$, are first randomly initialized. The algorithm then proceeds in two steps:

1. Each point x is assigned to a quantization level y^i that yields the smallest distortion $d(x, y^i)$.
2. Each quantization level is replaced by the average value of the points assigned to it.

The two steps are iterated until convergence.

We computed a set of optimal quantization levels (optimal response levels) for each of the 100 discretized values of $\hat{\theta}_{bin}$ (described in Section C.2). For each value of $\hat{\theta}_{bin}$, we generated a training dataset $\{x\}$ consisting of 50000 values drawn from the estimated stimulus distribution $p(x_t|\hat{\theta}_{bin})$. We determined the boundaries of each quantization level (i.e., the values $y^{i,L}$ and $y^{i,H}$ that bounded the set of stimuli that were mapped to the same quantization level) by assigning points in the training dataset to the quantization level y^i that minimized $d(x, y^i)$.

To compute optimal quantization levels for stimulus reconstruction, we used the standard distortion measure $d(x, y) = (x - y)^2$; in this case, the algorithm is guaranteed to converge to the global optimum. To compute optimal quantization levels for inference, we defined the distortion measure to be $d(x, y) = (\hat{\theta}_x - \hat{\theta}_y)^2$. The algorithm is not guaranteed to converge to a global optimum in this case, but we found empirically that the algorithm converged to a local optimum (Fig. 10). Moreover, the two distortion measures did not produce equivalent results.

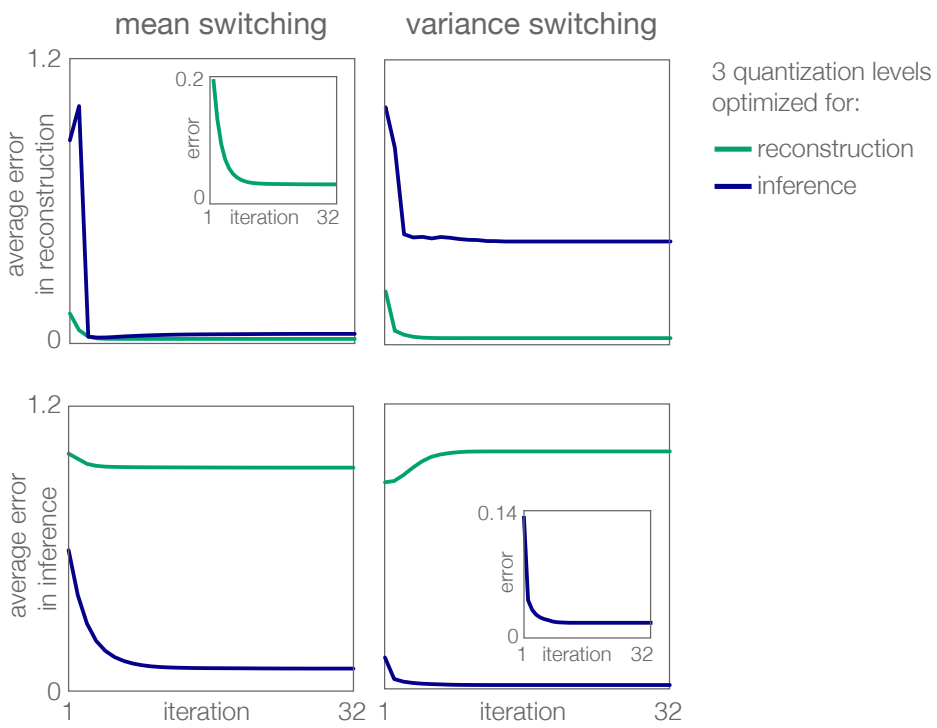


Figure 10: **Learning of optimal quantizers with Lloyd's algorithm.** Optimization of 3 quantization levels for reconstruction (green) differs from optimization for inference (blue). Regardless of the distortion measure, the algorithm converges to an optimum.

C.3.2 Limited gain and temporal acuity: encoding via temporal filtering

The optimal filtering coefficient was chosen to minimize the following objective function:

$$\alpha_t^* = \underset{\alpha_t}{\operatorname{argmin}} \langle E(x_t, y_t) \rangle_{p(x_t|\hat{\theta}_{t-1})} + \beta H(y_t, y_{t+1}|\hat{\theta}_{\tau<t}), \tag{42}$$

where as before, we choose $E(x, y) = (\hat{\theta}_x - \hat{\theta}_y)^2$ when optimizing for inference, and $E(x, y) = (x - y)^2$ when optimizing for reconstruction.

The joint entropy $H(y_t, y_{t+1}|\hat{\theta}_{\tau<t})$ can be determined analytically, as derived in Section B.5.3. We approximated the error term, $\langle E(x_t, y_t) \rangle_{p(x_t|\hat{\theta}_{t-1})}$, numerically. To do so, we first discretized α into 50 values evenly spaced between 0 and 1 (corresponding to 50 discrete values of α_{bin}). As described in Section B.5.3, we also discretized the posterior P_t^L into 100 values (corresponding to 100 discrete values of $\hat{\theta}_{bin}$). For each combination of α_{bin} and $\hat{\theta}_{bin}$, we generated 50000 pairs of stimulus samples (x_{t-1}, x_t) from the distribution $p(x_t|\hat{\theta}_{t-1})$. Each sample was used to compute values of the estimates $\hat{\theta}_x$ and $\hat{\theta}_y$. The errors $(\hat{\theta}_x - \hat{\theta}_y)^2$ and $(x_t - y_t)^2$ were then averaged over all 50000 stimulus pairs.

