Online Appendix for Coarse Q-learning: Indifference, Indeterminacy, and Instability (Jehiel & Satpathy (2026))

A Projection, step-sizes and stochastic approximation

This section supplies the technical details behind the discussion on stochastic approximation in Section 2. We fix a finite reduced tree $\mathcal{T}' = (\mathcal{S}, \Omega, p, \pi)$ and a payoff sensitivity parameter $\beta < \infty$. Throughout this section we state the stochastic-approximation arguments for the myopic case $\gamma = 0$. For $\gamma \in (0, 1)$, the continuation term enters the drift as a common scalar multiple of $\mathbf{1}$, so the relative-valuation dynamics are unchanged by Lemma B.9; accordingly, all results below extend verbatim after translating out the common continuation term.

For each class $s \in \mathcal{S}$, let $\Omega_s := \{\omega \in \Omega : \ s \in \omega\}$, $\Xi_s(\mathbf{v}) := \sum_{\omega \in \Omega_s} p(\omega)\, \sigma^s_\omega(\mathbf{v})$, so that $\Xi_s(\mathbf{v})$ is the probability (or propensity) that class s is selected under valuations $\mathbf{v}$. The conditional expected payoff of class s is

$$g_s(\mathbf{v}) := \frac{\sum_{\omega \in \Omega_s} p(\omega)\, \sigma^s_\omega(\mathbf{v})\, \pi_s(\omega)}{\sum_{\omega \in \Omega_s} p(\omega)\, \sigma^s_\omega(\mathbf{v})} = \frac{\sum_{\omega \in \Omega_s} p(\omega)\, \sigma^s_\omega(\mathbf{v})\, \pi_s(\omega)}{\Xi_s(\mathbf{v})}.$$

Because $\beta < \infty$, every available class is chosen with strictly positive probability in every menu in which it appears. Hence $\Xi_s(\mathbf{v}) > 0$ for all s and all $\mathbf{v} \in \mathbb{R}^{\mathcal{S}}$. We write $F(\mathbf{v}) := g(\mathbf{v}) - \mathbf{v}$, $\Lambda(\mathbf{v}) := \mathrm{diag}(\Xi_s(\mathbf{v}))_{s \in \mathcal{S}}$.

A.1 Projection and boundedness

For each $s \in \mathcal{S}$, let $m_s := \min_{\omega \in \Omega_s \cap \mathrm{supp}(p)} \pi_s(\omega)$, $M_s := \max_{\omega \in \Omega_s \cap \mathrm{supp}(p)} \pi_s(\omega)$, $K := \prod_{s \in \mathcal{S}} [m_s, M_s]$, with $m_s < M_s$. Let $\Pi_K : \mathbb{R}^{\mathcal{S}} \to K$ denote the coordinate-wise projection onto K.

Lemma A.1. *For every $\mathbf{v} \in \mathbb{R}^{\mathcal{S}}$, one has $g(\mathbf{v}) \in K$. Consequently,*

(i) the box K is positively invariant for the ODE $\dot{\mathbf{v}} = F(\mathbf{v})$;

(ii) if $\mathbf{v} \in K$, then $F(\mathbf{v}) \in T_K(\mathbf{v})$, where $T_K(\mathbf{v})$ is the tangent cone of K at $\mathbf{v}$;

(iii) the projected differential inclusion $\dot{\mathbf{v}} \in \Pi_{T_K(\mathbf{v})} F(\mathbf{v})$ coincides with the original mean-field ODE $\dot{\mathbf{v}} = F(\mathbf{v})$ on K.

The same conclusion holds for the diagonally preconditioned (asynchronous) field $\Lambda(\mathbf{v})F(\mathbf{v})$.

Proof. For fixed s, the weights $\omega \mapsto \dfrac{p(\omega)\sigma^s_\omega(\mathbf{v})}{\Xi_s(\mathbf{v})}$ form a probability vector on Ω_s, so $g_s(\mathbf{v})$ is

a convex combination of $\{\pi_s(\omega) : \omega \in \Omega_s\}$. Hence $g_s(\mathbf{v}) \in [m_s, M_s]$, proving $g(\mathbf{v}) \in K$. If $v_s = M_s$, then $F_s(\mathbf{v}) = g_s(\mathbf{v}) - M_s \le 0$; if $v_s = m_s$, then $F_s(\mathbf{v}) = g_s(\mathbf{v}) - m_s \ge 0$. This is exactly the inward-pointing condition for the box K, proving positive invariance and showing that $F(\mathbf{v}) \in T_K(\mathbf{v})$ for every $\mathbf{v} \in K$. Because $\Lambda(\mathbf{v})$ is diagonal with strictly positive entries, $\Lambda(\mathbf{v})F(\mathbf{v})$ has the same sign on every constrained coordinate as $F(\mathbf{v})$, so the same tangent-cone argument applies. For $\dot{\mathbf{v}} = F(\mathbf{v})$, by comparison with $\dot{x} = M_s - x$ and $\dot{y} = m_s - y$, for any solution and all $t \ge 0$, $m_s + (v_s(0) - m_s)e^{-t} \le v_s(t) \le M_s + (v_s(0) - M_s)e^{-t}$, $\forall\, s \in \mathcal{S}$. Finally, once the vector field already belongs to the tangent cone, projection onto that cone leaves it unchanged. Moreover, for each s, Ξ_s is continuous and strictly positive on the compact set K, hence $\underline{\Xi_s} := \min_{\mathbf{v}\in K} \Xi_s(\mathbf{v}) > 0$. Since $\mathbf{v} \mapsto \sigma^s_\omega(\mathbf{v})$ is C^∞, the map g is C^1 on a neighborhood of K. By compactness, $\sup_{\mathbf{v}\in K} \|Dg(\mathbf{v})\| < \infty$, so g is Lipschitz on K. □

Lemma A.1 justifies the usual projected recursion when primitive payoffs are not bounded. Indeed, if $\tilde{\mathbf{v}}_{k+1} = \mathbf{v}_k + \Delta_{k+1}$ is the tentative update under any of the gain specifications below, we may w.l.o.g. set $\mathbf{v}_{k+1} := \Pi_K(\tilde{\mathbf{v}}_{k+1})$. By construction $\mathbf{v}_k \in K$ for all k, and the limiting projected ODE coincides with the original mean field on K. Thus projection is therefore a purely technical device used to obviate any boundedness assumption on the discrete-time recursion; it does not alter the rest points, local Jacobians, or limit sets relevant for the analysis. This is the standard constrained-stochastic-approximation device treated in the ODE method and projected-recursion framework of (Kushner and Yin, 2003, Ch. 5).

A.2 Deterministic gains and the asynchronous mean field

We first consider the class-invariant calendar-time gain sequence $\{\alpha_k\}_{k\ge 0}$, with $\alpha_k > 0$, $\sum_k \alpha_k = \infty$, and $\sum_k \alpha_k^2 < \infty$.. For e.g., $\alpha_k = (k+1)^{-1}$. The projected recursion is

$$\mathbf{v}_{k+1} = \Pi_K\Big(\mathbf{v}_k + \alpha_k\, \mathbf{e}_{s_k}\big[r_k - v_k(s_k)\big]\Big). \tag{1}$$

Define the noise vector $H_{k+1} \in \mathbb{R}^{\mathcal{S}}$ by

$$H_{k+1}(s) := \mathbf{1}\{s_k = s\}\big[r_k - v_k(s)\big], \qquad s \in \mathcal{S}.$$

Then, conditional on $\mathcal{I}_k$,

$$\mathbb{E}\big[H_{k+1}(s) \mid \mathcal{I}_k\big] = \Xi_s(\mathbf{v}_k)\big(g_s(\mathbf{v}_k) - v_k(s)\big).$$

Thus the systematic drift of (1) is $\Lambda(\mathbf{v}_k)F(\mathbf{v}_k)$. To formulate the stochastic-approximation limit precisely, we introduce the standard continuous-time interpolations associated with the

gain sequences. For the slow valuation recursion, let

$$t_0 := 0, \qquad t_n := \sum_{j=0}^{n-1} \alpha_j, \qquad n \geq 1,$$

and define the affine interpolation $\bar{\mathbf{v}} : [0, \infty) \to K$ by

$$\bar{\mathbf{v}}(t_n + \theta\alpha_n) := (1-\theta)\mathbf{v}_n + \theta\mathbf{v}_{n+1}, \qquad \theta \in [0,1].$$

When the two-timescale recursion (3)–(4) is considered, we also define the fast clock

$$\tau_0 := 0, \qquad \tau_n := \sum_{j=0}^{n-1} \lambda_j, \qquad n \geq 1,$$

and the affine interpolation $\bar{\Xi} : [0, \infty) \to \mathbb{R}^{\mathcal{S}}$ by

$$\bar{\Xi}(\tau_n + \theta\lambda_n) := (1-\theta)\widehat{\Xi}_n + \theta\widehat{\Xi}_{n+1}, \qquad \theta \in [0,1].$$

All references below to "continuous-time interpolation" refer to these affine interpolants.

