Time cells might be optimized for predictive capacity, not redundancy reduction or memory capacity

Alexander Hsu and Sarah E. Marzen[*]

W. M. Keck Science Department, Claremont, California 91711, USA

(Received 19 May 2020; accepted 8 October 2020; published 2 December 2020)

Recently, researchers have found time cells in the hippocampus that appear to contain information about the timing of past events. Some researchers have argued that time cells are taking a Laplace transform of their input in order to reconstruct the past stimulus. We argue that stimulus prediction, not stimulus reconstruction or redundancy reduction, is in better agreement with observed responses of time cells. In the process, we introduce new analyses of nonlinear, continuous-time reservoirs that model these time cells.

DOI: 10.1103/PhysRevE.102.062404

I. INTRODUCTION

Recent experiments have revealed the presence of so-called time cells in the hippocampus, which seem to fire to signal the timing of a certain event [1]. Time cells fire even when location information or behavioral information is constant [2] and are thought to support episodic memory—memory of what, where, and when an event was experienced [1].

Reference [3] offers a novel explanation of time cells, which applies to not just temporal signals but also to spatial signals and others: They claim that time cells are computing a Laplace transform of the input and that the past input is linearly reconstructed from discrete samples of this Laplace transform. These model time cells are therefore linear continuous-time reservoirs, or linear echo state networks [4–6], which can simulate, predict, and remember limited types of input. Their nonlinear counterparts can simulate any type of input with enough nodes (neurons) [7].

Implicit in several descriptions of time cells [1,8,9] is that the goal of these cells is to reconstruct the past stimulus. This certainly seems like a worthwhile goal for an organism. However, some classic work suggests that neurons try to "efficiently code" their stimulus minimize redundancy [10], and some recent works have suggested that the goal of some biological subsystems is to predict the future, e.g., as in Refs. [11,12]. These goals might all sound similar, and to some extent they are—one needs memory to predict, for example. But it is also possible to have infinite memory and no predictive power [13]. Here we compare predictions of each of these normative principles to ascertain which are consistent with observed time cell properties. To do so, we extended the results and the methodologies of Ref. [13] to the case of some nonlinear and all linear continuous-time reservoirs, thus extending the work of Ref. [4].

Only maximization of predictive power of time cells when stimulated with naturalistic stimuli yields neuronal timescales that behave near to what is seen in experiment [3], suggesting that prediction—not reconstruction or redundancy reduction—may be key to understanding the properties of time cells. This conclusion assumes both that natural video's autocorrelation function does not have significant oscillatory components and that the brain also has "readout neurons" that simply communicate information about only the present stimulus. Prediction has already proven key for understanding other aspects of neural processing [11,12].

The paper starts by describing our setup, in which we specialize to a stationary stimulus and the normative principles listed above. We then describe the timescales of model neurons that minimize redundancy, maximize memory, or maximize prediction for both simple and more naturalistic stimuli and show that only maximization of predictive ability might match experiment.

II. SETUP

The organism is exposed to a continuously varying temporal signal $\overleftrightarrow{x}$, whose value at time t is x_t. For ease, we assume that the stimulus is a scalar with zero mean $\langle x_t\rangle = 0$ and unit variance $\langle x_t^2\rangle = 1$. This temporal signal is a realization of an ergodic stationary stochastic process with random variable $\overleftrightarrow{X}$ symbolizing the whole signal and $\overrightarrow{X}_t^T$ symbolizing the trajectory that starts at t and ends at $t+T$. Stationarity implies that $\Pr(\overrightarrow{X}_t^T)$ is independent of t, and ergodicity implies that different realizations have identical statistics.

We assume that the autocorrelation function of the input signal can be written as

$$R(t) = \int_0^\infty F(\lambda) e^{-\lambda|t|} d\lambda. \tag{1}$$

All autocorrelation functions can be written in this form if one extends the integral to exist over the complex plane. In this manuscript, we study exponential autocorrelation functions and oscillatory and exponentially decaying autocorrelation functions. In the latter case, we can use the formulas developed later by allowing $F(\lambda)$ to have support on imaginary numbers with negative real parts.

Three types of input are studied: a particle moving according to an overdamped Langevin equation, a particle moving

[*] smarzen@cmc.edu

according to an underdamped Langevin equation, and a particle whose position has statistics similar to that of natural video. In the first case, we approximate the autocorrelation function $R(t)$ as a single exponential,

$$R(t) = e^{-\lambda|t|}. \tag{2}$$

In the second case, we approximate the autocorrelation function $R(t)$ as a decaying exponential multiplied by an oscillatory function,

$$R(t) = e^{-\lambda|t|}\cos(\omega t). \tag{3}$$

In the second case, we turn to Ref. [14], in which it was found that the power spectrum of natural video is roughly $\frac{1}{|\omega|^{\alpha}}$ for α between 1 and 2, resulting in power-law autocorrelation functions with exponents between 1 and 2. In order to study model time cells under naturalistic conditions, we consider autocorrelation functions of the form

$$R(t) = \frac{1}{1+|t|^{\alpha}}. \tag{4}$$

Some of our results will hold more generally than for just these three conditions.

The organism is presumed to have model time cells whose activity changes as a function of sensory signal, so-called time cells. These neurons might have their response properties tuned based on one of many normative principles that we discuss below.

Finally, for what follows, we need to define the entropy of a random variable Y with realizations $y \in \mathcal{Y}$, and the mutual information of a random variable Y and another random variable Z with realizations $z \in \mathcal{Z}$. The entropy $H[Y]$ is given by $-\sum_y p(y)\log p(y)$, and the mutual information $I[Y;Z]$ is given by $\sum_{y,z} p(y,z)\log\frac{p(y,z)}{p(y)p(z)}$. The entropy is the uncertainty, while the mutual information captures the reduction in uncertainty we achieve by knowing one of the variables [15]. There are, of course, operational meanings to entropy and mutual information via Shannon's theorems, but we do not need these theorems for what follows.

A. Model of time cells

We are interested in two types of model time cells. The first type of time cell merely remembers what it saw a time s in the past. The second type of time cell follows the formulation of Ref. [3], as it computes a Laplace transform of the input. In the main text, we will only consider the second type. Results for the first type, which are qualitatively similar, are in the Appendix D.

The activity of time cell $f(t)$ with neuronal forgetting rate s (an inverse neuronal timescale) at time t is

$$f(t) = \int_0^{\infty} e^{-st'}x(t-t')dt', \tag{5}$$

which can be achieved via a leaky integrator,

$$\frac{df}{dt} = -sf + x. \tag{6}$$

This is a Laplace transform but sampled only at some values of s. The stimulus $\overleftarrow{x_t}$ can be inferred by an approximate inverse Laplace transform or (nearly equivalently) by an optimal linear estimate.

We imagine that there are N neurons, and that the ith neuron has a forgetting rate s_i. We order the neurons without loss of generality so that $\{s_i\}_{i=1}^N$ is monotonically increasing. The neural activity of time cell i at time t is denoted $f_{s_i}(t)$.

Although our setup might seem limited in that these recurrent networks are "simple"—that is, there are only self-loops and no connections between neurons—simple linear recurrent networks are just as powerful as the more complex linear recurrent networks with connections between different neurons. This fact comes from Ref. [13] and the formulas derived in the subsection below and is only true when recurrent networks are linear.

One might expect a qualitatively different story when the activities are nonlinear functions of past input, but in the Appendix C we show that linearity is desirable for maximal predictive capacity. Still, a full understanding of nonlinear reservoirs will be the subject of future work.

B. Variety of normative principles

There are at least four normative principles that could explain the properties of time cells: minimization of redundancy [10,16] between neighboring time cells; maximization of memory capacity, which is a metric for how well one can reconstruct the past stimulus from the present neuronal response [17–21]; maximization of the joint entropy of all neuronal responses, as derived from the efficient coding hypothesis [10], which is sometimes rephrased as redundancy reduction; and maximization of predictive capacity, which is a metric for how well one can predict the future stimulus from the present neuronal response [13].