The optimal value α_t^* was then chosen as the value of α_{bin} that minimized the objective in Eq. 42 for a given choice of the error function $E(x, y)$ and constraint strength β.

C.3.3 Limited neural activity: encoding via dynamic stimulus selection

The misalignment signal, derived in Section B.6.3, was defined in terms of the relative alignment between the true stimulus distribution, $p(x_t|\theta_t)$, and the estimated stimulus distribution, $p(x_t|\hat{\theta}_{t-1})$. When the parameters of the true stimulus distribution are known, the value of this signal can be computed analytically via Eq. 38. However, when the system does not have access to the stimulus distribution (as is the case here), this signal must be estimated directly from incoming stimulus samples. We consider a scenario in which the encoder can approximate Eq. 38 by computing a running-average of the stimulus surprise:

$$M_t = -\frac{1}{T} \sum_{\tau=0}^{T} \log p(x_{t-\tau}|\hat{\theta}_{t-1}) - H(x_t|\hat{\theta}_{t-1}), \tag{43}$$

where T specifies the number of timebins used to estimate the average surprise. All results in the main text were generated using $T = 10$ timebins.

C.4 Illustration of observer's sensitivity to surprise and uncertainty

Fig. 1B illustrated the relative impact of different stimuli on the observer's estimate of the mean of a Gaussian stimulus distribution, which is modulated by the observer's uncertainty and the surprise of incoming stimuli. To illustrate this, we used the optimal Bayesian estimator of the mean of a Gaussian stimulus distribution, as derived in [31]. We consider a snapshot of the inference process, when the observer's prior is centered around a particular mean estimate.

The observer's uncertainty is captured by the entropy of the prior distribution. In a Gaussian model, the prior is also a Gaussian distribution, and its entropy is fully determined by its variance. We discretized this variance into 100 values spaced uniformly between 0.1 and 5.

The surprise of a single stimulus observation is quantified by the negative log-likelihood of the stimulus value given the observer's estimate of the mean. In a Gaussian model, this surprise is a monotonically-increasing function of the distance between the stimulus and the observer's estimate of the mean. We

used the average value of the prior as a measure of the observer's estimate of the mean. We then considered an interval of length 2 measured from this estimate, and we discretized this interval into 100 uniformly-spaced values. We then computed the stimulus surprise for each discretized value along this interval.

For each combination of prior uncertainty and surprise, we computed the posterior over the mean (corresponding to a stimulus observation with a particular surprise value). We then computed the squared difference between the average value of the prior and the average value of the posterior, and we used this squared difference as a measure of the impact of a single stimulus observation on the observer's estimate of the mean.

C.5 Generating spike rasters

Fig. 4B showed simulated spike rasters for an encoding scheme with limited neural response levels. To generate these rasters, a stimulus sample x_t was randomly drawn from the true stimulus distribution $p(x_t|\theta_t)$. This stimulus was then mapped onto one of $N = 4$ neural response levels. Each response level was assigned a binary spike pattern from the set $\{[00], [10], [01], [11]\}$, where 1 or 0 correspond to presence or absence of a spike, respectively. Patterns were assigned to response levels $\{y_t^i\}$ according to the estimated probability, $p(y_t^i|\hat{\theta}_{t-1})$, that a particular level would be used to encode incoming stimuli. In this way, the pattern with the fewest spikes ([00]) was assigned to the response level with the highest estimated probability, and the pattern with the most spikes ([11]) was assigned to the level with the lowest estimated probability. This strategy (called "entropy coding") achieves the shortest average encoding of the input by using the fewest number of spikes [41]. We simulated spike patterns for 800 cycles of the probe environment using the set of response levels optimized for inference or stimulus reconstruction.

C.6 Computing metamer probabilities

Fig. 4D reported the probability of a metamer as a function of the alignment between the true state of the environment θ and the observer's prediction $\vec{\theta}$. We say that two stimuli x_t^1 and x_t^2 are metamers (i.e., they are indistinguishable to the observer) if in the process of encoding they become mapped on the same neural response level y^M (i.e., $y_t^1 = y_t^2 = y^M$). The probability of a metamer, $p(y_t^1 = y_t^2|\theta_t, \hat{\theta}_{t-1})$, depends on both the true and predicted states of the environment. We numerically estimated this probability for a mean-switching environment in the low state ($\theta = \theta^L$). We generated 100 values of $\hat{\theta}_{t-1}$, evenly spaced between θ^L and θ^H. For each value of $\hat{\theta}_{t-1}$, we drew 100000 pairs of samples from the stimulus distribution $p(x_t|\theta_t = \theta^L)$. We encoded each stimulus by mapping it onto the corresponding response level y_t (using an encoder with 8 response levels, optimized as described in Section C.3.1). If both stimuli in the pair were mapped on the same response level, we counted the trial as a metamer. The total probability of a metamer was computed as the proportion of all trials that resulted in metamers.

