Evolutionary Selection for Agent Populations under Comparable Evaluation

Anonymous • November 2026

Abstract. An agent population can allocate work and select proposals by comparing their measured performance. The validity of a comparison depends on the evaluation context, while allocation and selection normalize the resulting weights. We model the inner execution by a scored probability kernel and the outer selection by exponential reweighting followed by variation. We characterize exact aggregation when the scores and variation descend to a common quotient. A projective total-variation bound controls one-step errors from stale or mismatched scores. Under a contracting variation kernel we propagate that bound over repeated updates. These results give an interface for auditing the outer update separately from the evaluation assumptions that make it meaningful.

Keywords. Bilevel selection, agent populations, exponential reweighting, quotients, evaluation delay, Dobrushin contraction.

1 Execution and population selection

An agent system can learn at two levels. At the inner level, a proposed strategy carries out a task and returns a trace. At the outer level, a population controller assigns evaluation or deployment mass to available strategies. A useful outer algorithm should specify which inner measurements it can compare, how it constructs the next population, and how a change in those measurements affects that population. This paper addresses these three questions for evolutionary selection.

The model takes the initial population and the variation kernel as inputs. It applies to routing among fixed specialists, to allocating evaluation budgets to immutable proposals, and to selecting newly admitted variants. In each application the same selection rule can operate over a finite registry or a standard Borel proposal space. Finite registries admit an algebraic realization by sums. The measure-theoretic statements below describe the update independently of a specific registry.

1.1 An inner evaluation contract

Let S be a standard Borel space of strategy versions. A strategy version fixes the admissible behavior to be evaluated, including any state snapshot that affects its execution. A change of the behavior-defining data creates a different version. An evaluation context c specifies a task space D_c, a task law λ_c, an execution environment and budget, an outcome space Z_c, and a bounded measurable scoring function $r_c : D_c \times Z_c \to \mathbb{R}$. The outcome space includes statuses that the scoring policy uses to treat failure or noncompletion. The inner semantics is a probability kernel

$$K_c \colon S \times D_c \longrightarrow \mathcal{P}(Z_c).$$

Here and below, probability kernels include both deterministic maps and randomized evaluation. The expected score is

$$F_c(s) = \int_{D_c} \int_{Z_c} r_c(d, z)\, K_c(s, d, dz)\, \lambda_c(dd). \tag{1}$$

Measurability of F_c follows from the defining measurability of the kernel and of the bounded integrand.

A context identifier in an audit record stands for this collection of choices. Distinct realizations may represent the same contract when they induce the same relevant scored law. Equality of a score convention by itself need not imply equality of task laws or execution budgets. A stated comparison across contexts therefore uses transport carrying the score or a quantitative bound on the relevant discrepancy. A missing trace and a trace scored as failure may give different score estimands if they are pooled without an explicit inclusion rule.

Endogenous evaluation is also possible. Let $a(s, y)$ be a bounded measurable interaction payoff for evaluating s against y under a fixed contract, and let m be a probability measure on S. Set

$$F_m(s) = \int_S a(s, y)\, m(dy). \tag{2}$$

The opponent law or population at which the score was measured is now part of the comparison. In an empirical implementation, the exact integral and an estimate made from a sample are separate inputs to the error analysis. All statements about scores are relative to the declared bounded score function.

1.2 An outer update

For a probability measure m on S and a measurable weight $w : S \to (0, \infty)$ with $0 < \int_S w\, dm < \infty$, define the probability measure

$$\mathcal{S}_w(m)(B) = \frac{\int_B w(s)\, m(ds)}{\int_S w(s)\, m(ds)}. \tag{3}$$

In the score-based algorithm the weight is $w = \exp(\eta F)$, with $\eta \geq 0$ fixed by the controller in each round. The

score is held fixed for that transition, including when it was obtained from (2). A variation kernel $V : S \longrightarrow \mathcal{P}(S)$ follows selection. The combined transition $m \mapsto \mathcal{S}_w(m)V$ sends a selected strategy to a distribution over strategies for the next round. The identity kernel leaves the selected population unchanged. The semantics allows a controller to put exploration or externally admitted replacements in V.

Multiplication by a positive constant in w leaves (3) unchanged. Thus adding a common constant to every score of a round leaves its transition unchanged. A positive change of units $F \mapsto aF + b$ gives the same transition when the controller replaces η by η/a. The support of selection in the measure-theoretic sense is also preserved. A measurable set has selected mass zero exactly when its original mass is zero, because $w > 0$. Variation determines any movement to previously unoccupied regions. These identities help distinguish score conventions that can be changed by a reparametrization from differences that change allocations.