Proposition 1 (Asynchronous mean field)**.** *Assume that the projected recursion (1) is used and that primitive rewards satisfy the moment bound imposed in Section 2. Then the affine interpolation $\bar{\mathbf{v}}$ of $\{\mathbf{v}_k\}$ in the α-clock is an asymptotic pseudo-trajectory of the projected ODE $\dot{\mathbf{v}} = \Lambda(\mathbf{v})F(\mathbf{v}) + \mathbf{z}$, $\mathbf{z} \in -C_K(\mathbf{v})$, where $C_K(\mathbf{v})$ is the normal cone of K at $\mathbf{v}$. Equivalently, by Lemma A.1, the interpolation is an asymptotic pseudo-trajectory of*

$$\dot{v}_s = \Xi_s(\mathbf{v})\Big(g_s(\mathbf{v}) - v_s\Big), \qquad s \in \mathcal{S}. \tag{2}$$

Proof. Let $h(\mathbf{v}) := \Lambda(\mathbf{v})F(\mathbf{v})$, $H_{k+1}(s) := \mathbf{1}\{s_k = s\}\Big(r_k - v_k(s)\Big)$, and define the noise term

$$\eta_{k+1}(s) := H_{k+1}(s) - \mathbb{E}[H_{k+1}(s) \mid \mathcal{I}_k], \qquad s \in \mathcal{S}.$$

Then $\{\boldsymbol{\eta}_{k+1}\}_{k\geq 0}$ is an $(\mathcal{I}_k)$-martingale-difference noise sequence, and by construction

$$\mathbb{E}[H_{k+1}(s) \mid \mathcal{I}_k] = \Xi_s(\mathbf{v}_k)\Big(g_s(\mathbf{v}_k) - v_k(s)\Big) = h_s(\mathbf{v}_k).$$

Hence the projected recursion (1) can be written exactly as

$$\mathbf{v}_{k+1} = \Pi_K\Big(\mathbf{v}_k + \alpha_k\Big(h(\mathbf{v}_k) + \boldsymbol{\eta}_{k+1}\Big)\Big).$$

We now verify the standard hypotheses for projected stochastic approximation. First, by Lemma A.1, the iterates remain in the compact set K. Second, since K is compact and g and Ξ are continuous on K, the drift $h(\mathbf{v}) = \Lambda(\mathbf{v})F(\mathbf{v})$ is continuous, indeed Lipschitz, on K. Third, the payoff moment bound from Section 2 implies a uniform conditional second-moment bound for the noise. Indeed, because $\mathcal{S}$ is finite and $\mathbf{v}_k \in K$,

$$\|\boldsymbol{\eta}_{k+1}\|^2 \le 2\|H_{k+1}\|^2 + 2\Big\|\mathbb{E}[H_{k+1} \mid \mathcal{I}_k]\Big\|^2 \le C\Big(1 + |r_k|^2\Big)$$

for some deterministic constant $C < \infty$, and therefore

$$\sup_k \mathbb{E}\Big[\|\boldsymbol{\eta}_{k+1}\|^2 \mid \mathcal{I}_k\Big] < \infty \qquad \text{a.s.}$$

Finally, the gain sequence satisfies the Robbins-Monro conditions by assumption.

The projected-ODE theorem for stochastic approximation therefore applies to the affine interpolation $\bar{\mathbf{v}}$ in the α-clock, yielding that $\bar{\mathbf{v}}$ is an asymptotic pseudo-trajectory of the projected differential inclusion $\dot{\mathbf{v}} = h(\mathbf{v}) + \mathbf{z},\ \mathbf{z} \in -C_K(\mathbf{v})$, where $C_K(\mathbf{v})$ denotes the normal cone of K at $\mathbf{v}$. Substituting $h(\mathbf{v}) = \Lambda(\mathbf{v})F(\mathbf{v})$ gives $\dot{\mathbf{v}} = \Lambda(\mathbf{v})F(\mathbf{v}) + \mathbf{z},\ \mathbf{z} \in -C_K(\mathbf{v})$.

By Lemma A.1, the vector field $\Lambda(\mathbf{v})F(\mathbf{v})$ already lies in the tangent cone $T_K(\mathbf{v})$ for every $\mathbf{v} \in K$. Hence the projected differential inclusion coincides on K with the unprojected ODE $\dot{v}_s = \Xi_s(\mathbf{v})\Big(g_s(\mathbf{v}) - v_s\Big),\ s \in \mathcal{S}$, which is exactly (2). □

Equation (2) is the *asynchronous* mean field: in calendar time, class s moves at speed proportional to its current selection propensity $\Xi_s(\mathbf{v})$. As a result, rarely selected classes evolve more slowly than frequently selected ones. As a result, information accrues at very different rates in calendar time with the disparity growing in the payoff sensitivity parameter. Because choices depend on *relative* valuations, this imbalance can induce systematic miscalibration in cross-class comparisons over any finite horizon, even if payoffs are i.i.d. conditional on class in Alice's subjective model.

A natural behavioral response is therefore to reweight the impact of observations so as to keep the valuation vector more evenly calibrated in calendar time and thereby improve the accuracy of relative valuations and hence the quality of choices at any finite horizon. A related behavioral intuition is that rare outcomes can exert disproportionate influence on attention and memory. See, for e.g., the von Restorff (isolation) effect and related evidence on memory (Parker et al., 1998), and the amplification of dopaminergic responses to rare rewards (Rothenhoefer et al., 2021).

A.3 IP-normalized gains and the synchronous mean field

To synchronize effective learning speeds (sample sizes) across classes while lacking knowledge of the availability law f, Alice normalizes the step-size by the empirical selection propensity - an approach called inverse propensity weighting (IPW).[1] Intuitively, IPW compensates for the under-representation of rarely selected classes by up-weighting their observed rewards: when a class is chosen with low probability, its reward prediction error has greater influence on the valuation than when frequent updates occur. Formally, for each $s \in \mathcal{S}$, she sets $\alpha_k(s) = \alpha_k \cdot (\widehat{\Xi}_k(s))^{-1}$, where $\{\alpha_k\}_{k\geq 0}$ is a deterministic class-invariant sequence of step-sizes satisfying the Robbins-Monro conditions and $\widehat{\Xi}_k(s) \in (0,1)$ is Alice's period-k online estimate of the probability that class s is selected under the current policy induced by $\mathbf{v}_k$.

Since selection propensities drift as valuations evolve, $\widehat{\boldsymbol{\Xi}}_k$ must adapt on a faster timescale than $\mathbf{v}_k$, motivating a two-timescale online recursion. Behaviorally, the two-timescale separation captures that Alice calibrates how frequently each class is sampled faster than she revises her estimates about its payoffs.[2] Thus, Alice recursively estimates selection propensities $\widehat{\boldsymbol{\Xi}}_k$ on a faster timescale, and implements a normalized step-size $\alpha_k/\widehat{\Xi}_k(s)$ for updating her valuations à la Q-learning on the slower timescale. The two–timescale condition guarantees that the slow recursion for $v_k(s)$ sees $\widehat{\Xi}_k(s)$ approximately at its quasi–steady limit $\Xi_s(\mathbf{v}_k) = \sum_{\psi\in\Psi} f(\psi)\,\mathbf{1}\{s \in \mathcal{S}_\psi\}\,\sigma^s_\psi(\mathbf{v}_k) = \sum_{\omega\in\Omega_s} p(\omega)\,\sigma^s_\omega(\mathbf{v}) \in (0,1)$. This normalization equalizes effective speeds of information arrival (sample sizes) across classes in calendar time without altering the choice probabilities for a given valuation vector.

We now introduce inverse-propensity normalization. Let $\{\alpha_k\}_{k\geq 0}$ and $\{\lambda_k\}_{k\geq 0}$ satisfy

$$\sum_k \alpha_k = \infty, \qquad \sum_k \alpha_k^2 < \infty, \qquad \sum_k \lambda_k = \infty, \qquad \sum_k \lambda_k^2 < \infty, \qquad \lim_{k\to\infty} \frac{\alpha_k}{\lambda_k} = 0.$$

For instance, $\alpha_k = 1/(k+2)$, $\lambda_k = 1/(k+2)^\theta$ with $\frac{1}{2} < \theta < 1$. The fast recursion estimates selection propensities:

$$\widehat{\Xi}_{k+1}(s) = \widehat{\Xi}_k(s) + \lambda_k\Big(\mathbf{1}\{s_k = s\} - \widehat{\Xi}_k(s)\Big), \qquad s \in \mathcal{S}. \tag{3}$$

[1] We introduce inverse propensity weighting (IPW) following Fudenberg and Levine (1998) in the context of fictitious play learning with bandit feedback and Leslie and Collins (2005) for multi-agent Q-learning.

[2] The propensity process $\widehat{\Xi}_k$ is a running calibration of how often each class is selected under the current policy: updating it only requires registering the realized choice s_k, a binary and essentially noise-free signal that is observed every period. By contrast, the valuation process aggregates payoff realizations conditional on selection. Those signals are both noisier and effectively sparser (each class is informative only on the subsequence of periods in which it is chosen), so it is natural that valuation adjustment is more inertial.

We assume $\widehat{\Xi}_0(s) \in \Delta(\mathcal{S})$ and, for simplicity, $0 < \lambda_k \leq 1$ for all $k \geq 0$, so that the fast estimator remains in the simplex. Fix $\underline{\Xi} := \min_{s\in\mathcal{S}} \inf_{\mathbf{v}\in K} \Xi_s(\mathbf{v}) > 0$, which exists because K is compact and each Ξ_s is continuous and strictly positive on K. Choose any $\varepsilon \in (0, \underline{\Xi})$, and define the clipped estimator $\bar{\Xi}_k(s) := \max\{\widehat{\Xi}_k(s), \varepsilon\}$. The slow recursion is

$$\mathbf{v}_{k+1} = \Pi_K\Big(\mathbf{v}_k + \alpha_k \frac{\mathbf{e}_{s_k}}{\bar{\Xi}_k(s_k)}\big[r_k - v_k(s_k)\big]\Big). \tag{4}$$

Conditional on Alice's current information $\mathcal{I}_k$,

$$\mathbb{E}\left[\frac{\mathbf{1}\{s_k = s\}}{\bar{\Xi}_k(s)}\big(r_k - v_k(s)\big) \,\middle|\, \mathcal{I}_k\right] = \frac{\Xi_s(\mathbf{v}_k)}{\bar{\Xi}_k(s)}\big(g_s(\mathbf{v}_k) - v_k(s)\big).$$

The normalization is useful, since on the slow timescale, the ratio $\Xi_s(\mathbf{v}_k)/\bar{\Xi}_k(s) \to 1$.

