# Phase transitions in when feedback is useful

**Lokesh Boominathan**
Department of ECE
Rice University
Houston, TX 77005
lokesh.boominathan@rice.edu

**Xaq Pitkow**
Dept. of Neuroscience, Dept. of ECE
Baylor College of Medicine, Rice University
Houston, TX 77005
xaq@rice.edu

## Abstract

Sensory observations about the world are invariably ambiguous. Inference about the world's latent variables is thus an important computation for the brain. However, computational constraints limit the performance of these computations. These constraints include energetic costs for neural activity and noise on every channel. Efficient coding is one prominent theory that describes how such limited resources can best be used. In one incarnation, this leads to a theory of predictive coding, where predictions are subtracted from signals, reducing the cost of sending something that is already known. This theory does not, however, account for the costs or noise associated with those predictions. Here we offer a theory that accounts for both feedforward and feedback costs, and noise in all computations. We formulate this inference problem as message-passing on a graph whereby feedback serves as an internal control signal aiming to maximize how well an inference tracks a target state while minimizing the costs of computation. We apply this novel formulation of *inference as control* to the canonical problem of inferring the hidden scalar state of a linear dynamical system with Gaussian variability. The best solution depends on architectural constraints, such as Dale's law, the ubiquitous law that each neuron makes solely excitatory or inhibitory postsynaptic connections. This biological structure can create asymmetric costs for feedforward and feedback channels. Under such conditions, our theory predicts the gain of optimal predictive feedback and how it is incorporated into the inference computation. We show that there is a non-monotonic dependence of optimal feedback gain as a function of both the computational parameters and the world dynamics, leading to phase transitions in whether feedback provides any utility in optimal inference under computational constraints.

## 1 Introduction

A critical computation for the brain is to infer the world's latent variables from ambiguous observations. Computational constraints, including metabolic costs and noisy signals, limit the performance of these inferences. Efficient coding [1] is a prominent theory that describes how limited resources can be used best. In one incarnation, this leads to the theory of predictive coding [2], which posits that predictions are sent along feedback channels to be subtracted from signals at lower cortical areas; only the difference returns to the higher areas along feedforward channels, reducing the metabolic or informational cost of sending redundant signals already known to the higher areas. This theory does not, however, account for the additional costs or noise associated with the feedback. Depending on the costs for sending predictions and the reliability of signals encoding those predictions, we expect different optimal strategies to perform computationally constrained inferences. For example, if the feedback channel is too unreliable and expensive, we hypothesize that it is not worth sending any predictions at all. Here we offer a more general theory of inference that accounts for the costs and reliabilities of the feedback and feedforward channels, and the relative importance of good inferences

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

about the latent world state. We formulate the inference problem as control via message-passing on a graph, maximizing how well an inference tracks a target state while minimizing the message costs. Messages become control actions with their own costs to reduce while improving how well an inference tracks a target state. We call this method *inference as control*, as it flips the interesting perspective of viewing optimal control as an inference problem [3]. We solve this problem under Linear-Quadratic-Gaussian (LQG) assumptions: Linear dynamics, Quadratic state and control costs, and Gaussian noise for the process, observations, and messages. Our theory enables us to determine the optimal predictions and how are they are integrated into computationally constrained inference. This analysis reveals phase transitions in when feedback is helpful, as we change the computation parameters or the world dynamics. Finally, we connect our theory's constraints to biology by providing a simplified example of how feedback/feedforward pathways with different anatomical structures can yield different optimal strategies.

## 2 Related work

Our work brings together several related theories: Bayesian inference, efficient coding, and predictive coding. The idea that the brain performs Bayesian inference about latent variables amid uncertain sensory observation has been long studied in neuroscience [4–7]. Bayesian inference involves optimally combining prior information from the past or from surrounding context with the likelihood of current sensory observations [8–11]. The theory of efficient coding [1, 12–15] focuses on encoding the sensory observations themselves, capturing maximal information subject to limited biological resources. The theory of predictive coding [2] tackles such resource constraints by using top-down feedback predictions to suppress the part of sensory observations that is already known to the brain, and sending only the novel part of the observation back to the brain [16–21]. Understanding the utility of predictive feedback has also shown promise in gaining insights into deep learning algorithms and improving their performance, recently drawing wide interest in this topic [22–29].

Park *et al.* [30] brought together the ideas of a Bayesian brain and efficient coding, by identifying efficient neural codes as optimizing a posterior distribution while accounting for limited firing rates. Chalk *et al.* [18] unified theories of predictive coding, efficient coding, and sparse coding by showing how these regimes emerge in a three dimensional constraint space described by channel capacity, past information content about the world state, and the time point to which the estimate is targeted. Młynarski *et al.* [31] investigated how sensory coding can adapt over time to account for the trade-off of an inference cost against a metabolic cost for high-fidelity encoding. Aitchison *et al.* [32] argued that Bayesian inference can be achieved using predictive coding, but is not necessary. While these works made important contributions towards unifying different encoding schemes that are optimal under different circumstances, they all assume that information is costly but computation is free. In particular, none of them have explicitly accounted for biological constraints along the feedback channel as well. In our work, we discuss how the optimal strategy changes when balancing inference performance against energetic costs, when there is noise in both feedforward and feedback pathways.

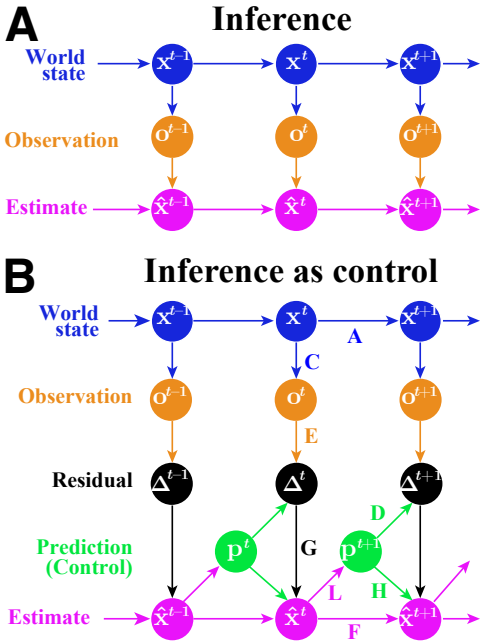


Figure 1: *Resource constrained inference modeled as a control problem.* ***A****: Graphical model of the inference problem, tracking a latent state in a Hidden Markov Model.* ***B****: Expanded inference problem, indicating prediction as control.*

## 3 Defining the problem

We consider an inference task in which the brain tracks a latent world state $x^t$ based on its noisy sensory observations $o^t$ (Fig 1A). The trajectory of the world state follows a known stochastic linear

dynamical system (Eq 1) with Gaussian process noise and Gaussian observations (Eq 2). At each time step, the brain sends a top-down prediction $p^t$ based on its best estimate $\hat{x}^{t-1}$ based on evidence up until the previous time step (Eq 3). The prediction is then sent through an additive Gaussian noise feedback channel to the sensory level (Eq 4). The noisy prediction $\tilde{p}^t$ is then combined with the new observations $o^t$ to form a residual $\Delta^t$ (Eq 5). The residual is then sent through an additive Gaussian noise feedforward channel to the brain (Eq 6). Based on this noisy residual $\tilde{\Delta}^t$ and the prediction it had just sent, the brain updates its estimate $\hat{x}^t$ (Eq 7). Fig 1B shows a trellis diagram of the dynamics.

Structurally, these dynamics are equivalent to a Kalman filter. The most common representation of this filter might even be viewed as predictive coding, as the update step uses the residual between the predicted and actual observation. However, the Kalman filter has no costs aside from the final inference, and no computational noise except the observation. Biological computation therefore may weigh fundamentally different tradeoffs in its inferences. Our goal is to find the parameters that minimize a weighted combination of inference loss, feedback energy cost, and feedforward energy cost (Eq 8) given limitations caused by computational noise.

We optimize the following parameters: the gain $L$ on the previous inference that is integrated into the new prediction; the prediction multiplier $D$ and observation multiplier $E$ that describe how the noisy prediction and the observation are weighted to form the residual; and the parameters $F$, $G$, and $H$ that determine how the inference is updated in light of the new noisy residual. The optimization is done at steady state, assuming the observer must continually update its estimate in a stationary dynamic environment. For mathematical tractability, we assume that all messages are linear functions of their inputs, the three losses are quadratic, the weight on the inference loss is a scalar, and all noise is independent Gaussian white noise. The equations governing this problem are:

$$x^t = A\, x^{t-1} + \eta_{\mathrm{p}}^t \qquad \text{state dynamics} \tag{1}$$

$$o^t = C\, x^t + \eta_{\mathrm{o}}^t \qquad \text{observation} \tag{2}$$

$$p^t = L\, \hat{x}^{t-1} \qquad \text{prediction} \tag{3}$$

$$\tilde{p}^t = p^t + \eta_{\mathrm{b}}^t \qquad \text{noisy prediction, feedback} \tag{4}$$

$$\Delta^t = D\, \tilde{p}^t + E\, o^t \qquad \text{residual} \tag{5}$$

$$\tilde{\Delta}^t = \Delta^t + \eta_{\mathrm{f}}^t \qquad \text{noisy residual, feedforward} \tag{6}$$

$$\hat{x}^t = F\, \hat{x}^{t-1} + G\, \tilde{\Delta}^t + H\, p^t \qquad \text{estimation} \tag{7}$$

$$\mathrm{Cost}_{\mathrm{tot}} = \lim_{T\to\infty} \frac{1}{T} \sum\nolimits_{t=1}^{T} \left( \underbrace{\left\| x^t - \hat{x}^t \right\|^2}_{\mathrm{Cost}_{\mathrm{inf}}} + \underbrace{\tilde{p}^{t\top} W_{\mathrm{b}}\, \tilde{p}^t}_{\mathrm{Cost}_{\mathrm{b}}} + \underbrace{\tilde{\Delta}^{t\top} W_{\mathrm{f}}\, \tilde{\Delta}^t}_{\mathrm{Cost}_{\mathrm{f}}} \right) \tag{8}$$

We consider how the optimal computational strategy varies with cost-weights ($W_{\mathrm{b}}, W_{\mathrm{f}}$) that determine the relative importance of feed**b**ack, feed**f**orward, and **inf**erence costs ($\mathrm{Cost}_{\mathrm{b}}$, $\mathrm{Cost}_{\mathrm{f}}$, and $\mathrm{Cost}_{\mathrm{inf}}$). $\eta_{\mathrm{p}}$, $\eta_{\mathrm{o}}$, $\eta_{\mathrm{b}}$, and $\eta_{\mathrm{f}}$ represent the process noise, observation noise, feedback noise, and feedforward noise with variances $\sigma_{\mathrm{p}}^2$, $\sigma_{\mathrm{o}}^2$, $\sigma_{\mathrm{b}}^2$, and $\sigma_{\mathrm{f}}^2$ respectively. We will see below that certain combinations of parameters determine the system's behavior at transition points.

## 4 Method: Inference as Control

We adopt a two-step approach to solve the optimization problem in Eq 8. First we fix the parameters $D$ and $E$ that determine how the feedback is used to subtract predictions, and solve in closed form for the optimal feedback gain ($L$) and optimal integration of residuals ($F$, $G$, $H$) as a function of the fixed parameters. We then numerically optimize for $D$ and $E$, given the optimal feedback. Mathematically, we write the two-step minimization as below, with an $\mathrm{argmin}$ determined analogously.

$$\min_{D,E,L,F,G,H} \mathrm{Cost}_{\mathrm{tot}} = \min_{D,E} \left( \min_{L,F,G,H} \mathrm{Cost}_{\mathrm{tot}} \right) \tag{9}$$

In order to find the closed form solution for fixed $D$ and $E$, we recast the minimization as an LQG control problem [33] where the prediction is treated as an internal control. The LQG dynamical Eqs 10-11 are obtained by concisely writing Eqs 1-2, and 4-6 in terms of an augmented state $z^t = \begin{bmatrix} x^t & \tilde{\Delta}^t \end{bmatrix}^\top$ (see Appendix A.1). Augmented dynamics, control, and measurement matrices $A_{\mathrm{aug}}$, $B_{\mathrm{aug}}$, $C_{\mathrm{aug}}$, and the noise vector $\eta_{\mathrm{aug}}$ are expressed in terms of $D$ and $E$. The feedforward and feedback energy costs are expressed as the LQG state and control costs respectively (Eq 12,

derived in Appendix A.2), where $Q = \begin{bmatrix} 0 & 0 \\ 0 & W_{\mathrm{f}} \end{bmatrix}$, and $R = W_{\mathrm{b}}$.

$$z^t = A_{\mathrm{aug}}\, z^{t-1} + B_{\mathrm{aug}}\, p^t + \eta_{\mathrm{aug}}^{t-1} \tag{10}$$

$$\tilde{\Delta}^t = C_{\mathrm{aug}}\, z^t \tag{11}$$

$$\min_p \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \left( \underbrace{{z^t}^\top Q\, z^t}_{\text{state cost}} + \underbrace{{p^t}^\top R\, p^t}_{\text{control cost}} \right) \tag{12}$$

Note that the inference cost is not explicitly added in the LQG objective function (Eq 12). However, using the *separation principle* [34], we show in Appendix A.3 that for a fixed $D$ and $E$ the LQG solution *automatically* minimizes the inference cost as well. The separation principle states that at each instant the observer needs to make an optimal estimate of the world state to form the optimal control. Furthermore, we show in Appendix A.4–A.6 that the LQG solution also provides the optimal $F, G, H, L$, and $\mathrm{Cost}_{\mathrm{tot}}$ in terms of $D$ and $E$. These are denoted as $F'$, $G'$, $H'$, $L'$, and $\mathrm{Cost}'_{\mathrm{tot}}$ respectively. Finally, we numerically optimize $\mathrm{Cost}'_{\mathrm{tot}}$ with respect to $D$ and $E$ (as in Eq 9).

