Goldstone Modes and the Coexistence Saddle: Spectral Bifurcation Analysis of Competing Ring Attractors Under Mean-Field Cross-Inhibition

Tomás P. Pavan[1] and **Claude**[2]

[1]Independent Researcher [2]Anthropic (Claude Opus 4.6, Large Language Model)

Correspondence: Tomás P. Pavan

Abstract

Persistent neural activity in working memory is often modeled as bump attractors on ring networks. When two such networks compete via cross-inhibition, the system must resolve which representation dominates – a winner-take-all (WTA) decision. We present a complete spectral bifurcation analysis of this transition in a coupled ring attractor model (two networks of $N = 48$ neurons each, cosine connectivity within, mean-field cross-inhibition between). We identify six key results.

First, the coexistence fixed point – where both bumps self-sustain simultaneously – exists only below a critical cross-inhibition strength $J_\times^* \approx 0.36$. Above this threshold, cross-inhibition is too strong for both representations to survive; the system admits only WTA solutions.

Second, the continuous rotational symmetry of each bump guarantees exactly two Goldstone modes (zero eigenvalues) that are protected against mean-field coupling. The first non-Goldstone eigenvalue – governing uniform amplitude competition – crosses zero at $J_\times^* \approx 0.3485$ via a pitchfork bifurcation, creating the coexistence saddle and two WTA stable states.

Third, the critical eigenvector projects maximally onto the spatially uniform (DC) direction, meaning the instability concerns total activity competition rather than spatial pattern rearrangement. This is a direct consequence of mean-field (spatially unstructured) cross-inhibition.

Fourth, large-scale stochastic simulations (128,000 trials across 256 parameter combinations) confirm the spectral predictions: swap errors emerge at the predicted $J_\times$ threshold, drive strength is secondary to cross-inhibition, and a non-monotonic valley at intermediate $J_\times$ identifies a functional operating regime for working memory.

Fifth, connectivity heterogeneity destroys the sharp pitchfork entirely, converting it into an imperfect bifurcation with no zero-crossing. The razor-thin instability window ($\Delta J_\times \approx 0.01$) is a symmetry artifact of the clean model; biological circuits operate in a regime of smooth crossover where no parameter precision is required.

Sixth, a Kramers escape analysis bridges the deterministic bifurcation and stochastic swap onset: because the barrier collapses quadratically ($\Delta V \propto |\lambda_{\text{dom}}|^2$), noise-driven escape becomes likely at $J_\times \approx 0.25$ – well below the pitchfork – for a normal-form coefficient $\gamma \approx 0.22$–0.36. We discuss implications for the behavioral cliff and argue that neural circuits operate in a valley regime where cross-inhibition and encoding drive are balanced, rather than near $J_\times^*$ itself.

---

1. Introduction

1.1 Working Memory and Competing Representations

Persistent neural activity in prefrontal and parietal cortex underlies the short-term maintenance of information in working memory (Goldman-Rakic, 1995; Funahashi et al., 1989). Ring attractor models capture a key feature of this activity: spatially tuned neurons form a localized "bump" of elevated firing that persists through recurrent excitation even after the sensory stimulus is removed (Compte et al., 2000; Ben-Yishai et al., 1995; Amari, 1977). These bumps encode continuous variables such as spatial location or orientation, and their precision is set by a balance between recurrent drive and noise-induced diffusion along the ring (Wimmer et al., 2014; Burak and Fiete, 2012). Such ring attractor dynamics have been observed experimentally in the Drosophila head direction system (Kim et al., 2017) and are reviewed in the broader context of attractor and integrator networks by Khona and Fiete (2022).

When multiple items must be stored simultaneously, as in multi-item visual working memory tasks, the standard approach posits multiple bump networks coupled through cross-inhibition (Edin et al., 2009; Wei et al., 2012). The cross-inhibition creates a competition: if it is weak, both bumps coexist and the system maintains multiple items; if it is strong, one bump suppresses the other in a winner-take-all (WTA) decision. The transition between these regimes determines the capacity limit of the working memory circuit (Edin et al., 2009).

Despite the importance of this transition, its spectral structure – the full set of eigenvalues and eigenvectors of the system's Jacobian – has not been characterized. Previous analyses have focused on one-dimensional (1D) mean-field reductions, projecting the high-dimensional dynamics onto a single dominance variable $D = \bar{r}_A - \bar{r}_B$ and characterizing the resulting cusp catastrophe (Thom, 1972; Zeeman, 1977). While this captures the topology of the bifurcation, it discards the 96-dimensional dynamics that include rotational modes, drift modes, and the full stability structure of the coexistence state.

1.2 The Behavioral Cliff

Psychophysical experiments reveal a striking feature of working memory performance: below a critical stimulus strength, accuracy does not degrade gradually but collapses abruptly – a "behavioral cliff" (Bays et al., 2009; Zhang and Luck, 2008). In the mixture model framework, this manifests as a sharp increase in the probability of reporting a non-target item (swap errors) or of random guessing, even for small changes in signal-to-noise ratio.

The standard theoretical account attributes the cliff to noise-driven escape from a metastable state (Kramers, 1940; Hanggi et al., 1990): when the cue is weak, the barrier between the correct attractor and competing attractors is low, and stochastic fluctuations cause the system to fall to a wrong state. This yields a cusp catastrophe potential $V(D) = D^4 + aD^2 + bD$, where a is controlled by the circuit's lateral inhibition and b by the cue strength. The cliff occurs at the cusp point where the barrier vanishes.

However, this account treats the cliff as a cue phenomenon – a consequence of weak sensory input. An alternative possibility, which we develop here, is that the cliff reflects a structural property of the circuit: the proximity of the effective cross-inhibition strength $J_\times$ to a critical value $J_\times^*$ where the coexistence state undergoes a spectral bifurcation.

1.3 From Mean-Field Reduction to Full Spectral Analysis

The 1D reduction $D = \bar{r}_A - \bar{r}_B$ captures the order parameter of the WTA transition but suppresses 95 of the 96 dynamical degrees of freedom. In particular, it cannot distinguish:

1. **Goldstone modes** – exactly-zero eigenvalues arising from the continuous rotational symmetry of each bump (Goldstone, 1961; Burak and Fiete, 2012). These modes govern bump drift and are protected by symmetry.
2. **Genuine instabilities** – eigenvalues that cross zero as parameters change, signaling structural reorganization of the attractor landscape.
3. **The character of the critical mode** – whether the instability that destroys coexistence projects onto the spatially uniform (DC) direction, is spatially patterned (cosine), or mixed.

Previous spectral approaches to ring networks have addressed non-Hermitian quasi-localization (Tanaka and Nelson, 2018) and the stability of persistent activity under short-term plasticity (Seeholzer et al., 2019), but the full eigenvalue structure of the *coupled* system has not been resolved. We present the first complete eigenvalue decomposition of the coupled ring attractor Jacobian, resolving all $2N = 96$ eigenvalues as a function of the cross-inhibition strength $J_\times$. By cleanly separating Goldstone modes from genuine instabilities, we identify the precise location, character, and consequences of the coexistence-to-WTA pitchfork bifurcation.

1.4 Summary of Contributions

Our main results are:

1. **Existence threshold.** The coexistence fixed point exists only for $J_\times < J_\times^{exist} \approx 0.36$. At the commonly used value $J_\times = 0.5$, coexistence is not a fixed point of the dynamics – both bumps cannot self-sustain under such strong cross-inhibition.
2. **Goldstone separation and pitchfork.** Two Goldstone modes (exactly-zero eigenvalues protected by rotational symmetry) persist at all $J_\times$ where coexistence exists. The first non-Goldstone eigenvalue λ_{dom} crosses zero at $J_\times^* \approx 0.3485$, creating a pitchfork bifurcation where the symmetric coexistence state becomes a saddle point and two WTA attractors are born.
3. **DC critical mode.** The critical eigenvector has its largest projection onto the uniform (DC) direction: an increase in network A's activity coupled with a decrease in network B's, localized to the active bump neurons by the gain mask $\sigma'(h_i)$. This reflects the mean-field character of the cross-inhibition and means the instability is about total activity competition, not spatial pattern rearrangement.
4. **Stochastic phase diagram.** A 128,000-trial parameter sweep confirms the spectral predictions and reveals a non-monotonic valley at intermediate $J_\times$ where swap error rates dip to 7–13% between two qualitatively different failure modes.
5. **Heterogeneity destroys the sharp bifurcation.** Connectivity heterogeneity breaks the exact A↔B exchange symmetry, converting the pitchfork into an imperfect bifurcation with no zero-crossing. The razor-thin instability window ($\Delta J_\times \approx 0.01$) is a symmetry artifact; biological circuits operate in a smooth crossover regime.

These results reframe the behavioral cliff as a $J_\times$-space phenomenon and identify a qualitative operating regime (the valley) where encoding drive and cross-inhibition are balanced for reliable

working memory.

---

2. Model

2.1 Single Ring Attractor

We consider a rate model with $N = 48$ neurons uniformly distributed on a ring. Each neuron i has a preferred angle $\theta_i = -\pi + 2\pi i/N$ and firing rate $r_i(t)$ governed by:

$$\tau \frac{dr_i}{dt} = -r_i + \sigma(h_i)$$

where $\tau = 10$ ms is the time constant and $\sigma(h) = r_{max}/(1 + e^{-\beta(h-h_0)})$ is a sigmoidal activation function with parameters $r_{max} = 1.0$, $\beta = 5.0$, $h_0 = 0.5$. The total input to neuron i is:

$$h_i = \sum_{j=1}^{N} W_{ij} r_j + I_i^{ext}$$

where the within-network connectivity has cosine tuning:

$$W_{ij} = \frac{1}{N} \left(-J_0 + J_1 \cos(\theta_i - \theta_j)\right)$$

with $J_0 = 1.0$ (uniform inhibition) and $J_1 = 6.0$ (tuned excitation). This connectivity supports a family of bump solutions at any angular position, forming a ring attractor (Amari, 1977; Ben-Yishai et al., 1995).

