Evolutionary Game Theory in Machine Learning and AI: Naturality, Regret, and Recurrence

Anonymous

Higher Categorical Semantics · Version 1.0, 17 November 2026

Abstract. The replicator equation of evolutionary game theory and the standard learning rules of machine learning are the same objects under different names: multiplicative weights is the replicator flow sampled in time, natural policy gradient is the replicator field, and Bayesian updating is the replicator map with log-likelihood as payoff. We organize these identifications around one structure. A learner is an open dynamical system with payoffs as input and a mixed strategy as output; a game is a wiring of learners; and a learning rule is *natural* if its closed-loop dynamics commutes with payoff-preserving maps of games, in particular with the duplication of actions. Each qualitative statement is paired with a quantitative one. Naturality: the replicator learner is natural, while softmax policy gradient, entropy regularization, score averaging and a mutation-chain ranking are not, with computed defects; duplicating an action k times multiplies the policy-gradient learning time by exactly $2k/(k+1)$. Regret: at most $\ln n$ in continuous time for every horizon, against $\sqrt{T \ln n/2}$ in discrete time. Recurrence: in zero-sum games the relative entropy to equilibrium is conserved by the flow, increases by exactly $\log \mathbb{E}e^{\eta f} \approx \eta^2 \operatorname{Var}(f)/2$ per step of multiplicative weights, and decays at rate at least α under entropy-regularized dynamics. Finite populations of learners obey the one-third law and the amplifier and suppressor phenomena of evolutionary graph theory. All numerical statements are recomputed by a companion script with 21 checks. No new theorem of learning theory is claimed; the contribution is the organization, the exact constants, and the naturality criterion.

1 Introduction

Evolutionary game theory studies populations in which strategies that earn more become more common. Its central object, the replicator equation

$$\dot{x}_i = x_i\big((Ax)_i - x^{\mathsf{T}}Ax\big), \tag{1}$$

was written down for animal conflicts [TaJo78, MSP73]. Machine learning arrived at the same equation several times: as the continuous-time limit of reinforcement learning [BoSa97] and of multiplicative weights [FrSc97, Sor09], as natural policy gradient on a softmax policy [Kak01], and implicitly in population-based training, where agents copy their betters [Jad17]. A compact account of the evolutionary side, written for philosophers and candid about the fragility of the models, is Alexander's Element [Alex23]; a companion volume [EGHC] retells it in categorical language. This paper carries the results of that volume to machine learning.

The organizing idea is to separate what the evolutionary literature fuses. In (1) a learning rule and a game are multiplied together. We take them apart (Section 2): a *learner* is an open dynamical system, a box with payoffs flowing in and a mixed strategy flowing out, and a *game* is a wiring of such boxes. Self-play, two-player training, leagues and populations on graphs are different diagrams filled with the same boxes. The question that the categorical setting makes precise is which learners are *natural*: whose behaviour does not depend on how the actions happen to be named, and in particular on whether one action has been listed twice.

Results. Each is stated with its status. Tags follow the conventions of the series: [classical] a classical result; [literature; source] a result of the cited literature; [reformulation; cf. source] a known result restated; [observation] an observation of this paper, proved here. Tags such as [B6] name checks in the companion script `verify-evolutionary-learning.py`.

- *Naturality* (Sections 2 and 5). The replicator learner is natural for all payoff-preserving maps. Softmax policy gradient is not natural for duplication, and the slowdown is the exact factor $2k/(k+1)$ (Theorem 2.6). Entropy regularization and two common ways of ranking agents fail the same test (Proposition 5.3).
- *Regret* (Section 3). An identity gives regret at most $\ln n$ in continuous time; the term that discretization adds is the one that reappears as divergence in games.
- *Recurrence* (Section 4). In zero-sum games

continuous-time learning cycles, discrete-time learning spirals out at a computable rate, and three repairs act at three different places of the wiring diagram.

- *Populations* (Section 6). Fixation probabilities, the one-third law, and the effect of the copying topology.

Table 1 lists the pairs of qualitative and quantitative statements. Section 9 states what is not claimed.

Related work. The link between reinforcement learning and replicator dynamics goes back to [BoSa97] and is surveyed in [BTHK15]; the behaviour of regularized learning in zero-sum games is the subject of [PiSh14, MPP18, BaPi18]; evolutionary methods for evaluating agents are developed in [BTPG18, OPP19]. On the categorical side, learners as open systems composed along wiring diagrams follow [SSV20, Mye21], and compositional accounts of gradient-based learning are given in [FST19, CGG22]. What we add to both is the question of naturality with respect to maps of games, which neither literature asks in this form.

2 Learners, games, and naturality

2.1 Games and their maps

A ***symmetric game*** (S, A) is a finite set S of actions and a matrix $A \in \mathbb{R}^{S\times S}$; A_{ij} is the payoff of i against j. Mixed strategies form the simplex $\Delta(S)$. A ***payoff-preserving map*** $f\colon (S, A) \to (T, B)$ is a function with $B_{f(i)f(j)} = A_{ij}$. These are the morphisms of a category Game [EGHC]. They are generated by three kinds: *inclusions* of a subset of actions, *clone maps*, which are surjections identifying actions with equal rows and columns, and *symmetries*. A map f induces the pushforward $f_*\colon \Delta(S) \to \Delta(T)$, which adds the weights of identified actions. The basic fact is that f_* preserves payoffs: $(Bf_*x)_{f(i)} = (Ax)_i$ and $(f_*x)^{\mathsf{T}}B(f_*x) = x^{\mathsf{T}}Ax$.

