A Hopfield network model of neuromodulatory arousal state

Mohammed Abdal Monium Osman[1,2] **Kai Fox**[1,2] **Joshua Isaac Stern**[1]
[1]Department of Neurobiology, Harvard Medical School
[2]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University
Boston, MA, USA
mohammed_osman@g.harvard.edu, kai_fox@hms.harvard.edu, sternj@g.harvard.edu

Abstract

Neural circuits display both input-driven activity that is necessary for real-time control of behavior and internally generated activity that is necessary for memory, planning, and other cognitive processes. A key mediator between these intrinsic and evoked dynamics is arousal, an internal state variable that determines an animal's level of engagement with its environment. It has been hypothesized that arousal state acts through neuromodulatory gain control mechanisms that suppress recurrent connectivity and amplify bottom-up input. In this paper, we instantiate this longstanding idea in a continuous Hopfield network embellished with a gain parameter that mimics arousal state by suppressing recurrent interactions between the network's units. We show that dynamics capturing some essential effects of arousal state at the neural and cognitive levels emerge in this simple model as a single parameter—recurrent gain—is varied. Using the model's formal connections to the Boltzmann machine and the Ising model, we offer functional interpretations of arousal state rooted in Bayesian inference and statistical physics. Finally, we liken the dynamics of neuromodulator release to an annealing schedule that facilitates adaptive behavior in ever-changing environments. In summary, we present a minimal neural network model of arousal state that exhibits rich but analytically tractable emergent behavior and reveals conceptually clarifying parallels between arousal state and seemingly unrelated phenomena.

1 Introduction

Neural circuits display both input-driven dynamics that are necessary for the real-time control of behavior and internally generated dynamics that are necessary for memory, planning, and other cognitive processes. These forms of activity often complement one another, as when a sensation triggers the recall of a relevant memory. However, they can also conflict, as when pleasant recollections leave a daydreamer oblivious to their surroundings. In animals, a potent regulator of this balance between intrinsic and evoked dynamics is arousal state [28, 17]. Arousal is an internal state variable that regulates the organism's overall level of activity and sensitivity to external input. During high arousal states, even weak sensory inputs can achieve system-wide control of perception and behavior. In contrast, during low arousal states, such as sleep, neural dynamics are decoupled from external input and dominated by spontaneous activity that is thought to reflect the querying of internal world models [28, 22]. It has been hypothesized that arousal state exerts these effects through gain control mechanisms that suppress the recurrent connections that generate spontaneous activity and amplify the feedforward connections that convey external input, an idea consistent with the known effects of acetylcholine, a potent neuromodulator of arousal state [9, 7, 6].

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

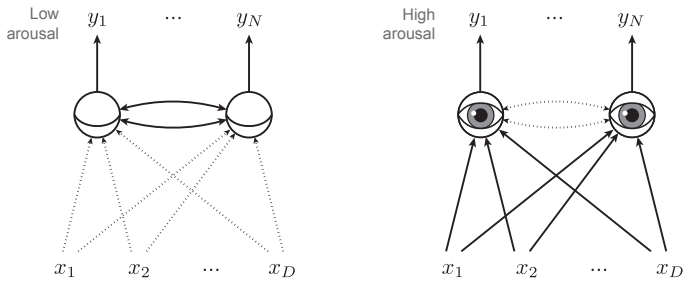

Figure 1: Network schematic illustrating low arousal state dominated by recurrent interactions (left) and high arousal state dominated by bottom-up sensory input (right).

In this paper, we study this interplay between arousal state, spontaneous activity, and stimulus-evoked dynamics in a continuous Hopfield network [13] embellished with a gain parameter that mimics arousal state by suppressing recurrent interactions between the network's units [9]. We show that varying the network's recurrent gain induces a phase transition from a low arousal state dominated by intrinsic dynamics, memory retrieval, and Bayesian priors, to a high arousal state dominated by evoked dynamics and sensory input. To build further intuition, we highlight parallels between our model's behavior and physical phenomena studied in statistical mechanics. Turning to biology, we then discuss how the dynamics of neuromodulator release flexibly balance a tradeoff between the reliable sensory responses characteristic of high arousal states and the cognitive functions enabled by strong recurrence. In summary, we present a minimal neural network model whose analytical tractability and emergent dynamics shed light on the functional role of arousal state in neural computation.