Proposition 2 (IPW-normalization and the synchronous mean field)**.** *Assume the moment condition of Section 2 and consider the coupled recursion (3)–(4). Then:*

(i) for each fixed $\mathbf{v} \in K$, the fast interpolation $\bar{\Xi}$ in the λ-clock tracks the globally asymptotically stable equilibrium $x = \Xi(\mathbf{v})$ of the frozen fast ODE $\dot{x}_s = \Xi_s(\mathbf{v}) - x_s,\ s \in \mathcal{S}$;

(ii) $\widehat{\Xi}_k - \Xi(\mathbf{v}_k) \to 0$ almost surely;

(iii) the affine interpolation $\bar{\mathbf{v}}$ of $\{\mathbf{v}_k\}$ in the α-clock is an asymptotic pseudo-trajectory of the projected ODE $\dot{\mathbf{v}} = F(\mathbf{v}) + \mathbf{z},\ \mathbf{z} \in -C_K(\mathbf{v})$, and hence, by Lemma A.1, of the unprojected mean field

$$\dot{\mathbf{v}} = g(\mathbf{v}) - \mathbf{v}. \tag{5}$$

Proof. Part (i) is immediate, since the fast ODE is linear and thus globally exponentially stable. For part (ii), the fast recursion (3) is a standard stochastic approximation to that ODE, with the slow state $\mathbf{v}_k$ frozen on the fast timescale because $\alpha_k/\lambda_k \to 0$. The usual two-time-scale stochastic-approximation theorem therefore yields $\widehat{\Xi}_k - \Xi(\mathbf{v}_k) \to 0$ almost surely (Kushner and Yin, 2003, Sec. 8.6.1.). Because $\varepsilon < \underline{\Xi}$, clipping is eventually inactive along the slow recursion. Substituting the quasi-steady relation $\bar{\Xi}_k(s) \approx \Xi_s(\mathbf{v}_k)$ into the conditional drift gives

$$\frac{\Xi_s(\mathbf{v}_k)}{\bar{\Xi}_k(s)}\big(g_s(\mathbf{v}_k) - v_k(s)\big) = \big(g_s(\mathbf{v}_k) - v_k(s)\big) + o(1),$$

uniformly on compact subsets of K. Thus the slow recursion has drift $F(\mathbf{v})$, up to an

asymptotically negligible error and the inward projection term. Applying the projected ODE method to the slow recursion yields the stated limit. □

Equation (5) is the *synchronous* mean field appearing in the main text. The normalization removes the endogenous calendar-time factor $\Xi_s(\mathbf{v})$ from each component of the drift and therefore equalizes effective learning speeds across classes. All results in the paper are proved for the synchronous formulation, and the same arguments extend to the asynchronous formulation after replacing F by ΛF, as we show below.

B Robustness under Asynchronous CQL dynamics

Part (i) of Proposition 3 below implies that the SVE correspondence and its high-sensitivity accumulation set are unchanged under asynchronous updating implying Theorem 1 is robust to asynchronous updating. Part (ii) implies that all results relying on equilibrium identities and local Jacobian stability in the cooperative regime extend verbatim to the asynchronous mean-field dynamics. Part (iii) yields the asynchronous analogue of the local stability of strict pure LSVE, and hence extends Theorems 3 and 6. Part (iv), together with part (i), yields the asynchronous analogue of the instability result in the cycling example, and hence extends Theorem 4. Part (v) shows that the global stability of the unique mixed equilibrium result proved in the cooperative regime in Theorem 5 also extends to asynchronous CQL. Thus, the only result not covered by this theorem is the *global* convergence for two classes in Theorem 2, which relies on the one-dimensional structure of the synchronous dynamics.

Proposition 3. *Fix a finite reduced tree $\mathcal{T}' = (\mathcal{S}, \Omega, p, \pi)$ and a finite sensitivity parameter $\beta < \infty$. Let $F_\beta(v) := g(v;\beta) - v$, $\Xi(v;\beta) := \operatorname{diag}\big(\Xi_s(v;\beta)\big)_{s\in\mathcal{S}}$, where $\Xi_s(v;\beta) := \sum_{\omega\in\Omega_s} p(\omega)\,\sigma^s_\omega(v;\beta)$, $\Omega_s := \{\omega \in \Omega : s \in \omega\}$. Consider the synchronous mean-field ODE*

$$\dot{v} = F_\beta(v), \tag{6}$$

and the (calendar-time) asynchronous ODE

$$\dot{v} = \Xi(v;\beta)\, F_\beta(v). \tag{7}$$

Let $K = \prod_{s\in\mathcal{S}}[m_s, M_s]$ where $m_s = \min_{\omega\in\Omega_s} \pi_s(\omega)$ and $M_s = \max_{\omega\in\Omega_s} \pi_s(\omega)$ with $m_s < M_s$ by assumption. Then $K \subset \mathbb{R}^{\mathcal{S}}$ is compact, convex, and $g(K) \subseteq K$, hence every SVE lies in K. Assuming that each class is available with positive probability, the following claims hold.

(i) ***Identical equilibria:*** *The synchronous and asynchronous systems have the same*

equilibrium set. In particular, for every finite β, the set of SVE $\mathcal{V}(\beta)$ is the same under (6) and (7). Additionally, K is positively invariant for both (6) and (7).

(ii) ***Identical local asymptotic stability in the cooperative regime:*** *Let v^* be an SVE and write $J_\beta(v^*) := DF_\beta(v^*) = Dg(v^*;\beta) - I$, $\Xi^* := \Xi(v^*;\beta)$. Then the asynchronous Jacobian $D\big(\Xi(\cdot;\beta)F_\beta(\cdot)\big)(v^*) = \Xi^* J_\beta(v^*)$. If $J_\beta(v^*)$ is Metzler, then $J_\beta(v^*)$ is Hurwitz $\iff$ $\Xi^* J_\beta(v^*)$ is Hurwitz. Hence v^* is locally asymptotically stable for the synchronous ODE if and only if it is locally asymptotically stable for the asynchronous ODE.*

(iii) ***Strict pure LSVE remain locally asymptotically stable for large β:*** *Let v^∞ be a strict pure LSVE and let $v^{(\beta)}$ denote the unique nearby SVE branch with $v^{(\beta)} \to v^\infty$ as $\beta \to \infty$. Then there exists $\hat{\beta} < \infty$ such that, for every $\beta \geq \hat{\beta}$, the asynchronous Jacobian at $v^{(\beta)}$ is Hurwitz. In particular, the local asymptotic stability conclusions of Theorems 2, 3 and 6 continue to hold for the asynchronous mean-field dynamics.*

(iv) ***Loss of asymptotic stability at the unique RPS equilibrium:*** *In the RPS example of Theorem 4, let $(x, y) = (0, 0)$ denote the unique symmetric SVE in relative coordinates, and let $A(\beta, z)$ be the Jacobian of the synchronous relative dynamics at the origin. Then the Jacobian of the asynchronous relative dynamics at the origin is $A^{\text{async}}(\beta, z) = \frac{1}{3} A(\beta, z)$. Hence the origin has the same stability type in the synchronous and asynchronous systems: it is a stable focus for $\beta < \beta_c$, non-hyperbolic at $\beta = \beta_c$, and an unstable focus for $\beta > \beta_c$, where $\beta_c = -16/(3z)$. Since the SVE set is the same for the synchronous and asynchronous systems, the eventual uniqueness result in the RPS example also carries over. Therefore, for all sufficiently large β, the asynchronous reduced RPS dynamics admit no asymptotically stable SVE.*

(v) ***Global asymptotic stability of unique mixed SVE in the cooperative regime:*** *The unique SVE v^* is globally exponentially asymptotically stable on K for the asynchronous ODE. More precisely, there exists $\varepsilon := \min_{s \in \mathcal{S}} \min_{v \in K} \Xi_s(v;\beta) > 0$ such that every solution of (7) with $v(0) \in K$ satisfies $\|v(t) - v^*\|_\infty \leq e^{-\varepsilon t} \|v(0) - v^*\|_\infty$, $\forall\, t \geq 0$.*

Proof. Because $\beta < \infty$, the softmax probabilities are strictly positive on available classes. Hence, for each $s \in \mathcal{S}$ and every v, $\Xi_s(v;\beta) > 0$ whenever class s appears in some supported menu. Under the standing assumption $\sum_{\omega \in \Omega_s} p(\omega) > 0$, this holds for every class s.

(i) If $F_\beta(v^*) = 0$, then trivially $\Xi(v^*;\beta)F_\beta(v^*) = 0$. Conversely, if $\Xi(v^*;\beta)F_\beta(v^*) = 0$, then each diagonal entry of $\Xi(v^*;\beta)$ is strictly positive, so $\Xi(v^*;\beta)$ is invertible and therefore

$F_\beta(v^*) = 0$. Thus the two systems have the same equilibrium set.

Since $g_s(v;\beta)$ is a convex combination of the expected payoffs $\{\pi_s(\omega) : \omega \in \Omega_s\}$, we have $g_s(v;\beta) \in [m_s, M_s],\ \forall\, s \in \mathcal{S},\ \forall\, v \in K$. Hence, on the lower face $\{v_s = m_s\}$, $g_s(v;\beta) - v_s \geq 0$, while on the upper face $\{v_s = M_s\}$, $g_s(v;\beta) - v_s \leq 0$. Since $\Xi_s(v;\beta) \geq 0$ for every s and every v, the asynchronous vector field $\dot{v}_s = \Xi_s(v;\beta)\Big(g_s(v;\beta) - v_s\Big)$ has the same sign on each boundary face as in the synchronous system. Therefore the vector field points inward on every face of ∂K, and K is positively invariant for the asynchronous CQL dynamics.