## 5 Results

In this paper we analyze our system for a one-dimensional world state to gain precise mathematical insight into the core computational problem. One useful perspective is that the feedback signal functions as a kind of self-control for the inference system. As we formulated the problem, this control has its own "action cost," and this allows us to use the known solutions for controllable systems to identify the optimal feedback. The control gain $L$ is therefore a fundamental parameter, as it indicates whether it is best to send a prediction ($L \neq 0$) or not ($L = 0$). The control gain always takes the opposite sign as the product of prediction-multiplier and observation-multiplier, ensuring that the prediction is subtracted from the observation, which thereby reduces the feedforward cost. However, when the optimal control gain is 0, the prediction-multiplier $D$ also becomes 0 so feedback noise does not corrupt the observation if no prediction is sent. Fig 2 shows how the optimal control gain, prediction-multiplier $D$ and observation-multiplier $E$ change with different parameters that define the constraints involved. We observe that there are cases where feedback and feedforward messages are useful, and others where messages are harmful.

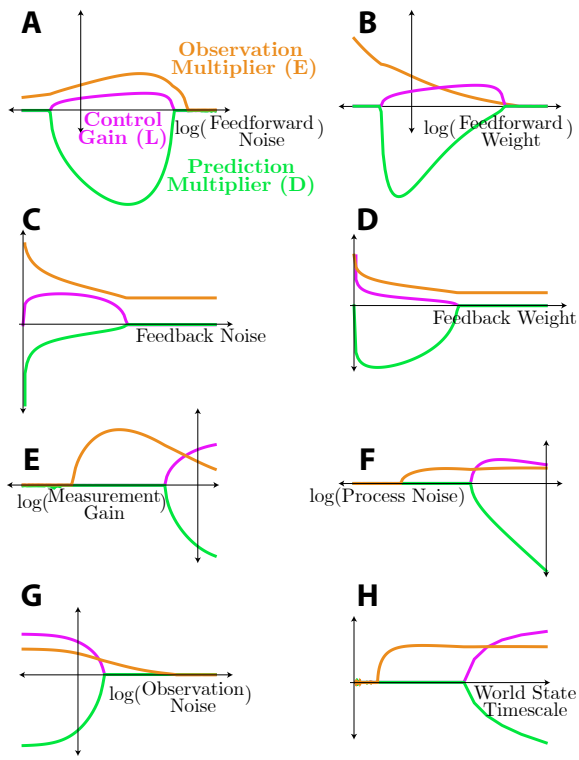


Figure 2: *The best use of predictions and observations depends on the world state dynamics and channels' noise costs. For tolerable feedback noise cost, as we increase either the* ***A****: feedforward noise, or the* ***B****: feedforward weight, the optimal strategy transitions from sending only feedforward, to predictive coding, back to only feedforward, and finally to no messages. For moderate feedforward noise cost, as we increase either the* ***C****: feedback noise, or the* ***D****: feedback weight, the optimal strategy transitions from predictive coding to feedforward messages only. When we increase either the* ***E****: measurement gain, or the* ***F****: process noise, or the* ***H****: world state timescale, the optimal strategy transitions from silence, to only feedforward messages, to predictive coding.* ***G****: Increasing the observation noise leads to the opposite sequence.*

### 5.1 Conditions for the Messages to be Useful

If the costs and noise are high, then it is not useful to send messages forward. To understand this quantitatively, we consider what happens without feedback ($D = 0$). In

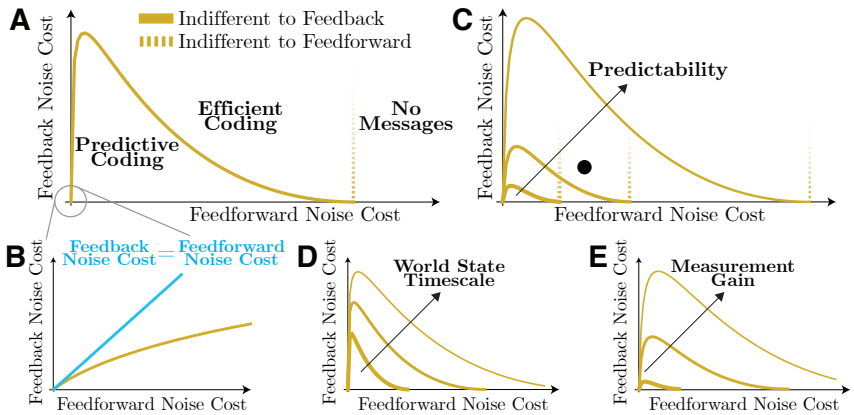


Figure 3: *Boundary curves divide the space of feedback and feedforward noise costs into regions of categorically different optimal strategies.* ***A****: Predictive coding is favored below the solid yellow line, feedforward efficient coding is favored above it, and no messages are favored to the right of the dashed line.* ***B****: The boundary curve determining whether feedback is useful lies below the line (cyan) along which the feedback and feedforward noise costs are equal, implying that for predictions to be useful the feedback channel must be cheaper and/or less noisy than the feedforward channel.* ***C****: With increasing predictability (thinner lines) of the world to be inferred, boundaries between coding strategies shift upwards and rightwards. As the predictability increases for a fixed value of channel parameters (dot), the optimal strategy transitions from sending no messages, to sending only feedforward messages, to sending and subtracting predictive feedback messages.* ***D****: An example of the shift in boundary curve with increasing predictability is shown for different world state timescales. As the timescale increases, the memory of the world state increases along with the SNR at the observation level, thereby increasing the predictability.* ***E****: A similar example is shown for different measurement gains. As the measurement gain increases, the observation SNR increases, thereby increasing the predictability.*

the absence of feedback, feedforward messages are worth its cost when the optimal observation multiplier $E$ at $D = 0$ is non-zero. We know that the derivative of the total cost at the optimal non-zero $E$, when it exists, should be 0. We set this derivative to 0 and then use Descartes' rule of signs to find when such a non-zero root exists. This yields the condition Eq 13. Sending even an infinitesimal feedforward message about the observation to the brain is worth its feedforward energy cost if and only if

$$(1 - A^2)\,\frac{U_{\mathrm{fn}}}{U_{\mathrm{s}}} < \frac{1}{1 + \frac{1}{\mathrm{SNR}_{\mathrm{o}}}} \tag{13}$$

where $U_{\mathrm{fn}} = W_{\mathrm{f}}\sigma_{\mathrm{f}}^2$ is the feed**f**orward **n**oise cost, i.e. the cost incurred by the noise variance in feedforward messages. $U_{\mathrm{s}} = \frac{\sigma_{\mathrm{p}}^2}{(1-A^2)}$ is the steady state signal power, and $\mathrm{SNR}_{\mathrm{o}} = \frac{\sigma_{\mathrm{p}}^2}{(1-A^2)}\frac{C^2}{\sigma_{\mathrm{o}}^2}$ is the observation Signal to Noise Ratio. When the left and right hand sides of Eq 13 are equal, the optimal strategy is indifferent between sending feedforward messages or not.

Next we consider the case where feedback is allowed. If the optimal prediction-multiplier is zero, then it is the same as no feedback, so by assumption the optimal observation-multiplier is non-zero if and only if feedforward is useful (Eq 13 holds). If the optimal prediction-multiplier is non-zero, then naturally the feedforward evidence must also be non-zero to have any signal worth predicting. Furthermore, when sending feedback is optimal, it must be worth sending some feedforward messages even if feedback is cut off. These purely feedforward messages might need to be smaller — as small as the residuals when we had feedback — but sending at least some information is evidently worth the cost, since it was worth paying that feedforward cost even when noisy feedback corrupted the inference. Thus, mathematically, feedforward messages are useful if and only if Eq 13 is true, both with and without feedback. Eq 13 implies that feedforward messages are worth their energy cost when either the observations are very reliable (high $\mathrm{SNR}_{\mathrm{o}}$), or when the relative cost of sending an arbitrarily small feedforward message is low (small $U_{\mathrm{fn}}/U_{\mathrm{s}}$).

Similarly, as shown in Appendix A.7, we find that feedback is valuable when

$$U_{\mathrm{bn}} < \Phi(U_{\mathrm{fn}}, A, \sigma_{\mathrm{p}}^2, \frac{C^2}{\sigma_{\mathrm{o}}^2}), \tag{14}$$

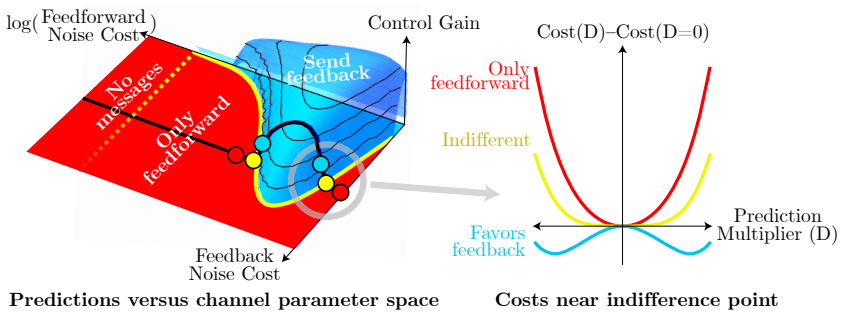


Figure 4: *Phase transition in the value of feedback.* Left*: Optimal control gain varies with the feedback and feedforward noise costs. As the feedforward noise cost increases (black line), the optimal strategy transitions non-monotonically from sending only feedforward messages, to predictive coding with suppressive feedback, back to sending only feedforward messages, and eventually to sending no messages. The yellow curve on the panel is the phase transition for feedback: the locus of points where the optimal strategy is indifferent to sending feedback (green area) or not (red area). Near the critical point (colored circles), the loss function goes through a phase transition [35], seen at* Right*: The cost function has a minimum at Prediction-multiplier $D = 0$ for low feedforward noise cost, favoring no feedback (red). But as the feedforward noise cost increases from low to moderate values, this extremum switches to a maximum, with nearby minima that favor feedback with nonzero $D$ (green). Right at the transition point, the system becomes indifferent to feedback (yellow). Another phase transition occurs in the opposite direction as the feedforward noise cost increases further from moderate to high.*

where $\Phi$ is a complicated function that involves solving for a quartic equation. Below in section 5.2 we will interpret the behavior of this function. $U_{\mathrm{bn}} = W_{\mathrm{b}}\sigma_{\mathrm{b}}^2$ is the feed**b**ack **n**oise cost, i.e. the cost incurred by the noise variance in feedback messages. In Fig 3A, $\Phi$ (solid yellow line) is plotted as a function of the feedforward noise cost. Where Eq 14 holds as an equality (feedback and feedforward noise costs that correspond to points on the yellow line), an optimal observer is indifferent about sending feed*back* messages. Sending feedback is useful below that line, and harmful above it.

As stated above, when sending feedback is optimal (Eq 14 satisfied), it must be worth sending feedforward messages even if feedback is cut off (Eq 13 holds true). This is indeed observed in Fig 3A, where the dashed line is the locus of points where the optimal strategy is indifferent to sending feed*forward* messages: to the left, sending feedforward messages is useful, and to the right, it is not. As expected, the region where feedback is useful is also where feedforward messages are useful. Also, for very small feedback noise cost, the only case where sending feedback becomes harmful is when sending even an infinitesimal feedforward message is not useful either. As a result, the feedback indifference curve ends exactly at the precise feedforward noise cost where the optimal strategy is indifferent to sending feedforward messages. Fig 3B shows a magnified version where feedforward and feedback products are near zero. At the origin, where the products equal zero, feedforward and feedback messages are free and/or noiseless. A noiseless message can be arbitrarily small and still convey information, so the cost can be arbitrarily small. In this case, there is nothing to gain or lose by subtracting predictions, so even free feedback makes no difference to costs or inferences. Thus the feedback indifference curve starts at the origin. Fig 3B also shows that for feedback to be useful, the feedback channel should be cheaper and/or less noisy than the feedforward channel. Note that this is true regardless of sensory parameter values (world state timescale, measurement gain, etc.).