For bounded F write $\bar{F}_m = \int_S F\, dm$. For any bounded measurable test function φ, dominated differentiation gives

$$\left.\frac{d}{d\eta}\right|_{\eta=0} \int_S \varphi\, d\mathcal{S}_{e^{\eta F}}(m) = \int_S \varphi(F - \bar{F}_m)\, dm. \tag{4}$$

Indeed, both integrals in the ratio are differentiable at zero. Their derivatives insert a factor F, and the denominator at zero is one. With $F = F_m$ this is the weak form of the replicator vector field. For finite strategy types, the right side tested against a coordinate indicator is $m_i(F_m(i) - \bar{F}_m)$. This links the outer controller to the continuous evolutionary dynamics treated by Alexander [1]. The finite-step rule belongs to multiplicative weights [2, 3]. When F is fixed across all rounds and V is the identity kernel, the derivative of the expected score along exponential reweighting is $\mathrm{Var}_{\mathcal{S}_{e^{\eta F}}(m)}(F)$, by the same differentiation at each η. Endogenous F_m may vary with the population, so that calculation is a statement about a frozen evaluation.

2 When a reduced population model is exact

A controller often compares populations across different registries or versions of an agent catalogue. Let $\pi : S \to Q$ be a measurable map to a second standard Borel space, and write $\pi_* m$ for pushforward. A reduced controller retains $\pi_* m$ rather than the original m. We first characterize exactly what happens to one weighting and then give a sufficient condition for repeated selection and variation.

Theorem 2.1 (Pushforward of a selection)**.** *With m and w as in (3), let $\bar{m} = \pi_* m$. Let $\bar{w}$ be a version of $d\pi_*(wm)/d\bar{m}$, where wm is the finite measure with density w with respect to m. Then*

$$\pi_* \mathcal{S}_w(m) = \mathcal{S}_{\bar{w}}(\bar{m}). \tag{5}$$

If $w = v \circ \pi$ for a positive measurable v, the same weight v can be used for every m.

Proof. For a $\bar{m}$-null set C, its inverse image is m-null and therefore wm-null. The Radon–Nikodym derivative in the statement exists. It is positive $\bar{m}$-almost everywhere since $w > 0$, and it can be given any positive value on the remaining null set. For $C \subseteq Q$ measurable, the numerator of the left side in (5) is $\pi_*(wm)(C) = \int_C \bar{w}\, d\bar{m}$. Both denominators equal $\int_S w\, dm$. If w factors through π, the latter derivative is v by the change-of-variables identity for pushforward. □

The reduced weight exists even when the original weight is inconsistent along fibres. In that situation $\bar{w}$ generally depends on the distribution within each fibre, information not supplied by $\bar{m}$. In particular, for exponential reweighting, the reduced weight is a conditional average of exponentials rather than the exponential of a conditional average of scores. The order of averaging and exponentiation matters for the resulting population.

There is a converse for a fixed w when the image of π contains at least two points. Suppose the same reduced reweighting gives exact descent for every population. If $\pi(s) = \pi(s')$, choose y with $\pi(y) \neq \pi(s)$. The laws $(\delta_s + \delta_y)/2$ and $(\delta_{s'} + \delta_y)/2$ push forward to the same law. Their selected masses at $\pi(s)$ are $w(s)/(w(s)+w(y))$ and $w(s')/(w(s') + w(y))$. As $w(y) > 0$, equality forces $w(s) = w(s')$. This proves pointwise constancy along fibres. In the degenerate case of a one-point image the reduced population always has unit mass at that point and the requirement imposes no constraint on w. A measurable reduced weight is part of the forward factorization hypothesis.

Proposition 2.2 (Exact reduced outer dynamics)**.** *Suppose V and a kernel $\bar{V} : Q \to \mathcal{P}(Q)$ satisfy*

$$V(s, \pi^{-1}C) = \bar{V}(\pi(s), C) \quad \text{for every measurable } C \subseteq Q. \tag{6}$$

If $w = v \circ \pi$, then

$$\pi_*(\mathcal{S}_w(m)V) = \mathcal{S}_v(\pi_* m)\bar{V}.$$

The result applies at every round for population-dependent scores when $F_m = \bar{F}_{\pi_* m} \circ \pi$.