2.2 Coupled System with Mean-Field Cross-Inhibition

We couple two identical ring networks A and B through mean-field cross-inhibition. The dynamics become:

$$\tau \frac{dr_i^A}{dt} = -r_i^A + \sigma\left(\sum_j W_{ij} r_j^A + I_i^{cue} - J_\times \bar{r}^B\right)$$

$$\tau \frac{dr_i^B}{dt} = -r_i^B + \sigma\left(\sum_j W_{ij} r_j^B - J_\times \bar{r}^A\right)$$

where $\bar{r}^X = \frac{1}{N}\sum_j r_j^X$ is the mean activity of network X and $J_\times \geq 0$ is the cross-inhibition strength. The external cue input is a von Mises tuning curve applied to network A only:

$$I_i^{cue} = c \cdot \frac{e^{\kappa \cos(\theta_i - \theta_{stim})}}{I_0(\kappa)}$$

with concentration parameter $\kappa = 2.0$ and cue gain $c \geq 0$.

The critical feature of mean-field cross-inhibition is that it depends only on the total activity $\bar{r}^X$ of the opposing network, not on the spatial pattern of its bump. This has profound consequences for the symmetry structure of the system (Fig. 1).

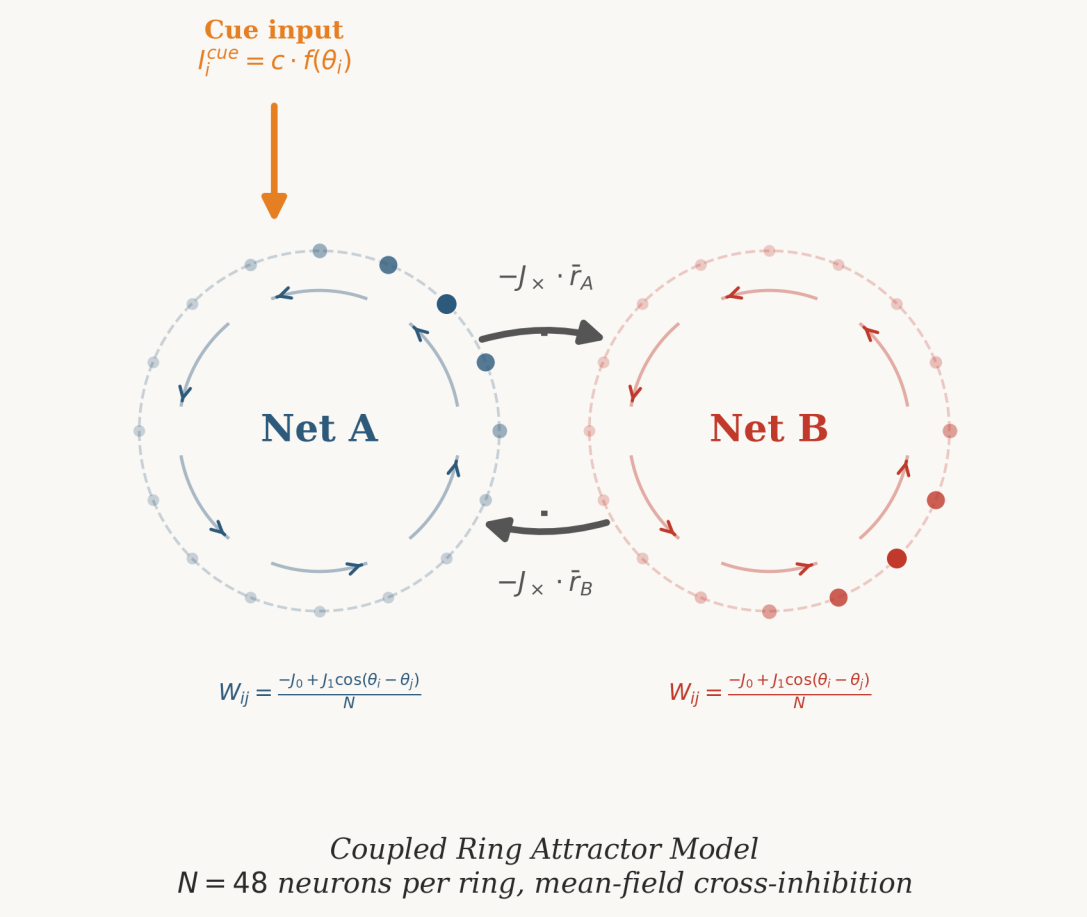


Figure 1. Model schematic. Two ring networks (A, B) of $N = 48$ neurons each, with cosine within-network connectivity ($J_0 + J_1 \cos \Delta\theta$) and mean-field cross-inhibition ($J_\times \bar{r}^X$). External cue input drives network A only. The cross-inhibition depends on mean activity, not bump position, preserving rotational symmetry.

2.3 Jacobian of the Coupled System

The steady-state condition $F(\mathbf{r}^*) = 0$ defines the fixed points, where $F_i^A = -r_i^A + \sigma(h_i^A)$ and similarly for B. The Jacobian $\mathbf{J} = \partial F / \partial \mathbf{r}$ evaluated at a fixed point $\mathbf{r}^*$ has a 2×2 block structure:

$$\mathbf{J} = \begin{pmatrix} -\mathbf{I} + \mathbf{S}_A \mathbf{W} & \mathbf{S}_A \mathbf{C} \\ \mathbf{S}_B \mathbf{C} & -\mathbf{I} + \mathbf{S}_B \mathbf{W} \end{pmatrix}$$

where $\mathbf{S}_X = \mathrm{diag}(\sigma'(h_i^X))$ is the diagonal matrix of sigmoid derivatives at the fixed point, and $\mathbf{C} = -\frac{J_\times}{N} \mathbf{1}\mathbf{1}^T$ is the rank-1 mean-field coupling matrix. The full Jacobian is $2N \times 2N = 96 \times 96$.

The block structure reveals that the cross-coupling enters only through the rank-1 matrix $\mathbf{C}$. This low-rank perturbation to the block-diagonal within-network dynamics is what makes the spectral

analysis tractable: the cross-inhibition can shift at most one eigenvalue per symmetry sector.

2.4 Symmetries

The coupled system possesses two symmetries at zero cue ($c = 0$):

Continuous rotational symmetry. The mean-field coupling $J_\times \bar{r}^X$ is invariant under any rotation of the bump profile: if r_i^X is a fixed point, so is r_{i+k}^X for any shift k. This gives a continuous family of fixed points parametrized by bump position, and by Goldstone's theorem (Goldstone, 1961), each such continuous symmetry produces an eigenvalue that is exactly zero. With two independent bumps, there are two Goldstone modes.

Discrete exchange symmetry. At $c = 0$, the system is invariant under $A \leftrightarrow B$. The coexistence fixed point (where both bumps are present with $\bar{r}^A = \bar{r}^B$) respects this symmetry; the WTA states ($\bar{r}^A \gg \bar{r}^B$ or vice versa) break it. The transition between these is governed by a pitchfork bifurcation.

The nonzero cue $c > 0$ breaks the exchange symmetry (favoring network A) and deforms the pitchfork into an imperfect bifurcation with hysteresis.

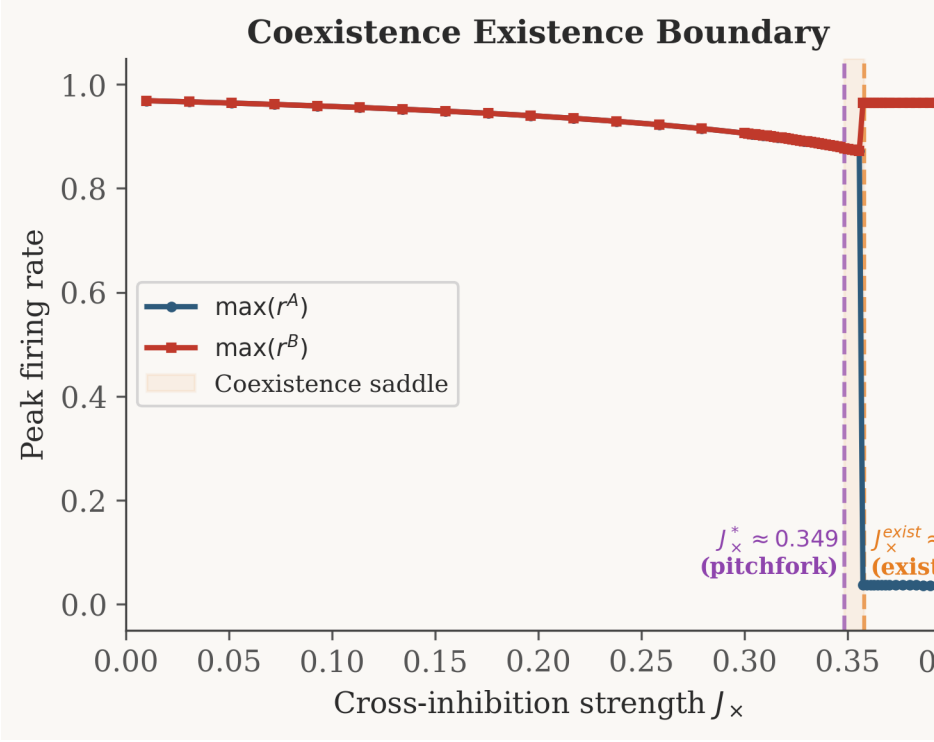
---

3. Results

3.1 Existence of the Coexistence Fixed Point

3.1.1 Numerical Method We locate fixed points using a two-phase approach. In Phase 1 (simulation), we establish both bumps using strong external drives ($I_{ext} = 5.0$), then remove the drives and simulate the coupled system for 5×10^4 time steps ($\Delta t = 0.1, \tau = 10$) under the target cross-inhibition $J_\times$. In Phase 2 (Newton), we polish the resulting state using Newton's method (scipy.optimize.fsolve) with the analytical Jacobian, achieving residuals $\|F(\mathbf{r}^*)\| < 10^{-10}$.

We verify convergence by checking: (i) the residual $\max_i |F_i(\mathbf{r}^*)| < 10^{-6}$; (ii) both bumps are active ($\max_i r_i^A > 0.3$ and $\max_i r_i^B > 0.3$); and (iii) the solution is not a WTA state ($|D| = |\bar{r}^A - \bar{r}^B| < 0.1$).



3.1.2 Critical Existence Threshold

Figure 2. Coexistence existence boundary. Peak firing rate of each network as a function of cross-inhibition strength $J_\times$. Below $J_\times^{exist} \approx 0.358$, both networks sustain bumps at matched amplitudes (coexistence). Above this threshold, one network collapses to baseline (WTA). The pitchfork bifurcation at $J_\times^* \approx 0.349$ (orange dashed) and existence boundary at $J_\times^{exist} \approx 0.358$ (purple dashed) delimit the narrow instability window $\Delta J_\times \approx 0.01$.