Each of these normative principles is quantified as follows. Redundancy, as is typical, is deemed to be the mutual information between the output of two neurons. We extend the definition of discrete-time memory capacity [13] and predictive capacity [13] to continuous-time via

$$\mathrm{MC} = \int_{-\infty}^{0} m(\tau)d\tau, \quad \mathrm{PC} = \int_{0}^{\infty} m(\tau)d\tau \tag{7}$$

where the memory function $m(\tau)$ is the squared correlation coefficient between the optimal linear estimate of $x(t+\tau)$ using $f(t)$, which one can show is

$$m(\tau) = \langle f(t)x(t+\tau)\rangle_t^{\top}\langle f(t)f(t)^{\top}\rangle_t^{-1}\langle f(t)x(t+\tau)\rangle_t. \tag{8}$$

We have assumed that the input is zero-mean and of unit variance. Although it seems unlikely that an organism is interested in arbitrarily long pasts, the infinite limit provides good intuition for the more biophysically reasonable, finite-time case. In the Appendix A, we provide a derivation of the following closed-form expression for MC:

$$\mathrm{MC} = 1^{\top}(C^{-1}\odot D_{\mathrm{MC}})1, \tag{9}$$

where

$$C_{ij} = \int_0^{\infty} F(\lambda)\frac{2\lambda+s_i+s_j}{(\lambda+s_i)(\lambda+s_j)(s_i+s_j)}d\lambda \tag{10}$$

and

$$(D_{\mathrm{MC}})_{i,j} = \int_{\lambda=0}^{\infty}\int_{\lambda'=0}^{\infty} F(\lambda)F(\lambda')\left\{\frac{1}{(\lambda^2-s_i^2)(\lambda'^2-s_j^2)}\right. \times\left[\frac{4\lambda\lambda'}{s_i+s_j} - \frac{2\lambda(\lambda'+s_j)}{s_i+\lambda'} - \frac{2\lambda'(\lambda+s_i)}{\lambda+s_j} \left.+\frac{(\lambda+s_i)(\lambda'+s_j)}{\lambda+\lambda'}\right]\right\}d\lambda'. \tag{11}$$

Furthermore, also in the Appendix A, we show that

$$\mathrm{PC} = 1^{\top}(C^{-1}\odot D_{\mathrm{PC}})1, \tag{12}$$

where C_{ij} is as before and

$$(D_{\mathrm{PC}})_{ij} = \int_0^{\infty}\int_0^{\infty}\frac{F(\lambda)F(\lambda')}{(\lambda+\lambda')(\lambda+s_i)(\lambda'+s_j)}d\lambda d\lambda'. \tag{13}$$

Note that these formulas also allow for complex λ, if the integrals or sums are appropriately extended. This is quite useful for oscillatory input.

These formulas are complicated, but they lead to two main points. First, the *only* relevant environmental statistics for MC and PC are the autocorrelation function of the input. This is true also for the discrete-time case. Hence, stimulating time cells with real natural video will, in theory, yield the same MC or PC as usage of the above formulas with the autocorrelation function of naturalistic input. Second, the formulas above may yield more accurate calculations of MC or PC, as we have traded difficulties associated with too little data for difficulties of accurate numerical integration and matrix inversion. Which of these difficulties is more pressing will depend on one's application.

We could also consider combinations of the above normative principles. For instance, one might try to maximize predictive power while minimizing memory, as in Refs. [11,22–24]. We discuss this possibility later but shy away from doing a combination of optimization principles in this paper because it is likely possible to achieve almost any desired optimal neural forgetting rate by appropriate choice of Lagrange multipliers.

III. RESULTS

In what follows, we derive the optimized neuronal timescales for each of the normative principles for three types of input: a particle moving according to an overdamped Langevin equation, a particle moving according to an underdamped Langevin equation, and a particle whose position has an autocorrelation function like that of natural videos.

For our linear time cells, as stated earlier, only the autocorrelation function of the input affects predictive capacity and memory capacity (see the Appendix A). This is a theoretical conclusion that greatly simplifies any effort to find optimal neuronal forgetting rates, as we only need to estimate the autocorrelation function of natural input and input such autocorrelation functions into the formulas given earlier and in our Appendices. In our two toy examples, the autocorrelation function takes the form of a single exponential (overdamped) and an oscillatory decaying exponential (underdamped) with Gaussian statistics, so that again, only the autocorrelation function determines memory capacity, predictive capacity, and also redundancy. Because the autocorrelation function uniquely determines memory and predictive capacity, the memory and predictive capacities given here for naturalistic input are the same as if we had simulated our model time cells being stimulated with natural video.

A. Redundancy equalization and minimization

It seems desirable to reduce redundancy between neurons [10]. Two simple examples will illustrate that redundancy minimization does not typically yield logarithmic scaling, as anticipated by Ref. [16]. Suppose that $x(t)$ is a Gaussian process, which is necessarily true for outputs of overdamped and underdamped Langevin equations. A straightforward calculation then gives

$$I[f_{s_i}(t); f_{s_{i+1}}(t)] = \log\sqrt{\frac{1}{1-\rho^2}}, \tag{14}$$

where $\rho^2 = \frac{\langle f_{s_i}(t)f_{s_{i+1}}(t)\rangle^2}{\langle f_{s_i}(t)^2\rangle\langle f_{s_{i+1}}(t)^2\rangle}$ is the correlation coefficient for zero-mean processes. Some straightforward algebra reveals that

$$\rho^2 = \frac{\left(\int_0^{\infty}\int_0^{\infty} e^{-s_i t}e^{-s_j t'}R(t-t')dtdt'\right)^2}{\left(\int_0^{\infty}\int_0^{\infty} e^{-s_i(t+t')}R(t-t')dtdt'\right)\left(\int_0^{\infty}\int_0^{\infty} e^{-s_j(t+t')}R(t-t')dtdt'\right)}. \tag{15}$$

This mutual information is a typical measure of redundancy [16]. Redundancy is minimized when the correlation coefficient is minimized.

In the overdamped case, a straightforward calculation gives, for $s_{i+1} = \Delta_i s_i$,

$$\rho^2 = \frac{\Delta_i}{(1+\Delta_i)^2}\frac{(2\lambda+s_i+\Delta_i s_i)^2}{(\lambda+s_i)(\lambda+\Delta_i s_i)}. \tag{16}$$

To equalize redundancy between two successive sets of neurons, we must set ρ^2 to be constant, which cannot be accomplished for this type of input. Some algebra reveals that equalized redundancy implies negative neuronal timescales, a biophysical impossibility. In fact, redundancy equalization is either unachievable or does not seem to imply logarithmic scaling *unless* the input has exactly power-law autocorrelation as was found in Ref. [16], based on calculations not shown here.

To minimize redundancy when the input moves according to an overdamped Langevin equation, we must make forgetting rates s_i, s_{i+1} as big as possible, while making Δ_i as large as possible as well. No matter the input, we tend to find that neurons should all forget past stimulus information as quickly as possible.

Typically, e.g., when performing independent components analysis [25], one finds that redundancy is reduced when different neurons pick up on orthogonal aspects of the stimulus. With this Laplace transform model of time cells, such decoupling is not possible, and reducing redundancy requires sending at least one of the neuronal forgetting rates to infinity. In particular, to minimize redundancy when the input moves according to an underdamped Langevin equation, or when the input's position has naturalistic statistics, some numerical experiments suggest that we must make forgetting rates as dissimilar as possible, i.e., $s_i \to 0$, $s_{i+1} \to \infty$. This intuitively makes some sense: to reduce correlation between neurons, we should make their responses as dissimilar as possible.

In all cases, to minimize redundancy, we desire to set at least one of the forgetting rates to be infinite, so as to decouple the neurons as much as possible.

B. Efficient coding

Usually the efficient coding hypothesis [10] is phrased as follows: We desire the channel $p(y|x)$ that maximizes mutual information subject to a capacity constraint, $p^*(y|x) := \arg\max_{p(y|x):I[X;Y]\leqslant C} I[X;Y]$. This, alone, is underdetermined, and so we also impose another constraint: that $p(y|x)$ be a deterministic mapping, so that $I[X;Y] = H[Y] - H[Y|X] = H[Y]$. Hence, we are searching for neural responses that maximize the joint entropy, $H[\{f_{s_i}(t)\}_{i=1}^N]$.

As it turns out, this objective function is directly related to the redundancy objective function described in the previous subsection. Repeatedly using the information theoretic identity $H[X;Y] = H[X] + H[Y|X]$ yields

$$H\big[\{f_{s_i}(t)\}_{i=1}^N\big] = H\big[f_{s_1}(t)\big] + H\big[f_{s_2}(t)|f_{s_1}(t)\big] + \dots \\ + H\big[f_{s_N}(t)|f_{s_1}(t), \dots, f_{s_{N-1}}(t)\big]. \quad (17)$$

An approximate Markovianity property holds, in that $f_s(t)$ is more strongly correlated with the far past when s is smaller, and so $H[f_{s_j}(t)|f_{s_1}(t), \dots, f_{s_{j-1}}(t)]$ is approximately $H[f_{s_j}(t)|f_{s_{j-1}}(t)]$. (This conditional entropy is an upper bound, achievable in the limit that $s_j - s_{j-1} \to \infty$, but which holds approximately when $s_j - s_{j-1}$ is very large.) Then, maximizing the joint entropy is approximately equivalent to maximizing $H[f_{s_j}(t)|f_{s_{j-1}}(t)]$, which is equivalent to $H[f_{s_j}(t)] - I[f_{s_j}(t); f_{s_{j-1}}(t)]$. To the extent that $H[f_{s_j}(t)]$ is roughly constant because s_j is so large that its statistics are governed mostly by the present input, we are left with a minimization of $I[f_{s_j}; f_{s_{j-1}}(t)]$– exactly the objective function of the previous section. Hence, the results about redundancy reduction hold for the efficient coding hypothesis, even though the objective functions are not exactly the same.

C. Recollecting the past

There are a number of ways to measure memory, but we focus on the simplest measure (memory capacity MC) that was invented to calibrate the performance of reservoir computers [13].

When the input has a single dominant timescale as in the overdamped Langevin equation, a glance at the expression for MC earlier suggests that MC will be maximized when the neuronal timescale is exactly matched to the input's timescale.