The exact optimal multipliers for the optimization in Eq 9 depends on all system parameters (Eqs 1–7): feedforward and feedback noise variances and weights, the measurement gain, world state timescale, the process and observation noise variances. However, from Eqs 13–14, we see that the optimal categorical *strategy* — which multipliers are nonzero — depends only on certain combinations of these parameters. For example, the phase transition depends on the measurement gain and observation noise only through their ratio, ($\mathrm{SNR_o}$), which determines the quality of observations. Specifically, in Appendix A.8–A.9, we prove that the feedforward (feedback) noise variance and feedforward (feedback) weights affect the phase transition only through their product, the feedforward (feedback) noise cost. The core reason is the following: If any (possibly suboptimal) multiplier is nonzero for one value of noise and cost weight, then there is a whole *family* of nonzero multipliers for noises and cost weights that are scaled inversely, keeping their product (the noise cost) fixed — where every member of this family has *identical* costs and signals. Thus if we optimize the multipliers for one noise and weight, and find it is nonzero, then the optimal multipliers will still be nonzero for this

entire family of inversely scaled noises and weights. In other words, whether feedforward or feedback is useful depends on the noise costs.

### 5.2 Transitions in the Optimal Strategy

Having seen above how there are categorically different optimal strategies for computationally constrained inference, we now examine how individual parameters move the system between these strategies. We broadly group parameters into three categories: feedforward, feedback, and sensory.

**Feedforward parameters:** We first consider the case of low to moderate feedback noise cost. Fig 4 shows the transition between optimal strategies as a function of the feedforward and feedback noise costs. The black line on Fig 4 left traces the transition as we increase the feedforward noise cost for a fixed feedback noise cost. For low values of the feedforward noise cost, feedforward messages are almost free, so the system does not save appreciable resources by sending predictions; even worse, noisy feedback would corrupt the signal with noise. Thus for low feedforward noise costs it is optimal to send no predictions. As this cost increases, at some point it becomes equally valuable to send or withhold a feedback message. For higher feedforward noise costs, we cross the point of indifference, to where feedforward messages are important yet their channel is not economical by itself. Predictive feedback then becomes preferable, even when accounting for additional feedback noise.

As feedforward noise cost increases, reliable transmission through the channel becomes less affordable. As a consequence, the inference degrades. Upon increasing the feedforward noise cost beyond a certain point, the inference becomes so poor that it is no longer possible to make a good prediction worthy of the cost it incurs. Thus the system resorts to sending only feedforward messages. For similar reasons, as the world becomes more predictable (Fig 3C), there is a wider range of feedforward and feedback noise costs for which sending predictions is optimal. The predictability can be increased either by lengthening the world state timescale, or by enhancing the observation SNR, enabling better inferences and thereby better predictions. Figs 3D–E demonstrates this for different world state timescales and measurement gains respectively. However, for large enough feedforward noise cost, it is best to remain silent as sending even feedforward messages alone becomes too expensive/noisy.

Fig 2A and B show how the optimal control gain $L$, prediction-multiplier $D$ and observation-multiplier $E$ change with feedforward noise and feedforward weight respectively. As only the product of noise variance and weight determines the transition in optimal strategy, we observe the same trend as either one of them increases. However, these parameters actually exhibit different effects away from the points of transitions. For example, when feedforward noise is close to $0$, the system can save costs by attentuating the signal ($E \to 0$) and still beat the negligible feedforward noise. In contrast, in the extreme case that the feedforward messages are nearly free ($W_{\mathrm{f}} \to 0$), the observation-multiplier increases arbitrarily ($E \to \infty$) to dominate any feedforward noise, improving the inference at no cost. These limiting cases provide helpful intuitions. Naturally, the challenging regime is in the middle, where there are real tradeoffs to make, and where feedback becomes relevant.

For the limiting case where feedback noise cost is extremely high, the feedback channel is too expensive/noisy to be used. Hence, as we slowly increase the feedforward noise cost from zero, we start by sending just the feedforward messages until the feedforward channel is too expensive/noisy, at which point we no longer send any messages.

**Feedback parameters:** Analogous to the analysis described for feedforward parameters, here we discuss how the optimal strategy changes as we increase the feedback noise cost, for different levels of feedforward noise cost. We first consider the interesting case of low to moderate feedfoward noise cost. For low values of feedback noise cost we are in the regime where feedback is cheap and/or noiseless, and hence it is beneficial to send both feedback and feedforward messages. But when the feedback noise cost becomes too high, sending predictions become too expensive/noisy, so it is best send only feedforward messages. This is shown in Fig 2C, D. Analogous to the feedforward case, although the same transition in optimal strategy is observed as we increase either the feedback noise or weight, these parameters actually exhibit different effects away from that transition. For example, $L \to 0$ when feedback noise is close to $0$, and $L \to -\infty$ when feedback weight is close to $0$. Finally, in the case of extremely high feedfoward noise cost, it is not worth sending even an infinitesimal feedforward message, so the optimal strategy is to not send any messages.

**Sensory parameters:** Fig 3C shows the boundary between strategies for three systems with different levels of predictability. For a fixed value of feedback and feedforward noise costs (dot), the optimal

strategy changes with the predictability. With unpredictable dynamics (thick curve), the dot lies in the region where sending no messages is optimal, since the feedforward channel is relatively poor (right of the thick dashed line). For a slightly higher predictability, the dot lies within the region where feedforward-only messages are optimal (above medium curve, but left of the corresponding dashed line). And for even higher predictability, the dot lies in the region where predictive coding is optimal (below the thin curve). Hence, as the predictability increases, there is a transition from sending no messages, to sending just feedforward, to sending both feedback and feedforward messages.

Similarly, if either the measurement gain or the process noise increases, it would yield a higher observation SNR, which would then improve the inference and thereby predictability. Figs 2E–F show the resultant transitions from no messages, to feedforward-only messages, then to predictive coding, with increasing measurement gain or process noise. Increasing observation noise has the opposite effect, making observations less reliable and thereby reducing the predictability. This is shown in Fig 2G. Finally, increasing the world state timescale increases the predictability since it increases both the memory of the system, and the observation SNR. Fig 2H reveals the familiar sequence of transitions in strategy, supporting our core intuitions about when feedback is valuable.

### 5.3 Anatomical and functional differences can lead to different channel noise costs

We characterized the properties of feedforward/feedback channel in terms of the channel noise and weight of energetic cost, and showed that the phase transition depends on the feedforward/feedback channel noise costs. But why would channels have different noise costs? Biologically, asymmetries in noise costs can arise from anatomical or functional constraints. For example, Dale's law [36] states that each neuron makes only excitatory or inhibitory synapses onto its targets. However, long-range feedback connections in the brain are excitatory [37, 38]. This implies that to subtract predictions using long-range inhibitory feedback, the signals must pass through two neurons: long-range excitation that is inverted by short-range inhibition. In contrast, feedforward signals can proceed directly to their targets. The sign inversion neurons induce additional noise in the feedback pathway. Below we review biological evidence that feedback and feedforward channels have distinct anatomical and functional properties.

To see how channels with different anatomical structures can yield different channel noise costs, contrast two versions of a channel with equal numbers of neurons, all having identical energetic weights and noise variances. One version has a single population of neurons that linearly encodes the message to be sent; the channel output simply decodes this population's representation. We call this channel the non-sequential case (Appendix Fig 9A). The second version is the sequential case (Appendix Fig 9B), and uses two populations sequentially connected in series, one after the other, like long-range inhibition comes from long-range excitation followed by short-range inhibition. The first population linearly encodes the message to be sent, which is then decoded and then encoded by the second population; the channel's output is a decoding of the second population.

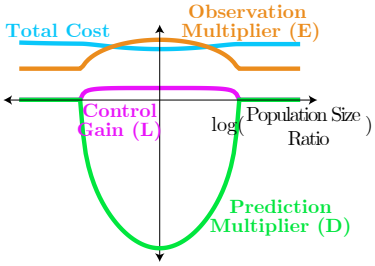


Figure 5: *Allocation of neurons between two sequentially connected populations of the feedback channel affects when sending feedback is useful.*

Appendix A.10.1 shows that for the non-sequential case with one population, as we scale up the number of neurons, the channel's weight scales up and the noise scales down inversely. Their product, the channel noise cost, does not change with the number of neurons, yielding the same optimal categorical strategy for a given noise and weight per neuron. For the sequential case with two populations, we show in Appendix A.10.2 that the channel noise cost again does not depend on the total number of neurons, but does depend on the ratio of population sizes. Furthermore, the total cost changes non-monotonically with population size ratio for the sequential case, and increases monotonically with total number of neurons for both sequential and non-sequential cases. Fig 5 shows the optimal strategy changing with the population size ratio of the feedback channel, for a fixed total number of neurons. We see that the feedback is not useful when the ratio is strongly skewed (away from 1).

From the above example, we saw that channels with different anatomical structures can yield different noise costs, and thereby different optimal strategies. Biologically, there exist differences between the

anatomy or functional properties of the feedback and feedforward channels. Anatomically, feedback projections generally outnumber feedforward ones by 2:1, but the feedforward pathways compensate with greater weight, dominating at short cortical distances and leveling off at longer distances [39]. Functionally, the sparser, stronger feedforward channels therefore have less ability to average away the noise, leading to a relatively higher amount of metabolically expensive activity caused by noise, and thus to a higher noise cost. This could establish conditions under which feedback is useful.

## 6 Discussion

In this paper, we define a new class of dynamic optimization tasks that reflect essential biological constraints on inference in the brain, by including cost and noise for each recurrently connected computational element. We solve this optimization problem by modeling inference as a control problem with prediction as self-control. The resultant optimization provides nontrivial predictions for when we expect suppressive feedback as a function of biological constraints, computational costs, and world dynamics.

Predictive coding is a promising theory for brain computations [40–44], and a variety of experimental studies have indeed shown predictive suppression effects in neural responses [45–52]. However, neural responses vary in how much they are suppressed depending on the context and SNR. Variants aiming to explain these effects all neglect the computational costs associated with predictive coding, even though those costs were part of the motivation for saving metabolic costs in the first place.

Past work described temporal decorrelation as optimal encoding strategy, subtracting a slow, predictable component from the recent measured signal, implementing a form of predictive coding [15, 53]. Like our results, they too observed that this strategy changes with the SNR. Efficient coding shows that mean responses are fully subtracted at all SNRs, but with different temporal kernels [15, 53]. Later studies of predictive coding observed that for low SNR, temporal decorrelation is no longer the best strategy [18]. Our work generalizes these results to account not only for feedforward noise and costs in sensory coding, but also for noise and costs in recurrent computation. In particular, the nonmonotonic dependence of optimal feedback on feedforward noise costs is a novel consequence of our framework with no analog in previous studies. We could also imagine generalizing our results to adaptive strategies, like [31], where optimal adaptation could implement our predicted changes in the optimal suppression as sensory statistics change. Conversely, by considering the computational costs of adaptation itself (and not just the activity resulting from adapted computation), we may find conditions where adapting is not beneficial.

Although we contend that it is costly to send predictions, if these predictions are based on inferences that are already being computed for a task, then is it really an extra cost to send them as predictions? Yes, because distinct neurons are used to send feedforward and feedback signals [54], so the brain might pay costs twice for the same inference. When beneficial, the brain could avoid the duplicate feedback costs by silencing the feedback neurons, and send its inferences to higher brain areas only through its feedforward neurons. A related question arises for feedback suppression, which is presumably implemented through inhibitory neurons. What is the benefit of turning on an inhibitory neuron just in order to turn off an excitatory neuron? Excitatory neurons outnumber inhibitory neurons by 4:1 [55], so few inhibitory neurons can suppress many excitatory ones, amounting to a potentially substantial savings. These biological constraints could be used in future elaborations of computationally constrained inference models.

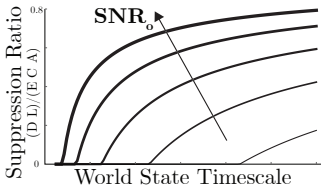


Figure 6: *World timescale and observation SNR affect predictive response suppression. SNR increases with line thickness.*

Our mathematical system's crucial theoretical parameters relate to biological quantities in three ways: as context-dependent parameters (time constant of the stimulus dynamics, sensory SNR), as neural parameters that could potentially be manipulated experimentally (feedforward and feedback SNR), and as developmentally-fixed parameters of the system (feedforward and feedback architecture). Testing our predicted dependence on these three types of parameters requires different considerations. First, the context-dependent parameters are by far the easiest to manipulate experimentally. One can control the stimulus to adjust the observations' SNR and the world state timescale, and measure whether any response suppression is modulated by these controlled parameters as shown in Fig 6. Second, the internal computational parameters like neural noise may be controllable through stimulus changes or experimental techniques

of causal manipulation. For example, electrical or optogenetic stimulation could directly inject neural variability. Another approach could be based on the fact that neural noise variance tends to increase with firing rate, so one could test our predictions about computational noise by elevating firing rates. Such methods could include providing a background sensory stimulus, or direct neural stimulation to increase baseline activity. In any case, it would be important to apply either natural manipulations or chronic unnatural ones to give the brain enough time and experience to optimize its computations. And finally, for the developmentally fixed parameters like architecture, we cannot easily alter the system experimentally, and it may be unreasonable to assume that properties in a real biological system can be unambiguously mapped to any particular value in our abstract theoretical system. However, we can compare *between* brain systems or species with architectural difference, and test whether functional properties covary with those differences as predicted by our theory. Thus it may be fruitful to compare between brain areas with different architectures, such as 3-layer paleocortex (e.g. olfactory cortex) versus 6-layer neocortex (e.g. auditory or visual cortex), or within neocortical areas with different properties. It may also be fruitful to compare between organisms with different architectures (e.g. mammals versus reptiles). These architectures have different microcircuits with distinct feedforward and feedback projection neurons. For example, feedback in different visual and motor cortical systems tends to target different inhibitory cells [56] which have different noise levels [57, 58]. We predict that systems with more feedback noise (e.g. SOM versus VIP) would have less predictive suppression, as observed experimentally [56].