We scan $J_\times$ from 0 to 0.5 (Fig. 2). Below $J_\times \approx 0.36$, Newton converges to a genuine coexistence fixed point with residual $< 10^{-10}$. Above this threshold, one bump suppresses the other during the simulation phase; Newton converges only to WTA solutions. At $J_\times = 0.50$ – a commonly used parameter value – coexistence does not exist as a fixed point of the deterministic dynamics.

The critical existence threshold lies between $J_\times = 0.355$ (both bumps survive, $\max r^A = 0.88, \max r^B = 0.88$) and $J_\times = 0.360$ (one bump collapses, $\max r^A = 0.97, \max r^B = 0.04$). The transition is sharp: a 1.4% increase in $J_\times$ converts stable coexistence into complete dominance.

3.1.3 Diagnostic: Fixed Point vs. Slow Manifold At $J_\times = 0.50$, a time-resolved diagnostic reveals that the system does not converge: residuals remain at $\sim 10^{-3}$ and the dominance variable D drifts monotonically toward ± 0.27. At $J_\times = 0.35$, residuals converge exponentially to machine precision ($\sim 10^{-15}$). This confirms that the coexistence state is a genuine fixed point below threshold and does not exist (even as a slow manifold) above it.

3.2 Goldstone Modes and the Protected Symmetry

3.2.1 Origin of the Goldstone Modes The mean-field cross-coupling $J_\times \bar{r}^B$ is a function of the mean activity $\bar{r}^B = \frac{1}{N}\sum_j r_j^B$ only. Any continuous rotation of the bump profile preserves this mean. We now prove that this protects the rotational modes as exact null vectors of the full coupled Jacobian.

Step 1: Rotational null vector of the uncoupled block. Let $\mathbf{r}^{A*}(\varphi)$ denote the bump solution of network A centered at phase φ. Because W_{ij} depends only on the angular difference $\theta_i - \theta_j$, the steady-state equation $-r_i^{A*} + \sigma(\sum_j W_{ij} r_j^{A*}) = 0$ holds for every φ. Differentiating both sides with respect to φ:

$$-\frac{\partial r_i^{A*}}{\partial \varphi} + \sigma'(h_i^{A*}) \sum_j W_{ij} \frac{\partial r_j^{A*}}{\partial \varphi} = 0$$

In matrix form, this is $(-\mathbf{I} + \mathbf{S}_A \mathbf{W}) \cdot \partial \mathbf{r}^{A*}/\partial \varphi = \mathbf{0}$, where $\mathbf{S}_A = \mathrm{diag}(\sigma'(h_i^{A*}))$. The rotational derivative is an exact null vector of the uncoupled Jacobian block. (The same holds for network B by identical argument.)

Step 2: Mean-field coupling annihilates the rotational mode. Consider the $2N$-dimensional perturbation $\mathbf{v}_A = (\partial \mathbf{r}^{A*}/\partial \varphi,\, \mathbf{0})^T$ corresponding to a shift of network A's bump alone. Multiplying by the full block Jacobian (Section 2.3) yields:

$$\mathbf{J} \cdot \mathbf{v}_A = \begin{pmatrix} (-\mathbf{I} + \mathbf{S}_A \mathbf{W}) \cdot \partial \mathbf{r}^{A*}/\partial \varphi + \mathbf{S}_A \mathbf{C} \cdot \mathbf{0} \\ \mathbf{S}_B \mathbf{C} \cdot \partial \mathbf{r}^{A*}/\partial \varphi + (-\mathbf{I} + \mathbf{S}_B \mathbf{W}) \cdot \mathbf{0} \end{pmatrix}$$

The upper block vanishes by Step 1. The survival of the zero eigenvalue depends entirely on the cross-coupling term $\mathbf{C} \cdot \partial \mathbf{r}^{A*}/\partial \varphi$ in the lower block. Recall that $\mathbf{C} = -\frac{J_\times}{N} \mathbf{1}\mathbf{1}^T$. Applying $\mathbf{C}$ to the rotational derivative:

$$\mathbf{C} \cdot \frac{\partial \mathbf{r}^{A*}}{\partial \varphi} = -\frac{J_\times}{N} \mathbf{1} \left(\mathbf{1}^T \cdot \frac{\partial \mathbf{r}^{A*}}{\partial \varphi} \right) = -\frac{J_\times}{N} \mathbf{1} \cdot \sum_{j=1}^{N} \frac{\partial r_j^{A*}}{\partial \varphi}$$

Because a rotation merely translates the bump profile around the periodic ring, the total activity (and thus the mean) is strictly conserved. Exchanging derivative and sum:

$$\sum_{j=1}^{N} \frac{\partial r_j^{A*}}{\partial \varphi} = \frac{\partial}{\partial \varphi} \sum_{j=1}^{N} r_j^{A*} = \frac{\partial}{\partial \varphi} (N \bar{r}^A) = 0$$

Since $\mathbf{1}^T \cdot \partial \mathbf{r}^{A*}/\partial \varphi = 0$, it follows that $\mathbf{C} \cdot \partial \mathbf{r}^{A*}/\partial \varphi = \mathbf{0}$. The rank-1 coupling matrix completely annihilates the rotational derivative, yielding $\mathbf{J} \cdot \mathbf{v}_A = \mathbf{0}$. By identical logic for network B, $\mathbf{v}_B = (\mathbf{0},\, \partial \mathbf{r}^{B*}/\partial \varphi)^T$ is also a null vector. □

Equivariance structure. The result follows from the SO(2) × SO(2) equivariance of the coupled system at $c = 0$: the dynamics commute with independent rotations of each ring. The mean-field coupling $J_\times \bar{r}^X$ is invariant under both SO(2) actions because it depends only on total activity, which is a rotation-invariant functional. Each SO(2) factor contributes one Goldstone mode to the kernel of the Jacobian. This protection is exact and holds at all $J_\times$ where the coexistence fixed point exists – it cannot be lifted by increasing cross-inhibition, only by breaking the rotational symmetry of either the within-network connectivity or the cross-coupling structure.

This is the neural circuit analog of the Goldstone theorem (Goldstone, 1961): a spontaneously broken continuous symmetry produces a massless (zero-energy) excitation. In our context, "massless" means neutrally stable – perturbations along the Goldstone direction neither grow nor decay. Because

mean-field coupling acts exclusively on the spatially uniform mode ($\mathbf{1}$), it is perfectly orthogonal to the zero-sum rotational modes, mathematically protecting positional memory from amplitude competition.

3.2.2 Numerical Identification We classify eigenvalues into Goldstone candidates ($|\lambda| < 10^{-3}$) and genuine modes ($|\lambda| \geq 10^{-3}$). For each eigenvector $\mathbf{v}$, we compute projections onto six basis directions:

- $\mathbf{d}_{dom}$: symmetric dominance (cosine envelope, $A \uparrow B \downarrow$)
- $\mathbf{d}_{drift,+}$: co-directional drift (sine envelope, both shift same way)
- $\mathbf{d}_{drift,-}$: anti-directional drift (sine envelope, shift opposite ways)
- $\mathbf{d}_{uni}$: uniform/DC ($A \uparrow B \downarrow$ flat)
- $\mathbf{d}_{gold,A}$: rotation of bump A (sine envelope, A only)
- $\mathbf{d}_{gold,B}$: rotation of bump B (sine envelope, B only)

The Goldstone eigenvectors project strongly onto $\mathbf{d}_{gold,A}$ and $\mathbf{d}_{gold,B}$, confirming their rotational character (Fig. 4, left panel).

3.2.3 Goldstone Count Across $J_\times$ Across the entire range $J_\times \in [0, 0.36)$ where coexistence exists, we find exactly two Goldstone modes. Their eigenvalues remain at $|\lambda| \sim 10^{-8}$ to 10^{-11} (machine precision for our iterative solver), and they are never lifted by increasing cross-inhibition. This confirms the symmetry protection: mean-field coupling cannot break rotational invariance.

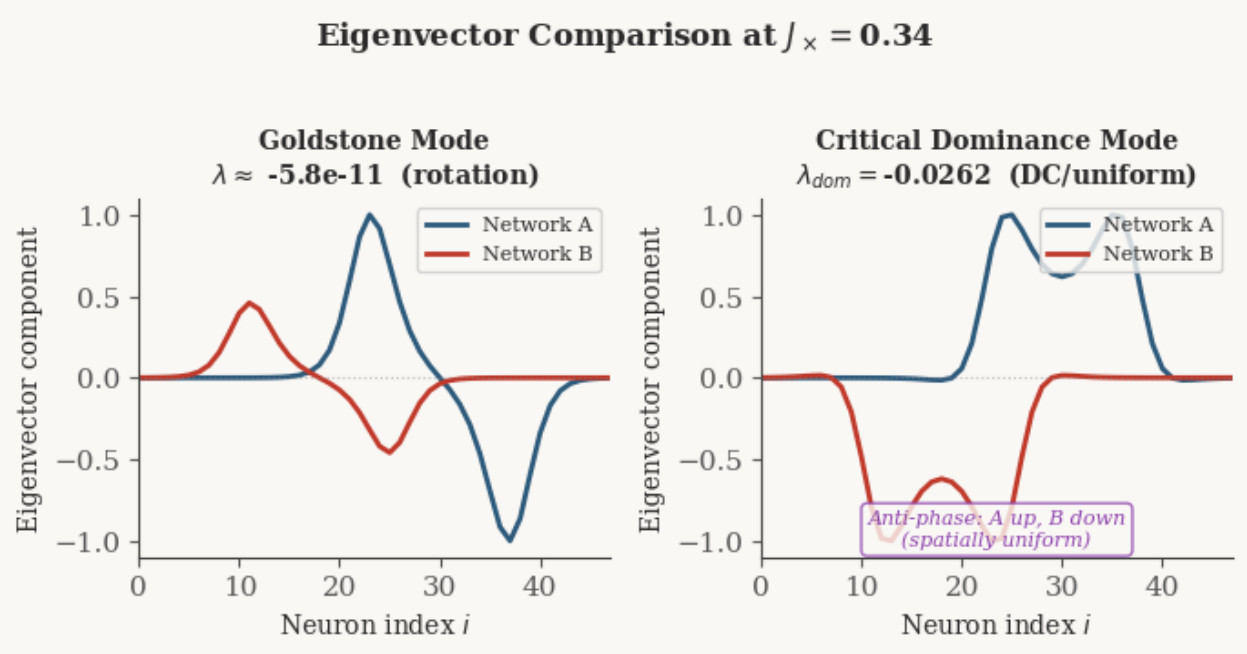


Figure 4. Eigenvector comparison at $J_\times = 0.34$ (near the pitchfork). Left: Goldstone mode ($\lambda \approx -5.8 \times 10^{-11}$, effectively zero), showing sinusoidal spatial structure in Network A – this is the rotational mode that slides the bump around the ring. Right: Critical dominance mode ($\lambda_{dom} = -0.026$, DC/uniform), showing anti-phase amplitude modulation localized to the active bump neurons. Because baseline neurons are strongly inhibited ($\sigma'(h_i) \approx 0$), they cannot participate in the linear instability; the mode is expressed only where the gain is nonzero. Despite this spatial localization, the mode projects maximally onto the uniform mean-field direction because the net

effect is a difference in total activity between networks. The two modes are qualitatively distinct: the Goldstone mode encodes *where* the bump sits; the critical mode encodes *which network wins*.

