Between Interface and Truth: Multi-Task Selection Drives Ecologically Veridical Perception

Giulio Valentino Dalla Riva
Baffelan OÜ
me@gvdallariva.net
www.baffelan.com

Abstract

When does optimisation for performance yield representations that track world structure? We develop a mathematical theory of agents with a single fixed encoding shared across tasks, and use it to resolve the broader debate over whether selection favors fitness-tuned interfaces or veridical perception. Selection favors *ecological veridicality*: preserving exactly those world-state distinctions required by the task ecology. The governing object is a separation condition on the task distribution μ: if a pair of states is distinguished on tasks with positive μ-measure, optimal encodings must separate it. In evolutionary terms, distinctions that systematically affect fitness across the ecology are selected to persist, while distinctions with zero aggregate fitness consequence can collapse neutrally without risk penalty. We prove static optimality results and deterministic evolutionary convergence (Price decomposition plus quasispecies recursion) to the best mutation-accessible optimum, with global convergence under primitive mutation and Wright–Fisher finite-population approximation on fixed horizons. As task diversity increases, resolved ecological complexity $k_T = |W/{\sim_T}|$ grows monotonically (graded cascade). The framework recovers both established poles of the debate: in the single-task limit, fitness-tuned interface encodings can dominate truth-tracking encodings, while in the fixed-encoding multi-task regime, selection favors ecological veridicality up to capacity limits.

1 Introduction

When an agent is optimised for performance, whether by natural selection, gradient descent, or Bayesian updating, does its internal representation come to reflect the true structure of the world? Or can high performance be achieved with representations that are systematically distorted, compressed, or even disconnected from reality? This question is fundamental across several disciplines. In evolutionary biology, it concerns whether natural selection produces accurate perception. In machine learning, it concerns whether deep networks trained on prediction tasks develop “truthful” internal representations. In philosophy of mind, it concerns the epistemic status of conscious experience. Despite its breadth, the question has lacked a precise mathematical treatment that identifies the *conditions* under which optimisation yields representational fidelity.

The sharpest formulation of the negative case is due to Hoffman et al. (2015). Their Fitness-Beats-Truth (FBT) theorem demonstrates that, in an evolutionary game where organisms act on the basis of perceptual information, encodings tuned to fitness generically dominate encodings that faithfully represent the world. The argument is elegant: a fitness-tuned "interface" encoding collapses world states that yield equal fitness into a single percept, using its limited channel capacity to mark only the fitness-relevant distinctions. A veridical encoding wastes capacity distinguishing states that are equivalent from the organism's perspective. In simulations and under broad conditions, the interface wins (Prakash et al., 2021). The case was further strengthened by Prakash et al. (2020), who proved that single-task payoff functions generically fail to preserve mathematical structures (total orders, permutation groups, cyclic groups, and measurable spaces), with the probability of structure-preservation approaching zero as the state space grows. Hoffman (2019) drew the philosophical conclusion that perception is fundamentally non-veridical: we see "icons on a desktop," not reality.

Berke et al. (2022) challenged this conclusion with a simple but consequential observation. Organisms do not face one task; they face many. A primate foraging for fruit also avoids predators, navigates terrain, selects mates, and monitors conspecifics. If the perceptual system is *cognitively impenetrable*, fixed across tasks, not re-tuned for each one, then the interface strategy fails. An encoding tuned to collapse fruit-irrelevant distinctions may merge precisely the states that predator-avoidance requires to be distinct. In evolutionary simulations with multiple tasks and a fixed encoding, Berke et al. found that the fittest organisms were those with veridical perception. The result was demonstrated by simulation only; the authors offered no analytical proof and explicitly called for "detailed mathematical models" to explain the phenomenon.