Our theory of *inference as control* could also be used in designing energy-efficient systems in other domains where the estimation takes place under constraints. For example, low-power applications of inference and control, such as drones or remote sensing, often place constraints on dynamic range and noise (e.g. quantization noise) for all signals, and could thus benefit from optimizations based on our model.

**Limitations.** There are several interesting avenues for generalizing our theory. The proposed *inference as control* method can be extended to the more general case of observer taking external actions in addition to the internal predictions, in order to maximize external rewards while minimizing both external action costs (e.g. movements) and internal computational costs. This could be modeled as a typical LQG control problem for external variables by including appropriate entries in matrices $Q$, $R$, and $B_{\mathrm{aug}}$ in addition to the internal controls we use for predictive coding. Though our results are limited to one-dimensional linear dynamics with uncorrelated noise, our future work will explore these computational constraints in more complex multivariate and nonlinear tasks. Accounting for graph-structured inferences [59] and controllability [60] may provide additional constraints on the brain's inference processes, and additional predictions for experiments.

Residuals between predictions and observations are useful not just for improving inferences, but also for learning, which this paper does not address. In principle, a multivariate generalization could explain such computations as hierarchical inferences or adaptation where slower changes are also subject to prediction. However, unlike adaptation, learning occurs out of equilibrium with the environment, and accounting for transient responses under nonstationary statistics will require an extension to our theory. Overall, this work points a way towards expanding theories of predictive coding and efficient coding to unify inference, learning, and control in biological systems.

**Acknowledgments.** The authors thank Itzel Olivos Castillo, Ann Hermundstad, Krešimir Josić, and Paul Schrater for useful discussions. This work was supported in part by NSF CAREER grant 1552868, the McNair Foundation, and AFOSR grant FA9550-21-1-0422 in the Cognitive and Computational Neuroscience program.

## References

[1] Horace B Barlow et al. Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1:217–234, 1961.

[2] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999.

[3] Hilbert J Kappen, Vicenç Gómez, and Manfred Opper. Optimal control as a graphical model inference problem. *Machine learning*, 87(2):159–182, 2012.

[4] Hermann Von Helmholtz. *Handbuch der physiologischen Optik: mit 213 in den Text eingedruckten Holzschnitten und 11 Tafeln*, volume 9. Voss, 1867.

[5] David C Knill and Alexandre Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719, 2004.

[6] Wei Ji Ma and Mehrdad Jazayeri. Neural coding of uncertainty and probability. *Annual review of neuroscience*, 37:205–220, 2014.

[7] Wei Ji Ma, Jeffrey M Beck, and Alexandre Pouget. Spiking networks for bayesian inference and choice. *Current opinion in neurobiology*, 18(2):217–222, 2008.

[8] Stephen W Kuffler. Discharge patterns and functional organization of mammalian retina. *Journal of neurophysiology*, 16(1):37–68, 1953.

[9] Haldan Keffer Hartline. The response of single optic nerve fibers of the vertebrate eye to illumination of the retina. *American Journal of Physiology-Legacy Content*, 121(2):400–415, 1938.

[10] Peter Kok, Janneke FM Jehee, and Floris P De Lange. Less is more: expectation sharpens representations in the primary visual cortex. *Neuron*, 75(2):265–270, 2012.

[11] Christopher Summerfield and Floris P De Lange. Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience*, 15(11):745–756, 2014.

[12] J Hans van Hateren. Real and optimal neural images in early vision. *Nature*, 360(6399):68–70, 1992.

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

[14] Simon Laughlin. A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung c*, 36(9-10):910–912, 1981.

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

[16] Michael W Spratling. Predictive coding as a model of response properties in cortical area v1. *Journal of neuroscience*, 30(9):3531–3543, 2010.

[17] Andre M Bastos, W Martin Usrey, Rick A Adams, George R Mangun, Pascal Fries, and Karl J Friston. Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711, 2012.

[18] Matthew Chalk, Olivier Marre, and Gašper Tkačik. Toward a unified theory of efficient, predictive, and sparse coding. *Proceedings of the National Academy of Sciences*, 115(1):186–191, 2018.

[19] Karl Friston and Stefan Kiebel. Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1211–1221, 2009.

[20] Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138, 2010.

[21] Yanping Huang and Rajesh PN Rao. Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(5):580–593, 2011.

[22] Robert Rosenbaum. On the relationship between predictive coding and backpropagation. *arXiv preprint arXiv:2106.13082*, 2021.

[23] Haiguang Wen, Kuan Han, Junxing Shi, Yizhen Zhang, Eugenio Culurciello, and Zhongming Liu. Deep predictive coding network for object recognition. In *International Conference on Machine Learning*, pages 5266–5275. PMLR, 2018.

[24] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

[25] Abdullahi Ali, Nasir Ahmad, Elgar de Groot, Marcel AJ van Gerven, and Tim C Kietzmann. Predictive coding is a consequence of energy efficiency in recurrent neural networks. *bioRxiv*, 2021.

[26] Kuan Han, Haiguang Wen, Yizhen Zhang, Di Fu, Eugenio Culurciello, and Zhongming Liu. Deep predictive coding network with local recurrent processing for object recognition. *arXiv preprint arXiv:1805.07526*, 2018.

[27] Michael W Spratling. A hierarchical predictive coding model of object recognition in natural images. *Cognitive computation*, 9(2):151–167, 2017.

[28] Michael W Spratling. A review of predictive coding algorithms. *Brain and cognition*, 112:92–97, 2017.

[29] Rakesh Chalasani and Jose C Principe. Deep predictive coding networks. *arXiv preprint arXiv:1301.3541*, 2013.

[30] Il Memming Park and Jonathan W Pillow. Bayesian efficient coding. *BioRxiv*, page 178418, 2017.

[31] Wiktor F Młynarski and Ann M Hermundstad. Adaptive coding for dynamic sensory inference. *Elife*, 7:e32055, 2018.

[32] Laurence Aitchison and Máté Lengyel. With or without you: predictive coding and bayesian inference in the brain. *Current opinion in neurobiology*, 46:219–227, 2017.

[33] Rudolf Emil Kalman et al. Contributions to the theory of optimal control. *Bol. soc. mat. mexicana*, 5(2):102–119, 1960.

[34] Mark Davis. *Stochastic modelling and control*. Springer Science & Business Media, 2013.

[35] Lev Davidovich Landau and Evgenii Mikhailovich Lifshitz. *Course of theoretical physics*. Elsevier, 2013.

[36] Piergiorgio Strata, Robin Harvey, et al. Dale's principle. *Brain research bulletin*, 50(5):349–350, 1999.

[37] Peter Somogyi, Gabor Tamas, Rafael Lujan, and Eberhard H Buhl. Salient features of synaptic organisation in the cerebral cortex. *Brain research reviews*, 26(2-3):113–135, 1998.

[38] György Buzsáki. Feed-forward inhibition in the hippocampal formation. *Progress in neurobiology*, 22(2):131–153, 1984.

[39] Nikola T Markov, Julien Vezoli, Pascal Chameau, Arnaud Falchier, René Quilodran, Cyril Huissoud, Camille Lamy, Pierre Misery, Pascale Giroud, Shimon Ullman, et al. Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. *Journal of Comparative Neurology*, 522(1):225–259, 2014.

[40] Karl Friston. Does predictive coding have a future? *Nature neuroscience*, 21(8):1019–1021, 2018.

[41] Xiao Liu, Xiaolong Zou, Zilong Ji, Gengshuo Tian, Yuanyuan Mi, Tiejun Huang, Michael Kwok Yee Wong, and Si Wu. Push-pull feedback implements hierarchical information retrieval efficiently. 2019.

[42] Janneke FM Jehee, Constantin Rothkopf, Jeffrey M Beck, and Dana H Ballard. Learning receptive fields using predictive feedback. *Journal of Physiology-Paris*, 100(1-3):125–132, 2006.

[43] Samuel J Gershman. What does the free energy principle tell us about the brain? *arXiv preprint arXiv:1901.07945*, 2019.

[44] Andreja Bubic, D Yves Von Cramon, and Ricarda I Schubotz. Prediction, cognition and the brain. *Frontiers in human neuroscience*, 4:25, 2010.

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

[46] Jan Homann, Sue Ann Koay, Alistair M Glidden, David W Tank, and Michael J Berry. Predictive coding of novel versus familiar stimuli in the primary visual cortex. *BioRxiv*, page 197608, 2017.

[47] Scott O Murray, Daniel Kersten, Bruno A Olshausen, Paul Schrater, and David L Woods. Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences*, 99(23):15164–15169, 2002.

[48] Ryszard Auksztulewicz and Karl Friston. Repetition suppression and its contextual determinants in predictive coding. *cortex*, 80:125–140, 2016.

[49] Tai Sing Lee and David Mumford. Hierarchical bayesian inference in the visual cortex. *JOSA A*, 20(7):1434–1448, 2003.

[50] Rajesh PN Rao and Dana H Ballard. Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural computation*, 9(4):721–763, 1997.

[51] David Mumford. On the computational architecture of the neocortex. *Biological cybernetics*, 66(3):241–251, 1992.

[52] Marta I Garrido, James M Kilner, Klaas E Stephan, and Karl J Friston. The mismatch negativity: a review of underlying mechanisms. *Clinical neurophysiology*, 120(3):453–463, 2009.

[53] J Hans Van Hateren. Spatiotemporal contrast sensitivity of early vision. *Vision research*, 33(2):257–267, 1993.

[54] João D Semedo, Anna I Jasper, Amin Zandvakili, Aravind Krishna, Amir Aschner, Christian K Machens, Adam Kohn, and Byron M Yu. Feedforward and feedback interactions between visual cortical areas use different population activity patterns. *Nature communications*, 13(1):1–14, 2022.

[55] Ryan A Shewcraft, Heather L Dean, Margaret M Fabiszak, Maureen A Hagan, Yan T Wong, and Bijan Pesaran. Excitatory/inhibitory responses shape coherent neuronal dynamics driven by optogenetic stimulation in the primate brain. *Journal of Neuroscience*, 40(10):2056–2068, 2020.

[56] Shan Shen, Xiaolong Jiang, Federico Scala, Jiakun Fu, Paul Fahey, Dimitry Kobak, Zhenghuan Tan, Jacob Reimer, Fabian Sinz, and Andreas S Tolias. Distinct organization of two cortico-cortical feedback pathways. *bioRxiv*, 2020.

[57] Wen-pei Ma, Bao-hua Liu, Ya-tang Li, Z Josh Huang, Li I Zhang, and Huizhong W Tao. Visual representations by cortical somatostatin inhibitory neurons—selective but with weak and delayed responses. *Journal of Neuroscience*, 30(43):14371–14379, 2010.

[58] Daniel J Millman, Gabriel Koch Ocker, Shiella Caldejon, Josh D Larkin, Eric Kenji Lee, Jennifer Luviano, Chelsea Nayan, Thuyanh V Nguyen, Kat North, Sam Seid, et al. Vip interneurons in mouse primary visual cortex selectively enhance responses to weak but specific stimuli. *Elife*, 9:e55130, 2020.

[59] Xaq Pitkow and Dora E Angelaki. Inference in the brain: statistics flowing in redundant population codes. *Neuron*, 94(5):943–953, 2017.

[60] Shi Gu, Fabio Pasqualetti, Matthew Cieslak, Qawi K Telesford, B Yu Alfred, Ari E Kahn, John D Medaglia, Jean M Vettel, Michael B Miller, Scott T Grafton, et al. Controllability of structural brain networks. *Nature communications*, 6(1):1–10, 2015.

## Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This work does not present any foreseeable societal consequence.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

# A Appendix

## A.1 Mapping the inference tracking dynamics onto an LQG control problem

Substituting Eqs 1- 2, and 4-5 in Eq 6.

$$\begin{aligned}
\tilde{\Delta}^t =& \Delta^t + \eta_{\mathrm{f}}^t \\
=& (D\,\tilde{p}^t + E\,o^t) + \eta_{\mathrm{f}}^t \\
=& D\,\tilde{p}^t + E\,(C\,x^t + \eta_{\mathrm{o}}^t) + \eta_{\mathrm{f}}^t \\
=& D\,\tilde{p}^t + E\,C\,x^t + E\,\eta_{\mathrm{o}}^t + \eta_{\mathrm{f}}^t \\
=& D\,\tilde{p}^t + E\,C\,(A\,x^{t-1} + \eta_{\mathrm{p}}^t) + E\,\eta_{\mathrm{o}}^t + \eta_{\mathrm{f}}^t \\
=& D\,\tilde{p}^t + E\,C\,A\,x^{t-1} + E\,C\,\eta_{\mathrm{p}}^t + E\,\eta_{\mathrm{o}}^t + \eta_{\mathrm{f}}^t \\
=& D\,(p^t + \eta_{\mathrm{b}}^t) + E\,C\,A\,x^{t-1} + E\,C\,\eta_{\mathrm{p}}^t + E\,\eta_{\mathrm{o}}^t + \eta_{\mathrm{f}}^t \\
=& E\,C\,A\,x^{t-1} + D\,p^t + E\,C\,\eta_{\mathrm{p}}^t + E\,\eta_{\mathrm{o}}^t + D\,\eta_{\mathrm{b}}^t + \eta_{\mathrm{f}}^t
\end{aligned} \tag{15}$$

Combining Eqs 1 and 15, we write the dynamics of augmented state $z^t = \begin{bmatrix} x^t & \tilde{\Delta}^t \end{bmatrix}^\top$ as in Eq 16. The noisy residual can now be written as a partial observation of the augmented state (Eq 17).

$$z^t = \begin{bmatrix} x^t \\ \tilde{\Delta}^t \end{bmatrix} = \overbrace{\begin{bmatrix} A & 0 \\ E\,C\,A & 0 \end{bmatrix}}^{A_{\mathrm{aug}}} \overbrace{\begin{bmatrix} x^{t-1} \\ \tilde{\Delta}^{t-1} \end{bmatrix}}^{z^{t-1}} + \overbrace{\begin{bmatrix} 0 \\ D \end{bmatrix}}^{B_{\mathrm{aug}}} p^t + \overbrace{\begin{bmatrix} I & 0 & 0 & 0 \\ E\,C & E & D & I \end{bmatrix} \begin{bmatrix} \eta_{\mathrm{p}}^t \\ \eta_{\mathrm{o}}^t \\ \eta_{\mathrm{b}}^t \\ \eta_{\mathrm{f}}^t \end{bmatrix}}^{\eta_{\mathrm{aug}}^{t-1}} \tag{16}$$

$$\tilde{\Delta}^t = \underbrace{\begin{bmatrix} 0 & I \end{bmatrix}}_{C_{\mathrm{aug}}} \underbrace{\begin{bmatrix} x^t \\ \tilde{\Delta}^t \end{bmatrix}}_{z^t} \tag{17}$$

Rewriting Eqs 16-17 using concise notation, we get the LQG dynamics equations where prediction $p$ is treated as the control.

$$\begin{aligned}
z^t =& A_{\mathrm{aug}}\,z^{t-1} + B_{\mathrm{aug}}\,p^t + \eta_{\mathrm{aug}}^{t-1} \\
\tilde{\Delta}^t =& C_{\mathrm{aug}}\,z^t
\end{aligned}$$

## A.2 Feedforward and feedback energy costs expressed as the LQG state and control costs respectively

We express the feedforward and feedback energy costs as the LQG state and control costs respectively. For the derivation, we introduce the matrices $Q = \begin{bmatrix} 0 & 0 \\ 0 & W_{\mathrm{f}} \end{bmatrix}$, and $R = W_{\mathrm{b}}$. We take the feedback noise $\eta_{\mathrm{b}}$ as i.i.d $\sim \mathcal{N}(0,\,\Sigma_{\mathrm{b}})$.

$$\begin{aligned}
\min_p \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\left(\tilde{\Delta}^{t\top} W_{\mathrm{f}}\,\tilde{\Delta}^t + \tilde{p}^{t\top} W_{\mathrm{b}}\,\tilde{p}^t\right) =& \min_p \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\left(\tilde{\Delta}^{t\top} W_{\mathrm{f}}\,\tilde{\Delta}^t + (p^t + \eta_{\mathrm{b}}^t)^\top W_{\mathrm{b}}\,(p^t + \eta_{\mathrm{b}}^t)\right) \\
=& \min_p \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\left(\tilde{\Delta}^{t\top} W_{\mathrm{f}}\,\tilde{\Delta}^t + p^{t\top} W_{\mathrm{b}}\,p^t + \eta_{\mathrm{b}}^{t\top} W_{\mathrm{b}}\,\eta_{\mathrm{b}}^t\right) \\
=& \min_p \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\left(\tilde{\Delta}^{t\top} W_{\mathrm{f}}\,\tilde{\Delta}^t + p^{t\top} W_{\mathrm{b}}\,p^t\right) + Tr(W_{\mathrm{b}}\,\Sigma_{\mathrm{b}}) \\
=& \min_p \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\left(z^{t\top} Q\,z^t + p^{t\top} R\,p^t\right) + Tr(W_{\mathrm{b}}\,\Sigma_{\mathrm{b}}) \\
=& \min_p \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\left(\underbrace{z^{t\top} Q\,z^t}_{\text{state cost}} + \underbrace{p^{t\top} R\,p^t}_{\text{control cost}}\right)
\end{aligned}$$

### A.3 Minimizing the total energy cost also minimizes the inference cost, for a fixed D& E

The LQG problem of interest is described by Eqs 10–12. The separation principle states that the solution for LQG problem includes an estimation part and a control part, both of which jointly form the optimal solution even if treated separately. The estimation part of LQG solution minimizes the average squared norm error between estimated state $\hat{z}$ and the true state $z$, as shown in the left side of equality in Eq 18. Note that at any time $t$, we get to observe all the evidence up until that time (i.e $\tilde{\Delta}^1$, $\tilde{\Delta}^2$,.., $\tilde{\Delta}^t$). This enables the simplification from Eq 18 to 19, because the optimal estimate $\hat{\tilde{\Delta}}^t$ given $\tilde{\Delta}^t$ is trivially $\tilde{\Delta}^t$ itself. Hence, from Eq 20, we conclude that although we define the LQG objective function to be just the total energy cost, the LQG solution that minimizes the total energy cost also minimizes the inference cost for a fixed $D$ and $E$.

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}(z^t - \hat{z}^t)^\top\ (z^t - \hat{z}^t) = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}(\begin{bmatrix} x^t \\ \tilde{\Delta}^t \end{bmatrix} - \begin{bmatrix} \hat{x}^t \\ \hat{\tilde{\Delta}}^t \end{bmatrix})^\top\ (\begin{bmatrix} x^t \\ \tilde{\Delta}^t \end{bmatrix} - \begin{bmatrix} \hat{x}^t \\ \hat{\tilde{\Delta}}^t \end{bmatrix}) \tag{18}$$

$$= \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}(x^t - \hat{x}^t)^\top\ (x^t - \hat{x}^t) \tag{19}$$

$$= \text{Cost}_{\text{inf}} \tag{20}$$

### A.4 LQG estimation solution to find $F$, $G$, and $H$ in terms of $D$ and $E$

The optimal estimate of the augmented state based on its noisy partial observations is given as

$$\begin{aligned} \hat{z}^t =& (I - K\ C_{\text{aug}})A_{\text{aug}}\ \hat{z}^{t-1} + K\ \tilde{\Delta}^t + (I - K\ C_{\text{aug}})B_{\text{aug}}\ p^t \\ K =& \check{\Sigma}\ C_{\text{aug}}^T(C_{\text{aug}}\ \check{\Sigma}\ C_{\text{aug}}^\top)^{-1}, \end{aligned} \tag{21}$$

where $\check{\Sigma}$ is the solution to a discrete algebraic Riccati equation (Eq 22), with $W$ as the covariance of noise vector $\eta_{\text{aug}}$.

$$\check{\Sigma} = A_{\text{aug}}\ \check{\Sigma}\ A_{\text{aug}}^\top + W - A_{\text{aug}}\ \check{\Sigma}\ C_{\text{aug}}^\top(C_{\text{aug}}\ \check{\Sigma}\ C_{\text{aug}}^\top)^{-1}C_{\text{aug}}\ \check{\Sigma}\ A_{\text{aug}}^\top \tag{22}$$

For brevity, we rewrite Eq 21 in terms of block matrics using the following substitutions,

$$\begin{aligned} F_{\text{aug}} =& (I - K\ C_{\text{aug}})A_{\text{aug}},\ H_{\text{aug}} = (I - K\ C_{\text{aug}})B_{\text{aug}} \\ \begin{bmatrix} \hat{x}^t \\ \tilde{\Delta}^t \end{bmatrix} =& \begin{bmatrix} (F_{\text{aug}})_{11} & (F_{\text{aug}})_{12} \\ (F_{\text{aug}})_{21} & (F_{\text{aug}})_{22} \end{bmatrix} \begin{bmatrix} \hat{x}^{t-1} \\ \tilde{\Delta}^{t-1} \end{bmatrix} + \begin{bmatrix} (K)_1 \\ (K)_2 \end{bmatrix} \tilde{\Delta}^t + \begin{bmatrix} (H_{\text{aug}})_1 \\ (H_{\text{aug}})_2 \end{bmatrix} p^t \end{aligned} \tag{23}$$

yielding

$$\begin{aligned} \hat{x}^t =& (F_{\text{aug}})_{11}\ \hat{x}^{t-1} + (F_{\text{aug}})_{12}\ \tilde{\Delta}^{t-1} + (K)_1\ \tilde{\Delta}^t + (H_{\text{aug}})_1\ p^t \\ \tilde{\Delta}^t =& (F_{\text{aug}})_{21}\ \hat{x}^{t-1} + (F_{\text{aug}})_{22}\ \tilde{\Delta}^{t-1} + (K)_2\ \tilde{\Delta}^t + (H_{\text{aug}})_2\ p^t. \end{aligned} \tag{24}$$

Since the states are Markovian, consequently the optimal estimates are Markovian as well. This implies that $\hat{x}^t$ given $\hat{x}^{t-1}$, does not depend on $\tilde{\Delta}^{t-1}$. Likewise, $\tilde{\Delta}^t$ given $\tilde{\Delta}^t$, does not depend on $\hat{x}^{t-1}$, $\tilde{\Delta}^{t-1}$, and $p^t$. The above two arguments can be used to deduce that

$$(F_{\text{aug}})_{12} = 0,\ (F_{\text{aug}})_{21} = 0,\ (F_{\text{aug}})_{22} = 0\ (K)_2 = I,\ (H_{\text{aug}})_2 = 0.$$

which we also verified empirically for the one dimensional case. Substituting $(F_{\text{aug}})_{12} = 0$ in Eq 24 and comparing it with Eq 7, we get the optimal solutions for $F$, $G$, and $H$ in terms of $D$ and $E$, denoted as $F'$, $G'$, and $H'$ respectively.

$$F' = (F_{\text{aug}})_{11},\ G' = (K)_1,\ H' = (H_{\text{aug}})_1$$

### A.5 LQG control solution to find $L$ in terms of $D$ and $E$

The optimal LQG control is given as

$$p^t = L_{\text{aug}}\ \hat{z}^{t-1} \tag{25}$$

$$L_{\text{aug}} = -\ (R + B_{\text{aug}}^\top\ P\ B_{\text{aug}})^{-1}B_{\text{aug}}^\top\ P\ A_{\text{aug}} \tag{26}$$

where $P$ is the solution of a discrete algebraic Riccati equation (Eq 27).

$$P = Q + A_{\text{aug}}^{\top}\, P\, A_{\text{aug}} - A_{\text{aug}}^{\top}\, P\, B_{\text{aug}}(R + B_{\text{aug}}^{\top}\, P\, B_{\text{aug}})^{-1}\, B_{\text{aug}}^{\top}\, P\, A_{\text{aug}} \tag{27}$$

As shown in Eq 16, the last set of columns in $A_{\text{aug}}$ are all zero vectors. As a consequence, since $A_{\text{aug}}$ is the last multiplier in the right side of equality in Eq 25, the last set of columns of $L_{\text{aug}}$ would also be zero vectors. More specifically, Eq 25 can be further simplified by writing $L_{\text{aug}}$ in terms of block matrices in the following way

$$p^t = [(L_{\text{aug}})_1 \quad (L_{\text{aug}})_2]\begin{bmatrix}\hat{x}^{t-1}\\ \tilde{\Delta}^{t-1}\end{bmatrix} = (L_{\text{aug}})_1\, \hat{x}^{t-1} + (L_{\text{aug}})_2\, \tilde{\Delta}^{t-1}, \text{ then } (L_{\text{aug}})_2 = 0$$
$$\implies p^t = (L_{\text{aug}})_1\, \hat{x}^{t-1}$$

Hence, the optimal prediction is linearly related to the estimate through $(L_{\text{aug}})_1$. For consistency in notations, we denote $(L_{\text{aug}})_1$ as $L'$, yielding

$$p^t = L'\, \hat{x}^{t-1}, \text{ where } L' = (L_{\text{aug}})_1.$$

### A.6 Closed form expressions for total cost in terms of $D$ and $E$

The optimal inference cost for fixed $D$ and $E$, which is the same as the augmented state's estimation error, is $\text{Cost}'_{\text{inf}} = Tr(\Sigma)$. Where $\Sigma$ is the augmented state's estimation error covariance matrix, such that

$$\Sigma = \check{\Sigma} - \check{\Sigma}\, C_{\text{aug}}^{\top}(C_{\text{aug}}\, \check{\Sigma}\, C_{\text{aug}}^{\top})^{-1}\, C_{\text{aug}}\, \check{\Sigma}. \tag{28}$$

The optimal total energy cost for a given $D$ and $E$, which is the optimal LQG cost, is given as $\text{Cost}'_{\text{energy}} = Tr(Q\, \Sigma) + Tr(P(\check{\Sigma} - \Sigma))$. As $\Sigma$ is the error covariance matrix in estimating $z^t$, and since the noisy residual $\tilde{\Delta}^t$ is fully observable at any time point, the corresponding lower block matrix in $\Sigma$ will be zero. Also, by definition in Eq 12, we know that the upper block matrix in $Q$ is zero. As a consequence, $Tr(Q\, \Sigma)$ would always be zero. Resulting in $\text{Cost}'_{\text{energy}} = Tr(P(\check{\Sigma} - \Sigma))$. Combining the inference cost and energy cost, we have the total cost in terms of $D$ and $E$ as

$$\text{Cost}'_{\text{tot}} = Tr(\Sigma) + Tr(P(\check{\Sigma} - \Sigma)) \tag{29}$$

### A.7 Condition for useful feedback

It can be algebraically shown that the total cost is an even function in both $D$ and $E$ and hence can be written as a function of $\check{D} = D^2$ and $\check{E} = E^2$. This limits the function domain to non-negative values, thereby reducing the search space for numerical optimization. We empirically observe that the cost function now has just one minimum when plotted with $\check{D}$ and $\check{E}$, and that the negative of gradient at any point always directs toward this minimum.