Proof. For any probability measure n, integration in (6) gives $\pi_*(nV) = (\pi_* n)\bar{V}$. Apply the equality with $n = \mathcal{S}_w(m)$ and use Theorem 2.1. For population-dependent scores, the reduced value is determined by $\pi_* m$, so induction proves equality at all rounds with the corresponding transition. □

The condition in (6) is strong lumpability, expressed using Markov kernels. It also follows from a square of kernels in the usual compositional language of probability [4]. If T denotes the full population transition and $\bar{T}$ the reduced transition, both vertical arrows in the following square are pushforwards π_*.

$$\begin{array}{ccc} \mathcal{P}(S) & \xrightarrow{\;T\;} & \mathcal{P}(S) \\ \downarrow & & \downarrow \\ \mathcal{P}(Q) & \xrightarrow{\;\bar{T}\;} & \mathcal{P}(Q). \end{array}$$

Reindexing the registry by a measurable isomorphism is a special case, with the scores and variation transported along the isomorphism. For strategy families, requiring the scores to factor is stronger than equality of a current leaderboard. A quotient that groups all proposals with the same measured score can still contain versions that generate different distributions of descendant strategies. A reduced controller is exact when both comparisons needed by Proposition 2.2 are in place.

The unnormalized measure also gives a rule for combining evaluation shards. For a convex mixture $m = \sum_i \alpha_i m_i$ evaluated under a common weight function, write $Z_i = \int w\, dm_i$. Linearity of integration gives

$$\mathcal{S}_w(m) = \frac{\sum_i \alpha_i Z_i\, \mathcal{S}_w(m_i)}{\sum_i \alpha_i Z_i}.$$

This formula holds for countably many shards with finite positive total normalizer. It identifies the sufficient local outputs for exact recombination. An implementation that averages local probability vectors with the original proportions α_i can change the global update, because it loses the unnormalized masses. Comparability of the local weights is a separate assumption.

3 Sensitivity to score and population mismatch

For probabilities m, n write $d_{\mathrm{TV}}(m,n) = \sup_B |m(B) - n(B)|$, where B ranges over measurable sets. Here d_{TV} is half the variation norm of the signed measure $m - n$. If a bounded measurable function has values in an interval of length L, the absolute difference of its expectations under m, n is at most $L d_{\mathrm{TV}}(m,n)$. For $L > 0$, subtract the lower endpoint, rescale to $[0,1]$, and integrate the differences of the probabilities of its superlevel sets.

Theorem 3.1 (Projective sensitivity)**.** *Let $w, \widehat{w}$ be positive m-integrable weights as in (3). Assume $\log(\widehat{w}/w) \in [\ell_-, \ell_+]$ m-almost everywhere, and set $\omega = \ell_+ - \ell_-$. Then*

$$d_{\mathrm{TV}}(\mathcal{S}_{\widehat{w}}(m), \mathcal{S}_w(m)) \le \tanh(\omega/4). \tag{7}$$

Proof. Set $p = \mathcal{S}_w(m)$, $r = \widehat{w}/w$, and $u = \int r\, dp$. The other selected measure has density r/u with respect to p, so the distance is $\int |r - u|\, dp/(2u)$. Put $a = \exp(\ell_-)$ and $b = \exp(\ell_+)$. If $a = b$, the distance is zero by proportionality. For $a < b$, the secant bound for the convex function $x \mapsto |x - u|$ on $[a,b]$ yields

$$\frac{\int |r-u|\, dp}{2u} \le \frac{(u-a)(b-u)}{u(b-a)} = \frac{a+b-u-ab/u}{b-a}.$$

The inequality $u + ab/u \ge 2\sqrt{ab}$ bounds this by $(\sqrt{b} - \sqrt{a})/(\sqrt{b} + \sqrt{a})$, which equals $\tanh(\omega/4)$. □

The essential oscillation of the log-density ratio is the Hilbert projective distance, since normalizing adds a common constant. Equation (7) is the specialization to selection of the projective control of total variation studied by Cohen and Fausti [5]. For $w = e^{\eta F}$ and $\widehat{w} = e^{\eta \widehat{F}}$, a uniform error $|\widehat{F} - F - b| \le \epsilon$ after any common shift b gives

$$d_{\mathrm{TV}}(\mathcal{S}_{e^{\eta \widehat{F}}}(m), \mathcal{S}_{e^{\eta F}}(m)) \le \tanh(\eta\epsilon/2). \tag{8}$$

A bound on the loss for one strategy cannot replace the uniform or oscillation bound on the relative scores of the strategies allocated mass. If a uniform bound is probabilistic, the displayed guarantee holds on the event on which the bound holds.

For the endogenous payoff in (2), suppose each function $a(s,-)$ has range in an interval of length at most L. The expectation bound gives

$$\sup_s |F_m(s) - F_n(s)| \le L d_{\mathrm{TV}}(m,n). \tag{9}$$

A stale score measured at n with uniform additional error ϵ yields an error of at most $\epsilon + L d_{\mathrm{TV}}(m,n)$ when applied at m. Thus freshness is determined by a change in the opponent distribution and its influence on the scores. The number of elapsed rounds by itself gives no such bound.