2 Related Work

A rich body of prior work has explored how neuromodulatory gain control mechanisms can alter the balance between intrinsic and evoked dynamics in neural circuits. For example, [9] put forth a circuit model of olfactory associative learning in piriform cortex (very similar to the Hopfield network analyzed in the present study) in which acetylcholine determines the multiplicative gain of the circuit's recurrent connections and explored how it interacts with synaptic plasticity. In addition, a recent study of auditory cortex has shown that modulating the gain on background inputs to excitatory cells can induce a transition from multistable to unistable attractor dynamics and explain the inverted-U relationship between arousal state and task performance in perceptual decision making [21]. Finally, in the active inference literature, previous work has argued that acetylcholine sets the balance between the likelihood and prior terms in Bayesian inference by modulating the response gain of pyramidal cells thought to convey sensory prediction errors to downstream brain areas [18].

Outside the context of arousal state, classic work on recurrent neural networks with random synaptic couplings has demonstrated a phase transition from unistable attractor dynamics to chaos as neural gain is increased, an insight relevant to our model despite its symmetric connectivity [26]. Finally, other work has employed deterministic annealing approaches, in which neural gain is gradually increased as some computation progresses to encourage the network to settle into an attractor state representing a good solution to an optimization problem, rather than getting trapped in a local minimum, another concept that will prove to be relevant [19, 2].

3 Network Dynamics

In this paper, we study a continuous Hopfield network with N recurrently coupled units whose time varying activations are governed by the differential equation

$$\frac{d\mathbf{y}}{dt} = -\mathbf{y} + f\left(\frac{1}{\alpha}\mathbf{M}\mathbf{y} + \mathbf{W}\mathbf{x}\right), \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^N$ denotes the vector of neural activations , $\mathbf{x} \in \mathbb{R}^D$ denotes the stimulus pattern, $\mathbf{M} \in \mathbb{R}^{N\times N}$ denotes the symmetric and zero-diagonal recurrent connectivity matrix, $\mathbf{W} \in \mathbb{R}^{N\times D}$

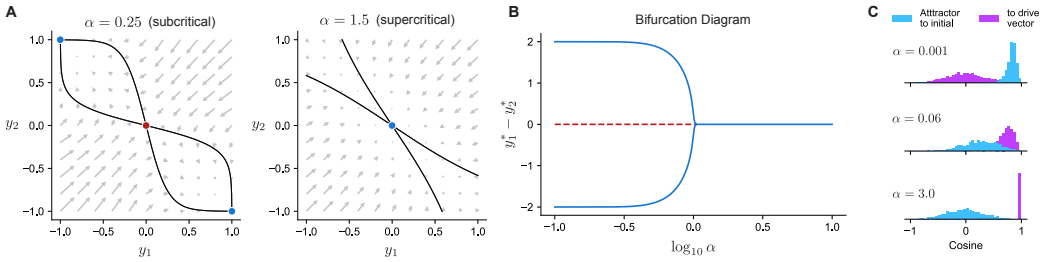

Figure 2: Network undergoes bifurcation from multistable associative dynamics to unistable input driven dynamics as alpha passes a critical point. (a) Phase portraits for the dynamics of 2-unit network under zero external stimulus. Black lines indicate nullclines of the dynamics, whose intersections result in the stable (blue) and unstable (red) fixed points. (b) Difference of unit activations at stable (blue, solid) and unstable (red, dashed) fixed points as a function of recurrence gain. (c) Distribution of angles between initialization point of integration and eventual attractor state for uniform random initializations and drive vectors $\mathbf{Wx}$ (purple) as well as between attractor and drive vector (blue).

denotes the feedforward input weights, and $f(\cdot)$ is an element-wise tanh nonlinearity. Due to their symmetric connectivity, networks of this form exhibit simple point attractor dynamics that are commonly used to model associative memory [13, 12].