2.2 Learners

Definition 2.1. A ***learner*** on S is an open dynamical system with input a payoff vector $\pi \in \mathbb{R}^S$, a state z, a readout $x = r(z) \in \Delta(S)$, and an update: $\dot z = V(z, \pi)$ in continuous time or $z^+ = U(z, \pi)$ in discrete time.

Five learners recur below. All read out by softmax, applied to Q/τ in the last case.

learner	state	update
replicator (R)	scores z	$\dot z = \pi$
Hedge, MWU (H)	scores z	$z^+ = z + \eta\pi$
optimistic (O)	scores z	$z^+ = z + \eta(2\pi_t - \pi_{t-1})$
policy grad. (PG)	logits θ	$\dot\theta_i = x_i(\pi_i - x\cdot\pi)$
Boltzmann Q (Q)	values Q	$\dot Q = \alpha(\pi - Q)$

(PG) is gradient ascent on expected payoff $x \cdot \pi$ in the logits; (Q) is the expected, all-actions form of Q-learning with Boltzmann exploration [TVL03, SaCr03].

Proposition 2.2 (Hedge is the replicator, sampled). [classical] *The replicator learner fed a payoff stream $\pi(\cdot)$ has*

$$x(t) = \operatorname{softmax}\Big(\log x(0) + \textstyle\int_0^t \pi\Big),$$

and in the variable x it obeys $\dot x_i = x_i(\pi_i - x\cdot\pi)$. One step of Hedge with step size η is the time-η map of this flow under a payoff held constant over the step. [B1]

Proposition 2.3 (Policy gradient and its natural form). [classical; [Kak01, Ama98]] *For $x = \operatorname{softmax}(\theta)$, $\partial(x\cdot\pi)/\partial\theta_i = x_i(\pi_i - x\cdot\pi)$. Multiplying this gradient by the inverse of the Fisher matrix $\operatorname{diag}(x) - xx^{\mathsf{T}}$, on the tangent space, gives $\pi - (x\cdot\pi)\mathbf{1}$: natural policy gradient on a softmax policy is the replicator learner.* [B2]

So (PG) differs from (R) by one factor of x_i in the update of the logits. Neural Replicator Dynamics makes exactly this change in a deep policy-gradient method, bypassing the gradient through the softmax, and reduces to Hedge in the single-state tabular case [HMO20]. In the variable x,

$$\text{(PG)}\qquad \dot x_i = x_i\Big(x_i(\pi_i - \bar\pi) - \sum_j x_j^2(\pi_j - \bar\pi)\Big), \quad (2)$$

with $\bar\pi = x\cdot\pi$. The extra factor has a cost that is easy to state.

Proposition 2.4 (Escape from a bad start). [classical; cf. [MXSS20]] *Two actions, payoff gap $D > 0$, the better action at initial probability ϵ. The time for it to reach $\frac12$ is*

$$t_{\mathrm R} = \frac{1}{D}\log\frac{1-\epsilon}{\epsilon},$$
$$t_{\mathrm{PG}} = \frac{1}{2D}\Big(\frac{1}{\epsilon} - \frac{1}{1-\epsilon} + 2\log\frac{1-\epsilon}{\epsilon}\Big).$$

Proof. With x the probability of the better action, (R) is $\dot x = Dx(1-x)$ and, by (2), (PG) is $\dot x = 2Dx^2(1-x)^2$. Integrate. □

For $D = 1$ and $\epsilon = 0.1$, 0.01, 0.001 the two times are 2.20 and 6.64, 4.60 and 54.1, 6.91 and 506: ratios 3.0, 11.8 and 73.3. [B3] The replicator learner recovers from a bad initialization in logarithmic time and policy gradient in time of order $1/\epsilon$. In evolutionary terms (PG) treats a rare good action as the drift game treats a neutral mutant [EGHC].

2.3 Closing the loop

A learner becomes a dynamical system when its input is supplied. ***Self-play*** in a symmetric game feeds back $\pi = A\,r(z)$; for (R) this is (1). Two learners in a bimatrix game (A, B) are wired crosswise, $\pi^1 = A\,r(z^2)$

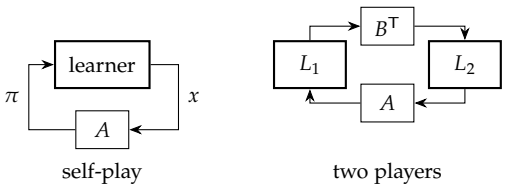


Figure 1: A game is a wiring of learners. The boxes can be changed without changing the diagram, and the diagram without changing the boxes.

and $\pi^2 = B^\mathsf{T} r(z^1)$. A league or a population on a graph is a larger diagram. Formally, open dynamical systems are algebras for an operad of wiring diagrams [SSV20, Mye21], and a multi-agent learning problem is a diagram together with a choice of box for each node (Figure 1).

2.4 Naturality

Definition 2.5. A family of learners, one for each finite set S, is ***natural*** for a class of payoff-preserving maps if for every $f\colon (S, A) \to (T, B)$ in the class, f_* carries self-play trajectories in $\Delta(S)$ to self-play trajectories in $\Delta(T)$.

Naturality for inclusions says that an action never tried is never missed, and the rest behaves as if it did not exist. Naturality for clone maps says that listing an action twice, or splitting it into indistinguishable variants, changes nothing about the total weight of the variants. Naturality for symmetries is equivariance.

Theorem 2.6 (Which learners are natural). [observation]

(a) (R) *and* (H) *are natural for all payoff-preserving maps.*

(b) (PG) *is natural for inclusions and symmetries and not for clone maps. If an action of total probability y is split into copies with probabilities $x_1, \dots, x_k$, the leading term of its drift is $\sum_c x_c^2$ where the unsplit learner has y^2.*

(c) In the two-action problem of Proposition 2.4, with either action split into k equal copies, the probability y of the better class obeys $\dot y = (1 + \frac{1}{k}) D y^2 (1 - y)^2$. Every learning time is multiplied by exactly $2k/(k+1)$.