Let $\check{E}'$ denote the optimal $\check{E}$ that minimizes the total cost at $\check{D} = 0$. $\check{E}'$ can be computed by equating the derivative of cost with respect to $\check{E}$ to 0, at $\check{D} = 0$ (Eq 30).

$$\left.\frac{\partial \text{Cost}'_{\text{tot}}}{\partial \check{E}}\right|_{\check{D}=0, \check{E}=\check{E}'} = 0 \tag{30}$$

We define $\psi$ as the derivative of total cost with respect to $\check{D}$, evaluated at $\check{D} = 0$ and $\check{E} = \check{E}'$ (Eq 31).

$$\psi = \left.\frac{\partial \text{Cost}'_{\text{tot}}}{\partial \check{D}}\right|_{\check{D}=0, \check{E}=\check{E}'} \tag{31}$$

Since the negative of gradient always points to the minimum, the following holds true: If $\psi$ is negative, then the negative of gradient points towards non-zero $\check{D}$, implying the optimal $\check{D}$ is non-zero. However, given that the domain of $\check{D}$ is non-negative, if $\psi$ is non-negative then the optimal $\check{D}$ is 0. Therefore, the optimal $\check{D}$ is non-zero if and only if $\psi < 0$. These conditions can be concisely written as

$$U_{\text{bn}} < \Phi(U_{\text{fn}}, A, \sigma_{\text{p}}^2, \frac{C^2}{\sigma_{\text{o}}^2}). \tag{32}$$

where $\Phi$ is a complicated function involving the roots $\check{E}'$ of the quartic equation Eq 30 substituted into Eq 31. Ultimately we use these analytic expressions to solve numerically for the remaining conditions on the optimal prediction-multiplier $D$ and observation-multiplier $E$.

### A.8 Channel communication in terms of pre and post multipliers

Let $s$ be a random variable describing the signal of interest, following the distribution $\mathcal{N}\left(0,\, \sigma^2_{\text{signal}}\right)$. The signal is pre-multiplied with $m_1$ before passing it through an additive Gaussian noise channel. The channel noise is described using the random variable $\eta \sim \mathcal{N}\left(0,\, \sigma^2_{\text{noise}}\right)$. The message sent over the channel would then be $m_1\, s + \eta$. The message is then post-multiplied with $m_2$ to form the channel output, $\boldsymbol{\pi}(s, m_1, m_2, \sigma^2_{\text{noise}}) = m_2\, (m_1\, s + \eta)$. Let $w$ denote the weight of channel's energetic cost. Then the energetic cost per message is given as,

$$\boldsymbol{\rho}\left(s, m_1, m_2, \sigma^2_{\text{noise}}, w\right) = w\, (m_1^2\, \sigma^2_{\text{signal}} + \sigma^2_{\text{noise}}). \tag{33}$$

**Lemma A.1.** *If the weight of energetic cost ($w$) is scaled up by an arbitrary non-zero factor $k$, and noise variance ($\sigma^2_{\text{noise}}$) is scaled down by $k$ times, then the channel output ($\boldsymbol{\pi}$) and energetic cost per message ($\boldsymbol{\rho}$) remain unchanged if the pre-multiplier is scaled down by $\sqrt{k}$ times and the post-multiplier is scaled up by $\sqrt{k}$ times.*

*Proof.* The notations used is the same as before, except that for the case of scaling we use overline. That is, $\overline{w} = k\, w$, $\overline{\sigma}^2_{\text{noise}} = \sigma^2_{\text{noise}}/k$, and $\overline{\eta} \sim \mathcal{N}\left(0,\, \overline{\sigma}^2_{\text{noise}}\right)$.

Below, we show that the channel output for the cases with and without scaling is the same.

$$\begin{aligned}
\boldsymbol{\pi}\left(s, \overline{m}_1, \overline{m}_2, \overline{\sigma}^2_{\text{noise}}\right) =& \overline{m}_2\, (\overline{m}_1\, s + \overline{\eta}) \\
=& \sqrt{k}\, m_2\, (\frac{m_1}{\sqrt{k}}\, s + \overline{\eta}) \\
=& m_2\, (m_1\, s + \sqrt{k}\, \overline{\eta}) \\
=& m_2\, (m_1\, s + \eta) = \boldsymbol{\pi}\left(s, m_1, m_2, \sigma^2_{\text{noise}}\right).
\end{aligned}$$

The energetic cost per message for the cases with and without scaling is the same as shown below,

$$\begin{aligned}
\boldsymbol{\rho}\left(s, \overline{m}_1, \overline{\sigma}^2_{\text{noise}}, \overline{w}\right) =& \overline{w}\, (\overline{m}_1^2\, \sigma^2_{\text{signal}} + \overline{\sigma}^2_{\text{noise}}) \\
=& k\, w\, (\frac{m_1^2}{k}\, \sigma^2_{\text{signal}} + \frac{\sigma^2_{\text{noise}}}{k}) \\
=& w\, (m_1^2\, \sigma^2_{\text{signal}} + \sigma^2_{\text{noise}}) = \boldsymbol{\rho}\left(s, m_1, \sigma^2_{\text{noise}}, w\right).
\end{aligned}$$

$\therefore$ We have,

$$\boldsymbol{\pi}\left(s, \frac{m_1}{\sqrt{k}}, \sqrt{k}\, m_2, \frac{\sigma^2_{\text{noise}}}{k}\right) = \boldsymbol{\pi}\left(s, m_1, m_2, \sigma^2_{\text{noise}}\right) \tag{34}$$

$$\boldsymbol{\rho}\left(s, \frac{m_1}{\sqrt{k}}, \frac{\sigma^2_{\text{noise}}}{k}, k\, w\right) = \boldsymbol{\rho}\left(s, m_1, \sigma^2_{\text{noise}}, w\right) \tag{35}$$

□

The interpretation of the above result is the following. In comparison to the case without scaling, for the case with scaling, the noise variance is lower $(\times \frac{1}{k})$ and the weight of energetic cost is higher $(\times k)$. Since the noise is lower, a lower amplification $(\times \frac{1}{\sqrt{k}})$ would suffice to retain the same information as for the case without scaling. But since the weight of energetic cost is higher $(\times k)$ for the case of scaling, the lower amplification would still cost the same energetic cost as in the case without scaling.

### A.9 Optimal categorical strategy depends on the feedforward/feedback noise costs

**Theorem A.2.** *If the feedback weight of energetic cost is scaled up by an arbitrary non-zero factor $k$, and the feedback noise variance is scaled down by $k$ times, then **(1)** the optimal total cost remains unchanged, **(2)** the optimal control gain $L$ scales down by $\sqrt{k}$ times, and the optimal prediction multiplier $D$ scales up by $\sqrt{k}$ times, **(3)** the optimal feedback energetic cost remains the same, **(4)** the optimal feedforward energetic cost, optimal inference cost, optimal multipliers $E$, $F$, and $G$ remain unchanged, while the optimal $H$ scales up by $\sqrt{k}$ times, and **(5)** the optimal categorical strategy remains unchanged.*

*Proof.* We layout the proof by first considering the *case without scaling* where the weight of feedback energetic cost is $w_b$, and noise variance is $\sigma_b^2$. We express the corresponding optimal control gain $L^\star$, optimal prediction multiplier $D^\star$, and optimal total cost $\mathrm{Cost}_{\mathrm{tot}}^\star$ in terms of the notations used in Lemma A.1. We then consider the *case with scaling*, where the weight of energetic cost is $\overline{w}_{\mathrm{b}} = k\, w_b$, and noise variance is $\overline{\sigma}_{\mathrm{b}}^2 = \frac{\sigma_{\mathrm{b}}^2}{k}$. We express the corresponding optimal control gain $\overline{L}^\star$, optimal prediction multiplier $\overline{D}^\star$, and optimal total cost $\overline{\mathrm{Cost}}_{\mathrm{tot}}^\star$ also in terms of the notations used in Lemma A.1. Finally, we prove the theorem by applying Lemma A.1 and deriving the required equality. For example, to prove the optimal total cost remains unchanged, we derive $\overline{\mathrm{Cost}}_{\mathrm{tot}}^\star = \mathrm{Cost}_{\mathrm{tot}}^\star$.

**Without scaling** ($w_b$**,** $\sigma_b^2$): We start by establishing the equivalence between the notations used in Appendix A.8 and that used for the feedback channel (Eqs 3-5, 8). For the feedback channel, the signal of interest $s$ is $\hat{x}^{t-1}$, the pre-multiplier $m_1$ is the control gain $L$, and the post-multiplier $m_2$ is the prediction multiplier $D$. The channel output is then given as $\boldsymbol{\pi}\left(\hat{x}^{t-1}, L, D, \sigma_{\mathrm{b}}^2\right) = D\,(L\,\hat{x}^{t-1} + \eta_b)$, and the energetic cost at time $t$ is $\boldsymbol{\rho}\left(\hat{x}^{t-1}, L, \sigma_{\mathrm{b}}^2, w_{\mathrm{b}}\right)$. The residual (Eq 5) and the feedback energetic cost (Eq 8) can then be expressed as

$$\Delta^t = \boldsymbol{\pi}\left(\hat{x}^{t-1}, L, D, \sigma_{\mathrm{b}}^2\right) + E\, o^t$$

$$\mathrm{Cost}_{\mathrm{b}} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{\rho}\left(\hat{x}^{t-1}, L, \sigma_{\mathrm{b}}^2, w_{\mathrm{b}}\right).$$

Note that $D$, $L$, $w_{\mathrm{b}}$, and $\sigma_{\mathrm{b}}^2$ are involved in the dynamics and the optimization Eqs 1-8 only through $\boldsymbol{\pi}\left(\hat{x}^{t-1}, L, D, \sigma_{\mathrm{b}}^2\right)$ and $\boldsymbol{\rho}\left(\hat{x}^{t-1}, L, \sigma_{\mathrm{b}}^2, w_{\mathrm{b}}\right)$ as shown above. Therefore, the minimization of total cost for a given $w_{\mathrm{b}}$, and $\sigma_{\mathrm{b}}^2$ with respect to $D$, and $L$ can be written as

$$\mathrm{Cost}_{\mathrm{tot}}^\star = \min_{L,D} \mathrm{Cost}_{\mathrm{tot}}(\boldsymbol{\pi}\left(\hat{x}^{t-1}, L, D, \sigma_{\mathrm{b}}^2\right), \boldsymbol{\rho}\left(\hat{x}^{t-1}, L, \sigma_{\mathrm{b}}^2, w_{\mathrm{b}}\right)) \tag{36}$$

$$L^\star, D^\star = \arg\min_{L,D} \mathrm{Cost}_{\mathrm{tot}}(\boldsymbol{\pi}\left(\hat{x}^{t-1}, L, D, \sigma_{\mathrm{b}}^2\right), \boldsymbol{\rho}\left(\hat{x}^{t-1}, L, \sigma_{\mathrm{b}}^2, w_{\mathrm{b}}\right)), \tag{37}$$

**With scaling** ($\overline{w}_{\mathrm{b}}$**,** $\overline{\sigma}_b^2$): Just as above, for the scaled case we have

$$\overline{\mathrm{Cost}}_{\mathrm{tot}}^\star = \min_{L,D} \mathrm{Cost}_{\mathrm{tot}}(\boldsymbol{\pi}\left(\hat{x}^{t-1}, L, D, \overline{\sigma}_b^2\right), \boldsymbol{\rho}\left(\hat{x}^{t-1}, L, \overline{\sigma}_b^2, \overline{w}_{\mathrm{b}}\right))$$

$$\overline{L}^\star, \overline{D}^\star = \arg\min_{L,D} \mathrm{Cost}_{\mathrm{tot}}(\boldsymbol{\pi}(\hat{x}^{t-1}, L, D, \overline{\sigma}_{\mathrm{b}}^2), \boldsymbol{\rho}(\hat{x}^{t-1}, L, \overline{\sigma}_{\mathrm{b}}^2, \overline{w}_{\mathrm{b}}))$$