APPENDIX I

Here, we provide a detailed derivation of the decoding distribution for the filtering encoder (described in Section B.5.2).

To simplify Eq. 29, we rewrite the first Gaussian as a function of $\alpha_t x_t$ (for notational simplicity, we will write $\sigma'^2 = (1 - \alpha_t)^2 \hat{\sigma}_{t-1}^2$:

$$
\begin{aligned}
\mathcal{N}(*) &= \mathcal{N}\left(y_t; \alpha_t x_t + (1-\alpha_t)\hat{\mu}_{t-1}, \sigma'^2\right) \\
&= \frac{1}{\sqrt{2\pi\sigma'^2}} \exp\left(-\frac{(y_t - (\alpha_t x_t + (1-\alpha_t)\hat{\mu}_{t-1}))^2}{2\sigma'^2}\right) \\
&= \frac{1}{\sqrt{2\pi\sigma'^2}} \exp\left(-\frac{(\alpha_t x_t - (y_t - (1-\alpha_t)\hat{\mu}_{t-1}))^2}{2\sigma'^2}\right) \\
&= \mathcal{N}\left(\alpha_t x_t; y_t - (1-\alpha_t)\hat{\mu}_{t-1}, \sigma'^2\right)
\end{aligned} \tag{44}
$$

We can now pull out the factor of α_t (again, for notational simplicity, we will write $\mu' = y_t - (1-\alpha_t)\hat{\mu}_{t-1}$):

$$
\begin{aligned}
\mathcal{N}(*) &= \mathcal{N}\left(\alpha_t x_t; \mu', \sigma'^2\right) \\
&= \frac{1}{\sqrt{2\pi\sigma'^2}} \exp\left(-\frac{(\alpha_t x_t - \mu')^2}{2\sigma'^2}\right) \\
&= \frac{1}{\alpha_t\sqrt{2\pi\sigma'^2/\alpha^2}} \exp\left(-\frac{\alpha_t^2(x_t - \mu'/\alpha_t)^2}{2\sigma'^2}\right) \\
&= \frac{1}{\alpha_t\sqrt{2\pi\sigma''^2}} \exp\left(-\frac{(x_t - \mu'')^2}{2\sigma''^2}\right) \\
&= \frac{1}{\alpha_t}\mathcal{N}\left(x_t; \mu'', \sigma''^2\right)
\end{aligned} \tag{45}
$$

where $\mu'' = \mu'/\alpha_t = (y_t - (1-\alpha_t)\hat{\mu}_{t-1}))/\alpha_t$ and $\sigma''^2 = \sigma'^2/\alpha_t^2 = (1-\alpha_t)^2\hat{\sigma}_{t-1}^2/\alpha_t^2$. Eq. 44 can now be written as a Gaussian over x_t:

$$
\mathcal{N}(*) = \frac{1}{\alpha_t}\mathcal{N}\left(x_t; (y_t - (1-\alpha_t)\hat{\mu}_{t-1})/\alpha_t, (1-\alpha_t)^2\hat{\sigma}_{t-1}^2/\alpha_t^2\right) \tag{46}
$$

This allows us to combine the two distributions in Eq. 29:

$$
\begin{aligned}
p(x_t|y_t, \vec{\theta}_t) &= \frac{1}{Z(y_t, \hat{\theta}_{\tau<t})}\frac{1}{\alpha_t}\mathcal{N}\left(x_t; \mu_A, \sigma_A^2\right)\mathcal{N}(x_t; \mu_B, \sigma_B^2) \\
&= \frac{1}{Z(y_t, \hat{\theta}_{\tau<t})}\frac{1}{\alpha_t}\frac{1}{\sqrt{2\pi(\sigma_A^2+\sigma_B^2)}}\exp\left(-\frac{(\mu_A-\mu_B)^2}{2(\sigma_A^2+\sigma_B^2)}\right)\mathcal{N}\left(x_t; \frac{\sigma_B^2\mu_A + \sigma_A^2\mu_B}{\sigma_A^2+\sigma_B^2}, \frac{\sigma_A^2\sigma_B^2}{\sigma_A^2+\sigma_B^2}\right) \\
&= \frac{f(y_t, \hat{\theta}_{\tau<t})}{Z(y_t, \hat{\theta}_{\tau<t})}\mathcal{N}\left(x_t; \frac{\sigma_B^2\mu_A + \sigma_A^2\mu_B}{\sigma_A^2+\sigma_B^2}, \frac{\sigma_A^2\sigma_B^2}{\sigma_A^2+\sigma_B^2}\right)
\end{aligned} \tag{47}
$$

where:

$$
\begin{aligned}
\mu_A &= (y_t - (1-\alpha_t)\hat{\mu}_{t-1})/\alpha_t \\
\mu_B &= \hat{\mu}_{t-1} \\
\sigma_A^2 &= (1-\alpha_t)^2\hat{\sigma}_{t-1}^2/\alpha_t^2 \\
\sigma_B^2 &= \hat{\sigma}_{t-1}^2
\end{aligned} \tag{48}
$$