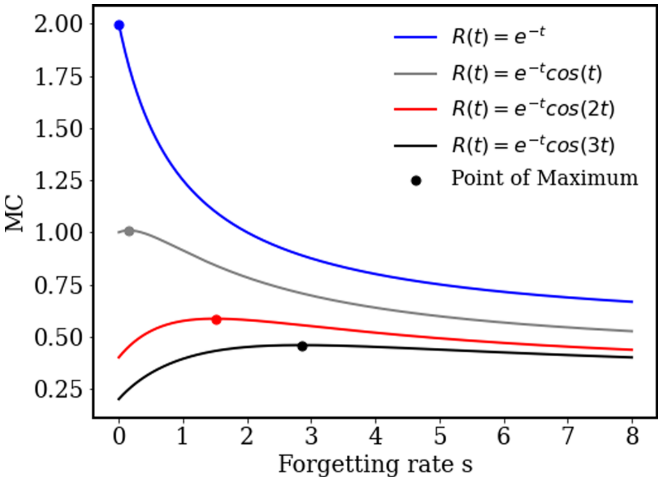

FIG. 1. A plot of memory capacity MC as a function of neuronal forgetting rate s for a single model time cell (a one-node linear reservoir) for some example autocorrelation functions. For inputs whose autocorrelation functions may be written as the sum (or integral) of exponentials, MC is maximized when the forgetting rate is 0 if all of the exponentials have sufficiently small oscillatory components. See series expansion in the Appendix E.

However, this is not the case. See the Appendix E. For input signals that do not have a significant oscillatory component, optimizing memory capacity means sending *all* forgetting rates to 0, so that at the limit, neurons are essentially estimates of the mean input symbol, even when the mean is zero. Such input includes both the overdamped Langevin equation and the naturalistic signals considered in this paper. A sketch of the argument is in the Appendix E.

Our finding here is similar to what was found for discrete-time reservoir computers [13]. An example is shown in Fig. 1, where we examine the behavior of memory capacity for some examples of overdamped and underdamped systems.

When the input has significant oscillatory components, then our argument for setting forgetting rates to 0 does not hold. For example, when the input moves according to an underdamped Langevin equation, MC is maximized at a nonzero s. As the frequency increases in the underdamped system, we find that the optimal forgetting rate generally increases as well [when $R(t) = e^t \cos(\omega t)$, the optimal s for MC as a function of ω is approximately piecewise linear]. Examining the values of $m(\tau)$ directly, this can perhaps be explained by the fact that remembering recent values very accurately is helpful in remembering the values in the period before. See Fig. 2.

The case of multiple neurons seems qualitatively similar to that of the single node case when examining scaling properties. We demonstrate this by examining the case of 10 neurons spaced equally between 0 and 1, scaling all of them by a factor α and examining the values of MC for networks generated in this manner. See Fig. 3. We still find that having a higher-frequency component in the underdamped system causes optimal forgetting rates that are greater than zero, unlike the overdamped case.

In conclusion, if an input has significant oscillatory components, then maximizing memory capacity MC may lead to nonzero forgetting rates. But if the input's

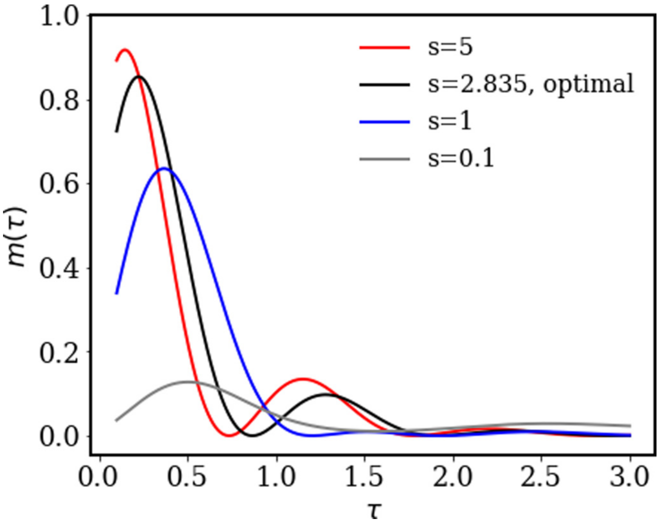

FIG. 2. A plot of the memory function $m(\tau)$, i.e., the squared correlation coefficient between network state and input a time τ in the past, for the autocorrelation function $R(t) = e^{-t}\cos(3t)$ and neuronal forgetting rates shown in the legend. Recall that MC $= \int_0^\infty m(\tau)d\tau$. By appropriately setting the neuronal forgetting rate, you can acquire information about both recent data and data farther in the past with some periodicity.

autocorrelation function seems to be the sum of decaying exponentials rather than a sum of oscillating and decaying exponentials—as seems to be true for naturalistic video [14]—then a series expansion in the Appendix E and numerical experiments presented here all suggest that maximizing memory capacity will yield forgetting rates that are as close to zero as possible.

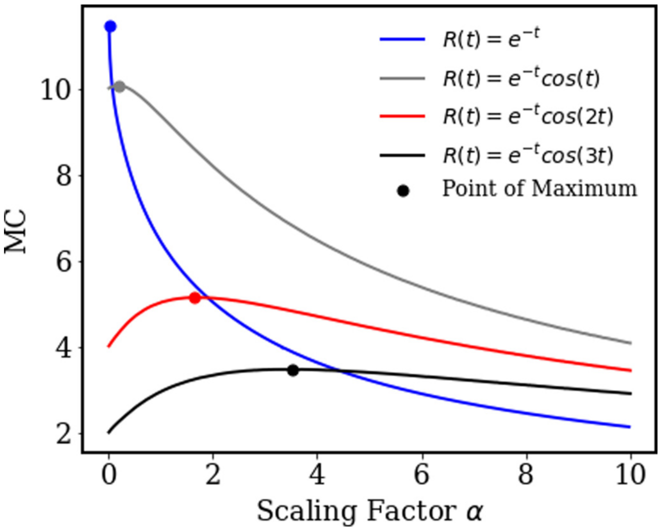

FIG. 3. Memory capacity as a function of scaling factor α for a network made up of neurons with forgetting rates $s_i = \alpha\frac{i}{10}$ for $i = 1, \dots, 10$, for various autocorrelation functions of the input. When oscillations are of high-enough frequency, the optimal α for maximal MC is nonzero.

D. Predicting the future

Finally, we might expect neurons to maximize something like a predictive capacity PC, as described earlier. As we detail in the Appendix C, perhaps surprisingly, linear recurrent neurons can beat nonlinear recurrent neurons at predicting input. As such, prediction already in part explains why time cells might want to perform an approximate Laplace transform.

Perhaps not surprisingly, PC is often optimized by setting $s \to \infty$, so that the current neuron acts best to only remember what it has just seen. This corresponds to the fact that the present signal usually has more information about future signals than past signals. To illustrate this phenomenon, we consider the impinging process to have an autocorrelation function of $R(t) = \frac{1}{2}e^{-\lambda_1|t|} + \frac{1}{2}e^{-\lambda_2|t|}$ and ask for the optimal forgetting rate of a single time cell s. There is a considerable region of values λ_1 and λ_2 for which this optimal forgetting rate is infinite. This corresponds to having a time cell that simply reads out the current value in the time series and has no memory.

This is not surprising from the perspective of understanding nearly Markovian signals. Recent stimuli convey more information than past ones, and so to predict optimally, one desires information about the most recent stimulus. But from another perspective, this is quite surprising. Earlier results in the static case [26] have shown that when predictive coding—minimization of error in predicting the stimulus—is used to optimize neuronal response properties, nontrivial neuronal weights without needing a time cell that has no recurrent connections. The key to the differences, in our opinion, are based in differences in setup. Rather than a supervised learning setting in which neuronal weights are tuned to send an input to a prescribed output, we consider a setting in which there are no weights *between* model time cells (based on Ref. [3]) and in which there is a learned mapping from infinite past inputs to a future input. The recurrent weight therefore represents not a connection to other neurons but a statement about about feature extraction: Which of the past inputs are most informative about the future input? And for many input time series, the most informative input is the most recent one.

Thus, we examine time cells with maximal predictive capacity in the presence of an additional cell which explicitly stores the present signal value. In other words, we imagine the situation shown in Fig. 4. Rather than having only cells that take an approximate Laplace transform by implementing a recurrent architecture, we allow for simply one cell to pass through all information about the present input. This second cell's architecture is entirely feedforward. It may be biologically relevant that time cells are more predictive when augmented by a single feedforward neuron. The optimal forgetting rate then is finite and increases with increasing λ_1, λ_2. See Appendix F.