A contract change can be handled in the same way when the tasks and outcomes can be compared. For two contexts on common spaces D, Z, use scoring maps taking values in $[0,1]$. Decompose the change in (1) by replacing the task law, execution kernel, and scoring map one at a time. The first term is bounded by $d_{\mathrm{TV}}(\lambda_c, \lambda_{c'})$. The second is bounded by $\sup_{s,d} d_{\mathrm{TV}}(K_c(s,d,\cdot), K_{c'}(s,d,\cdot))$, by the same expectation inequality. The final term is bounded by $\sup_{d,z} |r_c(d,z) - r_{c'}(d,z)|$. Their sum bounds $\sup_s |F_c(s) - F_{c'}(s)|$ and is an input to (8). A declared transport can first place different task or outcome representations in common spaces. The bound uses the transported quantities and therefore depends on the specified comparison.

4 Stability across repeated updates

Selection also changes when the population changes with the weights fixed. If $0 < a \le w \le b$, then

$$d_{\mathrm{TV}}(\mathcal{S}_w(m), \mathcal{S}_w(n)) \le (b/a) d_{\mathrm{TV}}(m,n). \tag{10}$$

For measurable B, put $c_B = \mathcal{S}_w(n)(B)$. The difference of its selected probabilities is

$$\frac{\int w(\mathbf{1}_B - c_B)\, d(m-n)}{\int w\, dm}.$$

The integrand takes values in $[-bc_B, b(1-c_B)]$, an interval of length b, and the denominator is at least a. Taking the supremum over B gives (10).

Suppose F_m has oscillation across S at most M uniformly in m and satisfies (9). Split a change of the selection at m to the selection at n into a change of base measure with weight fixed and a change of scores with base measure fixed. Combining (10) and (8), and using $\tanh x \le x$ for $x \ge 0$, gives the following Lipschitz constant for the endogenous selector:

$$C = e^{\eta M} + \eta L/2. \tag{11}$$

The ratio b/a for $w = e^{\eta F_m}$ can be taken to be at most $e^{\eta M}$. This use of score oscillation makes the constant unchanged by adding a common score offset. The stated L is sufficient; a bound on score differences modulo common offsets would likewise suffice in the second term.

For variation, write κ for the Dobrushin coefficient of V, the supremum of $d_{\mathrm{TV}}(V(s,\cdot), V(s',\cdot))$ over s, s'. Then $d_{\mathrm{TV}}(mV, nV) \le \kappa d_{\mathrm{TV}}(m,n)$. For each B, the function $s \mapsto V(s,B)$ has oscillation at most κ, so the assertion follows by the expectation bound. This is the usual coefficient for variation of probability measures [6].

Proposition 4.1 (Propagation of score errors)**.** *Let the reference transition at m use $e^{\eta F_m}$ and let the approximate transition at $\widetilde{m}_t$ use $e^{\eta \widetilde{F}_t}$, each followed by V. Set $d_t = d_{\mathrm{TV}}(\widetilde{m}_t, m_t)$ and assume that $\widetilde{F}_t - F_{\widetilde{m}_t}$ is uniformly bounded by ϵ_t in absolute value after some common shift. For C as in (11) and $\rho = \kappa C$,*

$$\begin{aligned} d_{t+1} &\le \rho d_t + \kappa \tanh(\eta\epsilon_t/2), \\ d_t &\le \rho^t d_0 \\ &\quad + \kappa \sum_{j<t} \rho^{t-1-j} \tanh(\eta\epsilon_j/2). \end{aligned} \tag{12}$$

In particular, if $\rho < 1$ and $\epsilon_t \le \epsilon$ for every t, the contribution of score errors is bounded by $\kappa \tanh(\eta\epsilon/2)/(1-\rho)$.

Proof. Insert the reference selection at $\widetilde{m}_t$ between the two selected populations. Bound the first distance by (8) and the second by Cd_t. Applying V contracts both by κ. Iterating the resulting one-step inequality proves (12). The uniform bound uses the sum of the geometric series. □

The recurrence applies along each realization on which the assumed error bounds hold. It can also use uniform bounds over round-dependent scores and a shared round-dependent variation kernel. The condition $\rho < 1$ is sufficient for uniform control over time. A controller with identity variation on a nontrivial strategy space has $\kappa = 1$ and cannot infer uniform control from this criterion. Strong mixing can make the condition easier to satisfy while changing the reference dynamics. The result bounds distance from that reference process and makes no assertion about the attained objective value.