(d) (Q) *is natural for symmetries and not for clone maps. Its rest points are the logit equilibria $x = \mathrm{softmax}(Ax/\tau)$, which depend on the number of copies of each action, and every face of the simplex is repelling: no action stays unused.*

Proof. (a) For (R), with $y = f_* x$, the pushforward identity gives $\dot y_t = \sum_{f(i)=t} x_i((Ax)_i - x^\mathsf{T} Ax) = y_t((By)_t - y^\mathsf{T} By)$ for t in the image and $y_t \equiv 0$ otherwise. (H) is a time-η map of the same flow with frozen payoffs, and frozen payoffs are also preserved. (b) In (2) the factor x_i gives invariance of faces. Summing (2) over the copies of one action gives $(\sum_c x_c^2)(\pi - \bar\pi) - y \sum_j x_j^2 (\pi_j - \bar\pi)$, and $\sum_c x_c^2 < y^2$ unless all but one copy is empty. (c) By symmetry the copies stay equal. Substituting $x_c = y/k$ for the split action in (2) and summing gives the stated equation in both cases; compare with $2Dy^2(1-y)^2$. (d) In the variable x, (Q) in self-play is

$$\dot x_i = x_i \Big(\tfrac{\alpha}{\tau} (\pi_i - \bar\pi) - \alpha \big(\log x_i - \textstyle\sum_j x_j \log x_j \big) \Big), \qquad (3)$$

whose rest points satisfy $\log x_i = \pi_i/\tau + \text{const}$. The entropy term makes $\dot x_i / x_i \to +\infty$ as $x_i \to 0$, and it distinguishes one action of weight y from two of weight $y/2$, since $2 \cdot \frac{y}{2} \log \frac{y}{2} \neq y \log y$. □

Computed defects. For Rock–Paper–Scissors with Rock duplicated, the largest discrepancy over fifty random states between the pushed-forward field and the field of the three-action game was 4×10^{-17} for (R) and 0.028 for (PG). [B19] From $\epsilon = 0.01$ the policy-gradient time to $\frac{1}{2}$ was 54.1, 72.1, 86.5 and 96.2 for $k = 1, 2, 4, 8$, and the replicator time 4.60 for every k. [B20] For (Q) the logit equilibrium of Rock–Paper–Scissors is uniform at every temperature; with Rock duplicated its mass on the three types is $(0.49, 0.29, 0.23)$ at $\tau = 2$, $(0.40, 0.36, 0.24)$ at $\tau = 0.5$ and $(0.34, 0.35, 0.31)$ at $\tau = 0.1$. [B21]

Remark 2.7. Large action spaces in practice contain near-duplicates: synonymous tokens, equivalent moves, redundant members of a league. Theorem 2.6 says which update rules are indifferent to them. It is the finite-dimensional shadow of the invariance of natural gradient under reparametrization [Ama98], stated for the maps that matter in games. Entropy bonuses are the clearest case of a method that counts names: at high temperature the regularized equilibrium gives duplicated Rock half the mass.

3 Regret

A learner has ***no regret*** if, against any stream of payoffs, its average payoff is asymptotically at least that of the best fixed action. For the replicator learner this is an identity.

Theorem 3.1 (Continuous time). [classical; [Sor09, KwMe17]] *For the replicator learner and any bounded measurable payoff stream,*

$$\int_0^T (\pi_i - x \cdot \pi)\, dt = \log \frac{x_i(T)}{x_i(0)} \le \log \frac{1}{x_i(0)}$$

for every action i and every horizon T. From the uniform initial state the regret is at most $\ln n$.

Proof. $\frac{d}{dt} \log x_i = \pi_i - x \cdot \pi$ by Proposition 2.2, and $x_i(T) \le 1$. □

The bound does not grow with T. Against a stream of six payoffs that alternates between random vectors

and vectors chosen to punish the current favourite, the regret at $T = 80$ was 0.24, against $\ln 6 = 1.79$, and the identity held to 5×10^{-8}. [B4]

Theorem 3.2 (Discrete time). [classical; [FrSc97, CBL06]] *For Hedge with payoffs in $[0,1]$ and $\eta = \sqrt{8 \ln n / T}$, the regret after T steps is at most $\sqrt{T \ln n / 2}$.*

Sketch. One step multiplies the normalizing constant by $\mathbb{E}_x e^{\eta\pi}$, and Hoeffding's lemma gives $\log \mathbb{E}_x e^{\eta(\pi - x\cdot\pi)} \le \eta^2/8$. Summing, the discrete learner satisfies the identity of Theorem 3.1 up to an excess of at most $\eta/8$ per step: the regret is at most $\ln n/\eta + \eta T/8$. □

Over forty runs with $n \in \{2, 5, 10\}$, $T \in \{200, 1000\}$, half against random payoffs and half against an adversary, the largest ratio of regret to bound was 0.50. [B5] The whole difference between $\ln n$ and $\sqrt{T}$ is the term

$$\Lambda_\eta(x, \pi) = \log \mathbb{E}_x \, e^{\eta(\pi - x\cdot\pi)} = \tfrac{1}{2}\eta^2 \operatorname{Var}_x(\pi) + O(\eta^3), \tag{4}$$

the price of holding the payoff fixed for a step. The same term decides the fate of discrete learning in games.

