NOISY REPLICATOR–MUTATOR DYNAMICS IN MULTI-AGENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Evolutionary game theory predicts where a population of learning agents ends up, but the classical replicator equation is derived for a payoff matrix that does not change. Two recent lines of work relax it in different ways. Wang et al. (2026) show that global payoff noise adds a multiplicative term $(\sigma_C - \sigma_D)^2(\frac{1}{2} - x)$, with σ the intensity of the payoff fluctuation, to the replicator equation, stabilising coexistence and generating stable limit cycles; Bauer et al. (2026) show that an additive mutation term $M(c - x)$ turns it into a replicator–mutator system whose interior equilibria are last-iterate attracting, which yields convergence guarantees for multi-agent reinforcement learning. The two perturbations have been analysed on their own; their interaction has not been studied. We derive the system that carries both, which we call Noisy Replicator–Mutator Dynamics (NRMD), from a stochastic policy-gradient update; the two terms act on different quantities, so they can be set independently. Our contributions are threefold: (i) we give a complete two-strategy theory of NRMD, consistent with its n-strategy form at $n = 2$; (ii) we validate the effect of noise in multi-agent experiments; and (iii) we propose a novel algorithm, Population-Replicator Policy Gradient (PRPG), that exploits NRMD. Code and data will be released.

1 INTRODUCTION

A population of learning agents that repeatedly play a game will, under many update rules, end up where evolutionary game theory predicts. The classical tool is the replicator equation $\dot{x}_i = x_i(\pi_i - \bar{\pi})$, which arises as the deterministic limit of imitation, reinforcement and a family of policy-gradient rules (Taylor & Jonker, 1978; Börgers & Sarin, 1997; Tuyls & Nowé, 2006; Bloembergen et al., 2015). Its prediction is however conditional on a payoff matrix that never changes, and real multi-agent systems violate that condition: rewards drift, and agents rarely behave deterministically.

Two recent papers relax the fixed-matrix assumption from different directions. Wang et al. (2026) let every payoff be perturbed by a global random noise shock drawn once per generation and shared by the whole population. It leaves the mean game unchanged, yet it enters the replicator equation through a *multiplicative* term that destabilises boundary states and can create stable limit cycles in rock–paper–scissors. Bauer et al. (2026) instead add a *mutation bias*, an additive term that pulls the strategy distribution towards a reference point; it leaves payoffs alone and acts geometrically, making interior equilibria last-iterate attracting and thereby supplying convergence guarantees for MARL in stable games.

Both perturbations have been characterised in isolation. Putting them in one system matters: the two terms differ in kind, one multiplicative and payoff-driven, the other additive and geometry-driven, so it is not obvious whether their effects add, cancel or interact, and the answer determines whether either analysis transfers to the regime of real multi-agent learning.

Contributions. Our contributions are threefold. *First*, we propose *noisy replicator–mutator dynamics* (NRMD), the ODE system that carries both perturbations, and analyse it as a single object; the noise term is due to Wang et al. (2026) and the mutation term to Bauer et al. (2026). What we add is a complete *two-strategy* theory in which the consequences are proved rather than assumed: the separation of shared from independent noise (Proposition 2), the identification of an entropy bonus with a mutation bias (Proposition 3), the exact noise threshold (Proposition 4), the muta-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tion condition (Proposition 5), the observability window (Corollary 1) and the convergence and rate statement (Proposition 6). The n-strategy form reduces to ours at $n = 2$ (Appendix G); extending the remaining statements to general n is open, and §3 marks where.

Second, we validate these predictions at three levels of abstraction (§4): a tabular benchmark that reproduces the growing mutation budget, with an ablation showing that the discrete offspring jump and the mutation term are both needed while the update order is not; a deep learner whose update *is* the NRMD update, which inherits the effect while a gradient learner of the same architecture and matched policy sets does not ($+0.410$ against $+0.020$), and a standard PPO control (Schulman et al., 2017) that shows nothing; and a population of LLM agents that recovers the mutation dimension quantitatively and is null in the noise dimension.

Third, building on the mutation–selection form derived above, we propose a new MARL algorithm, *Population-Replicator Policy Gradient* (PRPG, Algorithm 1): a population of learners whose update reproduces mutation–selection dynamics rather than following a reward gradient. Mainstream algorithms such as PPO (Schulman et al., 2017) are not of this form—their update is not affine in the sampled payoff—so neither Proposition 1 nor the mechanism behind it applies to them; §4.3 explains why, and §4.4 places PRPG in the matched comparison.

2 Related work

The replicator equation is the canonical population-level description of payoff-based learning (Taylor & Jonker, 1978; Hofbauer & Sigmund, 1998; Weibull, 1997); softmax and cross learning converge to it in the small-step-size limit (Börgers & Sarin, 1997; Tuyls & Nowé, 2006; Bloembergen et al., 2015), which is why its equilibria predict the behaviour of learning agents (Bowling & Veloso, 2001), and the correspondence has been carried to multi-agent Q-learning (Sato & Crutchfield, 2003; Kianercy & Galstyan, 2012) and to reinforcement with regularisation (Mertikopoulos & Sandholm, 2016). It is fragile under discretisation, where convergence can turn into instability or chaos (Falniowski & Mertikopoulos, 2024), a regime we take seriously in §4.2. Related mechanisms include fluctuating selection (Chesson & Warner, 1981; Wienand et al., 2017; Stollmeier & Nagler, 2018), environmental feedback and stochastic games (Hilbe et al., 2018), and mutation in classical replicator–mutator models (Hadeler, 1981; Page & Nowak, 2002; Bauer et al., 2019). Entropy bonuses are standard (Haarnoja et al., 2018) and shape the policy-gradient landscape (Ahmed et al., 2019; Mei et al., 2020; Cen et al., 2022); in games, entropy-like regularisation is the main device for last-iterate rather than average-iterate convergence (Daskalakis & Panageas, 2018; Mertikopoulos et al., 2019; Bravo et al., 2018), giving algorithms with last-iterate guarantees (Abe et al., 2022). The two perturbations we combine have nevertheless only been studied apart: prior work carries evolutionary dynamics into multi-agent learning with the mutation term but without the noise (Abe et al., 2022; Bauer et al., 2026), or analyses fluctuating payoffs with the noise but without the mutation (Wang et al., 2026). For convergence we use the ODE method for stochastic approximation (Benaïm, 1999; Borkar, 2008), taking care over the rate, polynomial rather than exponential for decreasing step sizes. Repeated games played by LLM agents depart systematically from rational play (Akata et al., 2023; Lorè & Heydari, 2023; Park et al., 2023).

3 Noisy replicator–mutator dynamics

3.1 Setting

We consider a well-mixed population of N agents playing a symmetric two-action game with payoff matrix

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \qquad \pi_C = ax + b(1-x), \qquad \pi_D = cx + d(1-x), \tag{1}$$

where $x \in [0,1]$ is the frequency of cooperation and rows (columns) index the focal (opponent) action. With $\Delta\pi := \pi_C - \pi_D$, the replicator equation is $\dot{x} = s\,x(1-x)\Delta\pi$, where $s \ll 1$ is the weak selection intensity (Wang et al., 2026).

Noise (Wang et al. 2026). One shock $\xi \sim \mathcal{N}(0,1)$ is drawn per generation and applied to every interaction, so the realised payoff matrix is $\mathbf{A} + \xi\mathbf{\Sigma}$ with $\mathbf{\Sigma}$ a fixed amplitude matrix and the mean

game left unchanged. We use *proportional* noise, $\mathbf{\Sigma} = k\mathbf{A}$, giving $\sigma_C = k\pi_C$, $\sigma_D = k\pi_D$ and

$$\Delta\sigma := \sigma_C - \sigma_D = k\,\Delta\pi. \tag{2}$$

Since $\mathbb{E}[\xi] = 0$ the shock has no first-order effect; at second order it modifies the replicator equation to

$$\dot{x} \;=\; s\,x(1-x)\Big[\Delta\pi \;+\; \big(\tfrac{1}{2} - x\big)(\Delta\sigma)^2\Big], \tag{3}$$

whose n-strategy form, with $\bar{\pi} = \sum_i x_i\pi_i$, $\bar{\sigma} = \sum_i x_i\sigma_i$ and $\mathrm{Var}(\sigma) = \sum_i x_i(\sigma_i - \bar{\sigma})^2$, is

$$\dot{x}_i \;=\; s\,x_i\Big[\pi_i - \bar{\pi} + \tfrac{1}{2}\big((\sigma_i - \bar{\sigma})^2 - \mathrm{Var}(\sigma)\big)\Big]. \tag{4}$$

Both are due to Wang et al. (2026) and are our starting point; for $n = 2$, Eq. (4) reduces to Eq. (3) (Appendix G).

Mutation (Bauer et al. 2026). An additive mutation bias turns the replicator equation into a replicator–mutator system,

$$\dot{x}_i \;=\; x_i(\pi_i - \bar{\pi}) \;+\; M\,(c_i - x_i), \tag{5}$$

with c a reference distribution and $M \geq 0$ the mutation strength. The second term depends on no payoff: it pulls the population towards c and, when c is interior, removes the boundary as a possible limit, which is what makes interior equilibria last-iterate attracting (Bauer et al., 2026). We take $c = \mathbf{1}/n$ throughout.

Combined system. Adding the two perturbations gives

$$\boxed{\dot{x}_i \;=\; s\,x_i\Big[\pi_i - \bar{\pi} + \tfrac{1}{2}\big((\sigma_i - \bar{\sigma})^2 - \mathrm{Var}(\sigma)\big)\Big] \;+\; M\,(c_i - x_i)} \tag{6}$$

which we call *noisy replicator–mutator dynamics* (NRMD), Eq. (6); in two strategies it reads

$$\dot{x} \;=\; s\,x(1-x)\Big[\Delta\pi + \big(\tfrac{1}{2} - x\big)(\Delta\sigma)^2\Big] \;+\; M\big(\tfrac{1}{2} - x\big). \tag{7}$$

The two terms are of different algebraic type: the noise term is *multiplicative* and quadratic in the payoff difference, the mutation term *additive* and independent of the payoff matrix. §3.2 shows that this distinction decides which learners inherit the mechanism.

3.2 Where the shock sits, and first-order immunity

Eq. (3) is usually obtained by expanding the individual-based process to second order in the payoff fluctuation. Our derivation exhibits *which coordinate* the shock enters, since that, and not its size, decides which learners inherit the mechanism.