(ii) At an equilibrium v^* we have $F_\beta(v^*) = 0$, so the product rule gives

$$D\Big(\Xi(\cdot;\beta)F_\beta(\cdot)\Big)(v^*) = D\Xi(v^*;\beta)\,[F_\beta(v^*)] + \Xi^* J_\beta(v^*) = \Xi^* J_\beta(v^*).$$

It remains to compare the spectra of $J_\beta(v^*)$ and $\Xi^* J_\beta(v^*)$. For Metzler matrices we use the standard criterion: M Hurwitz $\iff \exists\, x \gg 0$ such that $Mx \ll 0$. Suppose first that $J_\beta(v^*)$ is Hurwitz. Then there exists $x \gg 0$ such that $J_\beta(v^*)x \ll 0$. Since $\Xi^* \succ 0$ is diagonal, $\Xi^* J_\beta(v^*)x \ll 0$, so $\Xi^* J_\beta(v^*)$ is Hurwitz.

Conversely, if $\Xi^* J_\beta(v^*)$ is Hurwitz, there exists $x \gg 0$ such that $\Xi^* J_\beta(v^*)x \ll 0$. Multiplying component-wise by $(\Xi^*)^{-1} \succ 0$ preserves strict negativity, so $J_\beta(v^*)x \ll 0$. Hence $J_\beta(v^*)$ is Hurwitz. This proves the equivalence. By Lyapunov's indirect theorem, local asymptotic stability is the same for the synchronous and asynchronous ODEs.

(iii) By the proof of Theorem 3, there exist constants $\eta > 0$ and $C < \infty$ such that along the "strict pure" SVE branch $v^{(\beta)}$, $\|Dg(v^{(\beta)};\beta)\|_\infty \leq C\,\beta e^{-\beta\eta}$. Hence $J_\beta(v^{(\beta)}) = -I + E_\beta,\ E_\beta := Dg(v^{(\beta)};\beta)$, with $\|E_\beta\|_\infty \to 0$ exponentially fast. At the same equilibrium, the asynchronous Jacobian is $B_\beta := D\Big(\Xi(\cdot;\beta)F_\beta(\cdot)\Big)(v^{(\beta)}) = \Xi^*_\beta\, J_\beta(v^{(\beta)}) = \Xi^*_\beta(-I + E_\beta)$, where $\Xi^*_\beta := \Xi(v^{(\beta)};\beta)$ is diagonal with strictly positive entries.

For row s, the Gershgorin rightmost point of B_β is bounded by

$$(B_\beta)_{ss} + \sum_{r \neq s} |(B_\beta)_{sr}| \leq \Xi^*_{\beta,s}\Big(-1 + \sum_r |(E_\beta)_{sr}|\Big) \leq \Xi^*_{\beta,s}\Big(-1 + \|E_\beta\|_\infty\Big).$$

Since $\|E_\beta\|_\infty \to 0$, there exists $\hat{\beta} < \infty$ such that $\|E_\beta\|_\infty < 1$ for all $\beta \geq \hat{\beta}$. Then every Gershgorin disc of B_β lies strictly in the open left half-plane, so B_β is Hurwitz for all $\beta \geq \hat{\beta}$. Thus the nearby SVE is locally asymptotically stable for all sufficiently large β.

(iv) In the symmetric RPS tree, at the origin $(x, y) = (0, 0)$ all binary logit probabilities

equal 1/2. Each class appears in exactly two binary menus and one unary menu, each with probability 1/6. Therefore the class-selection propensities at the origin satisfy

$$\Xi_R(0) = \Xi_P(0) = \Xi_S(0) = \frac{1}{6}\Big(\frac{1}{2} + \frac{1}{2} + 1\Big) = \frac{1}{3}.$$

In reduced coordinates, the asynchronous vector field is

$$\dot{x} = \Xi_R(g_R - v_R) - \Xi_S(g_S - v_S), \qquad \dot{y} = \Xi_P(g_P - v_P) - \Xi_S(g_S - v_S).$$

At the equilibrium $(0, 0)$, the product-rule terms involving derivatives of Ξ vanish because $g_i - v_i = 0$ there, and since all three propensities equal 1/3, the Jacobian of the reduced asynchronous field is simply $A^{\text{async}}(\beta, z) = \frac{1}{3} A(\beta, z)$. Hence the asynchronous eigenvalues are exactly one third of the synchronous eigenvalues, so they have the same sign of real part and cross the imaginary axis at the same critical value β_c. This proves the stated stability classification. Finally, part (i) shows that synchronous and asynchronous CQL have the same SVE set for every finite β. Therefore any eventual uniqueness result established for the synchronous RPS dynamics holds verbatim for the asynchronous dynamics as well. In particular, for all sufficiently large β, the origin is the unique SVE and is linearly unstable; hence the asynchronous reduced RPS dynamics admit no asymptotically stable SVE.

(v) Under the hypotheses of Theorem 5, the synchronous CQL drift $F_\beta(v) = g(v; \beta) - v$ has a unique zero $v^* \in \operatorname{int} K$, the Jacobian $J_\beta(v) = DF_\beta(v)$ is Metzler and satisfies $J_\beta(v)\mathbf{1} = -\mathbf{1}\ \forall\, v \in K$, and K is positively invariant. As noted above, K is also positively invariant for the asynchronous dynamics $\dot{v} = \Xi(v; \beta)\, F_\beta(v)$. We will show v^* is globally exponentially asymptotically stable on K for the asynchronous CQL dynamics.

Since $\Xi(\cdot; \beta)$ is continuous and strictly positive component-wise on the compact set K, the quantity $\varepsilon := \min_{s \in \mathcal{S}} \min_{v \in K} \Xi_s(v; \beta)$ is well-defined and strictly positive. Let $z := v - v^*$. Because $F_\beta(v^*) = 0$ and K is convex, the fundamental theorem of calculus yields

$$F_\beta(v) - F_\beta(v^*) = \Big(\int_0^1 J_\beta\Big(v^* + \theta z\Big)\, d\theta\Big)\, z =: \overline{J}_\beta(v)\, z.$$

Thus the asynchronous dynamics can be written as $\dot{z} = \Xi(v; \beta)\, \overline{J}_\beta(v)\, z$. For every $v \in K$, the matrix $\overline{J}_\beta(v)$ is an average of Metzler matrices with row sum -1, hence it is itself Metzler and satisfies $\overline{J}_\beta(v)\mathbf{1} = -\mathbf{1}$. Define $M(v) := \Xi(v; \beta)\, \overline{J}_\beta(v)$. Because $\Xi(v; \beta)$ is diagonal and

positive and $\overline{J}_\beta(v)$ is Metzler, $M(v)$ is Metzler as well. Therefore its ℓ_∞ matrix measure is

$$\mu_\infty(M(v)) = \max_{s\in\mathcal{S}} \left(M_{ss}(v) + \sum_{r\neq s} |M_{sr}(v)| \right) = \max_{s\in\mathcal{S}} \Xi_s(v;\beta) \sum_{r\in\mathcal{S}} \overline{J}_{\beta,sr}(v).$$

Using the row-sum identity, $\mu_\infty(M(v)) = \max_{s\in\mathcal{S}} \Big(-\Xi_s(v;\beta)\Big) = -\min_{s\in\mathcal{S}} \Xi_s(v;\beta) \leq -\varepsilon$. Applying the standard logarithmic-norm to the linear time-varying system $\dot{z} = M(v(t))z$:

$$D_t^+ \|z(t)\|_\infty \leq \mu_\infty(M(v(t))) \, \|z(t)\|_\infty \leq -\varepsilon \, \|z(t)\|_\infty,$$

where D_t^+ denotes the upper Dini derivative with respect to time.[3] A scalar comparison argument yields $\|z(t)\|_\infty \leq e^{-\varepsilon t}\|z(0)\|_\infty, \ \forall\, t \geq 0$. Equivalently,

$$\|v(t) - v^*\|_\infty \leq e^{-\varepsilon t}\|v(0) - v^*\|_\infty, \qquad \forall\, t \geq 0.$$

Thus v^* is globally exponentially asymptotically stable on K for the asynchronous dynamics.

□

C RPS instability with full support on menus

We show that the instability result of Theorem 4 is robust to adding the ternary menu to the support. Consider the reduced RPS tree with three classes $\mathcal{S} = \{R, P, S\}$ and full menu support $\Omega = \Big\{\{R,P\},\{P,S\},\{R,S\},\{R\},\{P\},\{S\},\{R,P,S\}\Big\}$, $p(\omega) = \frac{1}{7} \ \forall\, \omega \in \Omega$. The binary-menu payoffs are $\pi_R(\{R,P\}) = -1$, $\pi_P(\{R,P\}) = 1$, $\pi_P(\{P,S\}) = -1$, $\pi_S(\{P,S\}) = 1$, $\pi_R(\{R,S\}) = 1$, $\pi_S(\{R,S\}) = -1$, the unary payoffs are $\pi_R(\{R\}) = \pi_P(\{P\}) = \pi_S(\{S\}) = z$, and the ternary-menu payoffs are symmetric: $\pi_R(\{R,P,S\}) = \pi_P(\{R,P,S\}) = \pi_S(\{R,P,S\}) = -z$. Throughout this subsection we assume $z \in (z_\star, 0)$, $z_\star := 12 - 4\sqrt{10}$.