Anderson (2015) identified the core logical issue from a philosophical standpoint. He argued that Hoffman et al. had "only considered the problem of adaptation over evolutionary time scales" and had "ignored the need for (and demonstrated capacity of) animals to adjust their behavior to achieve homeostasis within ontogenetic time scales." When organisms face multiple homeostatic demands whose payoff functions are nonmonotonic, "the perceptual response needs to track resources monotonically so that an animal can know how to adapt its behavior to achieve homeostasis" (Anderson, 2015, pp. 1508–1509). An interface encoding that maps perception directly onto payoffs provides no directional information; an animal can tell that its resource level is "off" but not *which direction it is off.* Anderson concluded that "it is possible to embrace the significance of fitness in shaping evolution without arriving at the epistemological and metaphysical conclusions of [interface theory]" (Anderson, 2015, p. 1511), but offered no formal framework to substantiate this. Indeed, Hoffman and Singh (2012, §8.3) themselves raised the multi-task question explicitly: does a general-purpose encoding serving multiple fitness functions become more veridical as the number of tasks grows? They conjectured that it would not: "there is no principled reason why maximizing the channel capacity for the best-fit fitness signal should automatically maximize channel capacity for the truth signal", and noted that "this must remain an open question until detailed mathematical models of this process are developed and studied."

This paper provides such models. The framework applies beyond the Hoffman/Berke debate to any system where a fixed representation must serve multiple objectives. We prove that multi-task optimisation with a fixed encoding selects for what we call *ecological veridicality*: the preservation of exactly those distinctions between world states that the task ecology demands. The key mathematical object is the *separation condition*, a property of the task distribution μ that determines which pairs of world states are distinguished on sets of tasks of positive μ-measure. Pairs that are separated must receive distinct internal representations; pairs that differ only on μ-null task sets may be freely merged without Bayes-risk cost. The result is not absolute truth but a graded, task-relative fidelity whose resolution is set by the agent's task ecology.

The core results are:

(i) A static optimality theorem: ecological veridicality minimises multi-task Bayes risk (uniquely up to percept-label symmetry), with explicit lower bounds in the full-separation regime (Theorem 4.1).

(ii) A deterministic mean-field evolutionary convergence proof. Price's equation identifies the direction of selection; quasispecies theory yields class-conditional convergence on quotient encoding space under constrained mutation; global convergence follows under primitive mutation with explicit spectral-rate bounds in terms of the fitness gap. A finite-population Wright–Fisher link completes the picture (Theorems 7.4, 7.6 and 7.10).

(iii) A graded separation cascade: resolved complexity k_T is monotone non-decreasing as the task ecology diversifies (Proposition 4.6).

(iv) Recovery of Hoffman's FBT as the single-task special case and Berke et al.'s simulations as a numerical instance, plus a heuristic FMB connection (Appendix C).

Our contribution is a synthesis built on these strands: we adapt the shared-representation insight from statistical learning theory (Baxter, 2000; Maurer et al., 2016) to the Hoffman/Berke evolutionary-perception setting, formalise it through the separation condition and graded cascade, prove deterministic mean-field convergence under evolutionary dynamics via quasispecies theory (with finite-population approximation guarantees), and show that the resulting framework resolves the Hoffman/Berke debate while providing a mathematical formalisation of aspects of von Uexüll's (2010) species-specific perceptual world (Umwelt).

The paper is organised as follows. Section 2 sets up the framework: world states, encodings, tasks, Bayes risk. Section 3 defines the separation condition and the task distance. Section 4 proves the static optimality theorems, including the main veridicality result, the lossy case, the phase transition, and the graded cascade. Section 5 provides finite-sample concentration bounds. Section 6 develops the Gaussian task model and spectral analysis. Section 7 proves mean-field evolutionary convergence via Price's equation and quasispecies theory to dominant mutation-accessible classes (global convergence under primitive mutation), and states the finite-population approximation on fixed horizons. Section 8 summarises the recovery of prior results as special cases, with full details in Appendix E. Section 9 discusses the implications, including the distinction between static optimality and evolutionary dynamics (§9.3), the ecological Umwelt interpretation (§9.4), and limitations (§9.5), with philosophy-of-evolution considerations woven into those sections. Appendix C gives a heuristic connection to Frank's FMB law, and Appendix D gives the full reducible-case proof for the primitive-block setting plus a periodic-class extension.