The bracket is a log-odds correction. Let a learner's logit follow the diffusion

$$\mathrm{d}z \;=\; \alpha\big[\Delta\pi\,\mathrm{d}t \;+\; \Delta\sigma\,\mathrm{d}W\big], \qquad x = \varsigma(z) = \frac{1}{1+e^{-z}}, \tag{8}$$

so that the shock enters the score, not the frequency. With $\varsigma'(z) = x(1-x)$ and $\varsigma''(z) = x(1-x)(1-2x)$, Itô's lemma gives $\mathrm{d}x = \varsigma'(z)\mathrm{d}z + \frac{1}{2}\varsigma''(z)\mathrm{d}\langle z\rangle$, and since $\frac{1}{2}(1-2x) = \frac{1}{2} - x$,

$$\mathrm{d}x \;=\; \alpha\,x(1-x)\Big[\Delta\pi + \alpha\big(\tfrac{1}{2} - x\big)(\Delta\sigma)^2\Big]\mathrm{d}t \;+\; \alpha\,x(1-x)\Delta\sigma\,\mathrm{d}W, \tag{9}$$

with diffusion coefficient $\alpha^2x^2(1-x)^2(\Delta\sigma)^2$. Setting $\alpha = s$ recovers Eq. (3) up to the normalisation of the shock amplitude—a per-generation increment of the payoff, not a Brownian one—and the $(\frac{1}{2} - x)$ prefactor is the ς'' term of the log-odds coordinate, not a modelling choice. The bracket is a genuine Itô correction, so it survives only while $\Delta\sigma = \Theta(1)$; if $\Delta\sigma = O(\sqrt{s})$ it is subleading and Eq. (3) collapses to the classical replicator equation, the caveat noted by Wang et al. (2026). The coordinate is not free either: applying the same shock to the frequency makes the two Itô corrections cancel exactly, leaving no bracket (Appendix A, Remark 1); the same boundary appears in the source analysis, where a pairwise-comparison update is unaffected by noisy payoffs (Wang et al., 2026).

First-order immunity of the payoff-linear class. The other side of the same observation is that a learner whose update is affine in the payoff cancels the shock exactly at first order.

Proposition 1 (First-order immunity)**.** *Let a per-agent update be* payoff-linear*,*

$$\Delta p_i = \alpha\, g(p_i)\,(a + b\,\tilde{r}_i), \tag{10}$$

with p_i the probability that agent i plays C, $g \geq 0$ a state-dependent weight, and g, a, b independent of the shock ξ. Then $\mathbb{E}[\Delta p_i \mid x] = \alpha g(p_i)(a + b r_i)$ does not depend on k; the replicator and softmax classes both have this form.

Proof. $\tilde{r}_i = r_i(1 + k\xi)$ with ξ independent of the state, so $\mathbb{E}[\tilde{r}_i \mid x] = r_i \mathbb{E}[1 + k\xi] = r_i$, and affineness in $\tilde{r}_i$ gives $\mathbb{E}[\Delta p_i \mid x] = \alpha g(p_i)(a + b r_i)$, in which k does not appear. See Appendix A.1. □

The cancellation is exact, not approximate. What remains is the second-order term $\frac{1}{2} f''(p)\mathrm{Var}[\Delta p_i] = O(\alpha^2 k^2)$ in the log-odds coordinate $f(p) = \log \frac{p}{1-p}$, one power of α below the drift (Lemma 1); for a learner that enters the payoff exponentially this same term *is* the mechanism of Proposition 4. The two classes differ not in whether a second-order term exists, but in whether it leads.

3.3 Shared versus independent noise

The word "global" in Wang et al. (2026) does real work, since most implementations of noisy MARL perturb each agent's reward separately.

Proposition 2 (Separation)**.** *Let N agents each receive a reward perturbation, and let Δx be the induced change in the population frequency and $\mathrm{Var}[\Delta x \mid x]$ its conditional variance. If the shocks are shared, $\xi_i = \xi$ for all i, then $\mathrm{Var}[\Delta x \mid x] = \alpha^2 x^2 (1-x)^2 (\Delta\sigma)^2 = \Theta(1)$. If the shocks are independent, $\xi_i \sim \mathcal{N}(0,1)$ i.i.d., then $\mathrm{Var}[\Delta x \mid x] = \alpha^2 x^2 (1-x)^2 (\Delta\sigma)^2 / N = O(1/N) \to 0$. Consequently the noise term of Eq.* (3) *survives in the mean-field limit if and only if the fluctuation has a component common to the population.*

Proof. See Appendix B. □

Proposition 2 is the formal reason for the protocol in §4, where one shock per round is shared: perturbing each agent's reward independently recovers ordinary replicator dynamics in the large-N limit, however large the per-agent noise is.

3.4 Entropy regularisation is exactly a mutation bias

The mutation term (5) is usually introduced as a modelling device. It is what an entropy bonus does.

Proposition 3 (Entropy bonus ⇒ mutation bias)**.** *Consider a softmax policy over two actions with logit z, so that $x = \varsigma(z)$ and $\dot{x} = x(1-x)\dot{z}$. Let the objective be the advantage plus an entropy bonus $\beta H(\pi)$ with $H(\pi) = -x \log x - (1-x)\log(1-x)$. Then the population-level limit of the update is*

$$\dot{x} \;=\; \eta\, x^2 (1-x)^2 \Big[\Delta\pi - \beta \log \tfrac{x}{1-x}\Big], \tag{11}$$

i.e. an entropy term is a mutation bias of state-dependent strength

$$M(x) \;=\; \eta\beta\, \psi(x), \qquad \psi(x) = \frac{x^2 (1-x)^2 \log \frac{1-x}{x}}{\frac{1}{2} - x}, \tag{12}$$

with $0 < \psi(x) \leq \frac{1}{4}$ on $(0,1)$ and $\psi(\frac{1}{2}) = \frac{1}{4}$.

Proof. See Appendix C. □

Three consequences. The mutation strength is $M \propto \eta\beta$, which Corollary 1 turns into a budget on $(\eta\beta, s)$. But $M(x) \to 0$ at both boundaries, so the entropy term does *not* remove the boundary degeneracy and the constant-M results below are a mid-interior approximation (Remark 2). Finally, Eq. (11) agrees with (7) on interior fixed points and stability and differs only in the time parametrisation, so we use the latter from here on.

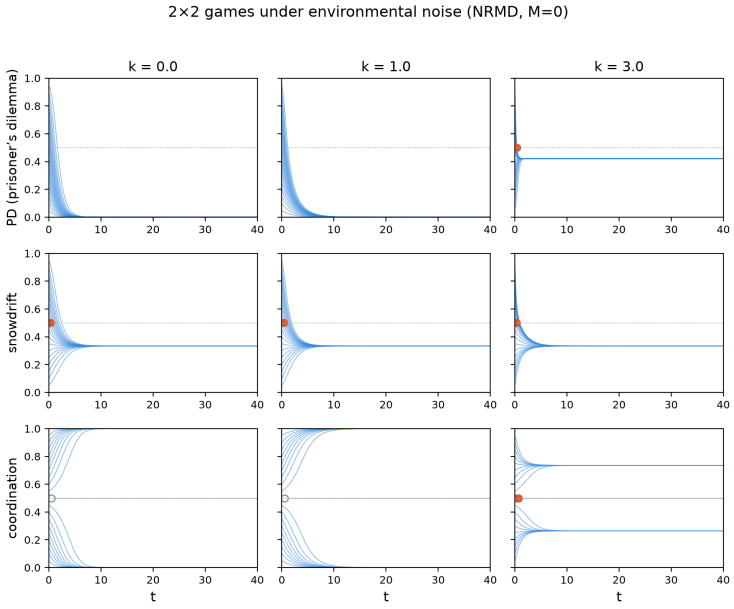

Figure 1: The threshold of Proposition 4 in 2×2 games (ODE (7), $M = 0$), nineteen trajectories per panel. Top: prisoner's dilemma $(3, 0; 5, 1)$; all-defection is the only attractor at $k = 0, 1$, while at $k = 3$ every trajectory reaches $x^*(3) = 0.422$ of Eq. (14). Middle: snowdrift, whose interior equilibrium needs no noise. Bottom: coordination, where noise replaces the boundary attractors by interior ones.

3.5 EQUILIBRIA

Noise alone: a sharp threshold. With $M = 0$, interior equilibria of Eq. (7) satisfy

$$H(x) := \Delta\pi(x) + \left(\tfrac{1}{2} - x\right)k^2\Delta\pi(x)^2 = 0 \qquad \Longleftrightarrow \qquad \left(\tfrac{1}{2} - x\right)k^2\,\Delta\pi(x) = -1 \quad (x \neq \tfrac{1}{2}). \tag{13}$$

Proposition 4 (Noise-induced coexistence in the prisoner's dilemma)**.** *Take the prisoner's dilemma* $(a, b, c, d) = (3, 0, 5, 1)$*, i.e.* $\pi_C = 3x$*,* $\pi_D = 4x + 1$*,* $\Delta\pi = -(1 + x)$*, and proportional noise* (2)*. Then Eq.* (7) *with* $M = 0$ *admits an equilibrium in* $(0, 1)$ *if and only if* $k > \sqrt{2}$*. When it exists it is unique, it lies in* $(0, \frac{1}{2})$*, it is attracting, and*

$$x^*(k) = -\frac{1}{4} + \frac{1}{2}\sqrt{\frac{9}{4} - \frac{4}{k^2}}, \tag{14}$$

which increases from 0 *at* $k = \sqrt{2}$ *to* $\frac{1}{2}$ *as* $k \to \infty$*. More generally, for a game with* $b < d$ *(so that all-defection is the stable boundary state of the classical equation), the all-defection state is destabilised by noise if and only if* $k^2 > 2/(d - b)$*.*

Proof. See Appendix D. □

The boundary value $H(0) = k^2/2 - 1$ is the whole content of the threshold: noise makes cooperation viable exactly when the noise-induced drift at the all-defection state turns positive. The prediction that the level of cooperation then *increases* with the noise scale is falsifiable and, in our experiments, holds; Fig. 1 shows the three 2×2 classes.

Adding mutation: the degeneracy is removed, and a growing budget is needed. With $M > 0$ the boundaries are no longer equilibria, since the mutation term contributes $\pm M/2$ there. Write $F(x) = s\,x(1 - x)H(x) + M(\frac{1}{2} - x)$ for Eq. (7).

Proposition 5 (Mutation removes the boundary degeneracy; the slope condition $M_{\mathrm{crit}}(k)$)**.** *For every* $k \geq 0$ *and* $M > 0$*,* $F(0) = M/2 > 0$ *and* $F(1) = -M/2 < 0$*: the boundaries are no longer equilibria,* F *has an odd number of roots in* $(0, 1)$*, and its smallest root* $x_1(k, M)$ *is attracting. The boundary slope satisfies* $F'(0) < 0$ *if and only if*

$$M \;>\; M_{\mathrm{crit}}(k) := \max\left\{0,\; s\left(\tfrac{k^2}{2} - 1\right)\right\}. \tag{15}$$

For $k \leq \sqrt{2}$ this holds at every $M > 0$; beyond $k = \sqrt{2}$ the slope turns negative only once $M \gtrsim sk^2/2$. It is a condition on the slope, not a bifurcation: the number of roots and their stability are the same on either side (Appendix E).

Proof. See Appendix E. □

Two remarks. M_{crit} is a condition on the boundary slope, not a threshold on the attractor: a bifurcation scan finds no change at $M = M_{\text{crit}}$. It is still the sharpest joint prediction to test, being the one quantity that mixes s, k and M. The combination is also not symmetric: mutation regularises the boundary for all k, whereas noise moves the interior equilibrium without bound.