As we have just seen, when the input moves according to an overdamped Langevin equation, the optimal neuronal timescale is some nontrivial function of the decay rate. Similarly, when the input moves according to an underdamped Langevin equation, we continue to see evidence of timescale matching. The optimal neuronal timescale is not the oscillatory timescale or the decay timescale but some nonobvious

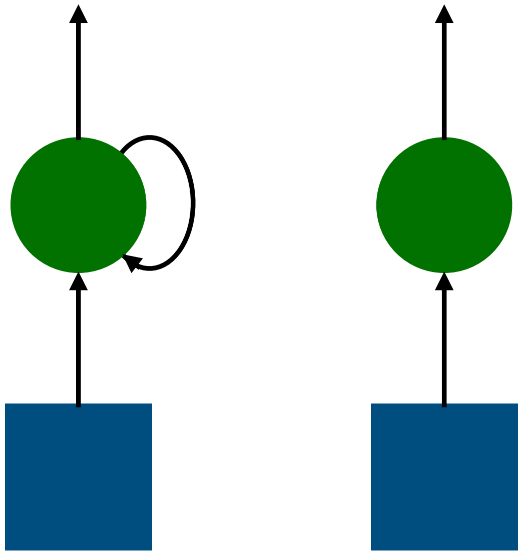

FIG. 4. A new biological setup that allows for increased predictive capacity. In both diagrams, the blue square represents the environment and the green circle represents the neuron's activation. At left, a recurrent neuron representing the current time cell model. At right, a feedforward neuron that we add to increase predictive capacity, which merely relays current environment information to the downstream region.

function of the two. The authors hope that the equations developed here might aid future efforts to discover this function.

Now we turn our attention to naturalistic signals that have power-law autocorrelation functions. We find that the optimal time constant of the additional time cell is a smoothly varying function of the power-law coefficient α, assuming that the input is naturalistic. For α between 1 and 1.789, PC has a local maximum for a relatively small s, i.e., ($s < 0.2$). For $\alpha < 1.56$, this local maximum is the global maximum. See Fig 5. Similarly, Fig. 6 in the Appendix F shows the optimal (PC maximizing) time constant of a neuron for the naturalistic stimulus. Roughly speaking, the optimal time constant of the neuron matches the time constant of the input, a form of timescale matching not seen in maximization of memory previously. At $\alpha = 1.6$, where the model time cell switches from an optimal intermediate forgetting rate to an optimally maximal forgetting rate.

IV. DISCUSSION

Surprisingly, the efficient coding hypothesis, maximization of memory, and redundancy reduction led to the same optimal model time cells when given naturalistic input—those that remembered the past as well as possible but were relatively useless for understanding the future. Predictive capacity favored time cells that forgot as much as possible, sans other constraints. When we included a hand-made neuron that stored the present input, time cells that maximized predictive capacity had timescales tuned to the environment. We considered the case of higher-dimensional input in the Appendix B, finding that our main conclusions were unaltered if spatial and temporal components of the spatiotemporal autocorrelation function were separable.

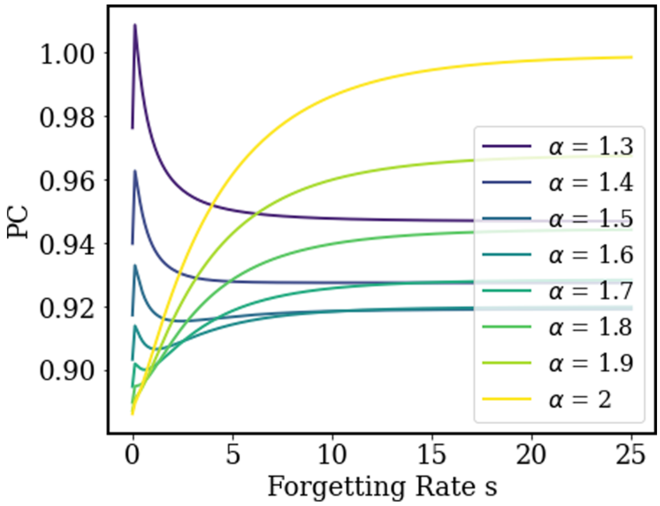

FIG. 5. Predictive capacity PC as a function of neuronal forgetting rate s for an input with an autocorrelation function $R(t) = \frac{1}{1+|t|^{\alpha}}$. PC($s$) is shown for various values of α. Intermediate forgetting rates s tend to maximize PC for small enough α, though there appears to be a phase transition in when an intermediate timescale is favored.

It is worth adding some cautionary words to these sweeping conclusions. These analyses depend on exactly what naturalistic input looks like. We followed Ref. [14]'s characterization of natural video. If the autocorrelation function of natural video were later found to be significantly oscillatory, then our results here suggest that maximization of memory capacity could explain observed neuronal forgetting rates. And some inputs to time cells might easily be oscillatory [27], and for such inputs, maximization of memory capacity would adequately explain nonzero and finite neuronal forgetting rates.

With that aside, to the best of our knowledge now, it seems as though prediction might be closest to the correct normative principle for time cells, as time cells have nonzero forgetting rates [3]. This may seem strange, as time cells are known for their ability to remember past events. However, one needs memory for prediction, and so optimizing for prediction does require memory of the "right" things [23]. For example, remembering what happened 100 days ago may provide far less useful information as to what will happen tomorrow than remembrance of the previous day's activities. A more reasonable objective function might be one that balances both memory and prediction, as memory has a coding cost, and prediction is desirable [11,22–24]. It would be difficult to find the appropriate objective function, however, without fitting to the data, and so we left this potentially thorny issue for future research.

It would also be interesting to see how our conclusions change when the predictive metric is no longer predictive capacity but predictive information [11,22,24], when memory is explicitly penalized while prediction is valued, when considering nonstationary stimuli, and when considering nonlinear reservoirs for which the central limit theorem does not hold (see Appendix C). In future research, we would also hope to better understand how redundancy, memory capacity, and predictive capacity vary with the number of neurons, as we ran into significant numerical integration difficulties here. Based on the work shown here, these changes would result in model time cells with nontrivial optimal time constants, as could be expected [22–24].

In conclusion, we have provided a quantitative framework for predicting optimal time constants of time cells that we hope will prove useful for those in neuroscience.

ACKNOWLEDGMENTS

The authors thank MIT Physics of Living Systems for its hospitality during visits and Marc Howard for enlightening conversations. A.H. and S.E.M. were funded by the Air Force Office of Scientific Research under Award No. FA9550-19-1-0411.

APPENDIX A: DERIVATION OF MC AND PC IN CONTINUOUS TIME

Say we have m time cells, where the ith time cell's activity is given in the main text:

$$f_i(t) = \int_{-\infty}^{t} x(t')e^{-s_i(t-t')}dt'. \tag{A1}$$

In this section, we calculate closed-form expressions for MC and PC, which were defined in the main text as well. Throughout, we assume stationarity, and we assume that the input's mean value is 0 and that its variance is 1.

Recall that, in this case, the memory function is

$$m(\tau) = p_\tau^\top C^{-1} p_\tau, \tag{A2}$$

where

$$p_\tau = \langle x(t-\tau)f(t)\rangle_t, \quad C = \langle f(t)f(t)^\top\rangle_t. \tag{A3}$$

We integrate this memory function from τ being ∞ to 0 to get MC, and from 0 to ∞ to get PC. Using our earlier expression for the activity f, we find that

$$\begin{aligned}
(p_\tau)_i &= \left\langle x(t-\tau)\int_{-\infty}^{t} x(t')e^{-s_i(t-t')}dt'\right\rangle_t, && \text{(A4)}\\
&= \int_{-\infty}^{t} \langle x(t-\tau)x(t')\rangle e^{-s_i(t-t')}dt', && \text{(A5)}\\
&= \int_{-\infty}^{t} R(t-\tau-t')e^{-s_i(t-t')}dt', && \text{(A6)}\\
&= \int_{0}^{\infty} R(t'-\tau)e^{-s_i t'}dt', && \text{(A7)}
\end{aligned}$$

and

$$\begin{aligned}
C_{ij} &= \langle f_i(t)f_j(t)\rangle_t, && \text{(A8)}\\
&= \left\langle \int_{-\infty}^{t} x(t')e^{-s_i(t-t')}dt' \int_{-\infty}^{t} x(t'')e^{-s_j(t-t'')}dt''\right\rangle_t, && \text{(A9)}\\
&= \int_{-\infty}^{t}\int_{-\infty}^{t} e^{-s_i(t-t')}e^{-s_j(t-t'')}R(t'-t'')dt'dt'', && \text{(A10)}\\
&= \int_{0}^{\infty}\int_{0}^{\infty} e^{-s_i t'}e^{-s_j t''}R(t'-t'')dt'dt''. && \text{(A11)}
\end{aligned}$$

At this point, we recall that

$$R(t) = \int_0^\infty F(\lambda)e^{-\lambda|t|}d\lambda. \tag{A12}$$

(One can also derive similar expressions by using the Fourier transform.) Plugging this in, we have

$$\begin{aligned}
C_{ij} &= \int_0^\infty\int_0^\infty e^{-s_i t'}e^{-s_j t''}\int_0^\infty F(\lambda)e^{-\lambda|t'-t''|}d\lambda dt'dt'', && \text{(A13)}\\
&= \int_0^\infty\int_0^\infty\int_0^\infty F(\lambda)e^{-s_i t'}e^{-s_j t''}e^{-\lambda|t'-t''|}dt'dt''d\lambda, && \text{(A14)}\\
&= \int_0^\infty F(\lambda)\frac{2\lambda+s_i+s_j}{(s+s_i)(s+s_j)(s_i+s_j)}d\lambda. && \text{(A15)}
\end{aligned}$$