Corollary 3.3. [classical] *If both players of a zero-sum game with payoffs in an interval of length 1 use Hedge with the step of Theorem 3.2, their time-averaged strategies form a $2\sqrt{\ln n/(2T)}$-equilibrium. In a general game the time-averaged joint play approaches the set of coarse correlated equilibria.*

In self-play of Rock–Paper–Scissors, whose payoffs have range 2, the time average over $T = 20000$ steps could be exploited for 0.0015, against the corresponding bound 0.0105. [B8]

4 Zero-sum games: recurrence, divergence, repair

Regret bounds speak of averages. What the strategies themselves do is another matter, and it is the matter that concerns adversarial training [Goo14], where the iterate and not the average is the model.

4.1 The flow cycles

Theorem 4.1 (Conservation and recurrence). [literature; [HoSi98, PiSh14, MPP18]] *Let A be antisymmetric with an interior equilibrium x^*. Along replicator self-play $D(x^* \,\|\, x(t))$ is constant. For two replicator learners in a zero-sum game with interior equilibrium (x^*, y^*) the sum $D(x^* \,\|\, x) + D(y^* \,\|\, y)$ is constant, and the flow is Poincaré recurrent: almost every trajectory returns arbitrarily close to its starting point infinitely often.*

Proof of conservation. $\frac{d}{dt} D(x^* \,\|\, x) = -\sum_i x_i^*(\pi_i - \bar\pi) = -(x^*)^{\mathsf{T}} A x + x^{\mathsf{T}} A x$. Antisymmetry gives $x^{\mathsf{T}} A x = 0$, and $(x^*)^{\mathsf{T}} A x = -x^{\mathsf{T}} A x^* = 0$ because Ax^* is constant on the support of an interior equilibrium and that constant is the value 0. □

Conservation says that such a game does no work on the learner. An antisymmetric matrix has no symmetric part, so the gradient component of the replicator field vanishes [EGHC]; and when the equilibrium is uniform the game is purely cyclic in the sense of the Hodge decomposition of Section 5.

4.2 The discretization spirals out

Proposition 4.2 (Exact drift of Hedge). [reformulation; cf. [BaPi18]] *In Hedge self-play of an antisymmetric game with interior equilibrium x^*, each step changes the relative entropy by exactly*

$$D(x^* \,\|\, x^+) - D(x^* \,\|\, x) = \Lambda_\eta(x, Ax) \;\ge 0,$$

with equality only if all actions in use earn the same. Hence $D(x^* \,\|\, x_t)$ is non-decreasing, and it increases by about $\frac{1}{2}\eta^2 \operatorname{Var}_x(Ax)$ per step.

Proof. $x_i^+ = x_i e^{\eta\pi_i}/Z$ with $Z = \mathbb{E}_x e^{\eta\pi}$, so the difference is $-\eta\, x^*\cdot\pi + \log Z$. As in Theorem 4.1, $x^*\cdot\pi = 0 = x\cdot\pi$, so the difference is $\log Z = \Lambda_\eta$, and Jensen's inequality gives the sign. □

At the state $(0.5, 0.3, 0.2)$ of Rock–Paper–Scissors the increase per step was 5.01×10^{-5}, 2.01×10^{-4} and 8.08×10^{-4} for $\eta = 0.05$, 0.1, 0.2, against $\frac{1}{2}\eta^2 \operatorname{Var} = 5.00 \times 10^{-5}$, 2.00×10^{-4}, 8.00×10^{-4}. [B6] With $\eta = 0.1$ the relative entropy increased at every step and some action fell below probability 0.01 after 1451 steps. [B7] The quantity (4) that costs a learner $\sqrt{T}$ in regret is, in a game, energy pumped into the cycle at every step. The time average still converges (Corollary 3.3); the iterate leaves.

4.3 Three repairs, in three places

Proposition 4.3 (Optimism). [literature; [RaSr13, DISZ18, DaPa19]] *The optimistic learner* (O), *which uses $2\pi_t - \pi_{t-1}$ as its prediction of the next payoff, converges in the last iterate in zero-sum games with a unique equilibrium, for small enough η.*

From the same initial state and $\eta = 0.1$ the relative entropy to equilibrium fell from 0.070 to 1.5×10^{-10} in 6000 steps. [B9]

Theorem 4.4 (Entropy regularization, with its rate). [observation] *Let A be antisymmetric with the uniform state u as equilibrium. Under the dynamics (3), for every $\tau > 0$,*

$$\frac{d}{dt} D(u \,\|\, x) = -\alpha\big(D(u \,\|\, x) + D(x \,\|\, u)\big),$$

and therefore $D(u \,\|\, x(t)) \le e^{-\alpha t} D(u \,\|\, x(0))$.

Proof. $\frac{d}{dt} D(u \,\|\, x) = -\sum_i u_i \dot{x}_i / x_i$. The payoff terms vanish as in Theorem 4.1. The entropy terms give $\alpha\big(\sum_i u_i \log x_i - \sum_j x_j \log x_j\big)$, and $D(u \,\|\, x) + D(x \,\|\, u) = \sum_j x_j \log x_j - \sum_i u_i \log x_i$. □

Convergence of Q-learning dynamics in competitive games is known in much greater generality [LPS21]; the point of Theorem 4.4 is the exact identity, in which the rate is the learning rate α and does not depend on the game or the temperature. With $\alpha = 0.7$ the relative entropy fell from 0.070 to 6.1×10^{-8} at $t = 10$, against the bound 6.4×10^{-5}. [B10] In general the limit is a logit equilibrium and not a Nash equilibrium [B11], and by Theorem 2.6(d) it is not natural.