3.3 The Pitchfork Bifurcation

3.3.1 The First Non-Goldstone Eigenvalue After removing the two Goldstone modes, we track the dominant genuine eigenvalue λ_{dom} as a function of $J_\times$ (Fig. 3). Key findings:

- At $J_\times = 0$: $\lambda_{dom} = -0.572$ (strongly stable). Without cross-inhibition, the coexistence state is deeply attractive.
- λ_{dom} increases monotonically with $J_\times$, crossing zero at $J_\times^* \approx 0.3485$.
- Above $J_\times^*$: $\lambda_{dom} > 0$ (saddle). The coexistence state acquires one unstable direction.
- At $J_\times = 0.356$: $\lambda_{dom} = +0.025$, and coexistence ceases to exist shortly after at $J_\times \approx 0.358$.

The crossing at $J_\times^*$ is a pitchfork bifurcation: the symmetric coexistence state ($D = 0$) loses stability, and two WTA states ($D > 0$ and $D < 0$) emerge as the new stable attractors. The $A \leftrightarrow B$ exchange symmetry is spontaneously broken.

3.3.2 Character of the Critical Eigenvector At $J_\times^*$, the critical eigenvector has the following projections:

Direction	$\lvert\langle \mathbf{v}_1, \mathbf{d}\rangle\rvert$
Uniform (DC)	0.43
Dominance (cosine)	0.34
Anti-drift (sine)	0.34
Co-drift (sine)	0.00

The largest projection is onto the uniform/DC direction (Fig. 4, right panel). Rather than a spatially flat baseline shift, the mode drives a gain-weighted amplitude modulation: a sharp increase in the active neurons of network A coupled with a sharp decrease in the active neurons of network B, while baseline neurons (where $\sigma'(h_i) \approx 0$) are effectively silent. Because this amplitude competition produces a net difference in total activity between networks, it projects maximally onto the mean-field coupling direction. The instability is about which network has more total activity, not about the spatial pattern of either bump.

This is a direct and falsifiable prediction of mean-field coupling. If cross-inhibition were spatially structured (depending on the relative positions of the two bumps), the critical eigenvector would acquire spatial structure (cosine or higher Fourier modes). The DC character is specific to coupling that "sees" only total activity.

3.3.3 The Narrow Existence Window The coexistence saddle – genuinely unstable, not merely Goldstone-neutral – exists only in the interval $J_\times \in [0.3485, 0.358]$, a width of $\Delta J_\times \approx 0.01$. Below $J_\times^*$, coexistence is a stable node. Above $J_\times^{exist} \approx 0.358$, it ceases to exist entirely.

This razor-thin window has two implications:

1. **Structural precision.** The bifurcation is sharp: a 3% change in $J_\times$ (from 0.348 to 0.358) takes the system from stable coexistence through saddle instability to complete collapse. The system is tuned near a critical point.

2. **Heterogeneity prediction.** In biological circuits with heterogeneous connectivity, the sharp boundary should be smeared into a broader regime where saddle-like dynamics persist (see Discussion).

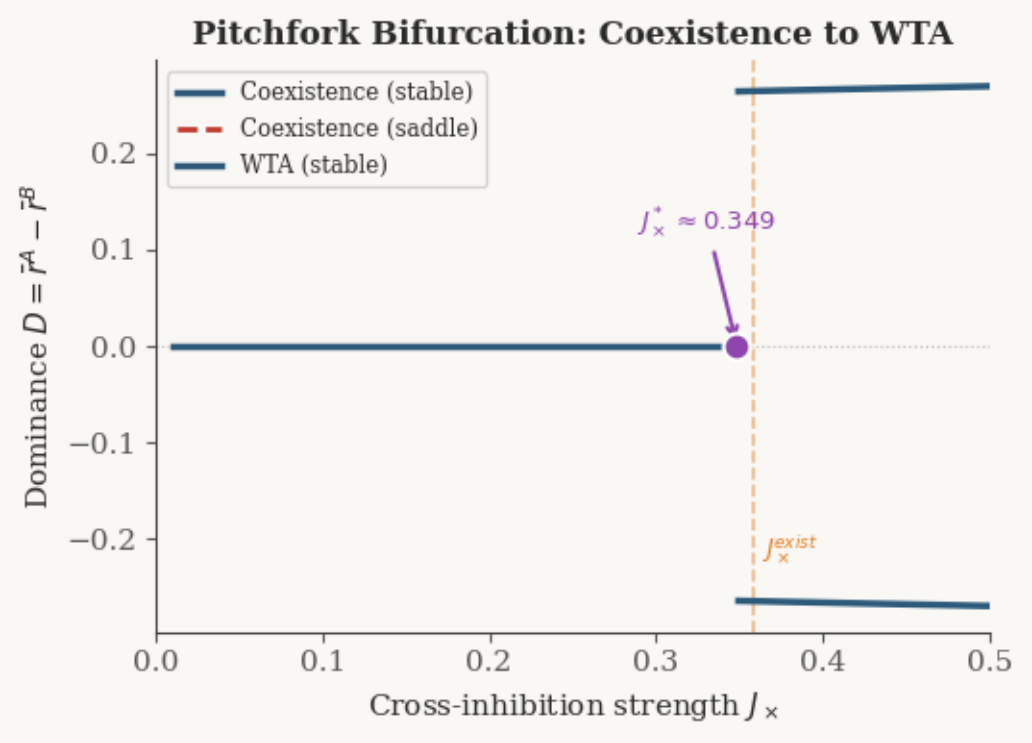


Figure 3. Pitchfork bifurcation diagram. Dominance $D = \bar{r}^A - \bar{r}^B$ vs. cross-inhibition strength $J_\times$. The coexistence branch ($D = 0$) is stable (solid) for $J_\times < J_\times^*$ and becomes a saddle (dashed red) above the pitchfork at $J_\times^* \approx 0.349$. Two WTA branches ($D > 0$ and $D < 0$, blue) emerge at $J_\times^{exist} \approx 0.358$ as stable attractors. The subcritical structure creates the narrow existence window $\Delta J_\times \approx 0.01$. Note: the unstable branches connecting the pitchfork to the saddle-node fold at $J_\times^{exist}$ are not shown; our continuation solver tracked stable and saddle fixed points only. The expected topology is a pair of unstable branches emerging from $J_\times^*$ at $D = 0$ and folding onto the stable WTA branches at $J_\times^{exist}$.

3.4 The Coexistence Saddle Under Cue

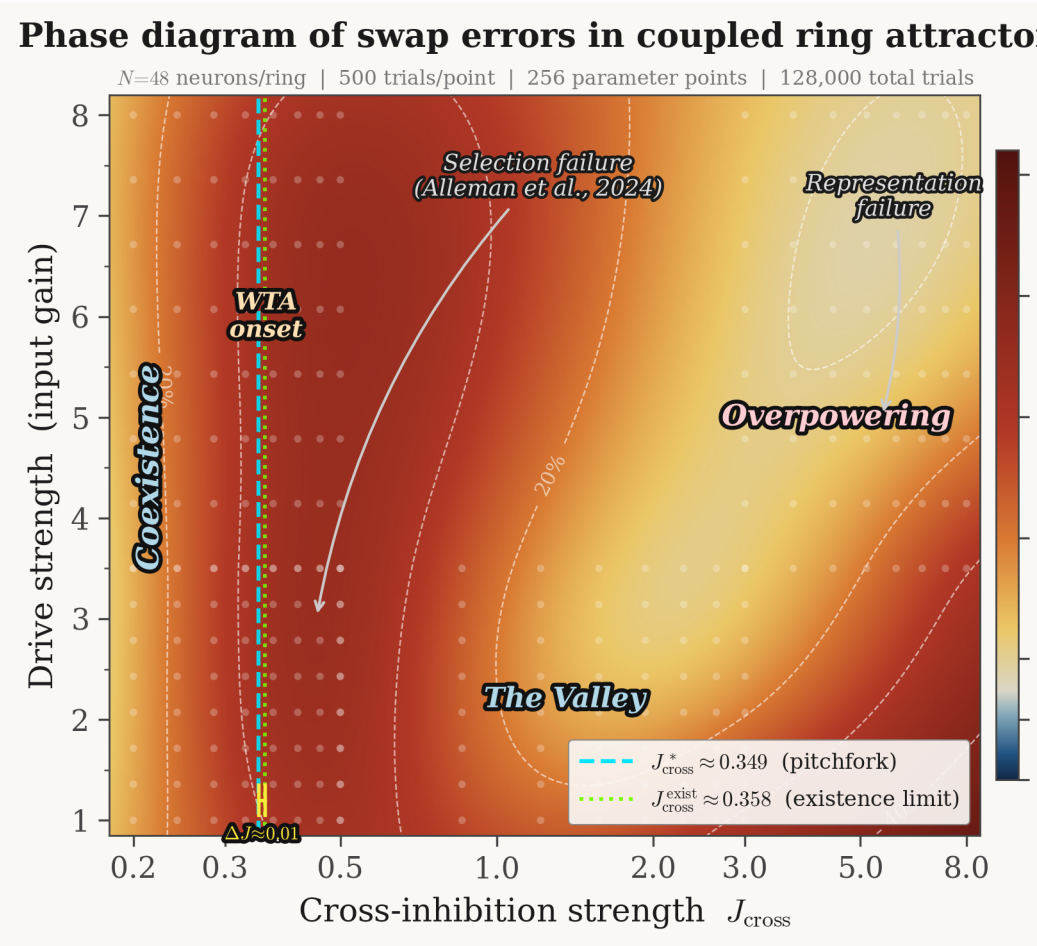
Newton continuation in cue gain c from 0 to 0.5 at $J_\times = 0.35$ (within the saddle window; see Fig. 3) reveals:

- The coexistence branch maintains both bumps across the full cue range, with D growing slowly as the cue favors network A.
- All 51/56 tracked solutions are unstable (saddle points), with $n_{positive} = 1$ for $c < 0.15$ and $n_{positive} = 2$ for $c > 0.15$ (a second eigenvalue crossing).
- The critical eigenvector's projection onto cosine and sine directions is $|\cos| \approx |\sin| \approx 0.47$ – a mixed dominance-drift mode at 45 degrees.