Because the function $f(y_t, \hat{\theta}_{\tau<t})$ does not depend on x_t, we can trivially obtain $Z(y_t, \hat{\theta}_{\tau<t})$ by integrating over x_t (as given by Eq. 5):

$$
\begin{aligned}
Z(y_t, \hat{\theta}_{\tau<t}) &= \int dx_t p(y_t|x_t, \hat{\theta}_{\tau<t})p(x_t|\hat{\theta}_{\tau<t}) \\
&= f(y_t, \hat{\theta}_{\tau<t})\int dx_t\,\mathcal{N}\left(x_t; \frac{\sigma_B^2\mu_A + \sigma_A^2\mu_B}{\sigma_A^2+\sigma_B^2}, \frac{\sigma_A^2\sigma_B^2}{\sigma_A^2+\sigma_B^2}\right) \\
&= f(y_t, \hat{\theta}_{\tau<t})
\end{aligned} \tag{49}
$$

The remaining terms in Eq. 47 are given by:

$$
\begin{aligned}
\sigma_A^2 + \sigma_B^2 &= \left(\frac{1 - 2\alpha_t + 2\alpha_t^2}{\alpha_t^2}\right)\hat{\sigma}_{t-1}^2 \\
\sigma_A^2\sigma_B^2 &= \left(\frac{(1-\alpha_t)^2}{\alpha_t^2}\right)\hat{\sigma}_{t-1}^4 \\
\frac{\sigma_A^2\sigma_B^2}{\sigma_A^2 + \sigma_B^2} &= \left(\frac{(1-\alpha_t)^2}{1 - 2\alpha_t + 2\alpha_t^2}\right)\hat{\sigma}_{t-1}^2 \\
\sigma_B^2\mu_A + \sigma_A^2\mu_B &= \left(\frac{\alpha_t y_t - (1-\alpha_t)(2\alpha_t - 1)\hat{\mu}_{t-1}}{\alpha_t^2}\right)\hat{\sigma}_{t-1}^2 \\
\frac{\sigma_B^2\mu_A + \sigma_A^2\mu_B}{\sigma_A^2 + \sigma_B^2} &= \frac{\alpha_t y_t - (1-\alpha_t)(2\alpha_t - 1)\hat{\mu}_{t-1}}{1 - 2\alpha_t + 2\alpha_t^2}
\end{aligned}
\tag{50}
$$

Putting everything together, the final form of Eq. 29 becomes:

$$
p(x_t|y_t, \hat{\theta}_{\tau<t}) = \mathcal{N}\left(x_t; \frac{\alpha_t y_t - (1-\alpha_t)(2\alpha_t - 1)\hat{\mu}_{t-1}}{1 - 2\alpha_t + 2\alpha_t^2}, \left(\frac{(1-\alpha_t)^2}{1 - 2\alpha_t + 2\alpha_t^2}\right)\hat{\sigma}_{t-1}^2\right) \tag{51}
$$

For $\frac{1}{2} \leq \alpha_t \leq 1$, we can see that: $0 \leq (1-\alpha_t)(2\alpha_t - 1) \leq \frac{1}{8}$ and $\frac{1}{2} \leq (1 - 2\alpha_t + 2\alpha_t^2) \leq 1$.

APPENDIX II

Here we provide a detailed derivation of the entropy of the output of filtering encoder (described in Section B.5.3).

To compute $H(y_t, y_{t+1}|\hat{\theta}_{\tau<t})$, we assume that the encoder has access to the history of estimates $\hat{\theta}_{\tau<t}$, and that it uses the most recent estimate $\hat{\theta}_{t-1}$ as an approximate prediction of future states (i.e., $\hat{\theta}_{t-1} \approx \vec{\theta}_t \approx \vec{\theta}_{t+1}$).

For reference, the entropy of a normal distribution is:

$$
H[\mathcal{N}(x; \mu, \sigma^2)] = \frac{1}{2}\log_e(2\pi e\sigma^2) \tag{52}
$$

We want to compute $H(y_t, y_{t+1}|\hat{\theta}_{\tau<t})$:

$$
H(y_t, y_{t+1}|\hat{\theta}_{\tau<t}) = H(y_{t+1}|y_t, \hat{\theta}_{\tau<t}) + H(y_t|\hat{\theta}_{\tau<t}) \tag{53}
$$

where $y_t = \alpha_t x_t + (1-\alpha_t)x_{t-1}$ is the output of the encoder, and $\alpha_t \in [0.5, 1]$ is the filtering coefficient.