5 Relationship to existing methods and audit

Population-based training interleaves improvement at the level of individual workers with population-level copying and hyperparameter changes [7]. Policy-space response oracles expand policy populations and compute a mixture from estimated empirical game payoffs [8]. Those methods supply meaningful settings for the inner/outer split. Automated Design of Agentic Systems searches over versions of agent programs using measured performance [9]. The outer selection studied here can govern allocation among such versions when the relevant common score law is specified. The journal's bilevel primer distinguishes the credit assigned to an inner adaptation from the method that proposes the next adaptation [10]. Here the contribution is to characterize when a particular outer selection respects reduced representations and how its results change when its evaluation inputs change. The selection rule is an established multiplicative update. Its use in this model neither chooses the policy search method nor fixes a downstream learning objective.

The outer computation can be replayed given its input measures, weights, and variation kernel. A defensible deployment record can associate each score with a strategy version, the evaluation contract, the opponent or task law, and a record of how uncertainty was estimated. A reduced run also specifies the quotient and the condition under which variation respects it. For distributed evaluation, each shard's original mixture mass and normalizer are needed when combining selected measures. A comparison between exact and approximate runs further states which parts of the transition are shared, since Proposition 4.1 isolates errors in scores. These are sufficient descriptions of the mathematical inputs. Artifacts that certify the implementation of the inner scorer or validate it on a task distribution address additional empirical questions.

A useful next empirical question is whether a declared uniform or relative error bound is attainable with real evaluation budgets. If a randomized evaluation is repeated adaptively, a confidence statement for a single, preselected comparison does not by itself give a simultaneous bound on all strategy scores used in the trajectory. A study can report the target population, sources of randomness, dependence among score estimates, and the baseline outer transition alongside any policy-search or routing metrics. The mathematical bounds can then be

applied at the same level at which the empirical guarantee is established.

5.1 Mechanized algebra

The companion Lean 4.19 project checks the finite-type algebra for the replicator drift and an affine reproductive update. For a population vector x of mass one and a payoff vector u, write $D_i = x_i(u_i - \sum_j x_j u_j)$. Choosing weights $1 + \eta u_i$ in (3), under the hypotheses that these weights are positive, gives a selected vector with increment

$$\mathcal{S}_{1+\eta u}(x)_i - x_i = \frac{\eta D_i}{1 + \eta \sum_j x_j u_j}.$$

The Lean identity also holds algebraically under the weaker hypothesis that the denominator is nonzero. `selectionStep_affine_sub` verifies the increment. The same module, `EvoGame.Replicator`, checks the simplex conditions under strictly positive reproductive factors, conservation of total mass for the drift, and invariance of the drift under a common shift in all scores.

The module `EvoGame.Aggregation` checks that summing over fibres intertwines drift and normalized selection when the fine payoff is pulled back from the quotient. It also checks composition of reindexings and the population-dependent and matrix versions of drift aggregation. These identities quantify over arbitrary finite strategy types. The measure-theoretic pushforward, exponential weights and stability results above are proved in the manuscript. The formal check addresses neither statistical score estimation nor the execution of agents.

References

[1] J. M. Alexander. *Evolutionary Game Theory*. Elements in Decision Theory and Philosophy, Cambridge University Press, 2023. doi:10.1017/9781108582063.

[2] S. Arora, E. Hazan, and S. Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8:121–164, 2012. doi:10.4086/toc.2012.v008a006.

[3] S. Sorin. Replicator dynamics: old and new. *Journal of Dynamics and Games*, 7(4):365–386, 2020. doi:10.3934/jdg.2020028.

[4] T. Fritz. A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, 2020. doi:10.1016/j.aim.2020.107239.

[5] S. N. Cohen and E. Fausti. Hyperbolic contractivity and the Hilbert metric on probability measures. Preprint, 2023, revised 2024. arXiv:2309.02413v2.

[6] S. Gaubert and Z. Qu. Dobrushin's ergodicity coefficient for Markov operators on cones. *Integral Equations and Operator Theory*, 81(1):127–150, 2015. doi:10.1007/s00020-014-2193-2.

[7] M. Jaderberg et al. Population based training of neural networks. Preprint, 2017. arXiv:1711.09846.

[8] M. Lanctot et al. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. Proceedings article.

[9] S. Hu, C. Lu, and J. Clune. Automated design of agentic systems. In *International Conference on Learning Representations*, 2025. Proceedings article.

[10] Meta Journal of Agentic Meta Algorithms. Self-improvement as a bilevel program: inner adaptation and outer credit. Editorial primer, 2026. Journal article.