The WTA branch (tracked simultaneously) is stable across all cue values, confirming that the saddle's unstable manifold connects to the two WTA basins.

3.5 Stochastic Phase Diagram of Swap Errors

3.5.1 Parameter Sweep To bridge the deterministic bifurcation analysis with behavioral predictions, we performed a large-scale stochastic simulation across the $(J_\times, c)$ parameter space. At each of 256 grid points (16 values of $J_\times$ from 0.05 to 8.0, 16 values of drive strength from 1.0 to 8.0), we ran 500 stochastic trials ($N_{total} = 128{,}000$). Two ring networks encoded items separated by $\pi/2$ radians, with independent Gaussian noise ($\sigma = 0.1$) added to each neuron during a 500-step maintenance period. Swap errors were classified as decoded responses within 0.3 rad of the non-target item's location. To assess sensitivity to this classification threshold, we repeated a focused sweep (16 values of $J_\times$ from 0.1 to 3.0, two drive levels, 200 trials each) with thresholds of 0.2, 0.3, 0.4, and 0.5 rad. The qualitative features of the phase diagram – the onset location ($J_\times \approx 0.20$–0.23 across thresholds), the peak near $J_\times \approx 0.3$–0.5, and the non-monotonic valley at $J_\times \approx 1.2$–1.5 – are robust to threshold choice. Wider classification windows increase absolute swap rates (e.g., 10% vs. 26% at $J_\times = 0.25$ for 0.2 vs. 0.5 rad) but preserve the ordering across $J_\times$ values and the location of qualitative transitions.



3.5.2 Onset of Swap Errors

Figure 5. Stochastic phase diagram in $(J_\times, c)$ space (128,000 trials across 256 parameter combinations). Color indicates swap error rate. Near-vertical isocontours show that swap probability depends primarily on $J_\times$, not drive strength. A non-monotonic valley at $J_\times \approx 1.2$–1.6 with strong drive identifies the functional operating regime where WTA dynamics and encoding strength are balanced.

Swap errors emerge at $J_\times \approx 0.25$, consistent with the spectral prediction of the pitchfork bifurcation at $J_\times^* \approx 0.3485$. The stochastic onset is lower than the deterministic bifurcation because noise-mediated escape from the metastable coexistence well occurs when the barrier height $\Delta V \sim \sigma^2$, which corresponds to a $J_\times$ slightly below the eigenvalue crossing. This is precisely the Kramers mechanism.

Between $J_\times \approx 0.25$ and 0.5, swap rates increase from 5% to approximately 45%. Above $J_\times \approx 1.0$, swap rates plateau near 50% – the system has become a noise-driven bistable switch with no memory of the initial encoding.

3.5.3 Drive Strength Is Secondary The phase diagram shows near-vertical isocontours of swap rate (Fig. 5): swap error probability depends primarily on $J_\times$ and only weakly on drive strength. This is a direct prediction of the spectral analysis: the critical eigenvector projects maximally onto the uniform (DC) direction, governing total activity competition rather than spatial encoding. Stronger drive does not protect against the dominance instability because the instability is orthogonal to the encoding direction.

This has a counterintuitive implication: increasing stimulus strength – the commonly proposed intervention for working memory failures – targets the wrong degree of freedom. The cliff is a $J_\times$ phenomenon, not a cue phenomenon.

3.5.4 The Non-Monotonic Valley and Two Failure Regimes At $J_\times \approx 1.2$–1.6 with moderate to strong drive, the phase diagram reveals a non-monotonic feature: swap rates dip to 7–13% between two distinct failure modes. Crucially, this valley lies well above the coexistence existence threshold ($J_\times^{exist} \approx 0.36$), meaning it operates in a pure WTA regime where deterministic coexistence does not exist. The two failure modes and the valley between them correspond to qualitatively different dynamical regimes:

1. **Near-critical swaps** ($J_\times \approx 0.3$–0.5): Near and just above the pitchfork, barriers separating coexistence from WTA are small and noise escapes freely. This is the *representation failure* regime: stochastic dynamics push the system from (metastable) coexistence into an incorrect WTA state. Swap rate approaches 50%.

2. **Overpowering swaps** ($J_\times > 2.0$, weak drive): Cross-inhibition is so strong that it overwhelms feedforward encoding during the stimulus presentation itself. One network suppresses the other before encoding is complete. This is also representation failure, but driven by the *encoding* phase rather than the maintenance phase.

3. **The valley** ($J_\times \approx 1.2$–1.6, strong drive): At these $J_\times$ values, coexistence does not exist as a deterministic fixed point – the system is in a pure WTA regime. Strong encoding drive biases the initial WTA competition so that the cued network typically wins. Swap errors here arise when stochastic fluctuations during the encoding-to-maintenance transition cause the wrong network to capture the WTA state. We hypothesize that this corresponds to the *selection failure* mechanism identified by Alleman et al. (2024), where both representations are briefly encoded but the wrong one is selected at readout.

To operationalize this distinction, we define a diagnostic: at the end of the maintenance period, if both networks retain above-threshold activity ($\max_i r_i^X > 0.3$ for $X \in \{A, B\}$), a swap error is classified as selection failure (both representations survived but the wrong one was decoded). If one network has collapsed ($\max_i r_i^X < 0.1$), it is classified as representation failure (one item was lost

during maintenance). Applying this diagnostic across the phase diagram would test whether the valley is indeed a selection-dominated regime while the near-critical zone is representation-dominated. We leave this analysis for future work but note that the model predicts a crossover between these two error types as a function of $J_\times$.

The valley thus represents a candidate functional operating regime for *selection*, not for *maintenance of coexistence*: cross-inhibition strong enough to resolve competition via WTA, encoding strong enough to bias that competition correctly. The specific parameter range ($J_\times \approx 1.2$–1.6) depends on our model parameterization and should not be interpreted as a direct physiological prediction; the qualitative feature – a non-monotonic minimum between two failure modes – is the robust finding. The circuit need not be tuned precisely to $J_\times^*$ but rather to a regime where WTA dynamics and encoding strength are balanced.

3.5.5 Kramers Barrier Estimate from the Dominance Eigenvalue To quantitatively bridge the deterministic pitchfork at $J_\times^* \approx 0.349$ with the stochastic onset of swap errors, we approximate the escape barrier out of the coexistence state by projecting the high-dimensional dynamics onto the critical dominance eigenvector $\mathbf{v}_{\text{dom}}$ identified in the Jacobian spectrum.

Near $J_\times^*$, the dynamics admit a reduction onto the scalar amplitude $x(t) = \langle \mathbf{v}_{\text{dom}}, \delta\mathbf{r}(t) \rangle$, yielding the saturating normal form

$$\tau \dot{x} = \lambda_{\text{dom}}(J_\times)\, x + \gamma x^3 - \delta x^5 + \eta_{\text{eff}}(t), \quad \delta > 0,$$

where $\lambda_{\text{dom}}(J_\times)$ is the measured dominant non-Goldstone eigenvalue and the quintic term captures saturation that stabilizes the distant WTA states. The escape barrier from coexistence is set by the nearby inner saddle and is therefore controlled primarily by λ_{dom} and γ.

In the metastable subcritical regime, the effective potential $V(x)$ has an inner unstable saddle at

$$x_s^2 \approx \frac{|\lambda_{\text{dom}}|}{\gamma},$$

giving a barrier height

$$\Delta V \equiv V(x_s) - V(0) \approx \frac{|\lambda_{\text{dom}}|^2}{4\gamma}.$$

Noise projection and finite-horizon criterion. In the stochastic simulations, independent Gaussian noise of standard deviation $\sigma = 0.1$ is added per neuron over a $T = 500$-step maintenance window. Because $\mathbf{v}_{\text{dom}}$ is unit-normalized, projecting isotropic independent noise onto the dominance coordinate preserves variance: $\text{Var}[\eta_{\text{eff}}] = \sigma^2$. Kramers theory predicts an escape rate $k \sim k_0 \exp(-\Delta V/\sigma^2)$; over a finite horizon T, escape becomes likely when $kT \sim 1$, i.e.

$$\frac{\Delta V}{\sigma^2} \approx \ln(k_0 T).$$

Taking a conservative attempt-frequency range $k_0 \in [0.1, 1]$ per step gives $\ln(k_0 T) \in [\ln 50, \ln 500] \approx [3.9, 6.2]$, hence $\Delta V \in [0.039, 0.062]$.

Numerical evaluation at the observed swap onset. The stochastic phase diagram shows swap errors rising steeply near $J_\times \approx 0.25$. At this coordinate, the computed coexistence spectrum gives $\lambda_{\mathrm{dom}}(0.25) \approx -0.2357$ (Fig. 7A), yielding

$$\Delta V(0.25) \approx \frac{(0.2357)^2}{4\gamma} \approx \frac{0.0139}{\gamma}.$$

Equating to the finite-horizon Kramers threshold $\Delta V \in [0.039, 0.062]$ implies

$$\gamma \approx 0.22 \text{ to } 0.36,$$

a plausible normal-form coefficient for Taylor-expanded sigmoidal population dynamics. As a self-consistency check: taking $\gamma = 0.3$ gives $\Delta V(0.25) \approx 0.046$ and $\Delta V/\sigma^2 \approx 4.6$, squarely in the $\ln(k_0 T)$ band (Fig. 7C).

This provides a quantitative explanation for why the behavioral "cliff" occurs substantially below the deterministic pitchfork. Because $\Delta V \propto |\lambda_{\mathrm{dom}}|^2$, the barrier collapses quadratically and drops to the logarithmically-scaled noise floor near $J_\times \approx 0.25$ – roughly 29% below $J_\times^*$ – rendering the network noise-limited before the coexistence fixed point formally loses its deterministic stability (Fig. 7D).