To compute each of the terms in Eq. 53, we need to compute $p(y_t|\hat{\theta}_{\tau<t})$ and $p(y_{t+1}|y_t, \hat{\theta}_{\tau<t})$. The first of these distributions is given by:

$$
p(y_t|\hat{\theta}_{\tau<t}) = \mathcal{N}\left(y_t; \alpha_t\hat{\mu}_{t-1} + (1-\alpha_t)\hat{\mu}_{t-1}, \alpha_t^2\hat{\sigma}_{t-1}^2 + (1-\alpha_t)^2\hat{\sigma}_{t-1}^2\right), \tag{54}
$$

whose entropy is given by:

$$H(y_t|\hat{\theta}_{\tau<t}) = \frac{1}{2}\log_e\left(2\pi e(\alpha_t^2\hat{\sigma}_{t-1}^2 + (1-\alpha_t)^2\hat{\sigma}_{t-1}^2)\right). \tag{55}$$

The second of these distributions can be written as:

$$p(y_{t+1}|y_t,\hat{\theta}_{\tau<t}) = \int dx_t p(y_{t+1}|x_t,\hat{\theta}_{\tau<t})p(x_t|y_t,\hat{\theta}_{\tau<t}) \tag{56}$$

Noting that $p(y_{t+1}|x_{t+1},x_t,\hat{\theta}_{\tau<t}) = \delta(y_{t+1} - (\alpha_t x_{t+1} + (1-\alpha_t)x_t))$, the first term in the integral in Eq. 56 is given by:

$$\begin{aligned} p(y_{t+1}|x_t,\hat{\theta}_{\tau<t}) &= \int dx_{t+1}\,\delta(y_{t+1} - (\alpha_t x_{t+1} + (1-\alpha_t)x_t))\mathcal{N}\left(x_{t+1};\hat{\mu}_{t-1},\hat{\sigma}_{t-1}^2\right) \\ &= \frac{1}{\alpha_t}\mathcal{N}\left(\frac{(y_{t+1}-(1-\alpha_t)x_t)}{\alpha_t};\hat{\mu}_{t-1},\hat{\sigma}_{t-1}^2\right) \\ &= \frac{1}{(1-\alpha_t)}\mathcal{N}\left(x_t;\frac{(y_{t+1}-\alpha_t\hat{\mu}_{t-1})}{(1-\alpha_t)},\frac{\alpha_t^2}{(1-\alpha_t)^2}\hat{\sigma}_{t-1}^2\right) \end{aligned} \tag{57}$$

The second term in the integral in Eq. 56 is given by:

$$p(x_t|y_t,\hat{\theta}_{\tau<t}) = \mathcal{N}\left(x_t;\frac{\alpha_t y_t - (1-\alpha_t)(2\alpha_t-1)\hat{\mu}_{t-1}}{\alpha_t^2+(1-\alpha_t)^2},\left(\frac{(1-\alpha_t)^2}{\alpha_t^2+(1-\alpha_t)^2}\right)\hat{\sigma}_{t-1}^2\right) \tag{58}$$

Combining the two terms, we have:

$$\begin{aligned} p(y_{t+1}|x_t,\hat{\theta}_{\tau<t})p(x_t|y_t,\hat{\theta}_{\tau<t}) &= \frac{1}{(1-\alpha_t)}\mathcal{N}\left(x_t;\mu_A,\sigma_A^2\right)\mathcal{N}\left(x_t;\mu_B,\sigma_B^2\right) \\ &= \frac{1}{(1-\alpha_t)}\frac{1}{\sqrt{2\pi(\sigma_A^2+\sigma_B^2)}}\exp\left(-\frac{(\mu_A-\mu_B)^2}{2(\sigma_A^2+\sigma_B^2)}\right) \\ &\mathcal{N}\left(x_t;\frac{\sigma_B^2\mu_A+\sigma_A^2\mu_B}{\sigma_A^2+\sigma_B^2},\frac{\sigma_A^2\sigma_B^2}{\sigma_A^2+\sigma_B^2}\right) \end{aligned} \tag{59}$$

where

$$\begin{aligned} \mu_A &= \frac{(y_{t+1}-\alpha_t\hat{\mu}_{t-1})}{(1-\alpha_t)} \\ \mu_B &= \frac{\alpha_t y_t - (1-\alpha_t)(2\alpha_t-1)\hat{\mu}_{t-1}}{\alpha_t^2+(1-\alpha_t)^2} \\ \sigma_A^2 &= \frac{\alpha_t^2}{(1-\alpha_t)^2}\hat{\sigma}_{t-1}^2 \\ \sigma_B^2 &= \frac{(1-\alpha_t)^2}{\alpha_t^2+(1-\alpha_t)^2}\hat{\sigma}_{t-1}^2 \end{aligned} \tag{60}$$

Putting these terms back into the integral in Eq. 56 gives:

$$
\begin{aligned}
p(y_{t+1}|y_t, \hat{\theta}_{\tau<t}) &= \frac{1}{(1-\alpha_t)} \frac{1}{\sqrt{2\pi(\sigma_A^2+\sigma_B^2)}} \exp\left(-\frac{(\mu_A-\mu_B)^2}{2(\sigma_A^2+\sigma_B^2)}\right) \\
&= \mathcal{N}\left(y_{t+1}; \alpha_t\hat{\mu}_{t-1} + \frac{(1-\alpha_t)(\alpha_t y_t - (1-\alpha_t)(2\alpha-1)\hat{\mu}_{t-1})}{1-2\alpha+2\alpha_t^2}, \alpha_t^2\hat{\sigma}_{t-1}^2 + \frac{(1-\alpha_t)^4}{1-2\alpha_t+2\alpha_t^2}\hat{\sigma}_{t-1}^2\right)
\end{aligned}
\tag{61}
$$