**(1)** By applying Eqs 34-35 of Lemma A.1, and Eq 36 from the case without scaling, we show the optimal costs for both the cases are equal.

$$\begin{aligned}
\overline{\mathrm{Cost}}_{\mathrm{tot}}^\star &= \min_{L,D} \mathrm{Cost}_{\mathrm{tot}}(\boldsymbol{\pi}(\hat{x}^{t-1}, L, D, \overline{\sigma}_b^2), \boldsymbol{\rho}(\hat{x}^{t-1}, L, \overline{\sigma}_b^2, \overline{w}_{\mathrm{b}})) \\
&= \min_{L,D} \mathrm{Cost}_{\mathrm{tot}}(\boldsymbol{\pi}(\hat{x}^{t-1}, L, D, \frac{\sigma_{\mathrm{b}}^2}{k}), \boldsymbol{\rho}(\hat{x}^{t-1}, L, \frac{\sigma_{\mathrm{b}}^2}{k}, k\, w_{\mathrm{b}})) \\
&= \min_{L,D} \mathrm{Cost}_{\mathrm{tot}}(\boldsymbol{\pi}(\hat{x}^{t-1}, \sqrt{k}\, L, \frac{D}{\sqrt{k}}, \sigma_{\mathrm{b}}^2), \boldsymbol{\rho}(\hat{x}^{t-1}, \sqrt{k}\, L, \sigma_{\mathrm{b}}^2, w_{\mathrm{b}})) \\
&= \min_{L,D} \mathrm{Cost}_{\mathrm{tot}}(\boldsymbol{\pi}(\hat{x}^{t-1}, L, D, \sigma_{\mathrm{b}}^2), \boldsymbol{\rho}(\hat{x}^{t-1}, L, \sigma_{\mathrm{b}}^2, w_{\mathrm{b}})) \\
&= \mathrm{Cost}_{\mathrm{tot}}^\star
\end{aligned}$$

**(2)** Similarly, we have

$$\begin{aligned}
\overline{L}^\star, \overline{D}^\star &= \arg\min_{L,D} \mathrm{Cost}_{\mathrm{tot}}(\boldsymbol{\pi}\left(\hat{x}^{t-1}, L, D, \overline{\sigma}_{\mathrm{b}}^2\right), \boldsymbol{\rho}\left(\hat{x}^{t-1}, L, \overline{\sigma}_{\mathrm{b}}^2, \overline{w}_{\mathrm{b}}\right)) \\
&= \arg\min_{L,D} \mathrm{Cost}_{\mathrm{tot}}(\boldsymbol{\pi}\left(\hat{x}^{t-1}, L, D, \frac{\sigma_{\mathrm{b}}^2}{k}\right), \boldsymbol{\rho}\left(\hat{x}^{t-1}, L, \frac{\sigma_{\mathrm{b}}^2}{k}, k\, w_{\mathrm{b}}\right)) \\
&= \arg\min_{L,D} \mathrm{Cost}_{\mathrm{tot}}(\boldsymbol{\pi}\left(\hat{x}^{t-1}, \sqrt{k}\, L, \frac{D}{\sqrt{k}}, \sigma_{\mathrm{b}}^2\right), \boldsymbol{\rho}\left(\hat{x}^{t-1}, \sqrt{k}\, L, \sigma_{\mathrm{b}}^2, w_{\mathrm{b}}\right)).
\end{aligned}$$

Combining above with Eq 37, we get $\overline{L}^\star = \frac{L^\star}{\sqrt{k}}$, and $\overline{D}^\star = \sqrt{k}\, D^\star$. See Fig 7 for the plot obtained using numerical optimization.

**(3)** Let $\overline{\text{Cost}}_{\text{b}}^{\star}$ be the optimal feedback energetic cost for the scaled case. Applying Eq 35 of Lemma A.1, we show that the optimal feedback energetic cost does not change for the scaled case.

$$\begin{aligned}
\overline{\text{Cost}}_{\text{b}}^{\star} &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{\rho}\left(\hat{x}^{t-1}, \overline{L}^{\star}, \overline{\sigma}_{\text{b}}^2, \overline{w}_{\text{b}}\right) \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{\rho}\left(\hat{x}^{t-1}, \frac{L^{\star}}{\sqrt{k}}, \frac{\sigma_{\text{b}}^2}{k}, k\, w_{\text{b}}\right) \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{\rho}\left(\hat{x}^{t-1}, L^{\star}, \sigma_{\text{b}}^2, w_{\text{b}}\right) \\
&= \text{Cost}_{\text{b}}^{\star}.
\end{aligned}$$

**(4)** As in Eq 36, the dynamics and total cost optimization depend on the feedback relevant terms ($D$, $L$, $w_{\text{b}}$, and $\sigma_{\text{b}}^2$) only through the feedback channel output and feedback energetic cost. Under optimal solution, we have that both the feedback channel output and the feedback channel energetic cost remain the same for the cases with and without scaling. That is, using Eqs 34-35 of Lemma A.1, we have $\boldsymbol{\pi}\left(\hat{x}^{t-1}, \overline{L}^{\star}, \overline{D}^{\star}, \overline{\sigma}_b^2\right) = \boldsymbol{\pi}\left(\hat{x}^{t-1}, L^{\star}, D^{\star}, \sigma_b^2\right)$, and $\boldsymbol{\rho}\left(\hat{x}^{t-1}, \overline{L}^{\star}, \overline{\sigma}_{\text{b}}^2, \overline{w}_{\text{b}}\right) = \boldsymbol{\rho}\left(\hat{x}^{t-1}, L^{\star}, \sigma_{\text{b}}^2, w_{\text{b}}\right)$. Therefore, the optimization problem and the dynamics for the cases with and without scaling remain to be equivalent. This results in $\overline{E}^{\star} = E^{\star}$, $\overline{F}^{\star} = F^{\star}$, $\overline{G}^{\star} = G^{\star}$, and the optimal feedforward energetic cost and inference cost to be unchanged. For the case with scaling, in addition to the changes in the optimal control gain and prediction multiplier shown above ($\overline{L}^{\star} = \frac{L^{\star}}{\sqrt{k}}, \overline{D}^{\star} = \sqrt{k}\, D^{\star}$), we also have $\overline{H}^{\star} = \sqrt{k}\, H^{\star}$. Note that $\overline{H}^{\star} = \sqrt{k}\, H^{\star}$ does not conflict the statement that optimization problem and the dynamics for the cases with and without scaling remain to be equivalent, but in fact is in agreement. This is because, for the dynamics and optimization to be exactly the same, the coefficient of $\hat{x}^{t-1}$ used in updating the estimate in Eq 7 should also be the same. The coefficient of $\hat{x}^{t-1}$ in Eq 7 is $F + H\, L$. So, when the optimal $L$ reduces by $\sqrt{k}$ factor, optimal $H$ has to increase by $\sqrt{k}$ to keep the coefficient unchanged ($\overline{L}^{\star} = \frac{L^{\star}}{\sqrt{k}}, \overline{H}^{\star} = \sqrt{k}\, H^{\star}$).

**(5)** From above, we have $\overline{L}^{\star} = \frac{L^{\star}}{\sqrt{k}}, \overline{D}^{\star} = \sqrt{k}\, D^{\star}, \overline{E}^{\star} = E^{\star}$. Using this we exhaustively show that whichever is the optimal strategy for the case without scaling, the same would be the optimal strategy for the case with scaling.

If for the case of without scaling is predictive coding, then $L^{\star},\ D^{\star},\ E^{\star} \neq 0$. For non-zero $k$, this would imply $\overline{L}^{\star},\ \overline{D}^{\star},\ \overline{E}^{\star} \neq 0$, resulting in predictive coding for the case of scaling.

If for the case of without scaling is efficient coding, then $L^{\star},\ D^{\star} = 0$, and $E^{\star} \neq 0$. This would imply $\overline{L}^{\star},\ \overline{D}^{\star} = 0$, and $\overline{E}^{\star} \neq 0$, resulting in efficient coding for the case of scaling.

If for the case of without scaling is to not send any messages, then $L^{\star},\ D^{\star},\ E^{\star} = 0$. For non-zero $k$, this would imply $\overline{L}^{\star},\ \overline{D}^{\star},\ \overline{E}^{\star} \neq 0$, resulting in sending no messages for the case of scaling. □

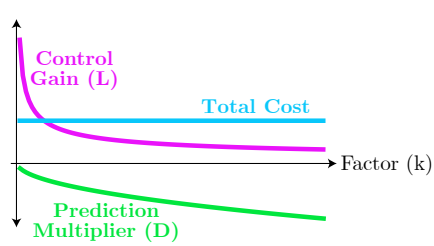


Figure 7: *When the feedback weight of energetic cost is scaled up by k times, and feedback noise variance is scaled down by k times, the optimal costs remain unchanged, and the magnitude of optimal control gain scales down by $\sqrt{k}$ times and the optimal prediction multiplier scales up by $\sqrt{k}$ times. While only the total cost is shown in the figure, we also observe the inference cost, feedback, and feedforward energetic costs to remain the same.*

**Theorem A.3.** *If the feedforward weight of energetic cost is scaled up by an arbitrary non-zero factor $k$, and the feedforward noise variance is scaled down by $k$ times, then **(1)** the optimal total cost remains unchanged, **(2)** the optimal observation multiplier $E$ scales down by $\sqrt{k}$ times, and the optimal multiplier for noisy residual $G$ scales up by $\sqrt{k}$ times, **(3)** the optimal feedforward energetic cost remains the same, **(4)** the optimal feedback energetic cost, optimal inference cost, optimal multipliers $F$, $H$, $L$ remain unchanged, while the optimal prediction multiplier $D$ scales down by $\sqrt{k}$ times, and **(5)** the optimal categorical strategy remains unchanged.*

*Proof.* The proof for the feedforward channel is equivalent to the proof for the feedback channel in Theorem A.2, with observation multiplier $E$ as the pre-multiplier $m_1$, and the multiplier for noisy residual $G$ as the post multiplier $m_2$. The theorem is verified using numerical optimization as shown in Fig 8. □

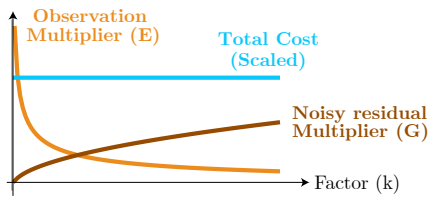


Figure 8: *When the feedforward weight of energetic cost is scaled up by $k$ times, and feedforward noise variance is scaled down by $k$ times, the optimal costs remain unchanged, and the observation multiplier scales down by $\sqrt{k}$ times, and the multiplier for noisy residual scales up by $\sqrt{k}$ times. While only the total cost is shown in the figure, we also observe the inference cost, feedback, and feedforward energetic costs to remain the same.*

**Theorem A.4.** *Under the optimal case, several quantities depend on the feedback weight, feedback noise variance, feedforward weight, and feedforward noise variance only through the feedback noise cost, and the feedforward noise cost. The quantities are **(1)** the total cost, **(2)** the optimal categorical strategy, **(3)** the feedback energetic cost, **(4)** the feedforward energetic cost, and **(5)** the inference cost.*

*Proof.* Below, we show the proof for the entity optimal total cost, but the same steps can be followed for proving for other entities listed as well. The layout of the proof is as follows. We consider four different cases of channel noise costs, each denoted by the sequence $(feedback\,weight,\ feedback\,noise,\ feedforward\,weight,\ feedforward\,noise)$. The four cases are $(w_b, \sigma_b^2, w_f, \sigma_f^2)$, $(w_b, \sigma_b^2, \overline{w}_f, \overline{\sigma}_f^2)$, $(\overline{w}_b, \overline{\sigma}_b^2, w_f, \sigma_f^2)$, and $(\overline{w}_b, \overline{\sigma}_b^2, \overline{w}_f, \overline{\sigma}_f^2)$. Where $\overline{w}_b = k_1\, w_b$, $\overline{\sigma}_b^2 = \sigma_b^2/k_1$, $\overline{w}_f = k_2\, w_f$, $\overline{\sigma}_f^2 = \sigma_f^2/k_2$, such that $k_1$ and $k_2$ arbitrary non-zero scaling factors. Since the feedback noise cost is same for $\overline{w}_b\, \overline{\sigma}_b^2 = w_b\, \sigma_b^2$, and feedforward noise cost is same for $\overline{w}_f\, \overline{\sigma}_f^2 = w_f\, \sigma_f^2$, in order to prove the theorem for total cost, it is sufficient to show

$$\text{Cost}^\star_{\text{tot}}\left(w_b, \sigma_b^2, w_f, \sigma_f^2\right) = \text{Cost}^\star_{\text{tot}}\left(w_b, \sigma_b^2, \overline{w}_f, \overline{\sigma}_f^2\right) = \text{Cost}^\star_{\text{tot}}\left(\overline{w}_b, \overline{\sigma}_b^2, w_f, \sigma_f^2\right) = \text{Cost}^\star_{\text{tot}}\left(\overline{w}_b, \overline{\sigma}_b^2, \overline{w}_f, \overline{\sigma}_f^2\right). \tag{38}$$