Corollary 1 (Observability window)**.** *Call the control admissible when the cooperation floor it already carries at $k = 0$ is at most κ_0, i.e. $x_1(0, M) \leq \kappa_0$. Near the boundary $x_1(0, M) = M/(2s)(1 + O(\kappa_0))$, so admissibility is $M \lesssim 2\kappa_0 s$. At noise level k the attracting state moves to $x_1(k, M) = x^*(k) + O(M/s)$, so a lift can be attributed to the noise rather than to that floor only if the noise-induced equilibrium clears it,*

$$x^*(k) \;>\; \frac{M}{2s}, \tag{16}$$

Proof. See Appendix A.2. □

At the admissible extreme $M = 2\kappa_0 s$ this is $x^*(k) > \kappa_0$, a *lower* bound on the noise: since $x^*(k) \to 0$ at $k = \sqrt{2}$ and $\approx k^2/2 - 1$ there, $k \gtrsim \sqrt{2(1 + \kappa_0)}$ to leading order ($\kappa_0 = 0.1$ gives $k > 1.48$, $\kappa_0 = \frac{1}{4}$ gives $k > 1.58$), so the attributable regime is the deep-noise one. The effects are indeed largest at $k = 4$ and $k = 5$, where x^* sits well above any admissible floor, and nearly absent for $k \leq 1$. Since M is what an entropy coefficient sets (Proposition 3), the trade-off is between $(\eta\beta, s)$ and the noise scale, not a bound on the noise alone.

3.6 Convergence

We now ask whether a learner reaches the equilibria above. Let $x_{t+1} = x_t + \alpha_t\big(F(x_t) + \zeta_{t+1}\big)$ with $\{\zeta_t\}$ a martingale-difference noise and F as in §3.5.

Proposition 6 (Almost-sure convergence and rate)**.** *Assume (i) F is locally Lipschitz on $[0, 1]$ and the scheme is projected onto $[0, 1]$; (ii) $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$; (iii) $\sup_t \mathbb{E}[\|\zeta_t\|^2] < \infty$. Then, almost surely, x_t converges to a compact invariant set of NRMD, which is a single stable equilibrium when the attractors are isolated (in particular for two strategies), whereas for more than two a limit cycle is not excluded—§4 exhibits one. If in addition $F'(x^*) =: -\lambda < 0$ at an isolated equilibrium and $\alpha_t = c/t$ with $\lambda c > \frac{1}{2}$, the averaged iterates satisfy $\mathbb{E}[\|x_t - x^*\|^2] = O(1/t)$, a polynomial rather than exponential rate.*

Proof. See Appendix F. □

For the prisoner's dilemma λ is available in closed form. At x^* of Eq. (14), with H as in (13), $F'(x^*) = s\, x^*(1 - x^*)H'(x^*) - M$ since $H(x^*) = 0$, and with $\Delta\pi = -(1 + x)$,

$$H'(x^*) = -\Delta\pi'(x^*) - k^2\Delta\pi(x^*)^2 = 1 - k^2(1 + x^*)^2 < 0 \quad \text{for every } k \geq \sqrt{2},$$

since then $k^2(1 + x^*)^2 \geq k^2 > 1$. Hence $F'(x^*) < 0$ and the equilibrium is attracting with $\lambda = M - s\, x^*(1 - x^*)H'(x^*) > 0$.

4 Experiments

We test the predictions at three levels of abstraction, from a tabular benchmark to a population of LLM agents. All settings use a global shock drawn once per round and shared by all agents (Proposition 2); data, code and exact commands are in Appendix O.

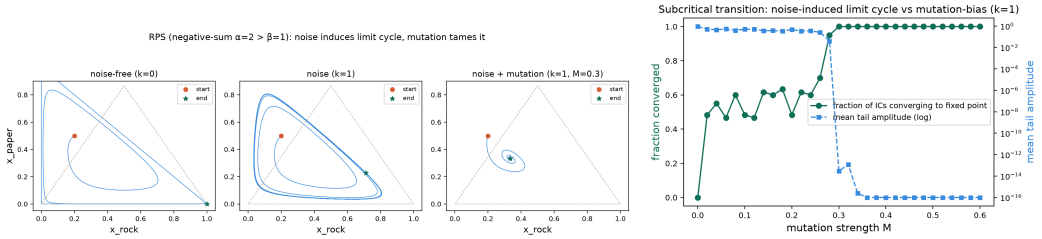

Figure 2: Left: noise induces a limit cycle in rock–paper–scissors ($\alpha = 2 > \beta = 1$); the trajectory reaches a boundary vertex at $k = 0$, winds onto a stable cycle at $k = 1$, and spirals into the interior fixed point once mutation is added ($k = 1$, $M = 0.3$). Right: the subcritical transition. Grey circles: fraction of 60 initial conditions reaching the fixed point; blue squares: mean tail amplitude (logarithmic axis), steady at order 10^0 throughout the coexistence window and collapsing by about thirteen orders of magnitude between $M = 0.26$ and $M = 0.30$.

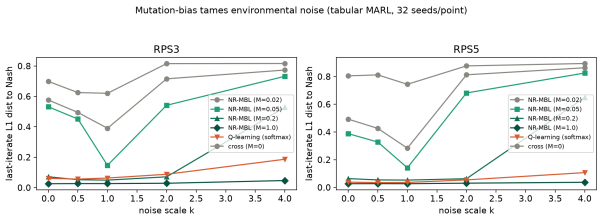

Figure 3: Tabular MARL: distance to Nash against k (flat: NR-MBL, $M = 1$; rising: softmax Q-learning); values in Table 4.

4.1 The ODE: phase diagram and subcritical transition

Phase diagram (reproduction). Solving Eq. (3) on 76,800 parameter points reproduces the seven dynamical regimes of Wang et al. (2026), including the one with two stable interior equilibria, which cannot occur without noise (Appendix I), as a check that (4) is implemented correctly.

Subcritical transition in rock–paper–scissors (new). With negative-sum payoffs ($\alpha = 2 > \beta = 1$) and $k = 1$ the noise-only dynamics has a stable limit cycle. Sweeping $M \in [0, 0.6]$ over 31 values with 60 initial conditions each, the cycle does not vanish at a single threshold: for $M \in [0.02, 0.28]$ the interior fixed point and the cycle *coexist*, and only for $M \geq 0.30$ do all initial conditions reach the fixed point (Fig. 2, right). Different initial conditions therefore give different long-run outcomes for identical parameters—a subcritical transition, and the sharpest qualitative difference between NRMD and either ingredient: the noise-only analysis cannot switch the cycle off, and the mutation-only analysis has no cycle to switch off.

4.2 Tabular MARL

We run a finite-population benchmark on rock–paper–scissors with $n = 3$ and $n = 5$ strategies, 30 algorithm–noise configurations per game and 32 seeds each ($\theta = 2 \times 10^{-4}$ for cross learning, $\alpha_Q = 0.01$; noise enters as $p \mapsto p + k|p|\xi$ with ξ shared). Table 4 (Appendix M) reports the L_1 distance from the uniform mixture, the unique Nash equilibrium, and the fraction of seeds that converge. The pattern matches Proposition 5 for two strategies and the n-strategy phase portrait of Appendix J. Cross learning ($M = 0$) converges at no noise level; softmax Q-learning degrades monotonically in success, reaching 0.187 and 0% at $k = 4$; NR-MBL at $M = 0.02$ and 0.05 is insufficient; $M = 0.2$ holds the distance just above 0.05 for $k \leq 2$ but fails at $k = 4$; and $M = 1.0$ keeps it below 0.05 throughout, with success 0.97–1.00 up to $k = 2$ and 0.56 at $k = 4$. The same ordering holds for $n = 5$, where Q-learning falls from 0.91 to 0.00 success while NR-MBL ($M = 1$) stays above 0.94. The mutation budget needed to maintain convergence therefore grows with the noise, consistently with the k^2 scaling of Eq. (15) (a slope condition, not a bifurcation, as Proposition 5 notes), in both dimensionalities (Appendix M); Fig. 3 plots two representative algorithms.

Table 1: Two update rules, one environment, matched policy sets ($P = 64$, 20 seeds). "Lift" is $x(k{=}5) - x(k{=}0)$ above each arm's own floor; sd columns are cross-seed; all runs reached an equilibrium.

Arm	**floor** $x(k{=}0)$	$x(k{=}5)$	**lift**	**sd**$(k{=}0)$	**sd**$(k{=}5)$
`repl` (replicator, $\|\theta\| \le 3$)	0.0200	**0.4300**	**+0.4101**	0.0030	0.0169
`grad`, $\theta_{\max} = 3$ (*same* set)	0.0025	**0.0219**	**+0.0195**	0.0000	0.0027
`grad`, $\theta_{\max} = 12$ (set escaped)†	0.0138	0.1078	+0.0940	0.0156	**0.0801**

4.3 Deep MARL: a theory-aligned learner and a negative control

Theory-aligned learner (NRL). PPO (Schulman et al., 2017) contains clipping, GAE and a centralised critic, none of which belongs to the class §3 covers. We therefore implement a learner whose update *is* the discrete NRMD update,

$$\Delta z_a \;=\; \eta\Big(\mathbf{1}[a = a_i]\hat{A}_i \;-\; M\,(z_a - \bar{z})\Big), \qquad \hat{A}_i = r_i - V(s), \tag{17}$$

where z are the logits of a network-parameterised softmax policy and r_i carries the global shock. In logit space $-M(z - \bar{z})$ is the contraction towards uniform that Proposition 3 derives from an entropy bonus. The environment is a four-agent pairwise rock–paper–scissors game whose unique Nash equilibrium is the uniform mixture, so a policy with no interior mass is a clear failure; we measure its entropy $H(\pi)$ (maximum $\log 3$) and $\|\pi - \mathbf{c}\|_1$ (maximum $4/3$). This learner validates the continuum limit rather than claiming algorithmic superiority.

The results in Fig. 4 follow the theory. With $M = 0$ the policy collapses onto a pure strategy: $H = 0.005$ and $\|\pi - \mathbf{c}\|_1 = 1.330$, essentially the maximum $4/3$. The entropy rises monotonically with M—0.005, 0.467, 0.769, 0.933 for $M = 0, 0.05, 0.2, 0.5$ pooled over noise—while the distance to uniform falls $1.330 \to 0.963 \to 0.674 \to 0.464$. A linear fit over the pooled values gives $H \approx 0.252 + 1.554M$, with a seed-to-seed standard deviation below 0.08 at every cell. At fixed mutation the entropy falls by 19.2% from $k = 0$ to $k = 3$ at $M = 0.2$ and by 13.3% at $M = 0.5$, the discrete analogue of the growing budget $M_{\mathrm{crit}}(k)$.

Negative control: standard deep MARL. We run the same protocol with a standard PPO-style learner (Schulman et al., 2017) (four variants: plain, an entropy bonus, a variance penalty, and both) on a cooperative coverage task, four noise levels and three seeds. Across all 16 cells the four variants differ by at most 0.36 in final return, while the seed-to-seed standard deviation within a cell is 1.1–2.7: the effect is below the noise floor. This is predicted, not incidental: PPO's update contains clipped ratios, generalised advantage estimation and a centralised critic, none of which is affine in the sampled payoff, so it is not of the form (10), Proposition 1 does not protect it, and it does not inherit the mechanism, which needs the payoff to enter an exponential. A task whose optimum is close to deterministic offers no role for an interior mixture either.