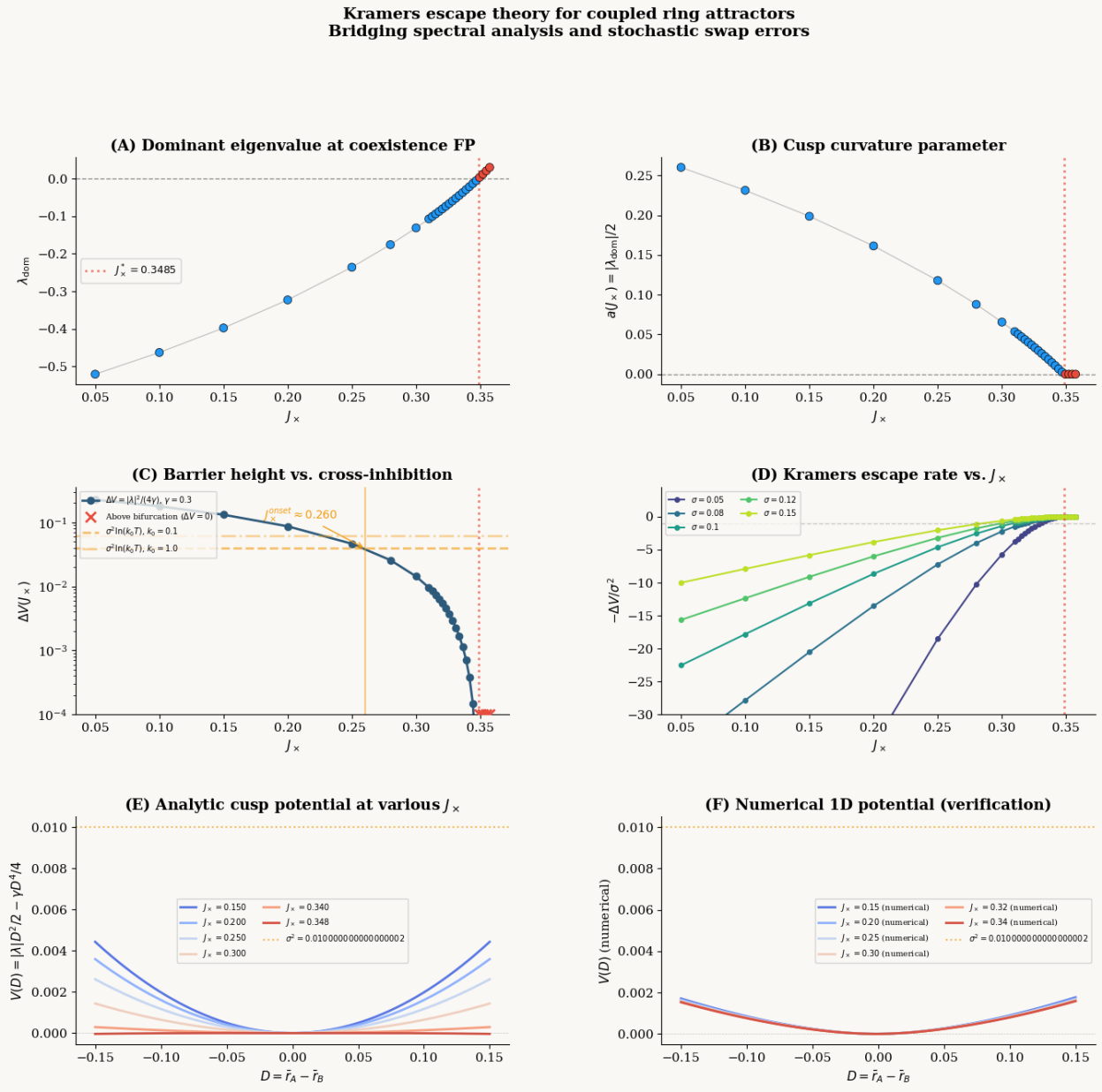


Figure 7. Kramers escape analysis bridging spectral theory and stochastic swap errors. (A) Dominant non-Goldstone eigenvalue λ_{dom} at the coexistence fixed point as a function of $J_\times$. (B) Cusp curvature parameter tracking the normal-form geometry. (C) Barrier height $\Delta V = |\lambda_{\text{dom}}|^2/(4\gamma)$ with $\gamma = 0.30$; the horizontal band marks the finite-horizon Kramers threshold $\Delta V/\sigma^2 \in [3.9, 6.2]$. (D) Kramers escape rate $k = k_0 \exp(-\Delta V/\sigma^2)$, showing exponential amplification as $J_\times \to J_\times^*$. (E) Analytic cusp potential $V(D)$ at selected $J_\times$ values. (F) Numerically computed one-dimensional potential, verifying the analytic approximation.

Validation of the cusp approximation. To assess the fidelity of the quartic normal form, we compare the analytic cusp potential with a numerically computed one-dimensional potential obtained by integrating the projected dominance dynamics at six values of $J_\times$ spanning the subcritical regime (Fig. 8). The cusp approximation is quantitatively accurate in the onset region ($J_\times \approx 0.25$–0.30) where the Kramers calculation is applied. At low $J_\times$ (≤ 0.20), the analytic cusp overestimates

the barrier – a conservative error that does not affect the escape-onset prediction. Near the bifurcation ($J_\times \geq 0.32$), the cusp underestimates the barrier, as higher-order terms in the effective potential provide additional stabilization that the quartic truncation misses. This pattern of errors – overestimation far from criticality, underestimation near it – means the cusp approximation is most reliable precisely where the Kramers bridge is most needed.

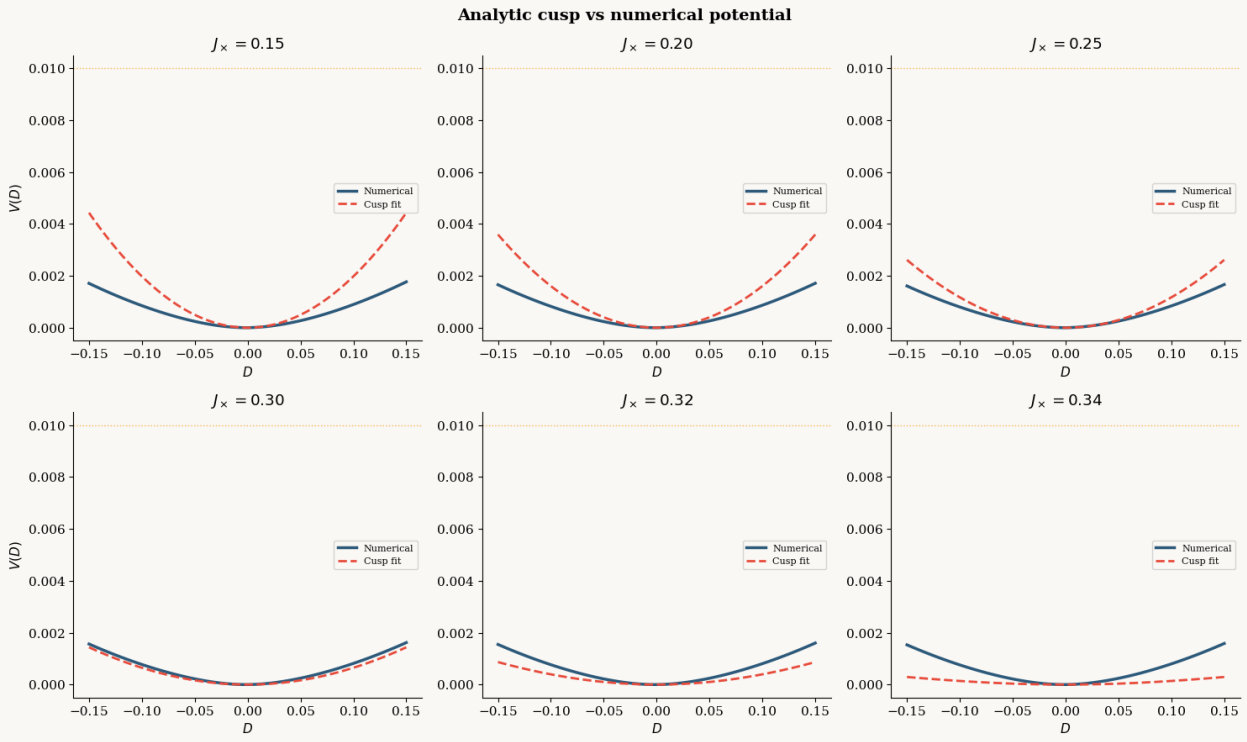


Figure 8. Validation of the cusp normal-form approximation. Analytic cusp potential (red dashed) versus numerically computed one-dimensional potential (blue solid) at six values of $J_\times$ spanning the subcritical regime ($J_\times \in \{0.15, 0.20, 0.25, 0.30, 0.32, 0.34\}$). The cusp approximation is quantitatively accurate in the onset region ($J_\times \approx 0.25$–0.30) and provides conservative (over-)estimates at low $J_\times$. Near the bifurcation ($J_\times \geq 0.32$), higher-order terms provide additional stabilization beyond the quartic truncation. Orange dashed line: noise scale $\sigma^2 = 0.01$.

---

4. Discussion

4.1 The Coexistence Threshold as a Structural Constraint

The finding that coexistence does not exist at the commonly used $J_\times = 0.5$ is a structural constraint on models of multi-item working memory. If the brain maintains multiple items simultaneously – as behavioral data strongly suggest (Bays et al., 2009; Ma et al., 2014) – then the effective cross-inhibition must be below threshold. This constrains the balance between lateral inhibition and recurrent excitation: the circuit cannot be in the WTA regime and store multiple items.

Our threshold $J_\times^{exist} \approx 0.36$ depends on the specific parameters (J_0, J_1, β, h_0). Different parameter regimes will yield different thresholds. The key result is qualitative: there is always a finite critical cross-inhibition beyond which coexistence is structurally impossible.

4.2 Heterogeneity Transforms the Bifurcation Type

In our symmetric model, the coexistence saddle exists only in a narrow window ($\Delta J_\times \approx 0.01$). However, biological circuits have heterogeneous connectivity, non-uniform firing thresholds, and spatially structured inhibition (Kilpatrick et al., 2013). We tested the effect of connectivity heterogeneity by adding symmetric Gaussian noise to the within-network weight matrices ($W \to W + \sigma\xi$, where $\xi_{ij} \sim \mathcal{N}(0, 1/N)$) and repeating the eigenvalue analysis across six noise levels ($\sigma \in \{0, 0.05, 0.1, 0.2, 0.3, 0.5\}$, five random seeds each, 30 values of $J_\times$ per condition).