In addition to the usual variables, the dynamics are determined by an "arousal" parameter $\alpha \in \mathbb{R}_+$ that divisively suppresses recurrent interactions between the network's units (Figure 1). In brief, when α is small, recurrence is strong and intrinsic dynamics dominate, while when α is large, recurrence is weak and the network state is controllable by bottom-up sensory input.

More formally, as α approaches 0, the dynamics approach

$$\lim_{\alpha \to 0} \frac{d\mathbf{y}}{dt} = -\mathbf{y} + \text{sign}(\mathbf{My}). \tag{2}$$

In this low arousal, sleep-like limit, sensory input has no effect on the network dynamics and the attractor states of the system correspond to the memory patterns stored in its recurrent weights. Conversely, as arousal diverges, the dynamics approach

$$\lim_{\alpha \to \infty} \frac{d\mathbf{y}}{dt} = -\mathbf{y} + f\left(\mathbf{Wx}\right). \tag{3}$$

In this high arousal limit, the dynamics are completely controlled by the input, the activity state of the network exponentially relaxes towards the stimulus-evoked fixed point at $f\left(\mathbf{Wx}\right)$, and the system effectively reduces to a feedforward logistic function of the input.

Note that the network is multistable in the low arousal limit and unistable in the high arousal limit. As this qualitative change in the fixed point structure entails, these regimes are separated by a bifurcation (or series of bifurcations). While a closed form expression for the initial bifurcation's critical point is unavailable when input is present, in the absence of input, it is solely determined by the largest eigenvalue of the connectivity matrix:

$$\alpha^* = \lambda_{\max}(\mathbf{M}), \tag{4}$$

which is proven in the Appendix.

Further intuition can be obtained by considering a simple instantiation of the model consisting of two mutually inhibitory neurons (Figure 2). At a given value of α, to visualize the dynamics of this two dimensional system, we display its flow field and overlay each unit's nullcline, the set of states for which the time derivative of that unit's activity is zero. Each nullcline is a sigmoidal curve whose steepness is determined by α and whose translational position is determined by the corresponding unit's net input. For a fixed input, at low α, the nullclines are steep and intersect at three points, resulting in two attractor states and a saddle point, while at high α the nullclines are shallow and intersect at only one point, which becomes the global attractor state of the system.

4 Energy Function

The dynamics of the network can also be construed as descent along an energy landscape whose local minima define the attractor states. This energy landscape is defined by the Lyapunov function

$$F(\mathbf{y}|\mathbf{x};\alpha) = -\frac{1}{2\alpha}\mathbf{y}^\top \mathbf{M}\mathbf{y} - \frac{1}{2}\mathbf{y}^\top \mathbf{W}\mathbf{x} - \sum_{i=1}^{N} H_2^{(e)}\left(\frac{y_i+1}{2}\right), \tag{5}$$

where $H_2^{(e)}(p) = -p\log p - (1-p)\log(1-p)$ [13]. This equation for the Lyapunov function of the network dynamics also equals the mean-field variational free energy for a probabilistic generative model called a Boltzmann machine [11, 1, 15]. While we will defer a more detailed overview of this connection to the Appendix, in brief, this equivalence allows us to cast the dynamics of the system as an approximate Bayesian inference procedure in which the (variational free) energy monotonically decreases as the system approaches an attractor state representing the locally optimal inference about the latent causes of its sensory input. In this framing, the network state $\mathbf{y}$ encodes a factorized probability distribution $q(\mathbf{z})$ over the generative model's latent variables and the energy $F(\mathbf{y}|\mathbf{x};\alpha)$ quantifies (or more precisely, upper-bounds) the discrepancy between $q(\mathbf{z})$ and the true Bayesian posterior $p(\mathbf{z}|\mathbf{x})$ [3, 5]. With these considerations in mind, we can equivalently write 5 as

$$\mathcal{F}[q|\mathbf{x};\alpha] = \frac{1}{\alpha}\langle E(\mathbf{z})\rangle_q + \langle E(\mathbf{x}|\mathbf{z})\rangle_q - \mathcal{S}[q] = F(\mathbf{y}|\mathbf{x};\alpha). \tag{6}$$