[3] The logarithmic norm (matrix measure) associated with the ℓ_∞ norm is defined, for any $M \in \mathbb{R}^{n\times n}$, by

$$\mu_\infty(M) \;\coloneqq\; \lim_{h\to 0^+} \frac{\|I + hM\|_\infty - 1}{h} \;=\; \max_i \left(M_{ii} + \sum_{j\neq i} |M_{ij}| \right).$$

A fundamental property is: for the linear time-varying system $\dot{\mathbf{z}} = A(t)\mathbf{z}$, one has

$$D_t^+ \, \|\mathbf{z}(t)\|_\infty \;\leq\; \mu_\infty\big(A(t)\big) \, \|\mathbf{z}(t)\|_\infty \,,$$

where D_t^+ denotes the upper Dini time derivative. Consequently,

$$\|\mathbf{z}(t)\|_\infty \;\leq\; \exp\!\left(\int_0^t \mu_\infty\big(A(\tau)\big)\, d\tau \right) \|\mathbf{z}(0)\|_\infty \,.$$

Proposition 4. *For the RPS tree with full support described above, the following hold:*

(i) Every VE valuation profile equalizes valuations across classes. Equivalently, in relative coordinates $x := v_R - v_S,\ y := v_P - v_S$, the only VE valuation profile is $(x, y) = (0, 0)$. Consequently, $\mathcal{V}(\infty) = \{(0, 0)\}$.

(ii) For every $\beta \geq 0$, $(x, y) = (0, 0)$ is an SVE of the reduced CQL dynamics with Jacobian

$$A(\beta, z) = \begin{pmatrix} -1 - \dfrac{27\beta z}{98} - \dfrac{3\beta}{28} & \dfrac{3\beta}{14} \\ -\dfrac{3\beta}{14} & -1 - \dfrac{27\beta z}{98} + \dfrac{3\beta}{28} \end{pmatrix}, \tag{8}$$

so $(0, 0)$ is a stable focus for $\beta < \beta_c$, non-hyperbolic at $\beta_c := -\dfrac{98}{27z}$, and an unstable focus for $\beta > \beta_c$. Moreover, the Hopf bifurcation at $\beta = \beta_c$ is supercritical, so for $\beta > \beta_c$ sufficiently close to β_c, a locally attracting periodic orbit bifurcates from $(0, 0)$.

(iii) There exists $\bar{\beta} < \infty$ such that $\mathcal{V}(\beta) = \{(0, 0)\}\ \forall\, \beta \geq \bar{\beta}$. Hence, for all sufficiently large β, the unique SVE is linearly unstable and the reduced CQL dynamics admit no asymptotically stable SVE.

Proof. We work throughout in relative coordinates $x := v_R - v_S,\ y := v_P - v_S$, which is without loss by translation invariance of the drift (Lemma B.9).

Step 1: VE valuation profiles. In the binary menus, the best-reply sets are

$$\{R, P\}: \quad \arg\max\{v_R, v_P\} = \begin{cases} \{R\}, & x > y, \\ \{P\}, & y > x, \\ \{R, P\}, & x = y, \end{cases}$$

$$\{P, S\}: \quad \arg\max\{v_P, v_S\} = \begin{cases} \{P\}, & y > 0, \\ \{S\}, & y < 0, \\ \{P, S\}, & y = 0, \end{cases}$$

$$\{R, S\}: \quad \arg\max\{v_R, v_S\} = \begin{cases} \{R\}, & x > 0, \\ \{S\}, & x < 0, \\ \{R, S\}, & x = 0. \end{cases}$$

In the ternary menu $\{R, P, S\}$, the best-reply set is the set of maximizers among (v_R, v_P, v_S).

We first rule out the six strict order cones. By cyclic symmetry it suffices to treat two. If $R \succ P \succ S$ (equivalently $x > y > 0$), then the best replies in the four nontrivial menus are

$$(\{R, P\}, \{P, S\}, \{R, S\}, \{R, P, S\}) = (R, P, R, R).$$

Hence

$$g_R = \frac{-1 + 1 + z - z}{4} = 0, \qquad g_P = \frac{-1 + z}{2}, \qquad g_S = z,$$

so

$$g_R - g_S = -z > 0, \qquad g_P - g_S = -\frac{1 + z}{2} < 0,$$

contradicting $y > 0$.

If $P \succ R \succ S$ (equivalently $y > x > 0$), then the best replies are

$$(\{R, P\}, \{P, S\}, \{R, S\}, \{R, P, S\}) = (P, P, R, P),$$

hence

$$g_R = \frac{1 + z}{2}, \qquad g_P = \frac{1 - 1 + z - z}{4} = 0, \qquad g_S = z.$$

Therefore

$$g_R - g_S = \frac{1 - z}{2} > 0, \qquad g_P - g_S = -z > 0,$$

but

$$(g_R - g_S) - (g_P - g_S) = \frac{1 + z}{2} > 0,$$

contradicting $y > x$. By cyclic relabeling, all strict cones are excluded. It remains to rule out tie faces away from the origin.

By symmetry it suffices to consider $x = 0$, $y \neq 0$. If $x = 0$ and $y > 0$, then $\{R, P\}$, $\{P, S\}$, and $\{R, P, S\}$ have unique best reply P, while $\{R, S\}$ is a tie between R and S. Let $\alpha \in [0, 1]$ be the probability of choosing R in $\{R, S\}$. Then

$$g_R = \frac{\alpha + z}{1 + \alpha}, \qquad g_S = \frac{z - 1 + \alpha}{2 - \alpha},$$

so

$$g_R - g_S = \frac{1 + z + 2\alpha(1 - z - \alpha)}{(1 + \alpha)(2 - \alpha)} > 0$$

for all $z \in (-1, 0)$ and $\alpha \in [0, 1]$, contradicting $x = 0$.

Now let $x = 0$ and $y < 0$. Then $\{R, P\}$ has unique best reply R, $\{P, S\}$ has unique best reply S, while both $\{R, S\}$ and $\{R, P, S\}$ are ties between R and S. Let $\alpha \in [0, 1]$ be the probability of choosing R in $\{R, S\}$, and let $\lambda \in [0, 1]$ be the probability of choosing R in $\{R, P, S\}$. Then

$$g_R = \frac{-1 + \alpha + z - \lambda z}{2 + \alpha + \lambda}, \qquad g_S = \frac{\alpha + \lambda z}{3 - \alpha - \lambda}.$$

The condition $x = 0$ requires $g_R = g_S$, i.e.

$$z = \zeta(\alpha, \lambda) := \frac{-2\alpha^2 - 2\alpha\lambda + 2\alpha + \lambda - 3}{\alpha + 6\lambda - 3}. \tag{9}$$

A direct computation gives

$$\partial_\lambda \zeta(\alpha, \lambda) = \frac{5(2\alpha^2 - \alpha + 3)}{(\alpha + 6\lambda - 3)^2} > 0,$$

so ζ is increasing in λ. Thus its maximum on $[0, 1]^2$ is attained at $\lambda = 1$, where

$$\zeta(\alpha, 1) = -\frac{2(\alpha^2 + 1)}{\alpha + 3}.$$

Differentiating in α,

$$\frac{d}{d\alpha}\zeta(\alpha, 1) = -\frac{2(\alpha^2 + 6\alpha - 1)}{(\alpha + 3)^2},$$

so the maximum on $[0, 1]$ occurs at $\alpha_\star = \sqrt{10} - 3$, with value

$$\max_{(\alpha,\lambda)\in[0,1]^2} \zeta(\alpha, \lambda) = \zeta(\alpha_\star, 1) = 12 - 4\sqrt{10} = z_\star.$$

Hence, if $z > z_\star$, equation (9) has no solution, so no VE lies on $\{x = 0,\ y < 0\}$. By cyclic relabeling, the same exclusion holds on $\{y = 0\} \setminus \{(0, 0)\}$ and $\{x = y\} \setminus \{(0, 0)\}$.

Therefore the only VE valuation profile is $(x, y) = (0, 0)$. Since every LSVE belongs to $\mathcal{VE}$ by Theorem 1, it follows that $\mathcal{V}(\infty) = \{(0, 0)\}$.

Step 2: Symmetric SVE and Hopf bifurcation. For finite β, the reduced CQL dynamics are

$$\dot{x} = U(x, y) := (g_R - g_S)(x, y; \beta) - x, \qquad \dot{y} = V(x, y) := (g_P - g_S)(x, y; \beta) - y.$$

The binary logit probabilities are

$$\sigma_R^{(RP)}(x, y) = \frac{1}{1 + e^{-\beta(x-y)}}, \qquad \sigma_P^{(PS)}(y) = \frac{1}{1 + e^{-\beta y}}, \qquad \sigma_R^{(RS)}(x) = \frac{1}{1 + e^{-\beta x}},$$

and the ternary logits are

$$\sigma_R^{(RPS)}(x,y) = \frac{e^{\beta x}}{e^{\beta x}+e^{\beta y}+1}, \quad \sigma_P^{(RPS)}(x,y) = \frac{e^{\beta y}}{e^{\beta x}+e^{\beta y}+1}, \quad \sigma_S^{(RPS)}(x,y) = \frac{1}{e^{\beta x}+e^{\beta y}+1}.$$

The class-conditional payoff signals are therefore

$$g_R = \frac{-\sigma_R^{(RP)} + \sigma_R^{(RS)} + z - z\,\sigma_R^{(RPS)}}{\sigma_R^{(RP)} + \sigma_R^{(RS)} + 1 + \sigma_R^{(RPS)}},$$

$$g_P = \frac{(1-\sigma_R^{(RP)}) - \sigma_P^{(PS)} + z - z\,\sigma_P^{(RPS)}}{(1-\sigma_R^{(RP)}) + \sigma_P^{(PS)} + 1 + \sigma_P^{(RPS)}},$$

$$g_S = \frac{(1-\sigma_P^{(PS)}) - (1-\sigma_R^{(RS)}) + z - z\,\sigma_S^{(RPS)}}{(1-\sigma_P^{(PS)}) + (1-\sigma_R^{(RS)}) + 1 + \sigma_S^{(RPS)}}.$$

At $(x,y) = (0,0)$, the binary logits equal $1/2$ and the ternary logits equal $1/3$, so

$$g_R(0,0;\beta) = g_P(0,0;\beta) = g_S(0,0;\beta) = \frac{2z}{7},$$

hence $(0,0)$ is an SVE for every finite β.