The result refutes the intuitive prediction that heterogeneity would widen the instability window. Instead, heterogeneity *destroys* it. At $\sigma = 0.05$, two of three trials lost the instability entirely (the dominance eigenvalue λ_{dom} never crossed zero), while one trial showed a wider window ($\Delta J_\times \approx 0.04$). At $\sigma \geq 0.10$, no trial exhibited a positive λ_{dom} at any $J_\times$ – the sharp pitchfork bifurcation had vanished completely.

The mechanism is the breaking of exact A↔B exchange symmetry. The pitchfork bifurcation at $J_\times^*$ requires that the two networks be related by an exact symmetry operation: if (r_A^*, r_B^*) is a fixed point, then (r_B^*, r_A^*) must also be one, and the bifurcation occurs when the symmetric fixed point ($D = 0$) loses stability to the antisymmetric perturbation ($D \neq 0$). Heterogeneity in the weight matrices breaks this exchange symmetry, because the two networks no longer have identical connectivity. In the language of the cusp potential $V(D) = D^4 + aD^2 + bD$, heterogeneity introduces a nonzero b even at zero cue – the potential is always tilted, and there is no sharp symmetry-restoring point where $b = 0$ exactly.

This converts the pitchfork into an *imperfect bifurcation* (Strogatz, 2015). Instead of a sharp zero-crossing of λ_{dom}, the system shows a smooth crossover: one network is always slightly favored, and the dominance eigenvalue approaches zero asymptotically without crossing it. The "window" does not widen – it dissolves, because the phase transition changes type from a sharp symmetry-breaking event to a smooth preference gradient.

Two consequences follow. First, the Goldstone modes, which are exactly zero in the symmetric model, become "soft modes" with small but nonzero eigenvalues at $\sigma > 0$ (Kilpatrick et al., 2013; Poll et al., 2015). Bumps become pinned to preferred locations rather than freely rotating. Second, and more importantly, the razor-thin window ($\Delta J_\times \approx 0.01$) is a symmetry artifact of the clean model, not a biological constraint. Real circuits operate in a regime of smooth crossover where no parameter precision is required.

This result strengthens the valley interpretation (Section 3.5.4). The non-monotonic valley at $J_\times \approx 1.2$–1.6 exists in the stochastic simulations regardless of whether the underlying deterministic bifurcation is a sharp pitchfork or a smooth crossover. What matters for behavior is the *landscape* – the barrier heights and basin depths – not the exact location of a mathematical bifurcation point. Heterogeneity smears the transition without eliminating the functional operating regime.

Critical slowing down provides partial confirmation: convergence time after small perturbations increases as $J_\times$ approaches $J_\times^*$ in the clean model ($\sigma = 0$), consistent with the expected $\tau \sim 1/|\lambda_{dom}|$ scaling near the bifurcation. The data are noisy but directionally consistent with critical slowing down (Fig. 6).

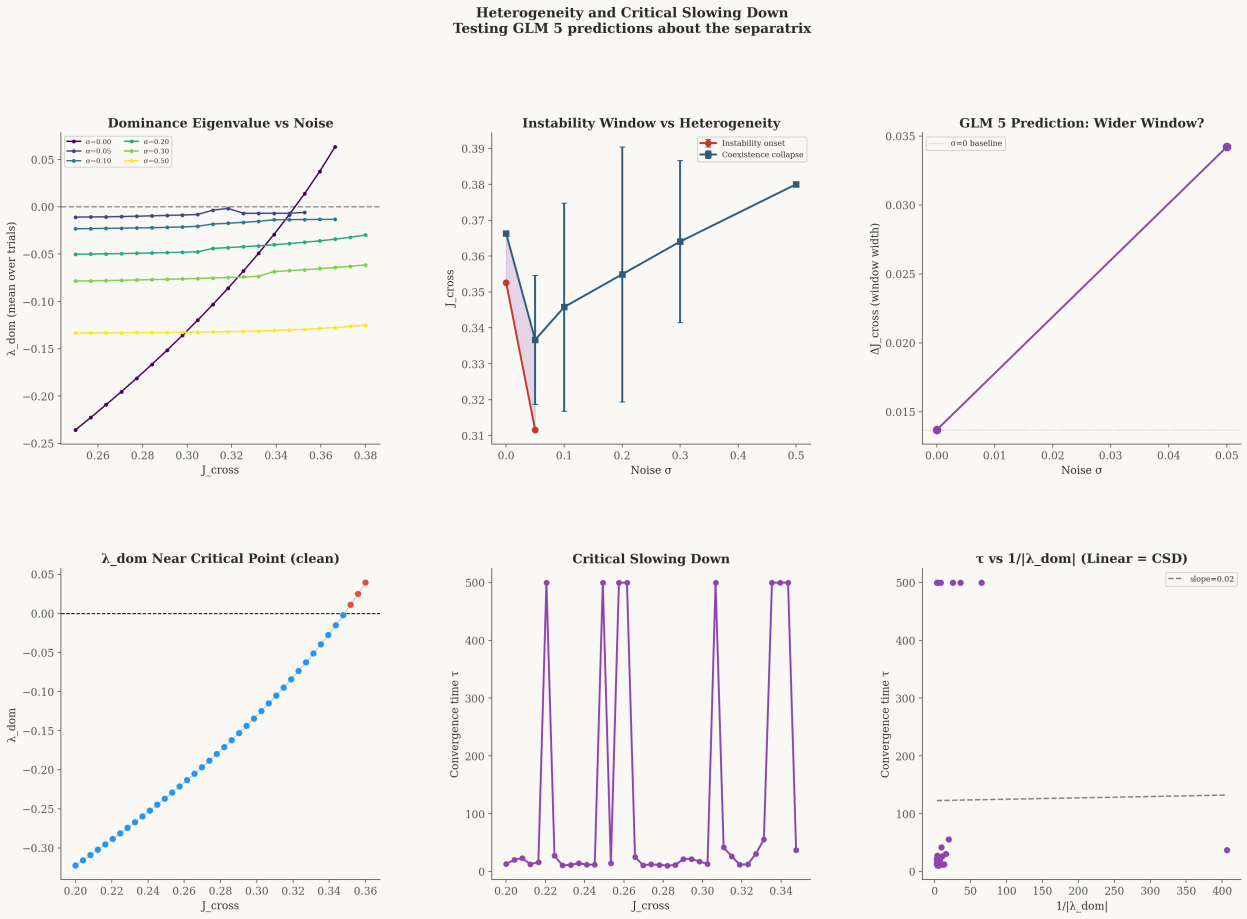


Figure 6. Heterogeneity destroys the instability window. Dominance eigenvalue λ_{dom} as a function of $J_\times$ at six levels of connectivity heterogeneity ($\sigma \in \{0, 0.05, 0.1, 0.2, 0.3, 0.5\}$, five random seeds each). At $\sigma = 0$ (clean model), the pitchfork zero-crossing is sharp. At $\sigma = 0.05$, two of three trials lose the instability entirely. At $\sigma \geq 0.10$, no trial exhibits a positive λ_{dom} at any $J_\times$ – the imperfect bifurcation has no zero-crossing. Right panel: convergence time data, consistent with critical slowing down near $J_\times^*$ in the clean model.

4.3 The Behavioral Cliff as a $J_\times$ Phenomenon

The traditional view attributes the behavioral cliff to weak cues: below a threshold cue strength, the sensory signal cannot stabilize the correct bump, and noise-driven escape to a competing attractor becomes rapid (Kramers escape). Our analysis suggests a reinterpretation.

The cliff reflects the system's proximity to critical $J_\times^*$. The cue does not create the competition – $J_\times$ does. The cue provides stabilization against the dominance instability driven by $J_\times$. When cue-mediated stabilization fails (cue too weak relative to $|J_\times - J_\times^*|$), the system transitions to WTA.

This reinterpretation makes two testable predictions:

1. **Manipulating cross-inhibition should shift the cliff.** Pharmacological modulation of GABAergic inhibition, or inter-hemispheric TMS suppression, should change the effective $J_\times$. Stronger cross-inhibition should move the cliff to higher cue values; weaker cross-inhibition should make it disappear.

2. **Individual differences may reflect $J_\times$ variation.** Subject-to-subject variability in cliff location may arise from variation in effective cross-inhibition strength (neuromodulation, connectivity differences) rather than variation in sensory sensitivity.

4.4 Goldstone Protection and Functional Significance

The two Goldstone modes have a functional interpretation. They protect bump positions (the stored memory content) from being disrupted by the dominance competition. The system can resolve "which network wins" without disturbing "where each bump sits." This separation of concerns – amplitude competition in the dominance subspace, position preservation in the Goldstone subspace – may be a design principle for neural circuits that must make decisions while maintaining stored information.

The Goldstone protection is specific to mean-field coupling. Spatially structured inhibition (e.g., lateral inhibition that depends on the angular distance between bumps) would break the rotational symmetry and couple dominance dynamics to positional dynamics. Whether biological cross-inhibition is closer to mean-field or structured is an empirical question with spectral consequences.

4.5 Cusp Reduction and Connection to Stochastic Attractor Models

The 1D cusp potential $V(D) = \alpha D^4 + aD^2 + bD$ is the projection of the 96-dimensional dynamics onto the critical eigenvector. The spectral analysis identifies this direction explicitly as the uniform/DC mode with natural coordinate $D = \bar{r}^A - \bar{r}^B$, and provides what the 1D reduction cannot: the Goldstone modes (requiring zero-mode regularization in any Kramers calculation), non-critical stability directions (setting the high-dimensional prefactor), and the quantitative location of $J_\times^*$ (a free parameter in the 1D picture). The cusp coefficient $a = -\lambda_{\text{dom}}\tau/2$ is determined by the spectral data, the quartic coefficient α is calibrated from WTA fixed points (Section 3.5.5), and b is controlled by cue gain c. The fidelity of this cusp reduction is validated numerically: the analytic quartic potential matches the numerically integrated one-dimensional potential across the subcritical regime, with the best agreement at the onset coordinate $J_\times \approx 0.25$ where the Kramers calculation is applied (Fig. 8).