4.4 One environment, two update rules, matched policy sets

Environments, not update rules, could explain a cross-environment screen. We therefore built a JAX implementation of *one* explicit N-player iterated prisoner's dilemma with the payoff matrix $[[3, 0], [5, 1]]$ used throughout (under random matching $\pi_D - \pi_C = x + 1 > 0$, so the all-defection corner exists by construction), and placed two update rules inside it sharing the environment, the shock process, the time budget and the seeds. In the replicator-class arm `repl` (the PRPG learner of §1, Algorithm 1) each individual carries a discrete type in $\{C, D\}$ that sets a scalar bias $\theta = 3(1 - 2 \cdot \text{type})$ on the defection logits, so the policy set is $|\theta| \le 3$; reproduction is exponential in payoff under the shared shock, and offspring types are copied with rare flips. In the gradient-class arm `grad`, REINFORCE (Williams, 1992) acts on the *same* scalar θ, with the same policy parameterisation and learning rate, but with θ continuous. We clip `grad` at $\theta_{\max} = 3$, exactly the policy set of `repl`, so both arms can express the same behaviours. The constraint favours the gradient arm, which then ranges over a continuum while the replicator arm uses only $\theta = \pm 3$; "more freedom and still flat" is therefore the strong statement. Table 1 reports the result ($P = 64$, horizon 20, 12000 generations, z-scored fitness, twenty seeds).

Within the matched policy set, the replicator arm lifts by $+0.4101$ with sd 0.003–0.017, while the gradient arm is flat at $+0.0195$ with sd 0.000–0.003: twenty-one times smaller. On one environment under one shock definition, the class of the update rule accounts for the difference.

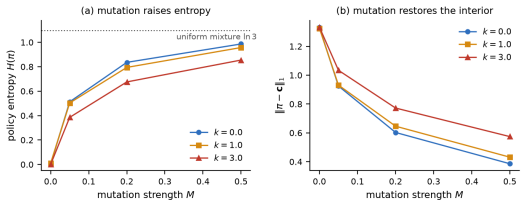

Figure 4: The theory-aligned deep learner of Eq. (17): (a) policy entropy $H(\pi)$ against M for three noise levels k (dotted: the uniform mixture $\log 3$) and (b) the L_1 distance to it. Both panels move as Propositions 3 and 5 predict.

Relaxing the clip raises the gradient arm's mean lift to $+0.0940$, but its cross-seed sd grows to 0.0801 and the per-seed values sit on multiples of $1/P$: what moves is the policy set, not a population equilibrium. Two caveats: the mutation term is a rare copy-flip rather than a controlled strength, so it fixes a floor but does not sweep M; and the noise scale is the implementation's raw k, not the dimensionless k of §3 under the z-scored normalisation. We therefore use Table 1 as a test of the class boundary of Proposition 1 and not of Corollary 1, whose window is calibrated in the theoretical normalisation.

4.5 LLM AGENTS IN A SOCIAL DILEMMA

Finally, as a proof of concept, we replace the learner with a population of LLM agents: $n = 6$ copies of Qwen2.5-7B-Instruct (Qwen Team, 2025) playing a repeated prisoner's dilemma for 20 rounds, with a global payoff shock of scale k and a mutation probability ε that overrides a decision with a uniformly random one. Because these aligned agents cooperate under any framing, M is this explicit mutation, not a temperature.

Pooled over the three noise levels, cooperation is 0.038, 0.145, 0.253, 0.331 and 0.478 for $\varepsilon = 0, 0.1, 0.2, 0.3, 0.5$, with coop $\approx 0.057 + 0.875\,\varepsilon$ ($R^2 = 0.91$); at $\varepsilon = 0.5$ the population sits at $0.478 \approx \frac{1}{2}$, precisely the mutation target $c = \frac{1}{2}$. The mutation dimension is thus recovered in agents never trained on the game. The noise dimension is not: the k-coefficient is 0.000 ± 0.005 per unit k, excluding shifts above 0.043 over $k \in [0, 4]$—a bounded null, not a failure to detect. The quantitative evidence for the mechanism rests on the twenty-seed two-arm comparison of Table 1; Tables 2 and 3 and Fig. 7 give the LLM dimensions in full.

5 CONCLUSION AND DISCUSSION

Escaping the defection trap. NRMD is a way out of a failure mode that recurs across multi-agent systems: wherever agents share a resource, a budget or a commons, each is tempted to maximise its own payoff at the collective's expense, and the system settles into the all-defection state—the attractor that noise destabilises. NRMD says how to leave it: a large enough fluctuation in the shared payoff environment makes cooperation viable again (Proposition 4, $k > \sqrt{2}$), and any mutation bias keeps the population interior (Proposition 5). Both hold analytically (§3) and experimentally (§4): a replicator-class learner raises cooperation from 0.020 to 0.430 while a gradient learner of the same size stays flat. Escape is not automatic: it needs the right update class, a noise scale above the threshold and the entropy coefficient that sets $M \propto \eta\beta$ (Proposition 3).

Scope. The limits are stated where they are derived and tested: the immune class (Proposition 1), the shared-shock requirement (Proposition 2), the observable window (Corollary 1), the normalisation caveat (§4.4) and the language-model limitation (§4.5); independent per-agent noise remains untested.

Future work. The most useful next step is a learner that does not need the noise scale chosen by hand; here k and M are fixed and §4.4 showed how much this matters. We therefore plan to learn the noise intensity online inside standard benchmarks such as SMAC and MPE—agents estimate the fluctuation level from their reward streams and a controller sets the effective noise and mutation strengths, with Eq. (15) and Corollary 1 supplying the set point. A second direction is to calibrate the two noise normalisations.

AI USE STATEMENT

Generative AI tools were used, and we disclose the categories of use. They assisted with the writing and editing of the manuscript, with retrieving bibliographic information, with writing the numerical code, and, in a limited way, with derivations. All AI-generated content was reviewed by the authors. The formulas, symbols and logic of the derivations were re-checked by hand and against numerical solutions of the ODE and of the stochastic process. The execution of the code and of the experiments was likewise reviewed by the authors, and its outputs were cross-validated against analytic predictions: the threshold $k = \sqrt{2}$ of Proposition 4, for instance, was confirmed both analytically and by simulation. Bibliographic information retrieved with AI assistance was checked against the publishers' records, and no synthetic datasets were generated. The authors take responsibility for the final content, including all text, the derivations of the propositions, the experiments, and the conclusions drawn from them.

ETHICS STATEMENT

This work is theoretical and computational. It does not involve human subjects, personal data or datasets of human behaviour; the experimental subjects are simulated agents and a publicly available language model. The most plausible misuse of the results is the deliberate injection of payoff noise into multi-agent systems in order to destabilise coordination; we note that the same analysis predicts the defence, since the mutation term restores interior stability, and Proposition 5 gives the strength at which the drift turns inward at the boundary.

REPRODUCIBILITY STATEMENT

All results in §4 were produced by the scripts released with this paper, which are self-contained apart from NumPy, SciPy, Matplotlib and PyTorch; the LLM experiments use vLLM to serve an open-weights model, through its OpenAI-compatible client. Appendix O lists the exact commands, hyperparameters, seeds and the mapping from each experiment to the figure or table it produces. Complete proofs of all statements are given in Appendices A–F, including the assumptions under which each holds. The raw result files (phase_diagram.csv, m_crit_scan.csv, phase2_results.csv, results_e4.csv, results_e4b.csv, results_e5.csv, results_e5_noise.csv, prpg_bin_seed20.json) are released with it.

REFERENCES

Kenshi Abe, Mitsuki Sakamoto, and Atsushi Iwasaki. Mutation-driven follow the regularized leader for last-iterate convergence in zero-sum games. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2022.

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. *International Conference on Machine Learning (ICML)*, 2019.

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *arXiv:2305.16867*, 2023.

Johann Bauer, Mark Broom, and Eduardo Alonso. The stabilization of equilibria in evolutionary game dynamics through mutation: mutation limits in evolutionary games. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 475(2231):20190355, 2019.

Johann Bauer, Sheldon West, Eduardo Alonso, and Mark Broom. Mutation-bias learning: an evolutionary game dynamics approach to convergence analysis in multi-agent reinforcement learning. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 482 (2329):20250449, 2026. doi: 10.1098/rspa.2025.0449.

Michel Benaïm. Dynamics of stochastic approximation algorithms. *Séminaire de Probabilités XXXIII*, pp. 1–68, 1999.

Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research*, 53:659–697, 2015.

Tilman Börgers and Rajiv Sarin. Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, 77(1):1–14, 1997.

Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

Michael Bowling and Manuela Veloso. Rational and convergent learning in stochastic games. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2001.

Mario Bravo, David Leslie, and Panayotis Mertikopoulos. Bandit learning in concave N-person games. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578, 2022.

Peter L. Chesson and Robert R. Warner. Environmental variability promotes coexistence in lottery competitive systems. *The American Naturalist*, 117(6):923–943, 1981.

Constantinos Daskalakis and Ioannis Panageas. Last-iterate analysis for primal-dual dynamics needs the right stepsize. *arXiv:1812.05144*, 2018.

Fryderyk Falniowski and Panayotis Mertikopoulos. On the discrete-time origins of the replicator dynamics: from convergence to instability and chaos. *arXiv:2402.09824*, 2024.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *International Conference on Machine Learning (ICML)*, 2018.

K. P. Hadeler. Stable polymorphisms in a selection model with mutation. *SIAM Journal on Applied Mathematics*, 41(1):1–7, 1981.

Christian Hilbe, Štěpán Šimsa, Krishnendu Chatterjee, and Martin A. Nowak. Evolution of cooperation in stochastic games. *Nature*, 559:246–249, 2018.

Josef Hofbauer and Karl Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1998.

Ardeshir Kianercy and Aram Galstyan. Dynamics of Boltzmann Q learning in two-player two-action games. *Physical Review E*, 85(4):041145, 2012.

Nunzio Lorè and Babak Heydari. Strategic behavior of large language models: Game structure vs. contextual framing. *arXiv:2309.05898*, 2023.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. *International Conference on Machine Learning (ICML)*, 2020.

Panayotis Mertikopoulos and William H. Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, 2016.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *International Conference on Learning Representations (ICLR)*, 2019.

Karen M. Page and Martin A. Nowak. Unifying evolutionary dynamics. *Journal of Theoretical Biology*, 219(1):93–98, 2002.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. *ACM Symposium on User Interface Software and Technology (UIST)*, 2023.

Qwen Team. Qwen2.5 technical report. *arXiv:2412.15115*, 2025.

Yuzuru Sato and James P. Crutchfield. Coupled replicator equations for the dynamics of learning in multiagent systems. *Physical Review E*, 67(1):015206, 2003.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017.

Frank Stollmeier and Jan Nagler. Unfair and anomalous evolutionary dynamics from fluctuating payoffs. *Physical Review Letters*, 120(5):058101, 2018.

Peter D. Taylor and Leo B. Jonker. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40(1-2):145–156, 1978.

Karl Tuyls and Ann Nowé. Evolutionary game theory and multi-agent reinforcement learning. *The Knowledge Engineering Review*, 20(1):63–90, 2006.

Guocheng Wang, Qi Su, Long Wang, and Joshua B. Plotkin. Noise-induced replicator dynamics in evolutionary games. *Physical Review Letters*, 137(9):097401, 2026. doi: 10.1103/3yby-qq2n.

Jörgen W. Weibull. *Evolutionary Game Theory*. MIT Press, 1997.

Karl Wienand, Erwin Frey, and Mauro Mobilia. Evolution of a fluctuating population in a randomly switching environment. *Physical Review Letters*, 119(15):158301, 2017. doi: 10.1103/PhysRevLett.119.158301.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992.