To minimize this variational free energy, the allocation of posterior probability—as represented by the activity state of the network—jointly favors states that have high prior probability (through $\langle E(\mathbf{z})\rangle_q = -\frac{1}{2}\mathbf{y}^\top\mathbf{M}\mathbf{y}$) and compatibility with the data (through $\langle E(\mathbf{x}|\mathbf{z})\rangle_q = -\frac{1}{2}\mathbf{y}^\top\mathbf{W}\mathbf{x}$), while also maximizing the entropy of the posterior (through $\mathcal{S}[q] = \sum_{i=1}^{N} H_2^{(e)}\left(\frac{y_i+1}{2}\right)$). As this decomposition reveals, the arousal level α determines the weakness of the prior used in inference. As α goes to infinity, the contribution of the prior term vanishes and the posterior probability of a state merely reflects its compatibility with the data. In this limit, inference reduces to maximum likelihood estimation regularized by an entropy term, and the underlying generative model reduces to a restricted Boltzmann machine. Conversely, as α goes to 0, the likelihood and entropy terms become negligible and the posterior probability of a state depends only on the prior. Varying α therefore allows the network to interpolate between sampling from its prior and performing inference over the data.

We can gain further intuition by considering some analogies to physical systems. In the absence of input, equation 6 reduces to $\mathcal{F}[q|\mathbf{x};\alpha] = \frac{1}{\alpha}\langle E(\mathbf{z})\rangle_q - \mathcal{S}[q]$ and α functions as a temperature parameter. Correspondingly, in the high arousal regime, the system minimizes variational free energy by maximizing the entropy of its belief distribution and thereby remaining agnostic about the unobserved state of the world. In contrast, in the low arousal regime, the system's strong priors instead favor the allocation of all belief to an arbitrary state that is likely under the generative model.

In the presence of input, temperature becomes an imperfect metaphor and α instead corresponds to the inverse of interaction strength in the Ising model. The Ising model is a lattice of interacting binary spin variables that represents an idealized magnetic solid experiencing the effects of temperature and an external field, and its interaction strength parameterizes the energetic cost of misalignment between neighboring spins [14, 20]. At subcritical interaction strengths, couplings within the lattice are too weak to buffer an arbitrary state against thermal fluctuations or an oppositely oriented applied field, so the solid's magnetization passively reflects that of the world around it. Above a critical point, in contrast, strong internal interactions lead to a global symmetry breaking event in which most of the spins adopt some arbitrary orientation that is subsequently hard to reverse, even if it is misaligned with the external field. Surprisingly, this suggests that the unreliable sensory responses and internally generated imagery of low arousal states like daydreaming are analogous to the hysteresis and spontaneous symmetry breaking properties of ferromagnetic solids.

5 Discussion

As these results highlight, while the recurrent dynamics that produce robust pattern completion, persistent activity, and structured spontaneous fluctuations enable cognitive functions like associative recall, working memory, and imagination, they do so at the expense of reliable responses to external

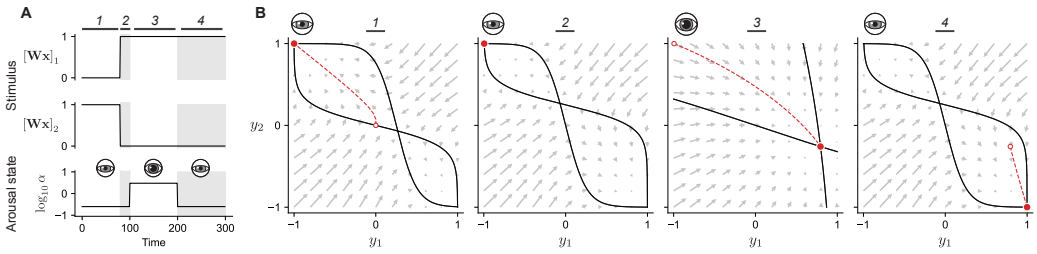

Figure 3: Dynamic annealing schematic: Momentary suppression of recurrence allows escape from suboptimal local minima, potentially explaining pupillary dynamics during perceptual switching. (a) Input drives to a 2-unit network and dynamic arousal state chosen to switch to a high value for a short period after stimulus switching. (b) Recurrence gain allows for dynamic control of hysteresis and annealing. Network state (red) during the example stimulus, first maintaining current activity invariant to stimulus switch (2), then adapting to new stimulus characteristics during high-arousal state (3) before collapsing to a corresponding intrinsic attractor (4).

input. To cope with this fundamental trade-off, neuromodulatory mechanisms must therefore match the balance between intrinsic and evoked dynamics to the current needs of the organism [24, 16, 25].