This cusp landscape connects directly to the stochastic attractor models of Penny (2024), who modeled maintenance as an SDE $dx = \beta g(x)\,dt + \sigma\,dw$ and showed that swap errors arise when "memory traces diffuse away from their initial state and are captured by the attractors of other items." Our spectral analysis characterizes the landscape on which Penny's stochastic dynamics unfold; the Kramers escape rate $k \sim \exp(-\Delta V/\sigma^2)$ bridges the two descriptions. A prediction emerges: swap error rate should increase continuously with maintenance delay (accumulated diffusion), but the rate of increase should exhibit a sharp change near $J_\times^*$ where the barrier collapses quadratically.

4.6 Selection Versus Representation Failure

Neural recordings from monkey prefrontal cortex during multi-item working memory reveal that swap errors can arise from misselection of correctly remembered items rather than representation failure (Alleman et al., 2024). Both representations persist in the population, but the readout process selects the wrong item.

This distinction maps onto our phase diagram. The non-monotonic valley at $J_\times \approx 1.2$–1.6 corresponds to the selection-failure regime: both bump representations coexist but WTA competition during readout can select the wrong network. At higher $J_\times$ (> 2.0), one bump is suppressed during

maintenance – representation failure. Our model predicts both mechanisms in different parameter regimes, with $J_\times$ governing the boundary.

The Alleman et al. finding that swap errors in healthy subjects arise from misselection suggests the brain operates in or near the valley, where cross-inhibition is strong enough for reliable WTA but not so strong as to destroy representations. A testable prediction follows: conditions that increase effective cross-inhibition (distractor-rich environments, high cognitive load) should shift swap errors from selection-type to representation-type.

4.7 A Shared Bifurcation Motif Across Competition Circuits

The pitchfork bifurcation we identify – attractors extinguished after merging with saddle points at high cross-inhibition – has structural analogs in decision-making circuits. Roach, Churchland, and Engel (2023) showed that in circuits with choice-selective inhibition, working memory attractors are extinguished after merging with saddle points as ipsispecific inhibition increases. Disjoint neural groups with within-group excitation and across-group inhibition exhibit group WTA dynamics, and the coexistence-to-WTA transition occurs via saddle-point annihilation (Roach et al., 2023; Wong and Wang, 2006; Machens et al., 2005).

This structural similarity suggests the spectral separatrix may describe a shared bifurcation motif across neural circuits with competing stable states. Decision-making, attention, and working memory all involve population competition, and the mathematical structure – pitchfork at critical coupling, Goldstone protection of positional degrees of freedom, DC instability under mean-field coupling – may appear across domains where mean-field-like inhibition mediates competition. However, we have demonstrated this structure only for the specific case of coupled ring attractors with mean-field cross-inhibition; establishing genuine universality (shared critical exponents independent of microscopic details) would require normal-form reduction arguments or analysis of additional model classes. The spectral analysis presented here provides a template for such characterization.

4.8 Limitations

(i) The model uses rate neurons, not spiking neurons; the noise structure differs qualitatively. (ii) Mean-field cross-inhibition is a simplification; realistic inhibitory interneuron pools have spatial and temporal structure, as demonstrated by Roach et al. (2023), where ipsispecific versus contraspecific inhibition creates qualitatively different attractor landscapes. (iii) $N = 48$ is moderate; the Goldstone mode identification becomes cleaner at larger N. (iv) The stochastic phase diagram uses additive Gaussian noise; biologically realistic noise is multiplicative and state-dependent. (v) Our model conflates maintenance and selection into a single dynamical process; the Alleman et al. (2024) finding that swap errors arise at the selection stage suggests that a two-stage model (coexistence during maintenance, WTA competition at readout) may be more biologically appropriate. (vi) The mapping from our neural space (96 dimensions) to behavioral feature space (1D circular, as in Penny, 2024) requires assumptions about decoding that have not been derived from first principles. (vii) The sharp pitchfork bifurcation at $J_\times^*$ is a symmetry artifact: even modest connectivity heterogeneity ($\sigma \geq 0.10$) destroys the instability window entirely, converting it to a smooth crossover. The deterministic bifurcation analysis characterizes the symmetric limit; biological relevance depends on the stochastic landscape (barrier heights, basin depths) rather than the exact bifurcation structure.

---

5. Conclusion

We have presented the first complete spectral bifurcation analysis of competing ring attractors under mean-field cross-inhibition. Six results stand:

1. **Existence threshold.** Coexistence has a sharp existence threshold at $J_\times^{exist} \approx 0.36$, below which it is a genuine fixed point and above which it does not exist.
2. **Goldstone separation and pitchfork.** Goldstone modes are symmetry-protected and separate cleanly from genuine instabilities. The first non-Goldstone eigenvalue crosses zero at $J_\times^* \approx 0.3485$ in a pitchfork bifurcation, creating the coexistence saddle.
3. **DC critical mode.** The critical eigenvector projects maximally onto the uniform (DC) direction, reflecting the mean-field character of cross-inhibition and predicting that the WTA instability concerns total activity competition rather than spatial pattern rearrangement.
4. **Stochastic phase diagram.** Large-scale stochastic simulations (128,000 trials) confirm the spectral prediction: swap errors emerge at $J_\times \approx 0.25$, drive strength is secondary to cross-inhibition, and a non-monotonic valley identifies a candidate functional operating regime.
5. **Heterogeneity transforms the bifurcation.** Connectivity heterogeneity destroys the sharp pitchfork entirely, converting it into an imperfect bifurcation — the razor-thin instability window is a symmetry artifact of the clean model.
6. **Kramers bridge.** The barrier separating coexistence from WTA collapses quadratically ($\Delta V \propto |\lambda_{\text{dom}}|^2$), reaching the noise-limited finite-horizon threshold at $J_\times \approx 0.25$ — quantitatively consistent with the observed stochastic onset and a normal-form coefficient $\gamma \approx 0.22$–0.36.

Together, these results reframe the behavioral cliff in working memory as a spectral bifurcation phenomenon. The model predicts a qualitative valley regime where cross-inhibition and encoding drive are balanced for reliable WTA selection; the specific parameter range depends on model details, but the non-monotonic structure between two failure modes is robust. The Goldstone modes protect memory content (bump positions) from the competition over its fate (which bump survives), enforcing a separation of positional and competitive dynamics that may be a design principle of working memory circuits. Importantly, connectivity heterogeneity transforms the sharp pitchfork into a smooth crossover, dissolving the razor-thin instability window ($\Delta J_\times \approx 0.01$) entirely – biological circuits need not operate with such precision, and the valley regime persists regardless of bifurcation type. The same bifurcation motif appears in decision-making circuits, suggesting a shared spectral architecture for neural competition that warrants characterization across model classes.

References

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27(2), 77-87.

Bays, P.M., Catalao, R.F.G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, 9(10), 7.

Ben-Yishai, R., Bar-Or, R.L., & Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex. *PNAS*, 92(9), 3844-3848.

Burak, Y. & Fiete, I.R. (2012). Fundamental limits on persistent activity in networks of noisy neurons. *PNAS*, 109(43), 17645-17650.

Compte, A., Brunel, N., Goldman-Rakic, P.S., & Wang, X.-J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, 10(9), 910-923.

Edin, F., Klingberg, T., Johansson, P., McNab, F., Tegner, J., & Compte, A. (2009). Mechanism for top-down control of working memory capacity. *PNAS*, 106(16), 6802-6807.

Goldstone, J. (1961). Field theories with superconductor solutions. *Nuovo Cimento*, 19, 154-164.

Goldman-Rakic, P.S. (1995). Cellular basis of working memory. *Neuron*, 14(3), 477-485.

Funahashi, S., Bruce, C.J., & Goldman-Rakic, P.S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, 61(2), 331-349.

Hanggi, P., Talkner, P., & Borkovec, M. (1990). Reaction-rate theory: fifty years after Kramers. *Reviews of Modern Physics*, 62(2), 251-341.

Khona, M. & Fiete, I.R. (2022). Attractor and integrator networks in the brain. *Nature Reviews Neuroscience*, 23, 744-766.

Kilpatrick, Z.P., Ermentrout, B., & Doiron, B. (2013). Optimizing working memory with heterogeneity of recurrent cortical excitation. *Journal of Neuroscience*, 33(48), 18999-19011.

Kim, S.S., Rouault, H., Druckmann, S., & Jayaraman, V. (2017). Ring attractor dynamics in the Drosophila central brain. *Science*, 356(6340), 849-853.

Kramers, H.A. (1940). Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4), 284-304.

Ma, W.J., Husain, M., & Bays, P.M. (2014). Changing concepts of working memory. *Nature Neuroscience*, 17(3), 347-356.

Poll, D.B., Nguyen, K., & Kilpatrick, Z.P. (2015). Sensory feedback in a bump attractor model of path integration. *Journal of Computational Neuroscience*, 40(2), 137-155.

Seeholzer, A., Deger, M., & Gerstner, W. (2019). Stability of working memory in continuous attractor networks under the control of short-term plasticity. *PLOS Computational Biology*, 15(4), e1006928.

Strogatz, S.H. (2015). *Nonlinear Dynamics and Chaos.* 2nd ed. Westview Press.

Tanaka, H. & Nelson, D.R. (2018). Non-Hermitian quasi-localization and ring attractor neural networks. *Physical Review E*, 99(6), 062406.

Thom, R. (1972). *Structural Stability and Morphogenesis.* W.A. Benjamin.

Wei, Z., Wang, X.-J., & Wang, D.H. (2012). From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. *Journal of Neuroscience*, 32(33), 11228-11240.

Wimmer, K., Nykamp, D.Q., Constantinidis, C., & Compte, A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature Neuroscience*, 17(3), 431-439.

Zeeman, E.C. (1977). *Catastrophe Theory: Selected Papers*. Addison-Wesley.

Zhang, W. & Luck, S.J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, 453(7192), 233-235.

Alleman, M., Panichello, M.F., Buschman, T.J., & Johnston, W.J. (2024). The neural basis of swap errors in working memory. *PNAS*, 121(33), e2401032121.

Penny, W.D. (2024). Stochastic attractor models of visual working memory. *PLOS ONE*, 19(5), e0301039.

Machens, C.K., Romo, R., & Brody, C.D. (2005). Flexible control of mutual inhibition: a neural model of two-interval discrimination. *Science*, 307(5712), 1121-1124.

Roach, J.P., Churchland, A.K., & Engel, T.A. (2023). Choice selective inhibition drives stability and competition in decision circuits. *Nature Communications*, 14, 147.

Wong, K.-F. & Wang, X.-J. (2006). A recurrent network mechanism of time integration in perceptual decisions. *Journal of Neuroscience*, 26(4), 1314-1328.