A Itô derivation of the noise bracket

Proof of the log-odds derivation of Eq. (3). Let z be the log-odds of cooperation and let the population-level dynamics in z carry both the deterministic drift and the shared shock,

$$\mathrm{d}z = \alpha\big[\Delta\pi\,\mathrm{d}t + \Delta\sigma\,\mathrm{d}W\big], \qquad x = \varsigma(z), \quad \varsigma'(z) = x(1-x), \quad \varsigma''(z) = x(1-x)(1-2x).$$

Itô's lemma gives $\mathrm{d}x = \varsigma'(z)\mathrm{d}z + \frac{1}{2}\varsigma''(z)\mathrm{d}\langle z\rangle$ with $\mathrm{d}\langle z\rangle = \alpha^2(\Delta\sigma)^2\mathrm{d}t$. The first term is $\alpha x(1-x)\Delta\pi\,\mathrm{d}t + \alpha x(1-x)\Delta\sigma\,\mathrm{d}W$; the second is

$$\tfrac{1}{2}\varsigma''(z)\,\alpha^2(\Delta\sigma)^2 = \tfrac{1}{2}x(1-x)(1-2x)\alpha^2(\Delta\sigma)^2 = \alpha^2 x(1-x)\big(\tfrac{1}{2}-x\big)(\Delta\sigma)^2,$$

using $\frac{1}{2}(1-2x) = \frac{1}{2}-x$. Adding the two and setting $\alpha = s$, $\Delta\sigma = k\Delta\pi$ recovers Eq. (3) up to the amplitude normalisation of §3.2, and identifies the diffusion coefficient of Eq. (9) as

$$\sigma^2(x) = \alpha^2 x^2(1-x)^2(\Delta\sigma)^2. \tag{18}$$

□

Remark 1 (The coordinate is not a free choice). If the same shock is applied to the frequency instead of the score, i.e. $\mathrm{d}x = \alpha x(1-x)\Delta\pi\,\mathrm{d}t + \alpha x(1-x)\Delta\sigma\,\mathrm{d}W$, and the log-odds $f(x) = \log\frac{x}{1-x}$ is then formed with Itô's lemma, one obtains $f'(x) = \frac{1}{x(1-x)}$ and

$$f''(x) = -\frac{1}{x^2} + \frac{1}{(1-x)^2} = \frac{-(1-x)^2 + x^2}{x^2(1-x)^2} = \frac{2x-1}{x^2(1-x)^2},$$

so that $\mathrm{d}f = \alpha\big[\Delta\pi - \alpha(\frac{1}{2}-x)(\Delta\sigma)^2\big]\mathrm{d}t + \alpha\Delta\sigma\mathrm{d}W$, that is, the bracket appears *with a minus sign*. Converting back to frequencies is not the inverse operation in the stochastic case: since $x = \varsigma(f)$ one must add the term $\frac{1}{2}\varsigma''(f)\mathrm{d}\langle f\rangle = \frac{1}{2}x(1-x)(1-2x)\alpha^2(\Delta\sigma)^2\mathrm{d}t$. The two corrections then cancel identically,

$$\begin{aligned}&-\alpha^2 x(1-x)\big(\tfrac{1}{2}-x\big)(\Delta\sigma)^2 + \tfrac{1}{2}\alpha^2 x(1-x)(1-2x)(\Delta\sigma)^2\\ &= \alpha^2 x(1-x)(\Delta\sigma)^2\Big[-\big(\tfrac{1}{2}-x\big) + \tfrac{1}{2}(1-2x)\Big] = 0,\end{aligned}$$

and the drift reduces to the bare $\alpha x(1-x)\Delta\pi$: no bracket survives. The bracket of Eq. (3) therefore requires the shock to act on the score. Both statements are elementary, but the pair matters, because it is exactly the distinction that separates the two classes of Proposition 1.

A.1 First-order immunity

Proof of Proposition 1. By Eq. (10), $\Delta p_i = \alpha g(p_i)(a + b\,\tilde{r}_i)$ with $\tilde{r}_i = r_i(1 + k\xi)$ and ξ a standard normal drawn independently of the state, so $\mathbb{E}[\tilde{r}_i \mid x] = r_i(1 + k\mathbb{E}[\xi]) = r_i$ and $g(p_i)$ is measurable with respect to the state. Linearity of the conditional expectation gives

$$\mathbb{E}[\Delta p_i \mid x] = \alpha g(p_i)\big(a + b\,\mathbb{E}[\tilde{r}_i \mid x]\big) = \alpha g(p_i)(a + br_i),$$

which contains no k. The class is exactly the updates that are affine in the realised payoff with shock-independent coefficients: the replicator update ($g = x(1-x)$, $a = 0$, $b = \Delta\pi$ normalised) and the fixed-temperature softmax update ($g = x(1-x)$, b the logit gain) are both in it. The residual k-dependence of the update is confined to $\mathrm{Var}[\Delta p_i \mid x] = \alpha^2 b^2 r_i^2 k^2$, and reaches the drift of the log-odds coordinate only at second order through $\frac{1}{2} f'' \mathrm{Var}[\Delta p_i] = O(\alpha^2 k^2)$. For the gated updating rule of §3.2 this is smaller by a factor α than the leading drift, whereas for the exponential rule of Proposition 4 the same second-order object is the leading term. □

A.2 Observability window

Proof of Corollary 1. At $k = 0$ the drift of Eq. (7) is $s\,x(1-x)\Delta\pi(x) + M(\frac{1}{2} - x)$. Expanding about the boundary as in Appendix E,

$$F(x) = \tfrac{M}{2} + x\big(s\Delta\pi(0) - M\big) + O(x^2), \qquad \Delta\pi(0) = b - d,$$

so the root nearest the boundary is

$$x_1(0, M) = \frac{M/2}{M - s\Delta\pi(0)} + O\big((M/s)^2\big) = \frac{M/2}{M + s(d-b)} + O\big((M/s)^2\big).$$

For the prisoner's dilemma $d - b = 1$, so $x_1(0, M) = M/(2s)$ to leading order, and the requirement $x_1(0, M) \le \kappa_0$ bounds the floor at $k = 0$ by $M \lesssim 2\kappa_0 s$. At noise level k the attracting root moves up to $x_1(k, M) = x^*(k) + O(M/s)$, so a lift is attributable to the noise rather than to that floor only while $x^*(k) > x_1(0, M)$; at the admissible extreme $M = 2\kappa_0 s$ this reads $x^*(k) > \kappa_0$. Since $x^*(k) \to 0$ at $k = \sqrt{2}$ and $x^*(k) = \frac{1}{2}k^2 - 1 + O((k^2-2)^2)$ just above it, the criterion is a *lower* bound on the noise,

$$k \gtrsim \sqrt{2(1 + \kappa_0)},$$

with no upper edge: the attributable regime is the deep-noise one. No condition on M_{crit} enters here, because by Proposition 5 the boundary ceases to be an equilibrium for *every* $M > 0$, and M_{crit} only fixes the sign of the boundary slope. □

Lemma 1 (Itô–Taylor truncation)**.** *Let $f \in C^3$ with f'' bounded on $[\delta, 1-\delta]$, and let $\Delta p = O(\alpha)$ with all moments finite. Then $\mathbb{E}[f(p + \Delta p) - f(p)] = f'(p)\mathbb{E}[\Delta p] + \frac{1}{2} f''(p)\mathbb{E}[(\Delta p)^2] + O(\alpha^3)$ uniformly on $[\delta, 1-\delta]$.*

Proof. Expand f to second order and bound the remainder by $\frac{1}{6}\sup_{[\delta,1-\delta]} |f'''|\,\mathbb{E}|\Delta p|^3$. For the updates of §3.2, $\Delta p = \alpha(a + b\,r(1 + k\xi))$ with payoffs bounded by 5 and ξ standard normal, so all moments of Δp are finite and $\mathbb{E}|\Delta p|^3 = O(\alpha^3)$ with a constant uniform in p. The factor $p(1-p) \le \frac{1}{4}$ keeps f'' bounded away from the boundary; the boundary itself is handled by the absorbing argument of Appendix D. □

B Separation of shared and independent noise

Proof of Proposition 2. Write $x = N_C/N$. To first order in the step size, and using $A_{ij} = a_{ij} + \xi_i$-type shocks with amplitude $\Delta\sigma$ in the relevant payoff differences,

$$\Delta x = \alpha\, x(1-x)\big(\Delta\pi + \Delta\sigma\,\bar{\xi}\big) + O(\alpha^2), \qquad \bar{\xi} := \frac{1}{N}\sum_{i=1}^{N} \xi_i.$$

The first-order term in $\Delta\pi$ is the classical drift, of second moment $\alpha^2 x^2 (1-x)^2 \Delta\pi^2$, and is M- and k-independent; the shock enters only through the population average $\bar{\xi}$, and it is the variance of

the increment that carries it. We therefore compute $\mathrm{Var}[\Delta x \mid x] = \mathbb{E}[(\Delta x)^2 \mid x] - (\mathbb{E}[\Delta x \mid x])^2$, which drops the classical term above.

Shared noise. If $\xi_i = \xi$ for all i then $\bar{\xi} = \xi$ exactly, and

$$\mathrm{Var}[\Delta x \mid x] = \alpha^2 x^2 (1-x)^2 (\Delta\sigma)^2 \mathbb{E}[\xi^2] + O(\alpha^3) = \alpha^2 x^2 (1-x)^2 (\Delta\sigma)^2,$$

using $\mathbb{E}[\xi^2] = 1$. This is $\Theta(1)$ in the population size and does not vanish as $N \to \infty$.

Independent noise. If the ξ_i are i.i.d. standard normal then $\bar{\xi}$ has mean zero and variance $1/N$, so by the strong law $\bar{\xi} \to 0$ almost surely and

$$\mathrm{Var}[\Delta x \mid x] = \alpha^2 x^2 (1-x)^2 (\Delta\sigma)^2 \, \mathbb{E}[\bar{\xi}^2] = \frac{\alpha^2 x^2 (1-x)^2 (\Delta\sigma)^2}{N} = O(1/N) \to 0.$$

Conclusion. The variance contribution of the shock to the mean-field dynamics is $\Theta(1)$ under a common shock and $O(1/N) \to 0$ under independent shocks, so the noise bracket of Eq. (3) is present in the large-population limit only for a shared (perfectly correlated) fluctuation. □

C Entropy regularisation and the mutation bias

Proof of Proposition 3. Let $z \in \mathbb{R}$ be the logit, $x = \varsigma(z) = 1/(1+e^{-z})$ the cooperation probability, and $H(\pi) = -x \log x - (1-x)\log(1-x)$ the entropy of the policy. Since $x'(z) = x(1-x)$ and $\frac{dH}{dx} = -\log \frac{x}{1-x}$, differentiating the entropy-regularised objective $J(z) = \mathbb{E}[\hat{A}] + \beta H$ with respect to z gives

$$\frac{\mathrm{d}J}{\mathrm{d}z} = x(1-x)\big(\hat{A}_C - \hat{A}_D\big) - \beta\, x(1-x) \log \frac{x}{1-x}.$$

Because x depends on z only through ς, $\dot{x} = x(1-x)\dot{z}$, and the population-level limit of the update $\dot{z} = \eta \frac{\mathrm{d}J}{\mathrm{d}z}$ is

$$\dot{x} = \eta\, x^2(1-x)^2\big(\hat{A}_C - \hat{A}_D\big) - \eta\beta\, x^2(1-x)^2 \log \frac{x}{1-x}.$$

Replacing the advantage difference by the population payoff difference $\Delta\pi$ (as in the replicator limit) gives Eq. (11).