The conditional entropy $H(y_{t+1}|y_t, \hat{\theta}_{\tau<t})$ is determined by the variance in this distribution:

$$
H(y_{t+1}|y_t, \hat{\theta}_{\tau<t}) = \frac{1}{2}\log_e\left(2\pi e\left(\alpha^2\hat{\sigma}_{t-1}^2 + \frac{(1-\alpha)^4}{1-2\alpha+2\alpha^2}\hat{\sigma}_{t-1}^2\right)\right) \tag{62}
$$

Combining the two entropy terms in Eqs. 55 and 62, we get:

$$
\begin{aligned}
H(y_t, y_{t+1}|\hat{\theta}_{\tau<t}) &= H(y_{t+1}|y_t, \hat{\theta}_{\tau<t}) + H(y_t|\hat{\theta}_{\tau<t}) \\
&= \frac{1}{2}\log_e\left(4\pi^2e^2\left(\alpha_t^2\hat{\sigma}_{t-1}^2 + \frac{(1-\alpha_t)^4}{1-2\alpha_t+2\alpha_t^2}\hat{\sigma}_{t-1}^2\right)\left(\alpha_t^2\hat{\sigma}_{t-1}^2 + (1-\alpha_t)^2\hat{\sigma}_{t-1}^2\right)\right)
\end{aligned}
\tag{63}
$$

References

[1] D. Kersten and P. Schrater, *Pattern inference theory: A probabilistic approach to vision.* na, 2002.

[2] D. H. Brainard, P. Longere, P. B. Delahunt, W. T. Freeman, J. M. Kraft, and B. Xiao, "Bayesian model of human color constancy," *Journal of Vision*, vol. 6, no. 11, pp. 10–10, 2006.

[3] Y. Weiss, E. P. Simoncelli, and E. H. Adelson, "Motion illusions as optimal percepts," *Nature Neuroscience*, vol. 5, no. 6, pp. 598–604, 2002.

[4] B. J. Fischer and J. L. Peña, "Owl's behavior and neural representation predicted by bayesian inference," *Nature Neuroscience*, vol. 14, no. 8, pp. 1061–1066, 2011.

[5] W. Młynarski, "The opponent channel population code of sound location is an efficient representation of natural binaural sounds," *PLOS Computational Biology*, vol. 11, no. 5, p. e1004294, 2015.

[6] D. Kersten, P. Mamassian, and A. Yuille, "Object perception as bayesian inference," *Annual Review of Psychology*, vol. 55, pp. 271–304, 2004.

[7] J. Fiser, P. Berkes, G. Orbán, and M. Lengyel, "Statistically optimal perception and learning: from behavior to neural representations," *Trends in Cognitive Sciences*, vol. 14, no. 3, pp. 119–130, 2010.

[8] R. P. Rao, B. A. Olshausen, and M. S. Lewicki, *Probabilistic models of the brain: Perception and neural function.* MIT press, 2002.

[9] R. Coen-Cagli, A. Kohn, and O. Schwartz, "Flexible gating of contextual influences in natural vision," *Nature neuroscience*, vol. 18, no. 11, p. 1648, 2015.

[10] B. Wark, A. Fairhall, and F. Rieke, "Timescales of inference in visual adaptation," *Neuron*, vol. 61, no. 5, pp. 750–761, 2009.

[11] M. DeWeese and A. Zador, "Asymmetric dynamics in optimal variance adaptation," *Neural Computation*, vol. 10, no. 5, pp. 1179–1202, 1998.

[12] M. R. Nassar, R. C. Wilson, B. Heasly, and J. I. Gold, "An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment," *Journal of Neuroscience*, vol. 30, no. 37, pp. 12366–12378, 2010.

[13] T. Lochmann, U. A. Ernst, and S. Deneve, "Perceptual inference predicts contextual modulations of sensory responses," *Journal of Neuroscience*, vol. 32, no. 12, pp. 4179–4195, 2012.

[14] S. Deneve, "Bayesian spiking neurons i: inference," *Neural Computation*, vol. 20, no. 1, pp. 91–117, 2008.

[15] S. B. Laughlin, R. R. d. R. van Steveninck, and J. C. Anderson, "The metabolic cost of neural information," *Nature Neuroscience*, vol. 1, no. 1, pp. 36–41, 1998.

[16] P. Mehta and D. J. Schwab, "Energetic costs of cellular computation," *Proceedings of the National Academy of Sciences*, vol. 109, no. 44, pp. 17978–17982, 2012.

[17] V. Balasubramanian, D. Kimber, and M. J. Berry II, "Metabolically efficient information processing," *Neural Computation*, vol. 13, no. 4, pp. 799–815, 2001.