From Theorem A.2, we directly get

$$\text{Cost}^\star_{\text{tot}}\left(w_b, \sigma_b^2, w_f, \sigma_f^2\right) = \text{Cost}^\star_{\text{tot}}\left(\overline{w}_b, \overline{\sigma}_b^2, w_f, \sigma_f^2\right) \tag{39}$$

$$\text{Cost}^\star_{\text{tot}}\left(w_b, \sigma_b^2, \overline{w}_f, \overline{\sigma}_f^2\right) = \text{Cost}^\star_{\text{tot}}\left(\overline{w}_b, \overline{\sigma}_b^2, \overline{w}_f, \overline{\sigma}_f^2\right). \tag{40}$$

From Theorem A.3, we get

$$\text{Cost}^\star_{\text{tot}}(w_b, \sigma_b^2, w_f, \sigma_f^2) = \text{Cost}^\star_{\text{tot}}(w_b, \sigma_b^2, \overline{w}_f, \overline{\sigma}_f^2) \tag{41}$$

$$\text{Cost}^\star_{\text{tot}}(\overline{w}_b, \overline{\sigma}_b^2, w_f, \sigma_f^2) = \text{Cost}^\star_{\text{tot}}(\overline{w}_b, \overline{\sigma}_b^2, \overline{w}_f, \overline{\sigma}_f^2). \tag{42}$$

Combining Eqs 39-42, we get Eq 38. Hence proved. □

### A.10 Channel noise and weight for non-sequential and sequential channels

Below, we derive the results for non-sequential and sequential feedback channels, but the derivation for the feedforward channel is the same. For simplicity, we assume that each neuron has gain $g_{\text{nb}}$, and is corrupted by i.i.d Gaussian noise $\mathcal{N}\left(0,\, \sigma_{\text{nb}}^2\right)$. Let $w_{\text{nb}}$ be the weight of the energetic cost of

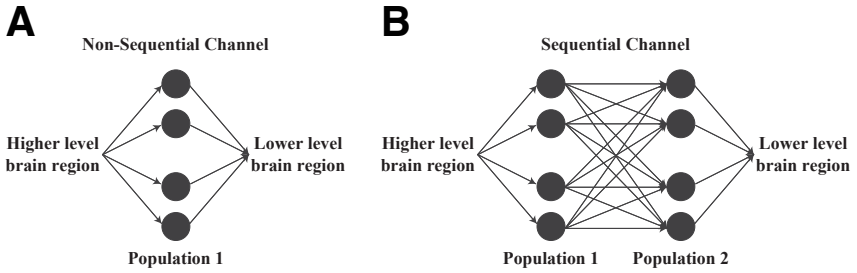


Figure 9: *Example of two different anatomical structures for feedforward/feedback channel.* ***A****: The non-sequential channel comprises a single population of neurons that linearly encodes the message that needs to be sent. The message received at the end of the channel is the decoding of the population's representation.* ***B****: The sequential channel comprises two populations of neurons sequentially connected one after the other. The first population linearly encodes the message to be sent, which is then decoded and then encoded by the second population. This is equivalently represented as each neuron in second population being connected to all the neurons in the first population since the weights are all equal. The message received at the end of the channel is a decoding of the second population's representation.*

a single neuron. The subscript $\mathrm{nb}$ indicates a single **n**euron in feed**b**ack channel. For the feedback channel, the message to be sent at any time $t$ is the noiseless prediction $p^t$. We compute the equivalent feedback noise variance and weight of the feedback channel in terms of the neuron noise and weight. The computed feedback noise and weight can then be plugged into the original optimization in Eq 8 to find the corresponding solution.

### A.10.1 Non-sequential feedback channel

Let $N_\mathrm{b}$ be the number of neurons in the population comprising the feedback channel.

**Feedback channel noise variance:** When the noiseless prediction $p^t$ is linearly encoded by the population of neurons, the decoded signal is $p^t + \eta_\mathrm{b}^t$ (as in Eq 4), where $\eta_\mathrm{b}^t$ is the feedback noise with feedback noise variance $\frac{\sigma_\mathrm{nb}^2}{N_\mathrm{b}\, g_\mathrm{nb}^2}$.

**Feedback channel energetic cost:** As each of the $N_\mathrm{b}$ neurons in the population encode $p^t$ with gain $g_\mathrm{nb}$, noise variance $\sigma_\mathrm{nb}^2$, and weight of energetic cost $w_\mathrm{nb}$, the total feed**b**ack **pop**ulation energy $\mathrm{Cost}_\mathrm{b}^\mathrm{pop}$ consumed is

$$\mathrm{Cost}_\mathrm{b}^\mathrm{pop} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} N_\mathrm{b}\, w_\mathrm{nb}\, g_\mathrm{nb}^2\, {p^t}^2 + N_\mathrm{b}\, w_\mathrm{nb}\, \sigma_\mathrm{nb}^2$$

Note that in Eq 8, the feedback energetic cost in the original formulation was written as

$$\begin{aligned}\mathrm{Cost}_\mathrm{b} &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} w_\mathrm{b}\, {\tilde{p}^t}^2 \\ &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} w_\mathrm{b}\, {p^t}^2 + w_\mathrm{b}\, \sigma_\mathrm{b}^2\end{aligned}$$

Therefore, expressing $\mathrm{Cost}_\mathrm{b}^\mathrm{pop}$ in terms of $\mathrm{Cost}_\mathrm{b}$, we have

$$\mathrm{Cost}_\mathrm{b}^\mathrm{pop} = \mathrm{Cost}_\mathrm{b} + (N_\mathrm{b} - 1)\, w_\mathrm{nb}\, \sigma_\mathrm{nb}^2 \tag{43}$$

where the weight of feedback channel energetic cost is $w_\mathrm{b} = N_\mathrm{b}\, w_\mathrm{nb}\, g_\mathrm{nb}^2$. Note that now there is an additional term to $\mathrm{Cost}_\mathrm{b}$ in Eq 43 when computing the total feedback energy, and therefore the total cost would also increase by the same amount in Eq 8. However, the optimization in Eq 9 still remains to be exactly the same because the additional term does not include any of the parameters we are optimizing over. Hence, we can solve for the case of non-sequential feedback channel by following the exact same procedure in Section 4.

**Feedback channel noise cost:** We computed the feedback noise variance as $\sigma_\mathrm{b}^2 = \frac{\sigma_\mathrm{nb}^2}{N_\mathrm{b}\, g_\mathrm{nb}^2}$, and the feedback weight of energetic cost as $w_\mathrm{b} = N_\mathrm{b}\, w_\mathrm{nb}\, g_\mathrm{nb}^2$. Therefore, their product, the feedback

channel noise cost is $w_{\mathrm{nb}}\ \sigma_{\mathrm{nb}}^2$. Note that the feedback channel noise cost is constant for a given noise variance and weight of energetic cost per neuron.

### A.10.2 Sequential feedback channel

Let $\frac{\gamma}{1+\gamma} N_{\mathrm{b}}$, and $\frac{1}{1+\gamma} N_{\mathrm{b}}$ be the number of neurons in the first and second population of the feedback channel. Where $N_{\mathrm{b}}$ is the total number of neurons in both the populations together, and $\gamma$ is the population size ratio. Population size ratio is the ratio between the number of neurons in the first population to the number of neurons in the second population.

**Feedback channel noise variance:** When the noiseless prediction $p^t$ is linearly encoded by the first population of neurons, $p^t + \eta_{1\mathrm{b}}^t$ would be the corresponding linearly decoded signal, where $\eta_{1\mathrm{b}}^t \sim \mathcal{N}\left(0,\ \frac{(1+\gamma)}{\gamma} \frac{\sigma_{\mathrm{nb}}^2}{N_{\mathrm{b}}\ g_{\mathrm{nb}}^2}\right)$. Then, when $p^t + \eta_{1\mathrm{b}}^t$ is linearly encoded by the second population of neurons, the corresponding decoded signal would be $p^t + \eta_{1\mathrm{b}}^t + \eta_{2\mathrm{b}}^t$, where $\eta_{2\mathrm{b}}^t \sim \mathcal{N}\left(0,\ \frac{(1+\gamma)\ \sigma_{\mathrm{nb}}^2}{N_{\mathrm{b}}\ g_{\mathrm{nb}}^2}\right)$. Expressing this noisy prediction in the form of Eq 4, we get $\eta_{\mathrm{b}}^t = \eta_{1\mathrm{b}}^t + \eta_{2\mathrm{b}}^t$, and $\sigma_{\mathrm{b}}^2 = \frac{(1+\gamma)^2}{\gamma} \frac{\sigma_{\mathrm{nb}}^2}{N_{\mathrm{b}}\ g_{\mathrm{nb}}^2}$ as the feedback channel noise variance.

**Feedback channel energetic cost:** As each of the $\frac{\gamma}{1+\gamma} N_{\mathrm{b}}$ number of neurons in the first population encode $p^t$ with gain $g_{\mathrm{nb}}$, noise variance $\sigma_{\mathrm{nb}}^2$, and weight of energetic cost $w_{\mathrm{nb}}$, the energy consumed by the first population of neurons is

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \frac{\gamma\ N_{\mathrm{b}}\ w_{\mathrm{nb}}\ g_{\mathrm{nb}}^2}{1+\gamma}\ {p^t}^2 + \frac{\gamma\ N_{\mathrm{b}}\ w_{\mathrm{nb}}\ \sigma_{\mathrm{nb}}^2}{1+\gamma} \tag{44}$$

And as each of the $\frac{1}{1+\gamma} N_{\mathrm{b}}$ neurons in the second population encode $p^t + \eta_{1\mathrm{b}}^t$, the energy consumed by the second population of neurons is

$$\begin{aligned} &\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \frac{N_{\mathrm{b}}\ w_{\mathrm{nb}}\ g_{\mathrm{nb}}^2}{1+\gamma}\ (p^t + \eta_{1\mathrm{b}})^2 + \frac{N_{\mathrm{b}}\ \ w_{\mathrm{nb}}\ \sigma_{\mathrm{nb}}^2}{1+\gamma} \\ &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \frac{N_{\mathrm{b}}\ w_{\mathrm{nb}}\ g_{\mathrm{nb}}^2}{1+\gamma}\ {p^t}^2 + (\frac{1}{\gamma} + \frac{N_{\mathrm{b}}}{1+\gamma})\ w_{\mathrm{nb}}\ \sigma_{\mathrm{nb}}^2 \end{aligned} \tag{45}$$

Combining energy of both the populations (adding Eqs 44 and 45), we get total feed**b**ack **pop**ulation energy $\mathrm{Cost}_{\mathrm{b}}^{\mathrm{pop}}$ as

$$\mathrm{Cost}_{\mathrm{b}}^{\mathrm{pop}} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} N_{\mathrm{b}}\ w_{\mathrm{nb}}\ g_{\mathrm{nb}}^2\ {p^t}^2 + (N_{\mathrm{b}} + \frac{1}{\gamma})\ w_{\mathrm{nb}}\ \sigma_{\mathrm{nb}}^2$$

Note that in Eq 8, the feedback energetic cost in the original formulation was written as

$$\begin{aligned} \mathrm{Cost}_{\mathrm{b}} &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} w_{\mathrm{b}}\ {\tilde{p}^t}^2 \\ &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} w_{\mathrm{b}}\ {p^t}^2 + w_{\mathrm{b}}\ \sigma_{\mathrm{b}}^2 \end{aligned}$$

Therefore, expressing $\mathrm{Cost}_{\mathrm{b}}^{\mathrm{pop}}$ in terms of $\mathrm{Cost}_{\mathrm{b}}$, we have

$$\mathrm{Cost}_{\mathrm{b}}^{\mathrm{pop}} = \mathrm{Cost}_{\mathrm{b}} + (N_{\mathrm{b}} - \gamma - 2)\ w_{\mathrm{nb}}\ \sigma_{\mathrm{nb}}^2 \tag{46}$$

where the weight of feedback channel energetic cost is $w_{\mathrm{b}} = N_{\mathrm{b}}\ w_{\mathrm{nb}}\ g_{\mathrm{nb}}^2$. Note that now there is an additional term to $\mathrm{Cost}_{\mathrm{b}}$ in Eq 46 when computing the total feedback energy, and therefore the total cost would also increase by the same amount in Eq 8. However, the optimization in Eq 9 still remains to be exactly the same because the additional term does not include any of the parameters we are optimizing over. Hence, we can solve for the case of sequential feedback channel by following the exact same procedure in Section 4.

**Feedback channel noise cost:** We computed the feedback noise variance as $\sigma_{\mathrm{b}}^2 = \frac{(1+\gamma)^2}{\gamma} \frac{\sigma_{\mathrm{nb}}^2}{N_{\mathrm{b}}\, g_{\mathrm{nb}}^2}$, and the feedback weight of energetic cost as $w_{\mathrm{b}} = N_{\mathrm{b}}\, w_{\mathrm{nb}}\, g_{\mathrm{nb}}^2$. Therefore, their product, the feedback channel noise cost is $\frac{(1+\gamma)^2}{\gamma}\, w_{\mathrm{nb}}\, \sigma_{\mathrm{nb}}^2$. Note that the feedback channel noise cost does not depend on the number of neurons $N_{\mathrm{b}}$, but depends on the ratio $\gamma$ that determines the anatomical structure.