For the second statement, define $\psi(x) = \dfrac{x^2(1-x)^2 \log \frac{1-x}{x}}{\frac{1}{2} - x}$ for $x \neq \frac{1}{2}$, so that the entropy term equals $\eta\beta\,\psi(x)(\frac{1}{2} - x)$. Put $t = 2x - 1 \in (-1, 1)$, so that $x(1-x) = (1-t^2)/4$, $\frac{1}{2} - x = -t/2$ and $\log \frac{1-x}{x} = -2\,\mathrm{artanh}\, t$. Then

$$\psi(x) = \frac{(1-t^2)^2}{4} \cdot \frac{\mathrm{artanh}\, t}{t},$$

which is positive and even in t. Since $\mathrm{artanh}\, t = \int_0^t \frac{\mathrm{d}s}{1-s^2} \leq \frac{t}{1-t^2}$ for $t \in (0,1)$,

$$\psi(x) \leq \frac{(1-t^2)^2}{4} \cdot \frac{1}{1-t^2} = \frac{1-t^2}{4} \leq \frac{1}{4},$$

with equality only at $t = 0$, i.e. $x = \frac{1}{2}$, where $\psi(\frac{1}{2}) = \frac{1}{4}$. Hence $0 < \psi(x) \leq \frac{1}{4}$ on $(0,1)$, and $M(x) = \eta\beta\,\psi(x)$ is a genuine mutation bias with $M \leq \eta\beta/4$. In particular $M(x) > 0$ for all interior x.

Remark 2 (ψ vanishes at the boundaries). The definition of ψ has a removable singularity at $x = \frac{1}{2}$, where the limit is $\frac{1}{4}$; and as $x \to 0^+$,

$$\psi(x) = \frac{x^2(1-x)^2 \log \frac{1-x}{x}}{\frac{1}{2} - x} = 2x^2 \log \frac{1}{x}\,(1 + o(1)),$$

so the induced bias itself is $M(x)\big(\frac{1}{2} - x\big) = \eta\beta\, x^2(1-x)^2 \log \frac{1-x}{x} = O(x^2 \log \frac{1}{x})$. This is $o(x)$, hence subleading to the selection term $s\, xH(0)$ of Eq. (7) near the boundary: the entropy bonus does *not* remove the boundary degeneracy, and the constant-M results of Proposition 5 and Corollary 1 are mid-interior approximations to it, not consequences of it. The bias is largest at the uniform distribution and dies at the boundary, which is the opposite of what a controlled constant mutation strength does.

□

D NOISE-INDUCED INTERIOR EQUILIBRIUM

Proof of Proposition 4. With $(a, b, c, d) = (3, 0, 5, 1)$ we have $\pi_C = 3x$, $\pi_D = 4x + 1$ and $\Delta\pi = -(1 + x)$, and by (2) $\Delta\sigma = k\Delta\pi = -k(1 + x)$. Interior equilibria of (7) with $M = 0$ satisfy

$$(\tfrac{1}{2} - x)k^2\Delta\pi(x) = -1.$$

Substituting $\Delta\pi = -(1 + x)$ and letting $u = x + 1 \in (1, 2)$, so that $\frac{1}{2} - x = \frac{3}{2} - u$, this becomes $k^2u(\frac{3}{2} - u) = 1$, i.e. $k^2u^2 - \frac{3}{2}k^2u + 1 = 0$, whose roots are

$$u_\pm = \frac{3}{4} \pm \frac{1}{2}\sqrt{\frac{9}{4} - \frac{4}{k^2}}.$$

For the root to lie in $(1, 2)$ we need $u \in (1, \frac{3}{2})$, which is the $+$ branch; the $-$ branch gives $u \le \frac{3}{4}$, hence $x = u - 1 < 0$ and no interior state. The condition $u_+ \ge 1$ is equivalent to

$$\frac{1}{2}\sqrt{\frac{9}{4} - \frac{4}{k^2}} \;\ge\; \frac{1}{4} \iff \frac{9}{4} - \frac{4}{k^2} \ge \frac{1}{4} \iff \frac{4}{k^2} \le 2 \iff k \ge \sqrt{2},$$

with $u_+ < 2$ automatic since $\sqrt{9/4 - 4/k^2} < 3/2$ for finite k; the root is then unique because $u \mapsto k^2u(\frac{3}{2} - u)$ is strictly decreasing on $(1, \frac{3}{2})$. This gives $x^*(k) = u_+ - 1 = -\frac{1}{4} + \frac{1}{2}\sqrt{9/4 - 4/k^2}$, which is increasing in k, equals 0 at $k = \sqrt{2}$ and tends to $\frac{1}{2}$ as $k \to \infty$.

For the general statement, $\Delta\pi(0) = b - d$ and $H(0) = \Delta\pi(0) + \frac{1}{2}k^2\Delta\pi(0)^2 = \Delta\pi(0)\big[1 + \frac{1}{2}k^2\Delta\pi(0)\big]$. With $b < d$ we have $\Delta\pi(0) < 0$, so $H(0) > 0$ iff $1 + \frac{1}{2}k^2\Delta\pi(0) < 0$ iff $k^2 > 2/(d - b)$.

Stability. For $M = 0$ the drift is $s\,x(1 - x)H(x)$ with H as in (13), so the interior root is attracting iff $H'(x^*) < 0$. Differentiating $H = \Delta\pi\big[1 + k^2(\frac{1}{2} - x)\Delta\pi\big]$ gives

$$H'(x) = \Delta\pi'(x)\Big[1 + 2k^2\big(\tfrac{1}{2} - x\big)\Delta\pi(x)\Big] - k^2\Delta\pi(x)^2,$$

and at the root, where $k^2(\frac{1}{2} - x^*)\Delta\pi(x^*) = -1$, the bracket equals -1, so

$$H'(x^*) = -\Delta\pi'(x^*) - k^2\Delta\pi(x^*)^2 = 1 - k^2(1 + x^*)^2 < 0$$

for every $k \ge \sqrt{2}$, since then $k^2(1 + x^*)^2 \ge k^2 > 1$. Hence $H'(x^*) < 0$ and x^* is attracting. For $k = 2$, $x^* = (\sqrt{5} - 1)/4 \approx 0.309$ and $H'(x^*) = 1 - 4(1.309)^2 \approx -5.85$. □

Remark 3. Because $\Delta\sigma = k\Delta\pi$ under proportional noise, $H(x)$ can be written as $\Delta\pi(x)\big[1 + k^2(\frac{1}{2} - x)\Delta\pi(x)\big]$, so the interior equilibria are exactly the solutions of $k^2(\frac{1}{2} - x)\Delta\pi(x) = -1$. This is the two-strategy case of the compound-parameter classification of Wang et al. (2026), which uses $x^* = (d - b)/(a - b - c + d)$ and $K = 1/[k^2(a - b - c + d)]$; for $(3, 0; 5, 1)$ we have $x^* = -1$ and $K = -1/k^2$, and the condition $k > \sqrt{2}$ corresponds to $|K| < 1/2$.

E THE MUTATION BUDGET $M_{\mathrm{crit}}(k)$

Proof of Proposition 5. Write $F(x) = s\,x(1 - x)H(x) + M(\frac{1}{2} - x)$, so that (7) is $\dot{x} = F(x)$. Since $s\,x(1 - x)H(x)$ vanishes at $x = 0$ and $x = 1$,

$$F(0) = \frac{M}{2} > 0, \qquad F(1) = -\frac{M}{2} < 0$$

for every $M > 0$ and every $k \ge 0$. By continuity F has at least one root in $(0, 1)$; the number of roots is odd, and since $F'(x_1) < 0$ at the first crossing the smallest root is attracting. (For $M = 0$ the boundary values are $F(0) = F(1) = 0$, and an interior root exists only under the conditions of Proposition 4; this is the degeneracy that the mutation term removes, and note that it is removed by *any* $M > 0$, because $F(0) = M/2$ already.)

For the sign of the drift at the boundary, expand there. Since $s\,x(1 - x)H(x) = s\,xH(0) + O(x^2)$ as $x \to 0^+$,

$$F(x) = \frac{M}{2} + x\big(sH(0) - M\big) + O(x^2), \qquad F'(0) = sH(0) - M.$$

Hence $F'(0) < 0$, i.e. the drift decreases through the boundary region, if and only if $M > sH(0)$. For the prisoner's dilemma of Proposition 4, $H(0) = k^2/2 - 1$, so

$$M > s\Big(\frac{k^2}{2} - 1\Big),$$

and taking the positive part (the bound is vacuous when the right-hand side is negative) gives Eq. (15). For $k \le \sqrt{2}$ any $M > 0$ therefore has $F'(0) < 0$; beyond $k = \sqrt{2}$ the boundary slope turns negative only at $M \approx sk^2/2$.

What M_{crit} is not. It is not a bifurcation. Solving $F = 0$ numerically on $[0, 1]$ with $s = 1$ for the prisoner's dilemma gives, across the threshold, an unchanged picture: for $k = 3$ ($M_{\mathrm{crit}} = 3.5$) the smallest root moves from 0.45594 at $M = 3.465$ to 0.45631 at $M = 3.535$, and for $k = 5$ ($M_{\mathrm{crit}} = 11.5$) from 0.48525 at $M = 11.385$ to 0.48539 at $M = 11.615$; the number of roots is one and the root is attracting on both sides in each case. Only the sign of $F'(0)$ changes. Nor is it needed for the boundary to be destabilised: $F(0) = M/2 > 0$ for every $M > 0$, so the boundary stops being an equilibrium as soon as any mutation is present, and the smallest root is macroscopic—above $x^*(k)$—for every $M > 0$ and every $k > \sqrt{2}$, since $F > 0$ on $(0, x^*)$ and $F(\frac{1}{2}) < 0$. What does change with M is where the attractor sits and how far it lies above the control floor, which is the object of Corollary 1. □

Remark 4. Removing the boundary degeneracy needs no condition—$F(0) = M/2 > 0$ for every $M > 0$—so the content of Eq. (15) is the sign of the boundary slope, $F'(0) = sH(0) - M$, and the k^2 growth of the mutation strength at which it turns negative. For the rock–paper–scissors experiments of §4 the analogous quantity is not available in closed form, and the mutation term there is a rare copy-flip rather than a controlled strength; the empirical pattern (the mutation strength required for convergence increases with the noise) is therefore consistent with, but does not test, the k^2 growth obtained here.

F Convergence

Proof of Proposition 6. Write the recursion as $x_{t+1} = \Pi\big(x_t + \alpha_t F(x_t) + \alpha_t \zeta_{t+1}\big)$, where Π is the projection onto $[0, 1]$ (NRMD keeps the simplex invariant, and in two strategies keeps $[0, 1]$ invariant, so the projection is inactive in the interior). Under (i)–(iii), F is locally Lipschitz with linear growth on the compact set, the martingale-difference noise has uniformly bounded second moment, and $\sum_t \alpha_t = \infty$, $\sum_t \alpha_t^2 < \infty$; the standard ODE method for stochastic approximation (Benaïm, 1999; Borkar, 2008) then applies: the interpolated process is an asymptotic pseudotrajectory of $\dot{x} = F(x)$, and its limit set is almost surely contained in an internally chain-transitive set of the flow.