[18] J. J. Harris, R. Jolivet, and D. Attwell, "Synaptic energy use and supply," *Neuron*, vol. 75, no. 5, pp. 762–777, 2012.

[19] D. Attwell and S. B. Laughlin, "An energy budget for signaling in the grey matter of the brain," *Journal of Cerebral Blood Flow & Metabolism*, vol. 21, no. 10, pp. 1133–1145, 2001.

[20] W. B. Levy and R. A. Baxter, "Energy efficient neural codes," *Neural Computation*, vol. 8, no. 3, pp. 531–543, 1996.

[21] S. B. Laughlin and T. J. Sejnowski, "Communication in neuronal networks," *Science*, vol. 301, no. 5641, pp. 1870–1874, 2003.

[22] A. M. Hermundstad, J. J. Briguglio, M. M. Conte, J. D. Victor, V. Balasubramanian, and G. Tkačik, "Variance predicts salience in central sensory processing," *eLife*, vol. 3, p. e03722, 2014.

[23] H. B. Barlow, "Possible principles underlying the transformations of sensory messages," 1961.

[24] J. Van Hateren, "A theory of maximizing sensory information," *Biological Cybernetics*, vol. 68, no. 1, pp. 23–29, 1992.

[25] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, p. 607, 1996.

[26] S. Laughlin, "A simple coding procedure enhances a neuron's information capacity," *Zeitschrift für Naturforschung c*, vol. 36, no. 9-10, pp. 910–912, 1981.

[27] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[28] S. E. Palmer, O. Marre, M. J. Berry, and W. Bialek, "Predictive information in a sensory population," *Proceedings of the National Academy of Sciences*, vol. 112, no. 22, pp. 6908–6913, 2015.

[29] W. S. Geisler, J. Najemnik, and A. D. Ing, "Optimal stimulus encoders for natural tasks," *Journal of Vision*, vol. 9, no. 13, pp. 17–17, 2009.

[30] J. Burge and W. S. Geisler, "Optimal speed estimation in natural image movies predicts human performance," *Nature Communications*, vol. 6, 2015.

[31] K. P. Murphy, "Conjugate bayesian analysis of the gaussian distribution," *def*, vol. 1, no. $2\sigma 2$, p. 16, 2007.

[32] A. L. Fairhall, G. D. Lewen, W. Bialek, and R. R. d. R. van Steveninck, "Efficiency and ambiguity in an adaptive neural code," *Nature*, vol. 412, no. 6849, pp. 787–792, 2001.

[33] R. C. Wilson, M. R. Nassar, and J. I. Gold, "A mixture of delta-rules approximation to bayesian inference in change-point problems," *PLOS Computational Biology*, vol. 9, no. 7, p. e1003150, 2013.

[34] A. E. Radillo, A. Veliz-Cuba, K. Josić, and Z. P. Kilpatrick, "Evidence accumulation and change rate inference in dynamic environments," *Neural Computation*, 2017.

[35] A. Veliz-Cuba, Z. P. Kilpatrick, and K. Josic, "Stochastic models of evidence accumulation in changing environments," *SIAM Review*, vol. 58, no. 2, pp. 264–289, 2016.

[36] C. Robert, *The Bayesian choice: from decision-theoretic foundations to computational implementation.* Springer Science & Business Media, 2007.

[37] P. Sterling and S. Laughlin, *Principles of neural design.* MIT Press, 2015.

[38] W. Bialek, F. Rieke, R. de Ruyter van Steveninck, and D. Warland, "Spikes: exploring the neural code," *MIT. Roddey, JC, Girish, B., & Miller, JP (2000). Assessing the performance of neural encoding models in the presence of noise. Journal of Computational Neuroscience*, vol. 8, no. 95, p. 112, 1997.

[39] J. Gjorgjieva, M. Meister, and H. Sompolinsky, "Optimal sensory coding by populations of on and off neurons," *bioRxiv*, p. 131946, 2017.

[40] V. Balasubramanian and M. J. Berry, "A test of metabolically efficient coding in the retina," *Network: Computation in Neural Systems*, vol. 13, no. 4, pp. 531–552, 2002.

[41] T. M. Cover and J. A. Thomas, *Elements of information theory.* John Wiley & Sons, 2012.

[42] J. Freeman and E. P. Simoncelli, "Metamers of the ventral stream," *Nature Neuroscience*, vol. 14, no. 9, pp. 1195–1201, 2011.

[43] J. H. McDermott, M. Schemitsch, and E. P. Simoncelli, "Summary statistics in auditory perception," *Nature Neuroscience*, vol. 16, no. 4, pp. 493–498, 2013.

[44] W. Bialek, F. Rieke, R. R. d. R. van Steveninck, and D. Warland, "Reading a neural code," in *Advances in Neural Information Processing Systems*, pp. 36–43, 1990.

[45] T. O. Sharpee, H. Sugihara, A. V. Kurgansky, S. P. Rebrik, M. P. Stryker, and K. D. Miller, "Adaptive filtering enhances information transmission in visual cortex," *Nature*, vol. 439, no. 7079, pp. 936–942, 2006.