When $\tau > 0$, we find that

$$(p_\tau)_i = \int_0^\infty \int_0^\infty F(\lambda) e^{-\lambda|t'-\tau|} d\lambda e^{-s_i t'} dt', \tag{A16}$$

$$= \int_0^\infty F(\lambda)\left(\int_0^\tau e^{-s_i t'} e^{-\lambda(\tau - t')} dt' + \int_\tau^\infty e^{-s_i t'} e^{-\lambda(t'-\tau)} dt'\right) d\lambda, \tag{A17}$$

$$= \int_0^\infty F(\lambda) \frac{2\lambda e^{-s_i\tau} - (\lambda + s_i)e^{-\lambda\tau}}{\lambda^2 - s_i^2} d\lambda. \tag{A18}$$

Otherwise, we find that

$$(p_\tau)_i = \int_0^\infty \int_0^\infty F(\lambda) e^{-\lambda|t'-\tau|} d\lambda e^{-s_i t'} dt', \tag{A19}$$

$$= \int_0^\infty F(\lambda) \frac{e^{-\lambda\tau}}{\lambda + s_i} d\lambda. \tag{A20}$$

Using our formula for the memory function $m(\tau)$ and for MC, PC, we have

$$\mathrm{MC} = \int_0^\infty m(\tau) d\tau, \tag{A21}$$

$$= \int_0^\infty \sum_{i,j} (p_\tau)_i (C^{-1})_{ij} (p_\tau)_j d\tau, \tag{A22}$$

$$= \sum_{i,j} (C^{-1})_{ij} \int_0^\infty \left[\int_0^\infty F(\lambda) \frac{2\lambda e^{-s_i\tau} - (\lambda + s_i)e^{-\lambda\tau}}{\lambda^2 - s_i^2} d\lambda\right]\left[\int_0^\infty F(\lambda) \frac{2\lambda' e^{-s_j\tau} - (\lambda' + s_j)e^{-\lambda'\tau}}{(\lambda')^2 - s_j^2} d\lambda'\right] d\tau$$

$$= \int_{\lambda=0}^\infty \int_{\lambda'=0}^\infty \left\{ \frac{F(\lambda)F(\lambda')}{\left(\lambda^2 - s_i^2\right)\left(\lambda'^2 - s_j^2\right)} \left[\frac{4\lambda\lambda'}{s_i + s_j} - \frac{2\lambda(\lambda' + s_j)}{s_i + \lambda'} - \frac{2\lambda'(\lambda + s_i)}{\lambda + s_j} + \frac{(\lambda + s_i)(\lambda' + s_j)}{\lambda + \lambda'}\right]\right\} d\lambda, d\lambda' \tag{A23}$$

$$= 1^\top (C^{-1} \odot D_{\mathrm{MC}}) 1, \tag{A24}$$

with D_{MC} having entries given in the main text, Eq. (11). Similarly,

$$\mathrm{PC} = \int_{-\infty}^0 m(\tau) d\tau, \tag{A25}$$

$$= \sum_{i,j} (C^{-1})_{ij} \int_{-\infty}^0 \left[\int_0^\infty F(\lambda) \frac{e^{-\lambda\tau}}{\lambda + s_i} d\lambda\right]\left[\int_0^\infty F(\lambda') \frac{e^{-\lambda\tau}}{\lambda' + s_i} d\lambda'\right] d\tau, \tag{A26}$$

$$= \sum_{i,j} (C^{-1})_{ij} \int_0^\infty \int_0^\infty \frac{F(\lambda)F(\lambda')}{(\lambda + \lambda')(\lambda + s_i)(\lambda' + s_j)} d\lambda d\lambda', \tag{A27}$$

$$= 1^\top (C^{-1} \odot D_{\mathrm{PC}}) 1, \tag{A28}$$

with D_{PC} given in Eq. (13) of the main text.

APPENDIX B: EXTENSION TO THE CASE OF MULTIDIMENSIONAL INPUT

Much of the sensory input that we receive, e.g., natural video, is high dimensional. To that end, we consider extending our analysis to the case of high-dimensional inputs, such that neuron i has activity f_i given by

$$\frac{df_i}{dt} = -s_i f_i + v_i^\top x. \tag{B1}$$

Now v_i is a vector such that the input x, also a vector, is converted into a scalar. In this way, it is relatively straightforward to alter the model of time cells.

However, the definitions of memory and predictive capacity need to be altered accordingly. We consider trying to predict $x_j(t + \tau)$ from $\vec{f}(t)$ and to remember $x_j(t - \tau)$ from $\vec{f}(t)$, and calculating PC_j and MC_j, respectively, by integrating the squared correlation coefficient over all τ. We then sum MC_j and PC_j over all dimensions j in order to get a final MC and PC.

Let $(p_\tau)_{i,j} = \langle x_j(t + \tau) f_i(t)\rangle_t$ and $C_{i,j} = \langle f_i(t) f_j(t)\rangle_t$, the latter as before, but the former with the additional index corresponding to the dimension of the input. Also, let $R_{j,k}(\tau) = \langle x_j(t) x_k(t - \tau)\rangle$ and $\overleftrightarrow{R}(\tau)$ be the matrix valued autocorrelation function with $R_{j,k}(\tau)$ as the entries. Some algebra similar to that of the Appendix above and not shown here gives

$$(p_\tau)_{i,j} = \int_0^\infty e^{-s_i t'} [\overleftrightarrow{R}(t' - \tau)\vec{v}_i]_j dt'. \tag{B2}$$

Note that $[\overleftrightarrow{R}(t' - \tau)\vec{v}_i]_j$ denotes the jth entry of the enclosed matrix-vector product. More straightforward algebra similar to that of the one-dimensional case in the previous Appendix gives

$$C_{i,j} = \int_0^\infty \int_0^\infty e^{-s_i t'} e^{-s_j t''} v_i^\top \overleftrightarrow{R}(t' - t'') v_j dt' dt'', \tag{B3}$$

Together, these determine the memory function for the jth element of the input:

$$m_j(\tau) = (\vec{p}^{\,j}_\tau)^\top C^{-1} \vec{p}^{\,j}_\tau \tag{B4}$$

and from there, the total memory capacity and predictive capacity:

$$\mathrm{MC} = \sum_j \int_{-\infty}^0 m_j(\tau) d\tau,\ \mathrm{PC} = \sum_j \int_0^\infty m_j(\tau) d\tau. \tag{B5}$$

In other words, we can understand the effect of spatial correlations on memory and predictive capacity by understanding its effects on p_τ and C.

We have just shown that only the (spatiotemporal) autocorrelation function is relevant for these metrics. And, furthermore, if the temporal component is constant or nearly constant across dimensions of the input, then we will find that $\overleftarrow{R}(\tau) = Sg(\tau)$, where S is the spatial covariance matrix and $g(\tau)$ represents the temporal component of autocorrelation function. In such a case, under some conditions specified below, the analysis of optimal forgetting rates will not be governed by spatial patterns but by $g(\tau)$. For instance, we find

$$(p_\tau)_{i,j} = (Sv_i)_j \int_0^\infty e^{-s_i t} g(t - \tau) dt, \tag{B6}$$

so that p_τ's τ dependence is strongly governed by $g(t)$, and

$$C_{i,j} = \int_0^\infty \int_0^\infty e^{-s_i t'} e^{-s_j t''} g(t' - t'') v_i^\top S v_j dt' dt''. \tag{B7}$$

Note that this splits into an element-wise product of a spatial component (with elements $v_i^\top S v_j$) and a temporal component [with elements $\int_0^\infty \int_0^\infty e^{-s_i t'} e^{-s_j t''} g(t' - t'') dt' dt''$]. Thus, when $\overleftarrow{R}(\tau)$ admits (or approximately admits) such a decomposition, we find that C and $(p_\tau)_j$ is roughly the same as that for a single pixel, and so analysis of one pixel is equivalent to an analysis of all pixels. Natural video may fall into this class of inputs after spatial processing by the visual cortex if receptive fields are sufficiently diffuse. More research will need to be conducted to elucidate the effects of the spatial component on maximization of MC or PC.

APPENDIX C: OPTIMALITY OF THE LAPLACE TRANSFORM

In this section, we consider a slightly more general model for how neuronal activity evolves:

$$\frac{df_i}{dt} = -\omega_i f_i + \sum_j J_{ij} \phi(f_j) + x(t). \tag{C1}$$

Due to the nonlinearity ϕ, the neuronal activities will no longer be Laplace transforms of the input.

This is intractable unless we make some assumptions. As such, we assume that there are a very large number of neurons N and that J_{ij} connections are randomly chosen from some distribution, where the mean is 0 and the variance is σ_J^2/N. Then $\eta_i(t) = \sum_j J_{ij}\phi(x_j)$ is normally distributed according to the central limit theorem. If this is the case, then the nonlinear term in the new evolution equation corresponds to Gaussian noise, and the now-linear system with Gaussian noise can still be analyzed.