Remark 4.5 (Where each repair acts). Averaging changes the *readout*: the box and the wiring are untouched and one reports $\frac{1}{T}\sum x_t$. Optimism changes the *box*: the learner carries one more step of memory. Regularization changes the *wiring*: (3) is the replicator learner in the game with the additional self-payoff $-\tau \log x_i$, a loop from each learner's output to its own input, up to a constant change of time scale. In evolutionary terms these are time-averaging of frequencies, anticipation, and mutation [TVL03]. The trade is visible in the table of Section 8: averaging keeps naturality and loses the iterate; regularization gains a rate and loses both naturality and the equilibrium.

5 Evaluating agents

Training by self-play, leagues and population-based methods all need to say which of a set of agents is better [Vin19, LZG17]. The data is an ***evaluation matrix***: an antisymmetric K with K_{ij} the advantage of agent i over agent j, for instance the log-odds of winning. This is a symmetric zero-sum game on the set of agents, and an ***evaluation method*** assigns to it a score vector or a distribution on agents.

Theorem 5.1 (Ratings and cycles). [literature; [JLYY11, BTPG18]] *Every antisymmetric K decomposes uniquely and orthogonally as*

$$K_{ij} = (r_i - r_j) + C_{ij}, \quad r_i = \tfrac{1}{n}\textstyle\sum_j K_{ij}, \quad \sum_j C_{ij} = 0.$$

The first part satisfies $G_{ij} + G_{jk} + G_{ki} = 0$ on every triple, and every matrix satisfying this is of that form. The second part carries the same curl $K_{ij} + K_{jk} + K_{ki}$ as K.

The ratings r are the least-squares fit of an Elo-like scale; the ***transitivity index*** $\|G\|^2/\|K\|^2$ measures how much of the data a scale can explain. For five agents with planted ratings and one planted cycle of strength 0.5 among three of them, the decomposition returned the ratings $(0.7, 0.3, 0, -0.3, -0.7)$, index 0.885 and curl 1.5 on the planted triple. [B12] That closed implies exact is the vanishing of the first cohomology of a simplex.

Proposition 5.2 (Incomplete comparisons). [literature; [JLYY11]] *If only some pairs of agents have been compared, the decomposition acquires a third, harmonic component: circulations around loops of the comparison graph that bound no triple of mutually compared agents. It is locally consistent and globally not a rating, and its dimension is the first Betti number of the comparison complex.*

Four agents compared only around a square, each beating the next by one unit, give the smallest example: zero divergence, no triangles, nothing to fit. [B15] Sparse tournament schedules create such classes by design.

The naturality criterion of Section 2 applies to evaluation unchanged. A league that contains two copies of an agent, or ten near-copies obtained from one checkpoint, should not be ranked differently for that reason.

Proposition 5.3 (Naturality of evaluation methods). [observation] *Let $f\colon (S, K) \to (T, K')$ be a clone map between evaluation games.*

(a) Mean score against the field, $s_i = \frac{1}{n}\sum_j K_{ij}$, is not natural: duplicating an agent raises the score of everything that beats it.

(b) The set of Nash equilibria is natural: x is an equilibrium of (S, K) if and only if $f_* x$ is one of (T, K'). When the equilibrium of (T, K') is unique, every rule that selects an equilibrium is therefore natural on types. Invariance to redundant agents is the stated aim of Nash averaging [BTPG18].

(c) The time average of replicator self-play is natural, and when (T, K') has a unique interior equilibrium it assigns to each type its equilibrium mass.

(d) The stationary distribution of the small-mutation chain in which a monomorphic population is invaded by a mutant drawn uniformly from the other names, *with transition probability given by the fixation probability, is not natural.*

Proof. (a) and (d) by example, below. (b) Payoffs against $f_* x$ at the types are the payoffs against x at the agents, and a clone map omits no type [EGHC, Prop. 1.10]. (c) By Theorem 2.6(a), f_* carries trajectories to trajectories and hence time averages to time averages; the time-average theorem for the replicator equation [HoSi98] identifies the limit. □

The example. Rock, a second copy of Rock, Paper, Scissors. Mean scores are $(0, 0, 0.25, -0.25)$: Paper looks best and Scissors worst, although the game is symmetric among the three types. The time-averaged replicator mass on the types is $(0.3334, 0.3333, 0.3333)$. [B13] The small-mutation chain, uniform on plain Rock–Paper–Scissors, gives $(0.2, 0.2, 0.4, 0.2)$ on the four names: Rock's type has mass 0.4, Paper 0.4 and Scissors 0.2. [B14] The reason is visible in the chain: a population of Scissors is now invaded twice as often.

Remark 5.4 (Chain components). The chain in (d) is in the manner of α-Rank [OPP19], which ranks agents by the stationary distribution of a Markov chain on strategy profiles built from fixation probabilities, and which grounds the construction in Conley's theory: the long-run behaviour of a dynamical system is carried by its chain recurrent set, and Markov–Conley chains are a finite model of the sink components. This is the right generalization of equilibrium for games that cycle [PaPi19], and the dynamical content is natural: chain components of a natural flow map into chain components. Statement (d) concerns only the mutation kernel of our small implementation, which counts names; we make no claim about α-Rank as published or as used. A kernel that draws mutants by type would restore invariance, at the price of knowing the types.

6 Finite populations of learners

Population-based training keeps N agents, lets better ones overwrite worse ones, and perturbs the copies [Jad17]. With copying in proportion to fitness and one replacement at a time this is the Moran process of population genetics, and its classical theory applies [Now06, EGHC].

Theorem 6.1 (Fixation and the one-third law). [literature; [NSTF04]] *Let two variants play a coordination game: each does best against its own kind, and the newcomer must exceed a fraction x^* of the population before it is favoured. Under weak selection in a large population a single newcomer takes over with probability greater than the neutral value $1/N$ if and only if $x^* < \frac{1}{3}$.*

With $N = 400$ and selection intensity 10^{-3}, $N\rho$ was 1.021, 1.009, 0.991 and 0.958 at $x^* = 0.25$, 0.30, 0.36 and 0.45. [B16] The deterministic dynamics says that a new convention, protocol or shared representation cannot invade from rarity at all when it needs partners; the finite population says it can, by drift, and when it is more likely than chance to do so.