[46] T. O. Sharpee, K. I. Nagel, and A. J. Doupe, "Two-dimensional adaptation in the auditory forebrain," *Journal of Neurophysiology*, vol. 106, no. 4, pp. 1841–1861, 2011.

[47] M. V. Srinivasan, S. B. Laughlin, and A. Dubs, "Predictive coding: a fresh view of inhibition in the retina," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 216, no. 1205, pp. 427–459, 1982.

[48] R. P. Rao and D. H. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nature Neuroscience*, vol. 2, no. 1, pp. 79–87, 1999.

[49] W. E. Vinje and J. L. Gallant, "Sparse coding and decorrelation in primary visual cortex during natural vision," *Science*, vol. 287, no. 5456, pp. 1273–1276, 2000.

[50] M. Weliky, J. Fiser, R. H. Hunt, and D. N. Wagner, "Coding of natural scenes in primary visual cortex," *Neuron*, vol. 37, no. 4, pp. 703–718, 2003.

[51] M. R. DeWeese and A. M. Zador, "Binary coding in auditory cortex," in *Advances in Neural Information Processing Systems*, pp. 117–124, 2003.

[52] B. A. Olshausen and D. J. Field, "Sparse coding of sensory inputs," *Current Opinion in Neurobiology*, vol. 14, no. 4, pp. 481–487, 2004.

[53] D. L. Rathbun, D. K. Warland, and W. M. Usrey, "Spike timing and information transmission at retinogeniculate synapses," *Journal of Neuroscience*, vol. 30, no. 41, pp. 13558–13566, 2010.

[54] L. M. Miller, M. A. Escabı, and C. E. Schreiner, "Feature selectivity and interneuronal cooperation in the thalamocortical system," *Journal of Neuroscience*, vol. 21, no. 20, pp. 8136–8144, 2001.

[55] P. Sabourin and G. S. Pollack, "Behaviorally relevant burst coding in primary sensory neurons," *Journal of Neurophysiology*, vol. 102, no. 2, pp. 1086–1091, 2009.

[56] A.-M. M. Oswald, M. J. Chacron, B. Doiron, J. Bastian, and L. Maler, "Parallel processing of sensory input by bursts and isolated spikes," *Journal of Neuroscience*, vol. 24, no. 18, pp. 4351–4362, 2004.

[57] N. A. Lesica and G. B. Stanley, "Encoding of natural scene movies by tonic and burst spikes in the lateral geniculate nucleus," *Journal of Neuroscience*, vol. 24, no. 47, pp. 10731–10740, 2004.

[58] G. Marsat and G. S. Pollack, "A behavioral role for feature detection by sensory bursts," *Journal of Neuroscience*, vol. 26, no. 41, pp. 10542–10547, 2006.

[59] W. S. Geisler, "Contributions of ideal observer theory to vision research," *Vision Research*, vol. 51, no. 7, pp. 771–781, 2011.

[60] E. P. Simoncelli, "Optimal estimation in sensory systems," *The Cognitive Neurosciences, IV*, pp. 525–535, 2009.

[61] W. J. Ma, J. M. Beck, P. E. Latham, and A. Pouget, "Bayesian inference with probabilistic population codes," *Nature Neuroscience*, vol. 9, no. 11, pp. 1432–1438, 2006.

[62] X.-X. Wei and A. A. Stocker, "A bayesian observer model constrained by efficient coding can explain'anti-bayesian'percepts," *Nature Neuroscience*, vol. 18, no. 10, pp. 1509–1517, 2015.

[63] D. Ganguli and E. P. Simoncelli, "Efficient sensory encoding and bayesian inference with heterogeneous neural populations," *Neural Computation*, 2014.

[64] T. Putzeys, M. Bethge, F. Wichmann, J. Wagemans, and R. Goris, "A new perceptual bias reveals suboptimal population decoding of sensory responses," *PLOS Computational Biology*, vol. 8, no. 4, p. e1002453, 2012.

[65] J. Rubin, N. Ulanovsky, I. Nelken, and N. Tishby, "The representation of prediction error in auditory cortex," *PLoS Computational Biology*, vol. 12, no. 8, p. e1005058, 2016.

[66] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information bottleneck for gaussian variables," *Journal of Machine Learning Research*, vol. 6, no. Jan, pp. 165–188, 2005.

[67] D. Strouse and D. J. Schwab, "The deterministic information bottleneck," *arXiv preprint arXiv:1604.00268*, 2016.

[68] M. Chalk, O. Marre, and G. Tkacik, "Relevant sparse codes with variational information bottleneck," in *Advances in Neural Information Processing Systems*, pp. 1957–1965, 2016.

[69] I. M. Park and J. W. Pillow, "Bayesian efficient coding," *bioRxiv*, p. 178418, 2017.

[70] M. Chalk, O. Marre, and G. Tkacik, "Towards a unified theory of efficient, predictive and sparse coding," *bioRxiv*, p. 152660, 2017.