A direct linearization yields the Jacobian (8). Its trace and discriminant are

$$\operatorname{Tr} A(\beta,z) = -2 - \frac{27z}{49}\beta, \qquad \operatorname{Tr}(A)^2 - 4\det(A) = -\frac{27}{196}\beta^2 < 0.$$

Thus the eigenvalues form a complex conjugate pair with real part

$$\Re\lambda = \frac{1}{2}\operatorname{Tr} A(\beta,z) = -1 - \frac{27z}{98}\beta,$$

so the origin is a stable focus for $\beta < \beta_c$, non-hyperbolic at

$$\beta_c := -\frac{98}{27z},$$

and an unstable focus for $\beta > \beta_c$. Moreover,

$$\frac{d}{d\beta}\Re\lambda\Big|_{\beta_c} = -\frac{27z}{98} > 0,$$

so the Hopf transversality condition holds. A direct symbolic computation of the first Lya-

punov coefficient at $\beta = \beta_c$ gives

$$\ell_1(\beta_c) = \frac{2681\sqrt{3}}{4374} \cdot \frac{1}{z} < 0,$$

hence the Hopf bifurcation is supercritical. Therefore, for $\beta > \beta_c$ sufficiently close to β_c, a locally attracting periodic orbit bifurcates from the origin.

Step 3: eventual uniqueness of SVE for large β. Fix $z \in (z_\star, 0)$. Let $(x_n, y_n) \in \mathcal{V}(\beta_n)$ with $\beta_n \uparrow \infty$. Since every accumulation point of (x_n, y_n) is an LSVE and $\mathcal{V}(\infty) = \{(0,0)\}$, we have $(x_n, y_n) \to (0,0)$. Define the scaled variables $X_n := \beta_n x_n,\ Y_n := \beta_n y_n$. We claim that (X_n, Y_n) remains bounded. Suppose instead that $\max\{|X_n|, |Y_n|\} \to \infty$. Passing to a subsequence, each of X_n, Y_n, and $X_n - Y_n$ converges in $[-\infty, \infty]$, with at least one infinite limit. The corresponding binary and ternary logits therefore converge to a greedy tie-breaking profile associated either with a strict order cone or with a tie face away from the origin. Since

$$x_n = (g_R - g_S)(x_n, y_n; \beta_n), \qquad y_n = (g_P - g_S)(x_n, y_n; \beta_n),$$

and $x_n, y_n \to 0$, the limiting greedy profile would yield VE payoff differences equal to zero. By Step 1, this is impossible away from the origin. Hence (X_n, Y_n) is bounded.

It follows that there exist $\rho > 0$ and $\beta_0 < \infty$ such that

$$\mathcal{V}(\beta) \subset B_{\rho/\beta}(0) \qquad \forall\, \beta \geq \beta_0. \tag{10}$$

Now define the reduced fixed-point map

$$h_\beta(x, y) := \begin{pmatrix} x - (g_R - g_S)(x, y; \beta) \\ y - (g_P - g_S)(x, y; \beta) \end{pmatrix} = -F_\beta(x, y).$$

Since each $g_i(\cdot; \beta)$ is a convex combination of payoffs in $\{-1, 1, z, -z\} \subset [-1, 1]$, we have

$$\|(g_R - g_S,\ g_P - g_S)\|_2 \leq 2\sqrt{2}.$$

Choose $R > 2\sqrt{2}$. For $u \in \partial B_R(0)$, define

$$H_t(u) := u - t\,(g_R - g_S,\ g_P - g_S)(u; \beta), \qquad t \in [0, 1].$$

Then $\|H_t(u)\|_2 > 0$ on $\partial B_R(0)$, so homotopy invariance gives

$$\deg(h_\beta, B_R(0), 0) = \deg(\mathrm{id}, B_R(0), 0) = 1 \qquad \forall\, \beta \geq 0.$$

Next, from (8),

$$Dh_\beta(0,0) = -A(\beta, z),$$

and

$$\det Dh_\beta(0,0) = \det A(\beta, z) = 1 + \frac{27z}{49}\beta + \Big(\frac{729z^2}{9604} + \frac{27}{784}\Big)\beta^2.$$

Since

$$\frac{729z^2}{9604} + \frac{27}{784} \geq \frac{27}{784} > 0,$$

there exist $\kappa > 0$ and $\beta_1 < \infty$ such that

$$\det Dh_\beta(0,0) \geq \kappa\beta^2 \qquad \forall\, \beta \geq \beta_1.$$

Moreover, each component of h_β is a rational combination of binary and ternary logits, and all denominators are uniformly bounded away from zero, so there exists $C_1 < \infty$ such that

$$\sup_{(x,y)\in\mathbb{R}^2} \|D^2 h_\beta(x,y)\| \leq C_1\beta^2 \qquad \forall\, \beta \geq 1.$$

Hence, for $u \in B_{\rho/\beta}(0)$,

$$\|Dh_\beta(u) - Dh_\beta(0)\| \leq C_1\beta^2\|u\| \leq C_1\rho\,\beta.$$

Using local Lipschitz continuity of the determinant on bounded subsets of 2×2 matrices, we obtain

$$|\det Dh_\beta(u) - \det Dh_\beta(0)| \leq C_2\rho\,\beta^2 \qquad \forall\, u \in B_{\rho/\beta}(0)$$

for some $C_2 < \infty$. Shrinking ρ if necessary so that $C_2\rho < \kappa/2$, we get

$$\det Dh_\beta(u) \geq \frac{\kappa}{2}\beta^2 > 0 \qquad \forall\, u \in B_{\rho/\beta}(0),\ \forall\, \beta \geq \beta_1.$$

Thus every zero of h_β in $B_{\rho/\beta}(0)$ is regular and has local index $+1$.

Fix $\beta \geq \bar{\beta} := \max\{\beta_0, \beta_1\}$. By (10), every zero of h_β lies in $B_{\rho/\beta}(0)$, so excision yields

$$\deg(h_\beta, B_{\rho/\beta}(0), 0) = \deg(h_\beta, B_R(0), 0) = 1.$$

Since every zero in $B_{\rho/\beta}(0)$ has local index $+1$, there can be exactly one such zero. Because $(0,0)$ is always an SVE, it follows that

$$\mathcal{V}(\beta) = \{(0,0)\} \qquad \forall\, \beta \geq \bar{\beta}.$$

This proves part (iii), and completes the proof. □

Proposition 4 shows that the instability mechanism of Theorem 4 survives after adding the ternary menu to the support. In particular, for $z \in (z_\star, 0)$ the symmetric SVE at the origin remains the unique high-sensitivity limit of SVE, undergoes a supercritical Hopf bifurcation at β_c, and is eventually the unique SVE for all sufficiently large β. Hence, with full menu support, the reduced RPS dynamics still admit a locally attracting periodic orbit for $\beta > \beta_c$ close to β_c, and admit no asymptotically stable SVE for all sufficiently large β. By the Poincaré–Bendixson theorem, any ω-limit set contained in M that is not an equilibrium must be a periodic orbit. Hence all non-trivial trajectories are repelled away from $(0,0)$ when $\beta > \beta_c$, and every bounded trajectory has a periodic orbit as its ω-limit set.

D Details of Theorem 1(d6)

Lemma D.1 (Generic Morse projection on the SVE graph)**.** *Let*

$$\widetilde{F}(\mathbf{v}, \beta, \theta) := \mathbf{v} - g(\mathbf{v}; \beta, \theta), \qquad \theta = (p, \pi) \in \Theta := \Delta(\Omega^+) \times \Pi^+,$$

and let $\mathcal{M} := \widetilde{F}^{-1}(0) \subset U \times (0,\infty) \times \Theta$ *for some open* $U \subset \mathbb{R}^{\mathcal{S}}$*. Write*

$$P : \mathcal{M} \to \Theta, \qquad P(\mathbf{v}, \beta, \theta) = \theta, \qquad \Xi : \mathcal{M} \to \mathbb{R}, \qquad \Xi(\mathbf{v}, \beta, \theta) = \beta.$$

Assume $\theta \in \Theta_{\mathrm{reg}}$*, so that the fiber* $\mathcal{G}_\theta := P^{-1}(\theta) = \{(\mathbf{v}, \beta) \in K \times (0,\infty) : \widetilde{F}(\mathbf{v}, \beta, \theta) = 0\}$ *is a one-dimensional embedded real-analytic submanifold. Then there exists a residual subset* $\Theta_{\mathrm{Morse}} \subset \Theta_{\mathrm{reg}}$ *such that, for every* $\theta \in \Theta_{\mathrm{Morse}}$*, the projection* $\xi_\theta : \mathcal{G}_\theta \to (0,\infty)$*,* $(\mathbf{v}, \beta) \mapsto \beta$*, is a Morse function. Consequently, every critical point of* ξ_θ *is a nondegenerate fold, the critical set is finite, and every connected component of* $\mathcal{G}_\theta$ *is an embedded real-analytic curve.*