Theorem 6.2 (The copying topology matters). [literature; [LHN05]] *Let offspring of agent i replace agent j with probability w_{ij}. The fixation probability of a variant of relative fitness r equals the well-mixed value $(1 - r^{-1})/(1 - r^{-N})$ for all r if and only if W is doubly stochastic. A star amplifies selection. On a directed line, in which each agent copies only its predecessor, fixation has probability $1/N$ whatever the fitness.*

For $N = 6$ and $r = 1.2$ the exact values are 0.167 on the line, 0.251 well-mixed and 0.287 on the star. [B17] The design reading is direct. Symmetric exchange between agents, of any shape, neither helps nor hurts selection. A hub that many agents copy from and that copies from them sharpens it. A strictly feed-forward pipeline of copying removes selection altogether: which variant ends up everywhere is decided by where it started.

Remark 6.3 (Populations of language-model agents). Recent work studies populations of agents built on large language models with exactly these tools: replicator dynamics over agent types in open-source games [OSG25], evolutionarily stable strategies as a design target for trust mechanisms [EvT25], group selection and replicator–mutator models for the evolution of cooperative prompts [GSP26], evolutionary game theory as a probe of strategic behaviour [SIL25], and a survey of the area [XLLS26]. The titles and identifiers of these five items were checked on 17 November 2026; we have not evaluated their results. Our contribution to that programme is a set of constraints. Two prompts that induce the same behaviour are clones, so an analysis of a prompt population should be natural in the sense of Section 2; stability should be asked of sets of prompts and not of single ones, because stability of a point is destroyed by cloning and stability of a set is not [EGHC, Thm. 2.8]; and conclusions drawn from the replicator equation about invasion by a single new agent are subject to the correction of Theorem 6.1.

7 Inference

Proposition 7.1 (Bayes is the replicator). [classical; [Sha09, Har09b]] *Let S be a set of hypotheses with likelihoods ℓ_h. Bayesian updating of a prior x on an observation o is one step of Hedge with $\eta = 1$ and payoff $\pi_h = \log \ell_h(o)$.* [B18]

Bayesian inference is the replicator learner fed log-likelihoods, and the dictionary is exact. Naturality for clone maps is the statement that splitting a hypothesis into two indistinguishable ones, with the prior mass divided, changes no posterior probability of any event. The regret identity of Theorem 3.1 is the statement that the log-loss of the Bayesian mixture exceeds that of the best hypothesis by at most $\log(1/x_h(0))$. And the elimination of dominated strategies with its rate [EGHC] is posterior concentration: if observations come from h^*, the log-odds of any other h against h^* fall at the expected rate $D(\ell_{h^*} \,\|\, \ell_h)$ per observation.

8 Summary of results

Table 1 collects the results. Read down the middle column it is a list of facts most of which are individually known; read across, each has a constant attached; and the recurring entry is naturality, which sorts the methods into those that depend on how actions and agents are named and those that do not.

Topic	Qualitative statement	Quantitative statement	Checks
Hedge and replicator	one is the other, sampled	exact at every step size	B1
Policy gradient	replicator with an extra factor x_i	escape time $\sim 1/(2D\epsilon)$ against $\log(1/\epsilon)/D$; ratios 3.0, 11.8, 73.3	B2, B3
Duplicated actions	(R), (H) natural; (PG), (Q) not	slowdown $2k/(k+1)$; defect 0.028	B19–B21
Regret	no regret	$\leq \ln n$ continuous; $\leq \sqrt{T \ln n/2}$ discrete; observed ratio ≤ 0.50	B4, B5
Zero-sum, flow	relative entropy conserved; recurrence	numerical drift 1.3×10^{-15} over time 40	R4 of [EGHC]
Zero-sum, Hedge	iterate leaves, average converges	$+\frac{1}{2}\eta^2$ Var per step; boundary after 1451 steps; exploitability 0.0015	B6–B8
Optimism	last iterate converges	$0.070 \to 1.5 \times 10^{-10}$ in 6000 steps	B9
Entropy regularization	converges; not natural; biased	rate at least α, by an identity; type masses $(0.40, 0.36, 0.24)$ at $\tau = 0.5$	B10, B11, B21
Evaluation	ratings plus cycles plus harmonic	index 0.885; curl 1.5	B12, B15
Redundant agents	averaging and name-mutation chain not natural; replicator average natural	$(0, 0, 0.25, -0.25)$; $(0.4, 0.4, 0.2)$; $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	B13, B14
New conventions	can invade by drift	iff barrier $< \frac{1}{3}$; $N\rho = 1.021, \ldots, 0.958$	B16
Copying topology	amplifiers and suppressors	$0.167 < 0.251 < 0.287$	B17
Inference	Bayes is Hedge at $\eta = 1$	exact	B18

Table 1: Qualitative statements and their quantitative forms. All numbers are produced by the companion script.

9 Limits

Scope. Everything here is tabular: a finite set of actions and exact payoffs. Function approximation, sampling noise, and partial observation are outside it; they are where methods such as [HMO20] do their work. The numbers come from games with at most six actions and populations of at most a few hundred, chosen to be exactly computable, and are illustrations of the statements and not benchmarks. No experiment on a trained model was run.