In two strategies the state space is $[0, 1]$, the flow is one-dimensional and therefore has no cycles, and Propositions 4 and 5 give finitely many equilibria of which the only attractors are the interior roots, so x_t converges almost surely to one of them. We do not claim this for $n > 2$: the n-strategy flow may carry a limit cycle, and the limit set of Benaïm (1999) may then be that cycle rather than an equilibrium. §4 exhibits such a cycle in rock–paper–scissors, where the relevant statement is convergence of the time average to the cycle rather than to a point.

For the rate, let x^* be an isolated equilibrium with $F'(x^*) = -\lambda < 0$ and $\alpha_t = c/t$. The recursion is then a linearised Robbins–Monro scheme with decreasing gain; under $\lambda c > \frac{1}{2}$ the iterates satisfy $\mathbb{E}[\|x_t - x^*\|^2] = O(1/t)$ (Fabian's theorem and the Polyak–Ruppert averaging argument), so the mean-square rate is polynomial. An exponential rate is not implied: the continuous-time linearisation $x - x^* \propto e^{-\lambda t}$ governs the ODE, but the discrete scheme accumulates noise of order α_t at time t, whose variance is $O(\alpha_t)$, and the resulting deviation is $O(\sqrt{\alpha_t})$, i.e. $O(t^{-1/2})$ without averaging and $O(t^{-1})$ in mean square with averaging. Reporting an exponential rate for a decreasing-step-size scheme is a common error, and we avoid it. □

For the closed form of λ used in §3.6, $F'(x) = s(1 - 2x)H(x) + sx(1 - x)H'(x) - M$, and at the root $H(x^*) = 0$, giving $\lambda = M - sx^*(1 - x^*)H'(x^*) > 0$ because $H'(x^*) < 0$.

G CONSISTENCY OF THE TWO-STRATEGY AND n-STRATEGY FORMS

With $n = 2$ and $i = C$, $\bar{\sigma} = x\sigma_C + (1-x)\sigma_D$ gives $\sigma_C - \bar{\sigma} = (1-x)\Delta\sigma$, hence $\frac{1}{2}\big[(\sigma_C - \bar{\sigma})^2 - \mathrm{Var}(\sigma)\big] = \frac{1}{2}(1-x)^2(\Delta\sigma)^2 - \frac{1}{2}x(1-x)(\Delta\sigma)^2 = \frac{1}{2}(1-x)(1-2x)(\Delta\sigma)^2$, while $\pi_C - \bar{\pi} = (1-x)\Delta\pi$. Multiplying by x in Eq. (4) gives $x(1-x)[\Delta\pi + (\frac{1}{2} - x)(\Delta\sigma)^2]$, which is Eq. (3). The n-strategy and two-strategy formulations are therefore consistent.

H THE GENERAL TWO-STRATEGY CASE

For a general 2×2 game with proportional noise, $H(x) = \Delta\pi(x)\big[1 + k^2(\frac{1}{2} - x)\Delta\pi(x)\big]$ and $\Delta\pi$ is affine, $\Delta\pi(x) = x(a-c) + (1-x)(b-d)$. Two cases arise.

$\Delta\pi$ *is constant* (i.e. $a - c = b - d$): the game is degenerate and H has a single root, $\frac{1}{2} - x = -1/(k^2\Delta\pi)$, so an interior equilibrium exists iff $0 < \frac{1}{2} + 1/(k^2\Delta\pi) < 1$.

$\Delta\pi$ *is non-constant*: H is a cubic, and interior roots must be obtained numerically; the number of equilibria and their stability are the object classified by the compound parameters of Wang et al. (2026). In our 76,800-point scan exactly seven stability signatures occur (Appendix I), and no eighth signature appears, which is consistent with their classification. Deriving the stability signature in closed form as a function of (x^*, K) for the noisy replicator–mutator system, rather than for the noisy replicator system, is left open; the mutation term destroys the two-parameter structure, since M enters the drift independently of (x^*, K).

I PHASE-DIAGRAM DETAILS

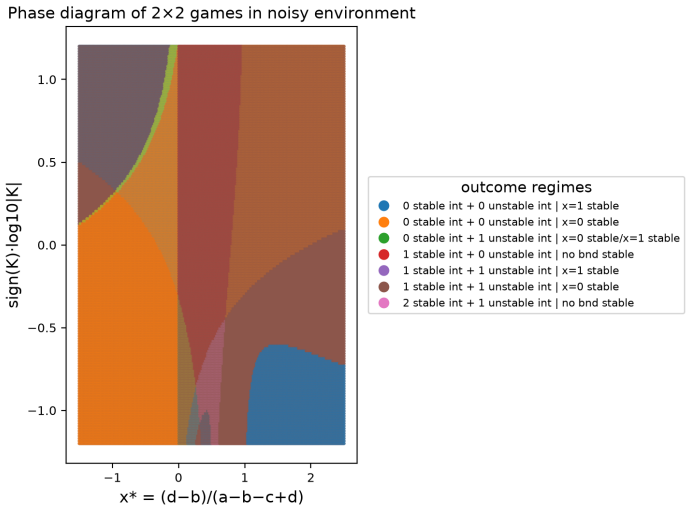

Figure 5: The 76,800-point phase diagram of Eq. (3), coloured by the stability signature of the equilibria. Exactly seven signatures occur, matching the seven regimes reported by Wang et al. (2026).

We solve Eq. (3) numerically on a grid of $x^* = (d-b)/(a-b-c+d) \in [-1.5, 2.5]$ (240 values, step ≈ 0.0167) and $\log_{10}|K| \in [-1.2, 1.2]$ (160 values for each sign of K), $K = 1/[k^2(a-b-c+d)]$, for a total of 76,800 parameter points. At each point the interior equilibria of the cubic drift are located by root finding and classified by the sign of the derivative, giving a stability signature (# stable interior equilibria, # unstable interior equilibria, whether $x = 0$ is stable, whether $x = 1$ is stable). Exactly seven signatures occur (Fig. 5), with frequencies 53.8%, 23.9%, 12.0%, 7.9%, 1.2%, 0.9% and 0.2%. The seven include signatures with two stable interior equilibria and with one stable interior equilibrium together with one stable boundary, neither of which is possible for a noise-free 2×2 game; the remaining four noise-free regimes are contained in the seven as special cases. This reproduces the classification of Wang et al. (2026); we report it as a reproduction, not as a contribution.

J THE TWO-STRATEGY (k, M) PLANE

The phase diagram of §4 is computed for the n-strategy form. For the two-strategy form (7) the drift $F(x) = s\,x(1-x)\big[\Delta\pi + (\frac{1}{2} - x)k^2\Delta\pi^2\big] + M(\frac{1}{2} - x)$ is a cubic, so its bifurcation structure can be computed exactly by root finding rather than by simulation. Figure 6 shows the result for the prisoner's dilemma $(3, 0; 5, 1)$. Two features are worth recording. First, the number of stable interior equilibria in the (k, M) plane is only ever 0 or 1: the bistable window of §4 requires the cyclic structure of a game with at least three strategies and does not occur here, so the two phenomena in the paper are genuinely different objects. Second, the mutation term moves the interior equilibrium monotonically upwards and flips the sign of the drift at the all-defection boundary once $M > s(k^2/2 - 1)$ (Proposition 5); below that value the equilibrium sits arbitrarily close to all-defection and is indistinguishable from it in a finite experiment, which is the same observability argument as Corollary 1 but read off the phase portrait directly.

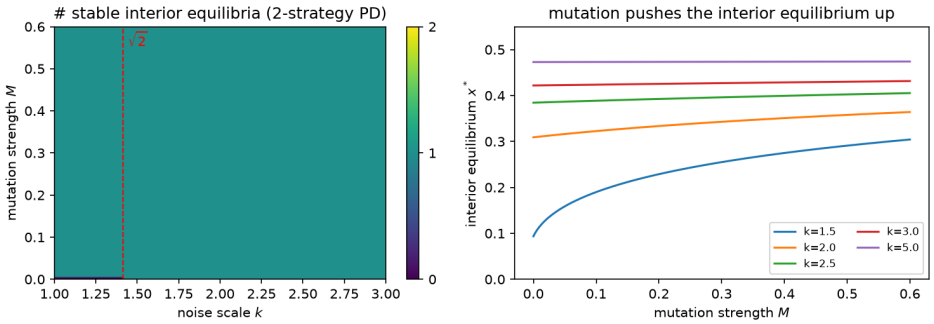

Figure 6: The two-strategy NRMD plane for the prisoner's dilemma $(3, 0; 5, 1)$, $s = 1$, computed by exact root finding of the cubic drift. Left: number of stable interior equilibria over (k, M); the red line is the threshold $k = \sqrt{2}$ of Proposition 4, below which no interior equilibrium exists for any M. Right: the interior equilibrium against M for several noise levels; mutation pushes it monotonically upwards, saturating near the mutation target $x = \frac{1}{2}$.

K LLM SOCIAL DILEMMA: FIGURE

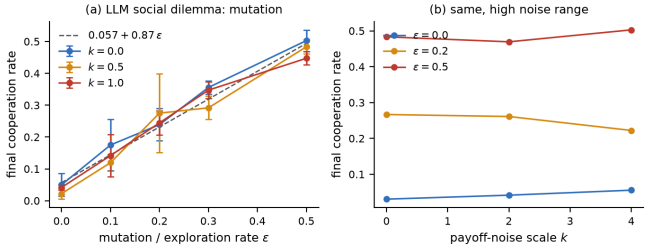

Figure 7: LLM social dilemma. (a) Final cooperation rate against the mutation/exploration rate ε for three noise levels, with the linear fit of Table 2; the dotted line at $\varepsilon = 0.5$ corresponds to the uniform mutation target. (b) The same agents over a fourfold larger noise range: the curves are flat, i.e. the noise dimension is not detectable with this population.

L LLM SOCIAL DILEMMA: COMPLETE TABLE

Table 2 gives the full cooperation-rate grid summarised in the text of §4.5.

M TABULAR MARL: COMPLETE TABLE

Table 4 gives the full grid summarised in §4.2; Fig. 3 shows two representative algorithms.

Table 2: LLM social dilemma: final cooperation rate after 20 rounds, mean over 3 seeds. Left: the mutation dimension, swept over low noise. Right: the same agents at high noise, where cooperation is essentially flat in k. The mutation dimension is clean and almost linear; the noise dimension is not detectable.

	cooperation rate, low noise			cooperation rate, high noise		
ε	k=0	k=0.5	k=1	k=0	k=2	k=4
0.0	0.050	0.022	0.042	0.031	0.042	0.056
0.1	0.175	0.119	0.142	—	—	—
0.2	0.239	0.275	0.244	0.267	0.261	0.222
0.3	0.356	0.292	0.347	—	—	—
0.5	0.503	0.483	0.447	0.483	0.469	0.503

Table 3: LLM social dilemma: the statistics behind Table 2. Top: mean over the three noise levels with a 95% confidence interval ($n = 9$ runs per value of ε). Bottom: ordinary least squares of the individual runs on (ε, k); the bound on the k-coefficient is set by the high-noise grid. The between-ε variance is $11.6\times$ the within-cell (seed) variance, which is why three seeds resolve the mutation dimension but cannot resolve a noise effect smaller than about 0.04.