Proof. Let $\mathcal{M}_{\mathrm{reg}} := P^{-1}(\Theta_{\mathrm{reg}}) \subset \mathcal{M}$. By part (c) of Theorem 1, for every $\theta \in \Theta_{\mathrm{reg}}$, the fiber $\mathcal{G}_\theta = P^{-1}(\theta)$ is a one-dimensional embedded real-analytic submanifold. Hence the restricted map $P|_{\mathcal{M}_{\mathrm{reg}}} : \mathcal{M}_{\mathrm{reg}} \to \Theta_{\mathrm{reg}}$ is a submersion. Let $V := \ker d(P|_{\mathcal{M}_{\mathrm{reg}}}) \subset T\mathcal{M}_{\mathrm{reg}}$ be its vertical tangent bundle. Since the fibers are one-dimensional, V has rank one. The vertical

differential of Ξ defines a smooth section $d^V\Xi \in \Gamma(V^*)$. For $\theta \in \Theta_{\text{reg}}$, the restriction $\xi_\theta = \Xi|_{\mathcal{G}_\theta}$ is Morse iff the section $d^V\Xi|_{\mathcal{G}_\theta}$ is transverse to the zero section of $V^*|_{\mathcal{G}_\theta}$. By a standard parametric transversality theorem (see Guillemin and Pollack (2010), p.68) applied to the restricted submersion $P|_{\mathcal{M}_{\text{reg}}} : \mathcal{M}_{\text{reg}} \to \Theta_{\text{reg}}$, there exists a residual subset $\Theta_{\text{Morse}} \subset \Theta_{\text{reg}}$ such that, for every $\theta \in \Theta_{\text{Morse}}$, the restricted section $d^V\Xi|_{\mathcal{G}_\theta}$ is transverse to the zero section. Hence ξ_θ is Morse. Since $\mathcal{G}_\theta$ is definable and one-dimensional, its critical set is a definable 0-dimensional subset and therefore finite. Thus every connected component of $\mathcal{G}_\theta$ is an embedded real-analytic curve, and over any bounded interval in β it is a finite union of real-analytic graphs separated by finitely many fold points. □

E Additional Illustrations

E.1 Example: The Good, the Bad, and the Unsteady

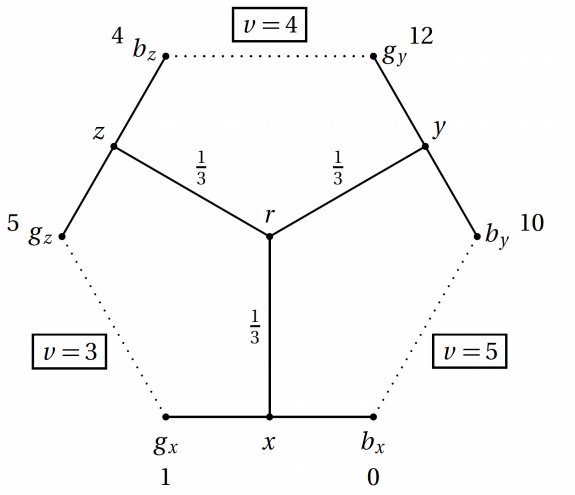


Figure 1: Good/Bad decisions

The decision tree is depicted by the solid lines in Figure 1. At the root r, Nature chooses one of the three nodes x, y, and z with equal probability. At each node, the decision-maker chooses between a good alternative and a bad alternative, with the former yielding the higher payoff. The three dotted lines indicate the similarity partition: $i = \{g_x, g_z\}$, $j = \{b_x, b_y\}$, $k = \{g_y, b_z\}$. This example is useful because it exhibits a counter-intuitive phenomenon: the strategy that selects the bad alternative at every node is a valuation equilibrium, whereas the strategy that selects the good alternative at every node is not.

Consider the pure policy that chooses b_x at x, b_y at y, and b_z at z. Under this policy, classes j and k are chosen on path, whereas class i is never chosen. Accordingly, valuation

consistency pins down the on-path coordinates at $v_j = 5,\ v_k = 4$, but it imposes no further restriction on v_i beyond the inequalities needed to keep i suboptimal in the menus in which it appears. Thus the bad pure policy is supported by a continuum of pure VE of the form $(v_i, 5, 4)$ with $v_i < 4$. In particular, $(3, 5, 4)$ is a representative strict VE.

This example also illustrates the refinement logic behind limiting SVE. In a pure VE, the valuation of a never-winner class is an off-path object and is therefore left unrestricted. By contrast, under logit choice with large but finite β, a never-winner class is still chosen through exponentially rare trembles. Conditional on such a tremble, the menu distribution is not arbitrary: it concentrates on the *closest-loss* menus, namely those in which the class loses by the smallest valuation gap. In the present example, for the representative bad VE $(3, 5, 4)$, class i loses to j at node x by gap 2, and loses to k at node z by gap 1. Hence, conditional on class i being chosen under rare logit trembles, the probability mass concentrates on node z, its closest-loss menu. Therefore the high-β consistency condition for i is governed by the expected payoff of i at node z which is 5, rather than by an arbitrary off-path value.

A direct computation confirms this. Evaluating the vector field at $(3, 5, 4)$ and then letting $\beta \uparrow \infty$, one obtains $\lim_{\beta\uparrow\infty} F_\beta(3, 5, 4) = (2, 0, 0)$. Equivalently, $\lim_{\beta\uparrow\infty} g_\beta(3, 5, 4) = (5, 5, 4)$. Thus the coordinates $v_j = 5$ and $v_k = 4$ are already stationary, but the off-path coordinate $v_i = 3$ is not: in the high-sensitivity limit, the dynamics push it upward toward the payoff generated in its closest-loss menu. Hence no bad pure VE of the form $(v_i, 5, 4)$ with $v_i < 4$ can be an LSVE. We conclude that the implication in Theorem 1(ii) is strict in general: every accumulation point of SVE is a VE, but not every VE is the limit of SVE as $\beta \uparrow \infty$. In this example, VE leaves the off-path valuation of class i free, whereas the high-sensitivity LSVE selects only those valuations that are consistent with closest-loss concentration.

Numerical simulations are consistent with this logic. Starting from the bad-VE representative $(3, 5, 4)$ and using a large sensitivity parameter, the trajectory quickly leaves that point: the i-coordinate moves upward, while the j- and k-coordinates remain close to 5 and 4, respectively. Repeating the simulation from many initial conditions with large β, we observe convergence to two distinct rest points. One lies near $(v_i, v_j, v_k) \approx (5, 5, 4)$, which corresponds to a partially mixed VE (near $(1, 1)$ in the relative coordinates), and the other at $(v_i, v_j, v_k) = (1, 0, 8)$, which corresponds to a strict pure VE (near $(-7, -8)$ in relative coordinates).

E.2 Example: Triangular cycling in RPS without unary menus

Consider the same RPS reduced tree as in Example 3.3, but now suppose that the unary menus are removed from the support and the menu distribution is uniform over the three

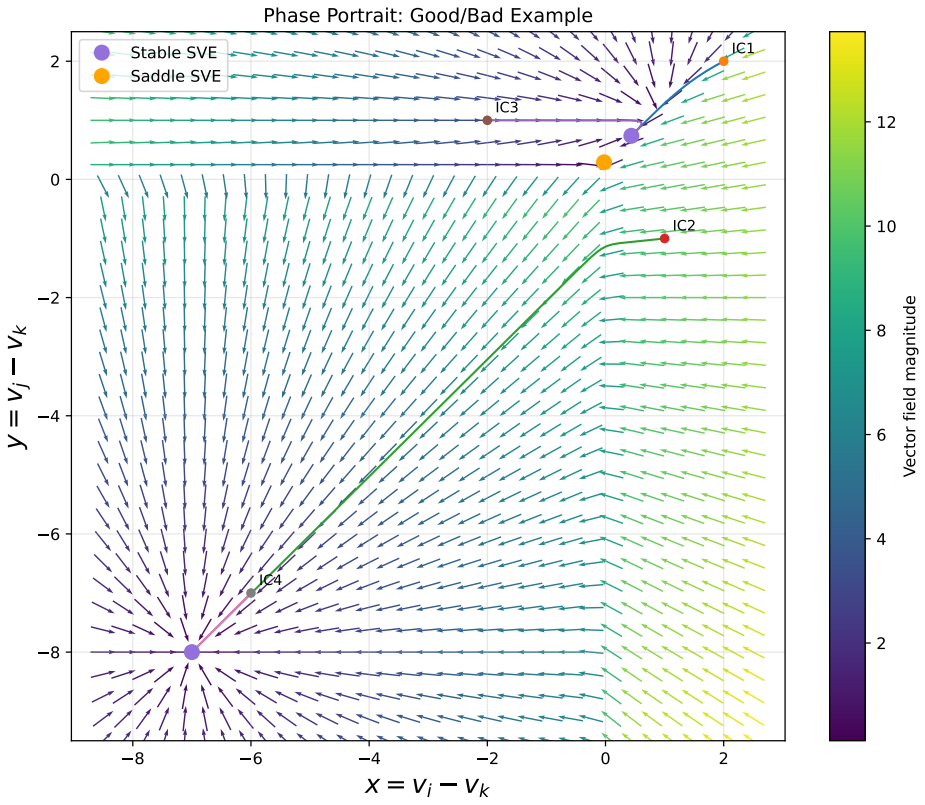


Figure 2: CQL dynamics with $\beta = 50$

binary menus and the ternary menu: $p(\omega_1) = p(\omega_2) = p(\omega_3) = p(\omega_7) = \frac{1}{4}$, $p(\omega_4) = p(\omega_5) = p(\omega_6) = 0$. The class-level expected payoffs remain unchanged. By symmetry, for every finite β there is a unique symmetric SVE, which in relative coordinates $(x, y) = (v_R - v_S, v_P - v_S)$ is located at the origin. For sufficiently large β, this equilibrium is linearly unstable, and the phase portrait displays a stable periodic orbit surrounding it.

Relative to the version with unary menus in support, however, the geometry of the cycle is qualitatively different: the orbit is triangular rather than hexagonal, with the three corners corresponding to the three cyclic valuation rankings. The attached phase portrait is consistent with exactly this picture: trajectories spiral away from the unstable symmetric SVE at the origin and approach a stable triangular limit cycle. The key difference is that, once unary menus are removed, a currently dominated class is no longer updated through any deterministic on-path menu. Thus off-path learning is driven entirely by logit trembles. This makes the high-β refinement logic particularly stark.