On slow time scales, this balance is manifest in sleep-wake cycles, and on intermediate time scales, the alternation between daydreaming and engaged task performance has a similar flavor. On fast time scales, moment-to-moment fluctuations in levels of acetylcholine and norepinephrine—as indexed by pupil dynamics—also have analogous effects [17, 23]. For example, pupil dilation accompanies perceptual flips in the interpretation of an ambiguous stimulus like the Necker cube, potentially indicating a momentary suppression of recurrence that permits switching between attractor states [27]. Conversely, pupil constriction accompanies hippocampal and cortical replay events, suggesting low neuromodulatory tone and strong recurrence during moments of memory recall [4]. Finally, models of the theta rhythm have argued that oscillating levels of acetylcholine in the hippocampus define a fast cycle of stimulus encoding and memory retrieval [8]. Taken together, these phenomena suggest that alternation between these modes structures neural computation across scales.

We propose that the dynamics of arousal state can be understood through the lens of yet another concept from statistical mechanics: namely, annealing. Annealing is the gradual reduction of a material or search algorithm's temperature as an optimization process unfolds, and it is premised on the idea that smoothly deforming a system's energy landscape in this manner can enable it to relax into a globally optimal configuration rather than getting stuck in a local minimum [2]. In neural circuits such as the continuous Hopfield model presented here, neural gain essentially plays the same role as temperature in annealing algorithms [19]. Unlike more rudimentary annealers, however, animals confront ever changing environments and therefore optimization problems. As such, the brain's neuromodulatory dynamics may constitute a remarkably rich and flexible annealing schedule that facilitates adaptive behavior and information processing under such circumstances (Figure 3).

6 Conclusion

In summary, we analyzed a minimal neural network model of arousal state premised on the longstanding idea that arousal state modulates cognitive and sensory processing by setting the balance between recurrent and feedforward connectivity. We showed that this model recapitulates several hallmarks of arousal state despite its simplicity and interpreted this behavior through the lenses of associative memory, Bayesian inference, and statistical mechanics. Our results shed light on the functional significance of arousal state and recurrent gain modulation in brain circuits, potentially suggesting that similar mechanisms could enhance the flexibility of neural computation in AI systems.

Acknowledgments

Many thanks to Mark Andermann, Ila Fiete, Akshay Jaggi, Cengiz Pehlevan, Jacob Zavatone-Veth, Stelios Smirnakis, Ganna Palagina, and Gord Fishell for enlightening discussions about these ideas.

References

[1] James R Anderson and Carsten Peterson. A mean field theory learning algorithm for neural networks. *Complex Systems*, 1(5):995–1019, 1987.

[2] Dimitris Bertsimas and John Tsitsiklis. Simulated annealing. *Statistical science*, 8(1):10–15, 1993.

[3] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[4] David J Foster. Replay comes of age. *Annual review of neuroscience*, 40(1):581–602, 2017.

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

[6] Ziv Gil, Barry W Connors, and Yael Amitai. Differential regulation of neocortical synapses by neuromodulators and activity. *Neuron*, 19(3):679–686, 1997.

[7] Michael E Hasselmo. Neuromodulation and cortical function: modeling the physiological basis of behavior. *Behavioural brain research*, 67(1):1–27, 1995.

[8] Michael E Hasselmo, Clara Bodelón, and Bradley P Wyble. A proposed function for hippocampal theta rhythm: separate phases of encoding and retrieval enhance reversal of prior learning. *Neural computation*, 14(4):793–817, 2002.

[9] Michael E Hasselmo and James M Bower. Cholinergic suppression specific to intrinsic not afferent fiber synapses in rat piriform (olfactory) cortex. *Journal of neurophysiology*, 67(5):1222–1229, 1992.