ε	0.0	0.1	0.2	0.3	0.5
mean cooperation	0.038	0.145	0.253	0.331	0.478
95% CI	±0.018	±0.045	±0.055	±0.030	±0.026

grid	regressor	coefficient	t	R^2
low noise ($n = 45$, $k \le 1$)	ε	0.875 ± 0.043	20.5	0.91
	k	-0.020 ± 0.018	-1.1	
high noise ($n = 27$, $k \le 4$)	ε	0.877 ± 0.042	21.2	0.95
	k	0.000 ± 0.005	0.0	
all runs ($n = 72$)	ε	0.876 ± 0.030	29.4	0.93
	k	-0.002 ± 0.004	-0.4	

Component ablation (which ingredients are necessary). We decompose the NR-MBL update into four variants, each removing one ingredient, on $n = 3$ and $n = 5$ rock–paper–scissors (Table 5). *Continuous* replaces the discrete action sample by the expected payoff of the mixed strategy, i.e. it removes the discrete offspring jump that Proposition 3 requires for a learner to inherit the noise-induced equilibrium; *no-mutation* sets $M = 0$; *sequential* updates the two populations one after the other instead of simultaneously. The result is a clean separation of the necessary ingredients. First, *continuous* is flat: the distance to Nash sits at 0.753–0.758 ($n = 3$) and 0.823–0.827 ($n = 5$) with no dependence on k whatsoever, so removing the discrete jump destroys the dose–response entirely—this is the sharpest evidence that only discrete-inheritance learners enjoy the mechanism, the boundary identified in §3. Second, *no-mutation* ($M = 0$) never converges at any k, its distance staying above 0.62, confirming that the mutation term is necessary. Third, *sequential* is statistically indistinguishable from *baseline*, so the update order is not a relevant ingredient. Finally *baseline* reproduces the predicted non-monotone shape: distance falls from $k = 0$ to a minimum at $k \approx 1$ (0.050 for $n = 3$, 0.052 for $n = 5$) and rises again at $k = 4$, the intermediate noise being beneficial and strong noise harmful.

N THE PRPG ALGORITHM

Algorithm 1 states PRPG, the population learner used as the `repl` arm of §4.4. It is Eq. (7) carried by a finite population of P agents: the reproduction step is fitness-proportional (the replicator part, whose exponential-in-payoff weight is what produces the Jensen drift), and the offspring inherit the parent's heritable parameter up to a rare flip (the mutation part). Only the scalar bias θ_i is heritable, and it is a deterministic function of a discrete type, so the heritable axis is exactly the "defection bias" of §4.4—a finite-population analogue of the two-strategy coordinate x of §3. Both ingredients are necessary: setting $\mu = 0$ deletes the mutation term and the population does not converge, while replacing the reproduction step by a gradient step on the *same* θ leaves the noise to cancel at first order, which is exactly the `grad` arm of Table 1 (Proposition 1). The two-strategy reduction of

Table 4: Tabular MARL (rock–paper–scissors, $n = 3$): L_1 distance from the uniform mixture after training, and success rate, for 32 seeds per cell. The mutation strength needed grows with the noise: $M = 0.2$ survives $k \leq 2$ but fails at $k = 4$, while $M = 1.0$ is almost flat. Q-learning degrades monotonically and cross learning does not converge at all. Lower is better.

	distance to Nash					success (fraction of seeds)
algorithm	k=0	k=0.5	k=1	k=2	k=4	k=0 → 4
cross (M=0)	0.699	0.625	0.620	0.816	0.816	0.00 → 0.00
Q-learning (softmax)	0.061	0.056	0.063	0.088	0.187	0.50 → 0.00
NR-MBL (M=0.02)	0.576	0.495	0.391	0.716	0.774	0.00 → 0.00
NR-MBL (M=0.05)	0.532	0.451	0.147	0.542	0.732	0.00 → 0.00
NR-MBL (M=0.20)	0.071	0.052	0.050	0.071	0.528	0.16 → 0.00
NR-MBL (M=1.0)	**0.026**	**0.027**	**0.027**	**0.029**	**0.047**	**0.97 → 0.56**

Table 5: Component ablation of the tabular NR-MBL update (rock–paper–scissors, $n = 3$ and $n = 5$): L_1 distance to the uniform Nash after training, 4 seeds per cell, $M = 0.2$ unless noted. *continuous* removes the discrete offspring jump, *no-mutation* sets $M = 0$, *sequential* updates populations in turn. Removing either the jump or the mutation destroys convergence; the update order does not matter.

	distance to Nash				
variant	k=0	k=0.5	k=1	k=2	k=4
$n = 3$					
baseline	0.079	0.056	**0.050**	0.070	0.529
continuous	0.753	0.753	0.754	0.755	0.758
no-mutation	0.713	0.653	0.625	0.816	0.816
sequential	0.074	0.048	0.048	0.070	0.530
$n = 5$					
baseline	0.063	0.053	**0.052**	0.064	0.648
continuous	0.823	0.823	0.823	0.824	0.827
no-mutation	0.795	0.781	0.732	0.874	0.894
sequential	0.065	0.053	0.051	0.064	0.642

the same update, with the type dimension as its only heritable axis, is also released and reproduces $x^*(k)$ across k (`prpg_validate.py`, §O).

Algorithm 1 Population-Replicator Policy Gradient (PRPG)

Require: P agents; G generations; horizon H; type space $\{C, D\}$; mutation rate μ; selection strength s; noise scale k; bias B

initialise $\tau_i \sim \text{Bernoulli}(1/2)$ for $i = 1, \ldots, P$ ▷ 1 = cooperate
for $g = 1$ to G **do**
  $\theta_i \leftarrow B\,(1 - 2\tau_i)$ for all i ▷ policy parameter, $|\theta_i| \leq B$
  $f_i \leftarrow \sum_{t=1}^{H} r_i(t)$ ▷ fitness of agent i in the shared environment
  $\tilde{f} \leftarrow (f - \bar{f})\,/\,\text{sd}(f)$ ▷ z-scored fitness
  draw one $\xi \sim \mathcal{N}(0, 1)$ shared by the population, and set $\pi_i \leftarrow \tilde{f}_i\,(1 + k\,\xi/\sqrt{s})$ ▷ the shock of Eq. (7)
  $w \leftarrow \text{softmax}(s\,\pi)$; sample parent $p \sim w$, victim $v \sim \text{Unif}(1, \ldots, P)$
  $\tau_v \leftarrow \tau_p$ ▷ reproduction: copy the heritable type
  if $\text{Unif}(0, 1) < \mu$ **then**
    $\tau_v \leftarrow 1 - \tau_v$ ▷ mutation: rare flip
  end if
end for
return tail-mean cooperation $\hat{x} = \frac{1}{P} \sum_{i=1}^{P} \mathbf{1}[\tau_i = C]$

Algorithm 1 is a *population* update, not a gradient update: no agent follows the gradient of its own payoff. That is the whole point of §4.4—two learners share one environment, one shock and one policy set, and the only difference is whether the reproduction is exponential in the payoff (Algorithm 1) or linear in it (REINFORCE on the same θ). The exponential makes the fitness–noise interaction survive the population average, while the linear update cancels it at first order.

O DATA, HYPER-PARAMETERS AND REPRODUCTION COMMANDS

All experiments are driven by released scripts. The ODE experiments (§4, Appendix I) use `phase1.py` (NumPy/SciPy, `solve_ivp` with LSODA, `rtol`=10^{-8}, `atol`=10^{-10} for the two-by-two trajectories and `rtol`=10^{-9}, `atol`=10^{-12} for the rock–paper–scissors runs); the tabular benchmark uses `phase2_mp.py` with 8 worker processes; the deep experiments use `E4_deep_marl.py` (standard PPO-style control) and `E4b_nr_l.py` (Eq. (17)); the LLM experiments use `E5_llm_social.py` against a local vLLM server. Aggregation and all figures are produced by `paper_figs.py`, which prints the tables used above.

- *Phase diagram.* `python phase1.py`; 76,800 points, $\approx$ 3 min on one CPU core.
- *Subcritical transition.* Same script, function `plot_rps_limit_cycle`; $\alpha = 2$, $\beta = 1$, $k = 1$, $M \in [0, 0.6]$ in 31 steps, integrated to $t = 1000$; the script uses 6 initial conditions per value of M, whereas the scan plotted here uses 60 (one fixed at $(0.2, 0.5, 0.3)$ plus 59 Dirichlet draws from `default_rng(0)`) and is reproduced end to end by the self-contained `fig4_subcritical_60ic.py`; convergence is recorded when the tail oscillation amplitude $|x_1|_{\max} - |x_1|_{\min}$ over the last 20% of the trajectory falls below 10^{-3}.
- *Tabular MARL.* `python phase2_mp.py`; games rock–paper–scissors with $n \in \{3, 5\}$; algorithms cross learning ($M = 0$), softmax Q-learning and NR-MBL with $M \in \{0.02, 0.05, 0.2, 1.0\}$; noise $k \in \{0, 0.5, 1, 2, 4\}$; 32 seeds per cell; step size 2×10^{-4} for cross learning, $\alpha_Q = 0.01$ and temperature schedule $\tau \downarrow$ for Q-learning; a seed counts as converged if the L_1 distance from the uniform mixture stays below 0.05 over the final 10% of training.
- *Theory-aligned deep learner.* `python E4b_nr_l.py --mut M --noise k --seed s --steps 20000`; four agents, pairwise rock–paper–scissors, $M \in \{0, 0.05, 0.2, 0.5\}$, $k \in \{0, 1, 3\}$, 3 seeds per cell (36 runs); the reported entropy and distance are averaged over the final 1000 steps.
- *Standard deep MARL (control).* `python E4_deep_marl.py --variant V --noise k --seed s`; a cooperative coverage task; $V \in$ {base, entropy, risk, nrmd}, $k \in \{0, 0.5, 1, 2\}$, 3 seeds.
- *Deep PRPG (Algorithm 1) and its negative control.* `python aggr_pd.py --arms repl,grad --th_clips 3 --ks 0,5 --seeds 0,...,19 --gens 12000 --fscale z`; $P = 64$, horizon 20, the same explicit N-player iterated prisoner's dilemma as §4.4; the reference implementation of Algorithm 1 ships as `code/cloud/aggr_pd.py`. The identical command with `--th_clips 12` is the set-escaped variant of Table 1.
- *LLM social dilemma.* `python E5_llm_social.py --game ipd --n 6 --rounds 20 --temp 0.7 --epsilon e --noise k --seed s`; Qwen2.5-7B-Instruct served by vLLM; $\varepsilon \in \{0, 0.1, 0.2, 0.3, 0.5\}$, $k \in \{0, 0.5, 1\}$ (45 runs) and a high-noise sweep $\varepsilon \in \{0, 0.2, 0.5\}$, $k \in \{0, 2, 4\}$ (27 runs); three seeds each. The reported cooperation rate is the fraction of cooperative actions in the final round.
- *PRPG desktop, seed hardening.* The bin arm was re-run with 20 seeds (was 5), P=64, 120,000 steps, $\mu = 0.02$, z-scored fitness: floor $x(k{=}0) = 0.0572 \pm 0.0055$, lift $x(5) - $floor $= 0.349$, monotone in k, correlation with $x^*(k)$ above the floor 0.951, bootstrap 95% CI ± 0.005. See `results/pd/prpg_bin_seed20.json`.
- *Aggregation.* `python paper_figs.py` reproduces every number in Tables 4 and 2 and Figures 4–7 from the raw CSV files.