Fix, for instance, a strict ranking $v_R > v_P > v_S$. Under the near-greedy logit rule, class R is chosen in $\omega_1 = \{R, P\}$, $\omega_3 = \{S, R\}$, and $\omega_7 = \{R, P, S\}$, while class P is chosen in $\omega_2 = \{P, S\}$, and class S is never chosen except through rare trembles. Accordingly, ignoring

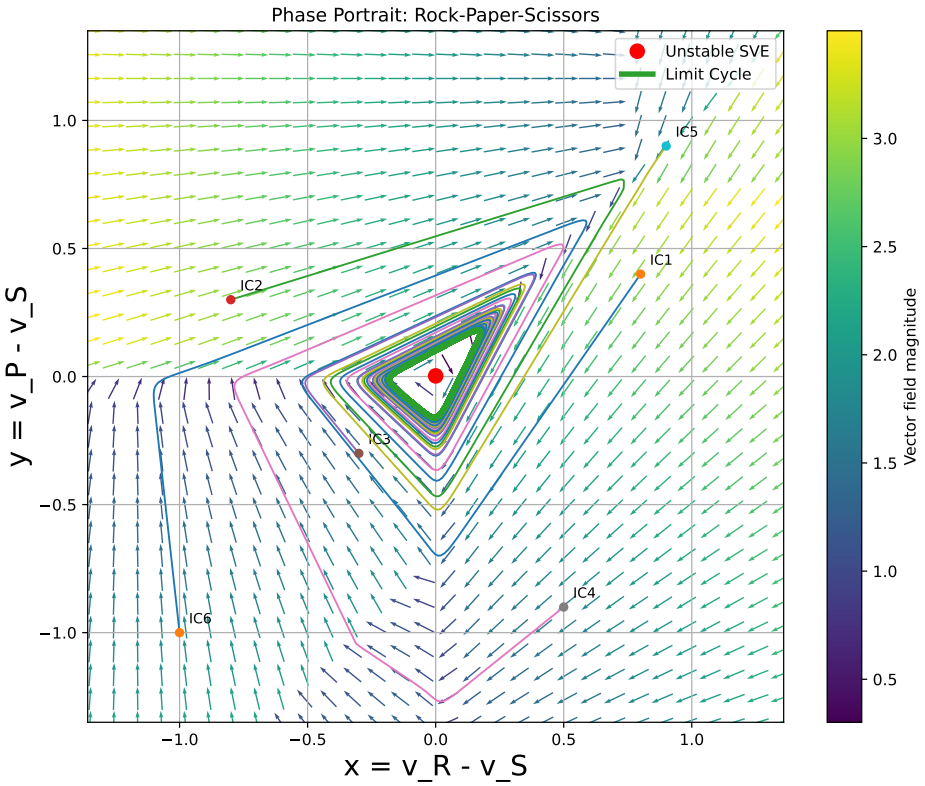


Figure 3: CQL dynamics with $\beta = 50$

exponentially small terms,

$$g_R \approx \frac{-1+1+0.4}{3} > 0, \qquad g_P \approx -1.$$

By contrast, the valuation of the off-path class S is pinned down by the menus in which it is *closest* to being chosen. Since S is below P by a smaller gap than it is below R, the closest-loss menu for S is $\omega_2 = \{P, S\}$, not ω_3 or ω_7. Conditional on a rare tremble selecting S, the menu distribution therefore concentrates on ω_2, so $g_S \approx 1$. Hence, in the region $v_R > v_P > v_S$, the current bottom class is pushed sharply upward, the current middle class is pushed downward, and the current top class drifts only mildly. This rotates the ranking directly to $v_S > v_R > v_P$. Repeating the same argument yields the cyclic succession

$$R > P > S \;\longrightarrow\; S > R > P \;\longrightarrow\; P > S > R \;\longrightarrow\; R > P > S.$$

Thus the dynamics visit only the three *cyclic* orderings, which explains why the periodic orbit has three corners and takes a triangular shape.

This should be contrasted with the version in which unary menus remain in support. There,

even a never-chosen class continues to be updated deterministically through its singleton menu. As a result, the lowest-ranked class does not jump directly from the bottom of the ranking to the top through rare trembles. Instead, the ranking changes by adjacent transpositions, and the trajectory visits all six strict orderings in succession. That is what produces the hexagonal cycle in the earlier example. Here, by contrast, the absence of unary menus makes off-path trembling (exploration) the only force that moves the dominated class, and the closest-loss refinement then pushes that class directly toward the payoff of the unique binary menu in which it would defeat the current middle class. This is the mechanism behind the triangular limit cycle. In this sense, the present example highlights even more sharply the role of smooth choice in CQL. At the exact greedy limit, the currently dominated class would receive no updates at all away from indifference, so the off-path coordinates would be left indeterminate. For large but finite β, logit noise regularizes the dynamics: rare trembles keep every class active, and the induced closest-loss concentration selects a definite off-path drift. The resulting regularization is precisely what sustains the persistent triangular cycling around the eventually unique unstable mixed SVE.

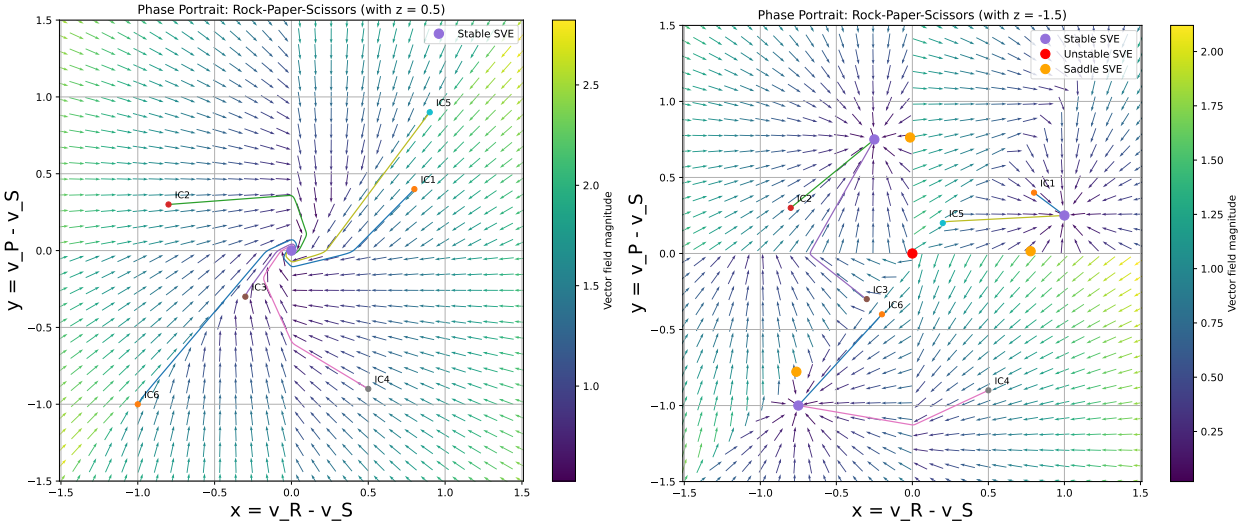
E.3 Example: Revisiting the RPS tree through singleton payoff shifts $z \in \mathbb{R}$

We revisit the symmetric RPS family from Theorem 4 to illustrate how varying singleton-menu payoffs organizes the transition between stability, cycling, and multiplicity. As in the proof of Theorem 4, set $z_R = z_P = z_S = z$, let $z \in \mathbb{R}$, and take $p(\omega) = 1/6$ for all $\omega \in \Omega \backslash \{\omega_7\}$, with $p(\omega_7) = 0$. For $z \in (-1, 0)$, the fully mixed SVE at the origin is unstable for large β, and trajectories converge to the hexagonal limit cycle discussed in the main text.

The phase portraits in Figure 4 show the two adjacent regimes. When singleton payoffs are high, the fully mixed SVE becomes stable and appears as the global attractor. When singleton payoffs are sufficiently low, strict valuation rankings become self-confirming: several SVE emerge near strict pure VE, each locally stable, while additional mixed SVE act as saddles. Thus the same RPS tree displays all three qualitative regimes emphasized in the paper: stable indifference, persistent cycling, and multiplicity of stable strict equilibria.

The stabilizing force in the high-z case is the same one isolated in Theorem 5: an overvalued class is sampled relatively more often in competitive menus, where its payoff is low relative to the singleton benchmark, pulling its valuation downward. Conversely, when z is low, singleton observations reinforce low-ranked classes rather than correcting them. This makes strict rankings locally self-confirming and generates the multiplicity predicted by Theorem

6. The intermediate region $z \in (-1, 0)$ is where neither force dominates meaning both *cooperative* and *competitive* forces co-exist among classes; in the RPS geometry, the resulting local instability of the unique mixed SVE produces the cyclic dynamics of Theorem 4.



(a) Unique stable fully-mixed SVE ($z = 0.5$) (b) Multiple locally stable strict SVE ($z = -1.5$)

Figure 4: Phase portraits for the symmetric RPS tree with $\beta = 10$.

References

Fudenberg, D. and Levine, D. (1998). *The Theory of Learning in Games*. MIT Press.

Guillemin, V. and Pollack, A. (2010). *Differential Topology*. AMS Chelsea Publishing. AMS Chelsea Pub.

Kushner, H. and Yin, G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*. Stochastic Modelling and Applied Probability. Springer New York.

Leslie, D. S. and Collins, E. J. (2005). Individual q-learning in normal form games. *SIAM J. Control. Optim.*, 44:495–514.

Parker, A., Wilding, E., and Akerman, C. (1998). The von restorff effect in visual object recognition memory in humans and monkeys. *Journal of Cognitive Neuroscience*, 10(6):691–703.

Rothenhoefer, K., Hong, T., Alikaya, A., and Stauffer, W. (2021). Rare rewards amplify dopamine responses. *Nature Neuroscience*, 24(4):465–469.