Novelty. The identifications of Propositions 2.2, 2.3 and 7.1, the regret bounds, recurrence, the Hodge decomposition, the one-third law and the isothermal theorem are from the literature. The observations are Theorem 2.6, the identity of Theorem 4.4, Proposition 5.3, and the placing of the three repairs in Remark 4.5. They are short, and parts of them are likely known in other forms; the invariance of Nash averaging under redundancy, for one, is a stated motivation of [BTPG18].

How categorical. Modestly. The category Game, the functor to dynamical systems, and the operad of wiring diagrams are ordinary category theory; the only cohomology is in degree one; the homotopy theory is in the Conley index behind Remark 5.4, which we invoke and do not compute. The language earns its place by one question, whether a construction commutes with the maps, which turned out to separate methods that the literature usually compares on other grounds.

References. The bibliography was assembled from memory, except for [OPP19, HMO20] and the five items of Remark 6.3, whose bibliographic details were checked online on 17 November 2026. It should be verified before it is relied on.

Authorship. This paper was drafted with the assistance of an AI system (Anthropic's Claude), which also wrote and ran the companion script.

Open problems. [programme] (1) Characterize the learners that are natural for clone maps among those derived from a regularizer; we expect entropy on types and not on names to be forced. (2) Give the rate in Theorem 4.4 for a non-uniform equilibrium, where the limit moves with τ. (3) Extend Proposition 5.3(c) from time averages to chain components, so that a ranking by recurrent classes is natural by construction. (4) Find the analogue of the isothermal theorem for populations of learners whose fitness depends on the population, where the graph and the game interact.

A The companion script

`verify-evolutionary-learning.py` uses only the Python standard library, fixes its random seed, runs in under half a minute, and exits with status zero if and only if its 21 checks pass. Flows are integrated by a fourth-order Runge–Kutta scheme; fixation probabilities are exact. The checks test instances and recompute the numbers quoted in the text. They do not prove the theorems.

B1	Hedge equals the time-η replicator flow, 30 random cases
B2	policy gradient in logits; Fisher preconditioning
B3	escape times, Proposition 2.4, against numerical integration
B4	continuous-time regret identity and bound
B5	discrete regret bound, 40 runs, random and adversarial
B6	exact drift of Hedge, Proposition 4.2, three step sizes
B7	monotone divergence to the boundary
B8	exploitability of the time average
B9	optimistic Hedge, last iterate
B10	identity and rate of Theorem 4.4
B11	rest points of (3) are logit equilibria
B12	Hodge decomposition of a planted example
B13	mean score and replicator average under cloning
B14	small-mutation chain: uniform, dominated, cloned
B15	a harmonic class on a square of comparisons
B16	one-third law, exact fixation, $N = 400$
B17	line, well-mixed, star: exact fixation, $N = 6$
B18	Bayes equals Hedge with $\eta = 1$
B19	naturality defect of (R) and (PG) under cloning
B20	slowdown $2k/(k+1)$ for $k = 1, 2, 4, 8$
B21	logit equilibrium with a cloned action, three temperatures

References

[Alex23] J. M. Alexander, *Evolutionary Game Theory*, Elements in Decision Theory and Philosophy, Cambridge University Press, 2023.

[Ama98] S. Amari, Natural gradient works efficiently in learning, *Neural Comput.* 10 (1998), 251–276.

[BaPi18] J. P. Bailey, G. Piliouras, Multiplicative weights update in zero-sum games, in *Proc. ACM Conf. Economics and Computation (EC)*, 2018, 321–338.

[BoSa97] T. Börgers, R. Sarin, Learning through reinforcement and replicator dynamics, *J. Econom. Theory* 77 (1997), 1–14.

[BTPG18] D. Balduzzi, K. Tuyls, J. Pérolat, T. Graepel, Re-evaluating evaluation, in *Advances in Neural Information Processing Systems* 31, 2018.

[BTHK15] D. Bloembergen, K. Tuyls, D. Hennes, M. Kaisers, Evolutionary dynamics of multi-agent learning: a survey, *J. Artificial Intelligence Res.* 53 (2015), 659–697.

[CBL06] N. Cesa-Bianchi, G. Lugosi, *Prediction, Learning, and Games*, Cambridge University Press, 2006.

[CGG22] G. S. H. Cruttwell, B. Gavranović, N. Ghani, P. Wilson, F. Zanasi, Categorical foundations of gradient-based learning, in *European Symposium on Programming (ESOP)*, 2022.

[DaPa19] C. Daskalakis, I. Panageas, Last-iterate convergence: zero-sum games and constrained min-max optimization, in *Innovations in Theoretical Computer Science (ITCS)*, 2019.

[DISZ18] C. Daskalakis, A. Ilyas, V. Syrgkanis, H. Zeng, Training GANs with optimism, in *Int. Conf. Learning Representations (ICLR)*, 2018.

[EGHC] Anonymous, *Evolutionary Game Theory in Higher Categories*, Higher Categorical Semantics, 2026. Companion volume.

[EvT25] Wang et al., Ev-Trust, a trust mechanism for multi-agent systems of language models built on evolutionary stability, preprint, arXiv:2512.16167, 2025. Full title and author list not recorded when the listing was checked.

[FrSc97] Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, *J. Comput. System Sci.* 55 (1997), 119–139.

[FST19] B. Fong, D. I. Spivak, R. Tuyéras, Backprop as functor: a compositional perspective on supervised learning, in *LICS 2019*, IEEE, 2019.

[Goo14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in *Advances in Neural Information Processing Systems* 27, 2014.

[GSP26] Group selection to evolve cooperative prompts in populations of LLM agents, listed for NeurIPS 2026; authors not recorded when the listing was checked.

[Har09b] M. Harper, The replicator equation as an inference dynamic, preprint, arXiv:0911.1763.