[10] Geoffrey E Hinton. Boltzmann machine. *Scholarpedia*, 2(5):1668, 2007.

[11] Geoffrey E Hinton, Terrence J Sejnowski, and David H Ackley. *Boltzmann machines: Constraint satisfaction networks that learn*. Carnegie-Mellon University, Department of Computer Science Pittsburgh, PA, 1984.

[12] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[13] John J Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the national academy of sciences*, 81(10):3088–3092, 1984.

[14] Ernst Ising. *Beitrag zur theorie des ferro-und paramagnetismus*. PhD thesis, Grefe & Tiedemann Hamburg, Germany, 1924.

[15] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[16] Eve Marder. Neuromodulation of neuronal circuits: back to the future. *Neuron*, 76(1):1–11, 2012.

[17] Matthew J McGinley, Martin Vinck, Jacob Reimer, Renata Batista-Brito, Edward Zagha, Cathryn R Cadwell, Andreas S Tolias, Jessica A Cardin, and David A McCormick. Waking state: rapid variations modulate neural and behavioral responses. *Neuron*, 87(6):1143–1161, 2015.

[18] Rosalyn J Moran, Pablo Campo, Mkael Symmonds, Klaas E Stephan, Raymond J Dolan, and Karl J Friston. Free energy, precision and learning: the role of cholinergic neuromodulation. *Journal of Neuroscience*, 33(19):8227–8236, 2013.

[19] Javier R Movellan. Contrastive hebbian learning in the continuous hopfield model. In *Connectionist models*, pages 10–17. Elsevier, 1991.

[20] Gordon F Newell and Elliott W Montroll. On the theory of the ising model of ferromagnetism. *Reviews of Modern Physics*, 25(2):353, 1953.

[21] Lia Papadopoulos, Suhyun Jo, Kevin Zumwalt, Michael Wehr, David A McCormick, and Luca Mazzucato. Modulation of metastable ensemble dynamics explains optimal coding at moderate arousal in auditory cortex. *bioRxiv*, 2024.

[22] Giovanni Pezzulo, Marco Zorzi, and Maurizio Corbetta. The secret life of predictive brains: what's spontaneous activity for? *Trends in cognitive sciences*, 25(9):730–743, 2021.

[23] Jacob Reimer, Matthew J McGinley, Yang Liu, Charles Rodenkirch, Qi Wang, David A McCormick, and Andreas S Tolias. Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nature communications*, 7(1):13289, 2016.

[24] Susan J Sara. The locus coeruleus and noradrenergic modulation of cognition. *Nature reviews neuroscience*, 10(3):211–223, 2009.

[25] James M Shine. Neuromodulatory influences on integration and segregation in the brain. *Trends in cognitive sciences*, 23(7):572–583, 2019.

[26] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural networks. *Physical review letters*, 61(3):259, 1988.

[27] Gabriel Wainstein, Christopher J. Whyte, Kaylena A. Ehgoetz Martens, Eli J. Müller, Brandon R. Munn, Vicente Medel, Britt Anderson, Elisabeth Stöttinger, James Danckert, and James M. Shine. Gain neuromodulation mediates perceptual switches: evidence from pupillometry, fmri, and rnn modelling. *Elife*, January 2024.

[28] Edward Zagha and David A McCormick. Neural control of brain state. *Current opinion in neurobiology*, 29:178–186, 2014.

A Deriving the Critical Point

Here we derive the critical arousal level α^* at which the network undergoes a phase transition (i.e. pitchfork bifurcation) from multistable to unistable attractor dynamics. In the absence of input, this occurs when the fixed point at the origin goes from being an attractor to a saddle point. To determine when this occurs, we will perform a linear stability analysis, linearizing the dynamics around the origin and determining when the largest eigenvalue of the Jacobian $\mathbf{J}|_{\mathbf{0}}$ becomes greater than 0, as a function of α. The entries of the Jacobian are given by

$$[\mathbf{J}(\alpha)]_{ij} = \frac{\partial}{\partial y_j}\frac{dy_i}{dt} = \begin{cases} -1 & i = j \\ \frac{1}{\alpha}[\mathbf{M}]_{ij}(1 - f([\mathbf{My}]_i)^2) & i \neq j, \end{cases}$$