We follow Ref. [28] in our treatment. We first characterize the noise properties:

$$\langle \eta_i(t) \rangle = \left\langle \sum_j J_{ij} \phi(x_j) \right\rangle = 0 \tag{C2}$$

and—assuming that N is so large that J_{ij} is roughly uncorrelated with $\phi(f_j)$, $\phi(f_i)$—we find

$$\langle \eta_i(t)\eta_j(t+\tau)\rangle = \left\langle \left(\sum_k J_{ik}\phi(x_k)\right)\left(\sum_{k'} J_{ik'}\phi(x_{k'})\right)\right\rangle, \tag{C3}$$

$$= \sum_{k,k'} \langle J_{ik}J_{jk'}\rangle \langle \phi(x_k)\phi(x_{k'})\rangle, \tag{C4}$$

$$= \sum_{k,k'} \frac{1}{N}\left(\delta_{i,j}\delta_{k,k'}\sigma_J^2\right)\langle \phi(x_k)\phi(x_{k'})\rangle, \tag{C5}$$

$$= \frac{\delta_{i,j}}{N}\sum_k \sigma_J^2 \langle \phi(x_k(t))\phi(x_k(t+\tau))\rangle, \tag{C6}$$

$$= \delta_{i,j}\sigma_J^2 \langle \phi(x(t))\phi(x(t+\tau))\rangle. \tag{C7}$$

Since the nonlinear term corresponds to Gaussian noise and the $\omega_i f_i$ term is linear, then given the input, f_i is normally distributed with mean 0 (since $\langle x\rangle = \langle \eta\rangle = 0$) and a covariance between $f_i(t)$ and $f_i(t+\tau)$ of $C(\tau)$:

$$C(\tau) := \langle f_i(t+\tau)f_i(t)\rangle. \tag{C8}$$

We then define

$$K(\tau) := \langle \phi(x(t))\phi(x(t+\tau))\rangle, \tag{C9}$$

which becomes

$$K(\tau) = \iint \phi(x)\phi(y)\frac{\exp\left[\frac{1}{2}\frac{C(0)x^2-2C(\tau)xy+C(0)y^2}{C(0)^2-C(\tau)^2}\right]}{2\pi\sqrt{|C(0)^2-C(\tau)^2|}}dxdy. \tag{C10}$$

If we can now find a relationship between $C(\tau)$ and $K(\tau)$, we will be able to solve for both. To do this, we return to the original evolution equation and solve explicitly for $x_i(t)$:

$$\dot{x}_i + \omega_i x_i = \eta_i(t) + f(t), \tag{C11}$$

$$e^{-\omega_i t}\frac{d}{dt}(e^{\omega_i t}x_i) = \eta_i + f, \tag{C12}$$

$$\frac{d}{dt}(e^{\omega_i t}x_i) = e^{\omega_i t}(\eta_i + f), \tag{C13}$$

$$e^{\omega_i t}x_i(t) = \int_{-\infty}^{t} e^{\omega_i s}[\eta_i(s) + f(s)]ds, \tag{C14}$$

$$x_i(t) = \int_{-\infty}^{t} e^{-\omega_i(t-s)}[\eta_i(s) + f(s)]ds. \tag{C15}$$

We know that x_i given f is normally distributed. Its mean is clearly 0. $C(\tau)$ is straightforwardly obtained:

$$\langle x_i(t)x_i(t+\tau)\rangle = \left\langle \left\{\int_{-\infty}^{t} e^{-\omega_i(t-s)}[\eta_i(s)+f(s)]ds\right\}\left\{\int_{-\infty}^{t+\tau} e^{-\omega_i(t+\tau-s)}[\eta_i(s)+f(s)]ds\right\}\right\rangle \tag{C16}$$

$$= \int_{-\infty}^{t}\int_{-\infty}^{t+\tau} e^{-\omega_i(t-s)}e^{-\omega_i(t+\tau-s')}\langle[\eta_i(s)+f(s)][\eta_i(s')+f(s')]\rangle dsds', \tag{C17}$$

$$C(\tau) = \int_{-\infty}^{t}\int_{-\infty}^{t+\tau} e^{-\omega_i(t-s)}e^{-\omega_i(t+\tau-s')}\left[\sigma_J^2 K(s-s') + R(s-s')\right]dsds', \tag{C18}$$

$$= \int_{-\infty}^{0}\int_{-\infty}^{\tau} e^{\omega_i s}e^{-\omega_i(\tau-s')}\left[\sigma_J^2 K(s-s') + R(s-s')\right]dsds', \tag{C19}$$

where R is the autocorrelation function of the input. In order to calculate PC and MC, we need

$$p_\tau = \langle x(t)f_i(t+\tau)\rangle, \tag{C20}$$

$$= \left\langle x(t)\int_{-\infty}^{t+\tau} e^{-\omega_i(t+\tau-s)}[\eta_i(s) + x(s)]ds\right\rangle, \tag{C21}$$

$$= \int_{-\infty}^{t+\tau} e^{-\omega_i(t+\tau-s)}\langle x(t)\eta_i(s)\rangle + \langle x(t)x(s)\rangle ds, \tag{C22}$$

$$= \int_{-\infty}^{\tau} e^{\omega_i(s-\tau)}R(s)ds. \tag{C23}$$

When $i \neq j$, we find that the activities of the two neurons are related via

$$\langle f_i(t) f_j(t+\tau)\rangle = \left\langle \left\{ \int_{-\infty}^{t} e^{-\omega_i(t-s)}[\eta_i(s)+x(s)]ds \right\} \left\{ \int_{-\infty}^{t+\tau} e^{-\omega_j(t+\tau-s')}[\eta_j(s')+x(s')]ds \right\} \right\rangle, \tag{C24}$$

$$= \int_{-\infty}^{t}\int_{-\infty}^{t+\tau} e^{-\omega_i(t-s)} e^{-\omega_j(t+\tau-s')}[\langle \eta_i(s)\eta_j(s')\rangle + \langle \eta_i(s)x(s')\rangle + \langle \eta_j(s)x(s')\rangle + R(s-s')]dsds'$$

$$= \int_{-\infty}^{t}\int_{-\infty}^{t+\tau} e^{-\omega_i(t-s)} e^{-\omega_j(t+\tau-s')} R(s-s')dsds', \tag{C25}$$

$$= \int_{-\infty}^{0}\int_{-\infty}^{\tau} e^{\omega_i s} e^{-\omega_j(\tau-s')} R(s-s')dsds'. \tag{C26}$$

This gives us our second relationship between $C(\tau)$ and $K(\tau)$.

Thus, we have

$$C(\tau) = \int_{-\infty}^{0}\int_{-\infty}^{\tau} e^{\omega_i s} e^{-\omega_i(\tau-s')}\big[R(s-s') + \sigma_J^2 K(s-s')\big]dsds', \tag{C27}$$

$$K(\tau) = \iint \phi(x)\phi(y)\frac{\exp\big[-\frac{1}{2}\frac{C(0)x^2-2C(\tau)xy+C(0)y^2}{C(0)^2-C(\tau)^2}\big]}{2\pi\sqrt{|C(0)^2-C(\tau)^2|}}dxdy, \tag{C28}$$

as the self-consistent equations.

In principle, that does it, but we seek some understanding from this math. To simplify things, we now assume that all the neurons have the same timescale ω, giving

$$(\vec{p}_\tau)_i = \int_{-\tau}^{\infty} e^{-\omega(s+\tau)}R(s)ds \tag{C29}$$

and

$$(\mathrm{Cov})_{ij} = \langle f_i(t) f_j(t)\rangle \tag{C30}$$

$$= \int_{-\infty}^{0}\int_{-\infty}^{0} e^{\omega_i s+\omega_j s'}\big[\sigma_J^2 K(s-s')\delta_{ij} + R(s-s')\big]dsds', \tag{C31}$$

$$= \int_{0}^{\infty}\int_{0}^{\infty} e^{-\omega(s+s')}R(s-s')dsds' + \sigma_J^2\delta_{i,j}\int_{0}^{\infty}\int_{0}^{\infty} e^{-\omega(s+s')}K(s-s')dsds', \tag{C32}$$

which is the covariance matrix for $f(t)$. Meanwhile, we still have

$$C(\tau) = \int_{-\infty}^{0}\int_{-\infty}^{\tau} e^{\omega(s+s'-\tau)}\big[R(s-s') + \sigma_J^2 K(s-s')\big]dsds', \tag{C33}$$

$$K(\tau) = \iint \phi(x)\phi(y)\frac{\exp\big[-\frac{1}{2}\frac{C(0)x^2-2C(\tau)xy+C(0)y^2}{C(0)^2-C(\tau)^2}\big]}{2\pi\sqrt{|C(0)^2-C(\tau)^2|}}dxdy, \tag{C34}$$

as the self-consistent equations.