[HMO20] D. Hennes, D. Morrill, S. Omidshafiei, R. Munos, J. Perolat, M. Lanctot, A. Gruslys, J.-B. Lespiau, P. Parmas, E. Duéñez-Guzmán, K. Tuyls, Neural replicator dynamics: multiagent learning via hedging policy gradients, in *Proc. AAMAS*, 2020; arXiv:1906.00190.

[HoSi98] J. Hofbauer, K. Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, 1998.

[Jad17] M. Jaderberg, V. Dalibard, S. Osindero, W. M. Czarnecki, J. Donahue, A. Razavi, O. Vinyals, T. Green, I. Dunning, K. Simonyan, C. Fernando, K. Kavukcuoglu, Population based training of neural networks, preprint, arXiv:1711.09846.

[JLYY11] X. Jiang, L.-H. Lim, Y. Yao, Y. Ye, Statistical ranking and combinatorial Hodge theory, *Math. Program.* 127 (2011), 203–244.

[Kak01] S. Kakade, A natural policy gradient, in *Advances in Neural Information Processing Systems* 14, 2001.

[KwMe17] J. Kwon, P. Mertikopoulos, A continuous-time approach to online optimization, *J. Dyn. Games* 4 (2017), 125–148.

[LHN05] E. Lieberman, C. Hauert, M. A. Nowak, Evolutionary dynamics on graphs, *Nature* 433 (2005), 312–316.

[LPS21] S. Leonardos, G. Piliouras, K. Spendlove, Exploration-exploitation in multi-agent competition: convergence with bounded rationality, in *Advances in Neural Information Processing Systems* 34, 2021.

[LZG17] M. Lanctot, V. Zambaldi, A. Gruslys, A. Lazaridou, K. Tuyls, J. Pérolat, D. Silver, T. Graepel, A unified game-theoretic approach to multiagent reinforcement learning, in *Advances in Neural Information Processing Systems* 30, 2017.

[MPP18] P. Mertikopoulos, C. Papadimitriou, G. Piliouras, Cycles in adversarial regularized learning, in *Proc. ACM-SIAM Symp. Discrete Algorithms (SODA)*, 2018, 2703–2717.

[MSP73] J. Maynard Smith, G. R. Price, The logic of animal conflict, *Nature* 246 (1973), 15–18.

[MXSS20] J. Mei, C. Xiao, C. Szepesvári, D. Schuurmans, On the global convergence rates of softmax policy gradient methods, in *Proc. Int. Conf. Machine Learning (ICML)*, 2020.

[Mye21] D. J. Myers, Double categories of open dynamical systems (extended abstract), in *Applied Category Theory 2020*, EPTCS 333, 2021, 154–167.

[Now06] M. A. Nowak, *Evolutionary Dynamics: Exploring the Equations of Life*, Harvard University Press, 2006.

[NSTF04] M. A. Nowak, A. Sasaki, C. Taylor, D. Fudenberg, Emergence of cooperation and evolutionary stability in finite populations, *Nature* 428 (2004), 646–650.

[OPP19] S. Omidshafiei, C. Papadimitriou, G. Piliouras, K. Tuyls, M. Rowland, J.-B. Lespiau, W. M. Czarnecki, M. Lanctot, J. Perolat, R. Munos, α-Rank: multi-agent evaluation by evolution, *Sci. Rep.* 9 (2019); arXiv:1903.01373.

[OSG25] Evaluating LLMs in open-source games, preprint, arXiv:2512.00371, 2025; authors not recorded when the listing was checked.

[PaPi19] C. Papadimitriou, G. Piliouras, Game dynamics as the meaning of a game, *ACM SIGecom Exchanges* 16 (2019), no. 2, 53–63.

[PiSh14] G. Piliouras, J. S. Shamma, Optimization despite chaos: convex relaxations to complex limit sets via Poincaré recurrence, in *Proc. ACM-SIAM Symp. Discrete Algorithms (SODA)*, 2014, 861–873.

[RaSr13] A. Rakhlin, K. Sridharan, Optimization, learning, and games with predictable sequences, in *Advances in Neural Information Processing Systems* 26, 2013.

[SaCr03] Y. Sato, J. P. Crutchfield, Coupled replicator equations for the dynamics of learning in multiagent systems, *Phys. Rev. E* 67 (2003), 015206.

[Sha09] C. R. Shalizi, Dynamics of Bayesian updating with dependent data and misspecified models, *Electron. J. Stat.* 3 (2009), 1039–1074.

[SIL25] Strategic intelligence in large language models: evidence from evolutionary game theory, preprint, arXiv:2507.02618, 2025; authors not recorded when the listing was checked.

[Sor09] S. Sorin, Exponential weight algorithm in continuous time, *Math. Program.* 116 (2009), 513–528.

[SSV20] P. Schultz, D. I. Spivak, C. Vasilakopoulou, Dynamical systems and sheaves, *Appl. Categ. Structures* 28 (2020), 1–57.

[TaJo78] P. D. Taylor, L. B. Jonker, Evolutionarily stable strategies and game dynamics, *Math. Biosci.* 40 (1978), 145–156.

[TVL03] K. Tuyls, K. Verbeeck, T. Lenaerts, A selection-mutation model for Q-learning in multi-agent systems, in *Proc. AAMAS*, 2003, 693–700.

[Vin19] O. Vinyals et al., Grandmaster level in StarCraft II using multi-agent reinforcement learning, *Nature* 575 (2019), 350–354.

[XLLS26] Xie, Liu, Lan, Szolnoki, Evolutionary games in LLM-agent societies: early milestones and perspectives, *Chaos Solitons Fractals* 213 (2026), 119158. Initials not recorded when the listing was checked.