which evaluated at the origin yields

$$\mathbf{J}|_{\mathbf{0}}(\alpha) = \frac{1}{\alpha}\mathbf{M} - \mathbf{I}.$$

Then we see that the critical point is simply a function of this matrix's eigenvalues:

$$\begin{aligned} 0 = \lambda_{\max}(\mathbf{J}|_{\mathbf{0}}(\alpha^*)) &= \lambda_{\max}\left(\frac{1}{\alpha^*}\mathbf{M} - \mathbf{I}\right) \\ 1 &= \lambda_{\max}\left(\frac{1}{\alpha^*}\mathbf{M}\right) \\ \alpha^* &= \lambda_{\max}\left(\mathbf{M}\right). \end{aligned}$$

Finally, we note that since α is a non-negative scalar, this critical point will only lie within the relevant parameter range if $\lambda_{\max}(\mathbf{M}) \geq 0$. As we will now show, this condition always holds because of the symmetry and zero-diagonal constraints we have imposed on $\mathbf{M}$. First, since $\mathbf{M}$ is symmetric, all of its eigenvalues are real, and second, since the diagonal of $\mathbf{M}$ is zero,

$$\text{tr}(\mathbf{M}) = \sum_{i=1}^{N} \lambda_i = 0.$$

Since the eigenvalues sum to zero, either all of them equal zero or at least one is positive and one is negative. In either case, $\lambda_{\max}\left(\mathbf{M}\right) \geq 0$, completing the proof.

B Boltzmann Machines and Variational Inference

Here we review Boltzmann machines and variational inference as they pertain to the problem at hand (though see [10] and [3] for more general overviews). The Boltzmann machine [11] is a generative model consisting of N binary latent variables that are dubbed "causes" and collectively encoded in the "world state" $\mathbf{z} \in \{\pm 1\}^N$. The couplings between these variables are summarized in the matrix $\frac{1}{\alpha}\mathbf{M}$, such that alignment between z_i and z_j is energetically favorable if $\frac{1}{\alpha}[\mathbf{M}]_{ij}$ is positive and energetically unfavorable if $\frac{1}{\alpha}[\mathbf{M}]_{ij}$. While the world state $\mathbf{z}$ cannot be accessed directly, it gives rise to an observation $\mathbf{x} \in \mathbb{R}^D$ via a noisy generative process parameterized by the matrix $\mathbf{W}$. Given this observation and the parameters of the model, the goal of inference is compute $p(\mathbf{z}|\mathbf{x};\alpha)$, the posterior distribution over world states. This probability is related to the internal energy

$$E(\mathbf{z}|\mathbf{x};\alpha) = -\frac{1}{2}\mathbf{z}^\top\left(\frac{1}{\alpha}\mathbf{M}\right)\mathbf{z} - \frac{1}{2}\mathbf{z}^\top\mathbf{W}\mathbf{x} = \frac{1}{\alpha}\underbrace{E(\mathbf{z})}_{-\frac{1}{2}\mathbf{z}^\top\mathbf{Mz}} + \underbrace{E(\mathbf{x}|\mathbf{z})}_{-\frac{1}{2}\mathbf{z}^\top\mathbf{Wx}}$$

via the Boltzmann distribution

$$p(\mathbf{z}|\mathbf{x};\alpha) = \frac{1}{Z(\mathbf{x};\alpha)}\exp\left(-E(\mathbf{z}|\mathbf{x};\alpha)\right),$$

where $Z(\mathbf{x};\alpha) = \sum_{\mathbf{z}}\exp\left(-E(\mathbf{z}|\mathbf{x};\alpha)\right)$ is a normalizing constant called the partition function.