Notice that

$$\mathrm{Cov} = R_0 1_N + \sigma_J^2 K_0 I_N, \tag{C35}$$

where 1_N is a $N \times N$ matrix of all 1's, and I_N is the $N \times N$ identity matrix. We also have

$$\vec{p}_\tau = R_\tau 1_N, \tag{C36}$$

where now 1_N is the length N vector of all 1's. Then we have

$$\mathrm{PC}_\tau = R_\tau^2 1_N^\top \big(R_0 1_N + \sigma_J^2 K_0 I_N\big)^{-1} 1_N, \tag{C37}$$

$$= R_\tau^2 1_N^\top \left[\sigma_J^{-2} K_0^{-1}\left(I_N + \frac{R_0}{\sigma_J^2 K_0}1_N\right)^{-1}\right]1_N, \tag{C38}$$

$$= \frac{R_\tau^2}{\sigma_J^2 K_0} 1_N^\top \left[\sum_{k=0}^{\infty}(-1)^k\left(\frac{R_0}{\sigma_J^2 K_0}1_N\right)^k\right]1_N, \tag{C39}$$

$$= \frac{R_\tau^2}{\sigma_J^2 K_0} 1_N^\top \left[\sum_{k=0}^{\infty} \left(-\frac{R_0 N}{\sigma_J^2 K_0} \right)^k 1_N \right] 1_N \tag{C40}$$

$$= \frac{R_\tau^2 N^2}{\sigma_J^2 K_0} \left[\sum_{k=0}^{\infty} \left(-\frac{R_0 N}{\sigma_J^2 K_0} \right)^k \right] = \frac{R_\tau^2 N^2}{\sigma_J^2 K_0} \left(1 + \frac{R_0 N}{\sigma_J^2 K_0} \right)^{-1}, \tag{C41}$$

where

$$R_\tau = \int_0^\infty e^{-\omega s} R(s-\tau) ds, \tag{C42}$$

$$R_0 = \int_0^\infty \int_0^\infty e^{-\omega(s+s')} R(s-s') ds ds', \tag{C43}$$

$$K_0 = \int_0^\infty \int_0^\infty e^{-\omega(s+s')} K(s-s') ds ds'. \tag{C44}$$

Clearly, we can increase N arbitrarily and arbitrarily increase $m(\tau)$. To get total PC, we integrate over τ and find the following:

$$\mathrm{PC} = \frac{N^2}{\sigma_J^2 K_0} \left(1 + \frac{R_0 N}{\sigma_J^2 K_0} \right)^{-1} \int_0^\infty R_\tau^2 d\tau, \tag{C45}$$

$$= \frac{N^2}{\sigma_J^2 K_0} \left(1 + \frac{R_0 N}{\sigma_J^2 K_0} \right)^{-1} \int_0^\infty \int_0^\infty \int_0^\infty e^{-\omega(s+s')} R(s-\tau) R(s'-\tau) ds ds' d\tau. \tag{C46}$$

If we look at how to maximize this, then we see that there is a critical parameter

$$\rho = N/\sigma_J^2 K_0, \tag{C47}$$

which gives

$$\mathrm{PC} = \frac{N\rho}{1 + R_0 \rho} \int_0^\infty \int_0^\infty \int_0^\infty e^{-\omega(s+s')} R(s-\tau) R(s'-\tau) ds ds' d\tau. \tag{C48}$$

PC is clearly maximized when $\rho \to \infty$, which can be achieved by: the number of nodes N going to infinity, the nonlinearity weight variances $\sigma_J^2 \to 0$, or the nonlinearity-controlled $K_0 \to 0$. When we are in that limit, we find

$$\mathrm{PC}_{\max} = N \frac{\int_0^\infty \int_0^\infty \int_0^\infty e^{-\omega(s+s')} R(s-\tau) R(s'-\tau) ds ds' d\tau}{\int_0^\infty \int_0^\infty e^{-\omega(s+s')} R(s-s') ds ds'}. \tag{C49}$$

Given that $\sigma_J \to 0$ is optimal, to maximize PC in this admittedly limited setup, we should opt to minimize nonlinearities.

APPENDIX D: RESULTS FOR A DIFFERENT MODEL TIME CELL

In this Appendix, the activity of time cell $f(s)$ at time t is a direct readout of $x(t-s)$. Assuming continuity of $x(t)$, this is equivalent to assuming that the activity of time cell $f(s)$ at time t is a direct readout of $\frac{1}{s_1+s_2} \int_{t-s_1}^{t-s_2} x(s') ds'$ for some s by the intermediate value theorem. When we refer to this neuron's timescale, we mean the delay time s.

If stationarity holds, then

$$I[f(s_i); f(s_{i+1})] = I[x(t-s_i); x(t-s_{i+1})], \tag{D1}$$

$$= I[x(0); x(s_{i+1}-s_i)]. \tag{D2}$$

Hence, equalizing redundancy between two neighboring neurons implies keeping $s_{i+1} - s_i$ a constant. This is emphatically *not* the logarithmic scaling of Ref. [16]. One can extend this argument to any measure of redundancy, as any measure of redundancy as described in Ref. [16] is a function of the joint probability distribution $P(f(s_i), f(s_{i+1}))$ and hence subject to the restrictions of stationarity. Note also that for almost all processes, $I[f(s_i); f(s_{i+1})]$ will tend to 0 as $s_{i+1} - s_i$ increases to infinity.

The efficient coding hypothesis argument in the main text applies equally well, and in some ways more rigorously, to these model time cells. Hence, redundancy reduction and efficient coding are equivalent for these model time cells as well.

In order to remember the entire past as well as possible, one would want to place the receptive fields of neurons as far back as possible, assuming that remembering what happened 3 years ago was as important as remembering what happened 1 day ago. We would therefore expect that optimally, $s_i \to \infty$.

Finally, for most signals, the recently observed signal is a better clue to the future than a previously observed signal, as discussed in the main text. We would therefore expect $s_i \to 0$ optimally.

APPENDIX E: ADDITIONAL ANALYSIS OF MEMORY CAPACITY

In this Appendix, we analyze memory capacity of optimal time cells for a few different types of input statistics. In all situations, we consider the one-node (one neuron, one time cell) case. For all these input types, we find that MC is maximized as $s \to 0$.

Let us start with the simplest possible input: a Markovian signal with timescale λ_0:

$$R(t) = e^{-\lambda_0 |t|}.$$

In this case,

$$\text{MC}(s) = \frac{4\lambda_0 + s}{2\lambda_0^2 + 2\lambda_0 s}.$$

The derivation of this is somewhat tricky, but achievable by using $F(\lambda) = \delta(\lambda - \lambda_0)$ and carefully keeping track of singularities. One can check that there is a maximum of MC as $s \to 0$.

When $R(t)$ is instead a mixture of timescales

$$R(t) = \tfrac{1}{2} e^{-\lambda_0 |t|} + \tfrac{1}{2} e^{-\lambda_1 |t|}$$

then

$$\begin{aligned}\text{MC}(s) = &\frac{4\lambda_0\lambda_1(\lambda_0+\lambda_1)^3 + \left(\lambda_0^4 + 17\lambda_0^3\lambda_1 + 36\lambda_0^2\lambda_1^2 + 17\lambda_0\lambda_1^3 + \lambda_1^4\right)s + 2(\lambda_0+\lambda_1)\left(\lambda_0^2 + 10\lambda_0\lambda_1 + \lambda_1^2\right)s^2}{4\lambda_0\lambda_1(\lambda_0+\lambda_1)(\lambda_0+s)(\lambda_1+s)(\lambda_0+\lambda_1+2s)} \\ &+ \frac{\left(\lambda_0^2 + 6\lambda_0\lambda_1 + \lambda_1^2\right)s^3}{4\lambda_0\lambda_1(\lambda_0+\lambda_1)(\lambda_0+s)(\lambda_1+s)(\lambda_0+\lambda_1+2s)}.\end{aligned}$$

And for the case that $F(\lambda) = \{ {\frac{1}{b} \atop 0} \quad {0 < x < b \atop \text{otherwise}}$, we find that memory is maximized as $s \to 0$, $\text{MC} \to \infty$. This is a special case of the class of $F(\lambda)$ which takes on the form

$$F(\lambda) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases},$$

which produces autocorrelation functions of the form

$$R(t) = \frac{2e^{-(a+b)|t|}\left[e^{a|t|}(1 + b|t|) - e^{b|t|}(1 + a|t|)\right]}{(a^2 - b^2)t^2}.$$

These correspond to autocorrelation functions produced by averaging over an interval of characteristic timescales. In this case, $\lim_{s \to 0^+} \text{MC}$ is available in closed form:

$$\lim_{s \to 0^+} \text{MC} = \frac{2\log\left(\frac{a}{b}\right)}{a - b}.$$

Setting $a \to 0$ shows the logarithmic divergence of MC.

For an argument as to why MC is optimized by sending $s \to 0$ in general when $F(\lambda)$ is supported on the real numbers, consider the one-node case. Then D_{MC} reduces to

$$D_{\text{MC}} = \int_0^\infty \int_0^\infty F(\lambda)F(\lambda') \frac{2\lambda + 2\lambda' + s}{s(\lambda + \lambda')(\lambda + s)(\lambda' + s)} d\lambda d\lambda'$$

and has the series expansion centered at $s = 0$

$$\int_0^\infty \int_0^\infty \left\{ F(\lambda)F(\lambda')\left[\frac{2}{\lambda\lambda' s} - \frac{2\lambda^2 + 3\lambda\lambda' + 2\lambda'^2}{\lambda^2\lambda'^2(\lambda + \lambda')} + O(s)\right]\right\} d\lambda d\lambda'.$$

Multiplying by $C(s)^{-1}$, we therefore have that

$$\text{MC} = \int_0^\infty \int_0^\infty \left\{ \frac{F(\lambda)F(\lambda')}{C(s)}\left[\frac{2}{\lambda\lambda' s} - \frac{2\lambda^2 + 3\lambda\lambda' + 2\lambda'^2}{\lambda^2\lambda'^2(\lambda + \lambda')} + O(s)\right]\right\} d\lambda d\lambda'.$$

The two terms which, in s, have the largest contributions $[\frac{2}{\lambda\lambda' s} - \frac{2\lambda^2+3\lambda\lambda'+2\lambda'^2}{\lambda^2\lambda'^2(\lambda+\lambda')}]$, are both maximized by setting $s \to 0$. It is clear that for $\lambda, \lambda' > 0$, $\frac{2}{\lambda\lambda' s}$ increases unboundedly by decreasing $s \to 0$. The constant coefficient in this expansion $-\frac{2\lambda^2+3\lambda\lambda'+2\lambda'^2}{\lambda^2\lambda'^2(\lambda+\lambda')}$ is negative. Clearly, $\lambda^2\lambda'^2(\lambda + \lambda') > 0$, and $2\lambda^2 + 3\lambda\lambda' + 2\lambda'^2$ is a positive definite quadratic form, making

$$\frac{2\lambda^2 + 3\lambda\lambda' + 2\lambda'^2}{\lambda^2\lambda'^2(\lambda + \lambda')}$$

positive. Hence, MC is maximized as $s \to 0$.

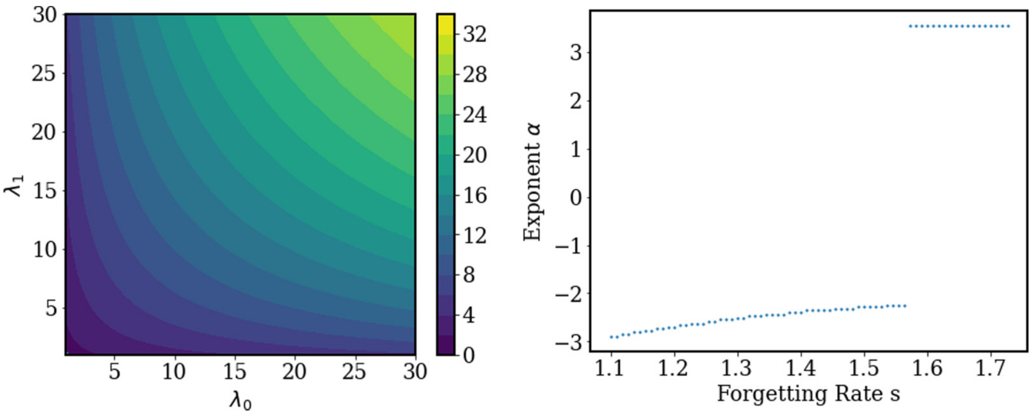

FIG. 6. Left: A plot of the optimal forgetting rate s for an environment with $R(t) = \frac{1}{2}(e^{-\lambda_1|t|} + e^{-|\lambda_2|t})$, with λ_1 and λ_2 on the x axis and y axis, respectively, and the value of s which maximizes PC denoted by color. Right: A plot of the log of the optimal forgetting rate, $\log s$, as a function of the parameter α for input autocorrelation functions of the form $R(t) = \frac{1}{1+|t|^\alpha}$. Note the sudden increase at $\alpha = 1.6$ from optimal s being finite to optimal s being the maximal s we searched over.

APPENDIX F: MAXIMIZING PREDICTIVE CAPACITY

In Fig. 6 (left), we show the optimal forgetting rate of a model time cell when impinged on by an input with autocorrelation function $R(t) = \frac{1}{2}(e^{-\lambda_1|t|} + e^{-|\lambda_2|t})$. This model time cell was augmented with another cell that stored the present value.

In Fig. 6 (right), we show the optimal forgetting rate of a model time cell when impinged on by an input with autocorrelation function $R(t) = \frac{1}{1+|t|^\alpha}$. This model time cell was augmented with another cell that stored the present value. Note that the optimal forgetting rate attains some intermediate, nontrivial value for most α, indicating timescale matching. Furthermore, note that there appears to be a phase transition at $\alpha \approx 1.6$ at which point the model time cell desires to have a maximal forgetting rate. The α's in our environment tend to be between 1 and 2 [14].

[1] H. Eichenbaum, Time cells in the hippocampus: A new dimension for mapping memories, Nat. Rev. Neurosci. **15**, 732 (2014).

[2] B. J. Kraus, R. J. Robinson II, J. A. White, H. Eichenbaum, and M. E. Hasselmo, Hippocampal "time cells": Time versus path integration, Neuron **78**, 1090 (2013).

[3] M. W. Howard, C. J. Macdonald, Z. Tiganj, K. H. Shankar, Q. Du, M. E. Hasselmo, and H. Eichenbaum, A unified mathematical framework for coding time, space, and sequences in the hippocampal region, J. Neurosci. **34**, 4692 (2014).

[4] M. Hermans and B. Schrauwen, Memory in linear recurrent neural networks in continuous time, Neural Netw. **23**, 341 (2010).

[5] M. Lukoševičius and H. Jaeger, Reservoir computing approaches to recurrent neural network training, Comput. Sci. Rev. **3**, 127 (2009).

[6] H. Jaeger, The "echo state" approach to analyzing and training recurrent neural networks-with an erratum note. German National Research Center for Information Technology GMD Technical Report, Vol. 148 (2001), Bonn, Germany.

[7] L. Grigoryeva and J.-P. Ortega, Echo state networks are universal, Neural Networks **108**, 495 (2018).

[8] C. J. Macdonald, K. Q. Lepage, U. T. Eden, and H. Eichenbaum, Hippocampal "time cells" bridge the gap in memory for discontiguous events, Neuron **71**, 737 (2011).

[9] M. Jazayeri and M. N. Shadlen, A neural mechanism for sensing and reproducing a time interval, Curr. Biol. **25**, 2599 (2015).

[10] H. B. Barlow, Possible principles underlying the transformation of sensory messages, in *Sensory Communication*, edited by W. Rosenblith (MIT Press, Cambridge, Mass., 1961), pp. 217–234, Ch. 13.

[11] S. E. Palmer, O. Marre, M. J. Berry, and W. Bialek, Predictive information in a sensory population, Proc. Natl. Acad. Sci. U.S.A. **112**, 6908 (2015).

[12] P. W. Glimcher, Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis, Proc. Natl. Acad. Sci. U.S.A. **108**, 15647 (2011).

[13] S. Marzen, Difference between memory and prediction in linear recurrent networks, Phys. Rev. E **96**, 032308 (2017).

[14] D. W. Dong and J. J. Atick, Statistics of natural time-varying images, Netw., Comput. Neural Syst. **6**, 345 (1995).

[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley-Interscience, New York, 1991).

[16] M. W. Howard and K. H. Shankar, Neural scaling laws for an uncertain world, Psychol. Rev. **125**, 47 (2018).

[17] H. Jaeger, *Short Term Memory in Echo State Networks*, Vol. 5 (GMD-Forschungszentrum Informationstechnik, Bremen, Germany, 2001).

[18] O. L. White, D. D. Lee, and H. Sompolinsky, Short-Term Memory in Orthogonal Neural Networks, Phys. Rev. Lett. **92**, 148102 (2004).

[19] J. Boedecker, O. Obst, J. T. Lizier, N. M. Mayer, and M. Asada, Information processing in echo state networks at the edge of chaos, Theory Biosci. **131**, 205 (2012).

[20] I. Farkaš, R. Bosák, and P. Gergel, Computational analysis of memory capacity in echo state networks, Neural Netw. **83**, 109 (2016).

[21] P. Barančok and I. Farkaš, Memory capacity of input-driven echo state networks at the edge of chaos, in *Proceedings of the International Conference on Artificial Neural Networks* (Springer, Berlin, 2014), pp. 41–48.

[22] S. Still, J. P. Crutchfield, and C. J. Ellison, Optimal causal inference: Estimating stored information and approximating causal architecture, Chaos **20**, 037111 (2010).

[23] S. E. Marzen and J. P. Crutchfield, Predictive rate-distortion for infinite-order markov processes, J. Stat. Phys. **163**, 1312 (2016).

[24] M. Chalk, O. Marre, and G. Tkačik, Toward a unified theory of efficient, predictive, and sparse coding, Proc. Natl. Acad. Sci. U.S.A. **115**, 186 (2018).

[25] G. D. Brown, S. Yamada, and T. J. Sejnowski, Independent component analysis at the neural cocktail party, Trends Neurosci. **24**, 54 (2001).

[26] J. C. R. Whittington and R. Bogacz, An approximation of the error backpropagation algorithm in a predictive coding network with local Hebbian synaptic plasticity, Neural Comput. **29**, 1229 (2017).

[27] K. Mizuseki, A. Sirota, E. Pastalkova, and G. Buzsáki, Theta oscillations provide temporal windows for local circuit computation in the entorhinal-hippocampal loop, Neuron **64**, 267 (2009).

[28] K. Rajan, L. F. Abbott, and H. Sompolinsky, Stimulus-dependent suppression of chaos in recurrent neural networks, Phys. Rev. E **82**, 011903 (2010).