In practice, computing the true Bayesian posterior via equation (3) is infeasible because evaluating the partition function involves summing over all 2^N possible world states. This motivates the use

of variational methods, a class of inference algorithms that seeks to approximate the true posterior using a family of distributions that is tractable to parameterize and optimize over [3, 5]. In this case, we turn to the mean-field approximation, which assumes that the true posterior factorizes into the product of an independent distribution for each latent variable: i.e. that

$$p(\mathbf{z}|\mathbf{x}) \approx q(\mathbf{z}) = \prod_{i=1}^{N} q_i(z_i),$$

where each of the N factors can be parameterized by the probability it assigns a value of $+1$ for its associated latent, i.e. $q_i(z_i = +1)$. While the mean field approximation makes a very strong independence assumption that does not hold, one can nonetheless optimize over this family of distributions (i.e. through gradient descent) to obtain a reasonable approximation of the true posterior. This optimization amounts to minimizing a quantity called the variational free energy, which is equivalent to minimizing an upper bound on the KL divergence from the approximating distribution to the true Bayesian posterior [15, 3]. In this case, the mean field variational free energy is given by

$$\mathcal{F}[q|\mathbf{x};\alpha] = \langle E(\mathbf{z}|\mathbf{x};\alpha)\rangle_q - \mathcal{S}[q].$$

Since the variables are binary and independent, the expected world state under a candidate approximating distribution $\mathbf{y} = \langle \mathbf{z} \rangle_q$ uniquely specifies it. We can therefore equivalently write equation (4) as a function of $\mathbf{y}$ and simplify to recover the Lyapunov function of the continuous Hopfield network.

C Numerical Simulations

To generate the plots in figures 2 and 3, we used the system's governing equation to determine the flow fields and nullclines and simulated its dynamics using Euler's method, a basic and widely used iterative, discrete time update rule of the form

$$\mathbf{y}(t + \Delta t) = \mathbf{y}(t) + \Delta t \frac{d\mathbf{y}(t)}{dt},$$

where the initial condition $\mathbf{y}(0)$ is specified by the user and the time derivative $\frac{d\mathbf{y}(t)}{dt}$ depends on the current state, arousal level, external input, and network parameters, as described in the main text 1. Dynamics for phase diagrams were calculated using two-unit networks with inhibitory connections of strength -1 between the units. Nullclines were computed analytically.

To generate angle distributions from initializations and stimuli to trajectories' limiting behavior, we created a 10-unit network with $\mathbf{W} = \mathbf{I}$ and $\mathbf{M} = \mathbf{I} - \mathbf{1}\mathbf{1}^\top$ to generate a densely connected inhibitory recurrent structure. Pairs ($n = 1000$) of vectors $\mathbf{x}, \mathbf{y}_0$ were generated with elements uniformly distributed on $[-1, 1]$ and stimulated dynamics were run until convergence for each arousal state.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, the abstract succinctly describes the motivation for the study, the model we explored, what our analyses of the model revealed, and the conclusions we drew from these analyses. It does not include claims outside the scope of the study's content.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Throughout the paper, we stress that the model we analyze is a "minimal" model of the phenomenon of interest designed to yield conceptual insight and recapitulate its essential features, rather than reproduce it in considerable detail.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The only novel mathematical claim in the paper requiring proof is the existence of a phase transition separating the two regimes discussed, which we prove in the Appendix. We also clearly state our assumption that the network's connectivity matrix is symmetric and zero-diagonal, both in said proof and the introduction of the model.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We fully describe the network model in the main text and describe the details of our simulations in the Appendix and the relevant figure captions.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Given the simplicity of the model and the analyses performed, which are straightforward to implement and described in detail, we have not included the very modest amount of code used for our simulations.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The relevant parameters and hyperparameters for the few simulations in the paper are described in the Appendix and figure captions.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper primarily relies on analytical approaches and uses simulations to illustrate qualitative results, rather than to make quantitative claims requiring statistical analysis.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Our simulations were very simple and would be very easy to run on any personal computer, so we do not think a detailed description of the hardware would be informative or beneficial to readers.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: None of the potential harms discussed in the Code of Ethics pertain to our submission.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: While we mention in passing that the neural mechanism we study in this paper could be interesting to incorporate into deep learning models, this claim is a bit too speculative and removed from our study to warrant a detailed consideration of its potential social impacts in this brief workshop paper, which is more focused on a neuroscientific phenomenon.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper analyzes a fairly simple mathematical model that is primarily intended to develop intuition and lacks such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: The modest amount of code used in this study was implemented by the authors.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper involves neither crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper involves neither crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.