1.1 Related Work and Intellectual Genealogy

The qualitative motivation for our main result is consistent with statistical learning theory: shared representations can improve cross-task generalisation. Baxter (2000) analysed inductive bias learning in environments of related tasks and derived sample-complexity improvements from multi-task training. Maurer et al. (2016) derived finite-sample excess-risk bounds characterising regimes where multi-task representation learning outperforms independent task learning. In the information-theoretic literature, related ideas appear as optimal data reduction via sufficient statistics.

Our contribution is not the discovery of this principle but its embedding in the evolutionary perception framework of Hoffman et al. (2015) and the resolution of specific open questions within that debate. The following table clarifies the correspondence:

This paper	Multi-task learning theory
World states W	Input space $\mathcal{I}$
Encoding $p\colon W \to X$	Shared representation $\varphi\colon \mathcal{I} \to Z$
Task $f\colon W \to \mathbb{R}$	Task-specific loss function ℓ_t
Separation condition (Definition 3.3)	Shared-representation condition
Equivalence classes $[w]_\mu$	Learned invariances / null space of task family
Ecological veridicality	Sufficiency of shared representation
Separation margin δ_μ	Task diversity / eigenvalue gap

What the learning theory literature does *not* provide, and what we contribute here, is:

1. **The connection to Hoffman's FBT theorem.** The FBT result is formulated in a specific evolutionary game with fitness functions, perception maps, and reproductive dynamics. Prakash et al. (2020) further proved that single-task payoff functions generically fail to preserve mathematical structures. Showing that multi-task optimality *reverses* this, that the composite selective pressure across tasks restores exactly the structure the task ecology demands, requires reformulating the learning-theoretic principle within this game, which involves different definitions, different loss structures, and different notions of optimality.

2. **Evolutionary convergence guarantees.** The learning theory results concern statistical risk minimisation by an algorithm. Our Theorems 7.4 and 7.6 analyse the deterministic mean-field natural-selection recursion (replicator-mutator on quotient encoding space), proving convergence to the dominant mutation-accessible optimum and, under primitive mutation, to the global optimum. Theorem 7.10 links this to finite-population Wright–Fisher dynamics via a law-of-large-numbers limit. Rate is controlled by the spectral ratio and hence by the fitness gap; in full-separation feasible-veridical primitive regimes

this yields explicit δ_μ-dependent bounds. This requires quasispecies theory and spectral analysis, not PAC-Bayes bounds.

3. **The graded cascade and ecological Umwelt.** The learning theory does not address the evolutionary question of how veridicality increases as the task ecology diversifies over phylogenetic time. Proposition 4.6 and §9.4 provide this developmental-evolutionary picture.

4. **A heuristic bridge to Frank's (2025) FMB law.** Identifying Price's equation as a conceptual bridge between evolutionary dynamics and Frank's force-metric-bias decomposition is specific to the evolutionary setting and has no analogue in the learning theory.

We present this paper as a focused synthesis drawing on statistical learning theory, evolutionary game theory, and philosophy of evolution, with new results specific to the fixed-encoding multi-task setting analyzed here. Philosophically, we treat the theorems as conditional adaptation claims in the sense emphasized by Sober's evidential and methodological analyses (Sober, 2008, 2024): evidential support is model-comparative, and population-level laws require initial conditions and auxiliaries before yielding concrete predictions. This also aligns with Pigliucci and Kaplan's argument that adaptationist explanation must be assessed against alternative causal pathways and developmental structure, not inferred directly from statistical fit alone (Pigliucci and Kaplan, 2006, prelude; chs. 2 and 5). Likewise, following Godfrey-Smith's model-based treatment of Darwinian populations (Godfrey-Smith, 2009, 2024), the framework is an idealized map from ecological structure to selective pressure, not a complete biological reconstruction.

2 Framework

2.1 World, Encoding, Tasks

We begin by fixing the basic objects of the theory. The setup follows Hoffman et al. (2015) and Berke et al. (2022): there is a finite world that the agent cannot observe directly, a finite set of internal percepts through which it represents the world, and a collection of tasks that the world imposes. The key structural assumption, cognitive impenetrability, is that the mapping from world to percept is fixed; only the downstream response to percepts can vary across tasks.

Definition 2.1 (World). *Let $W = \{w_1, \ldots, w_N\}$ be a finite set of* world states *equipped with a prior distribution π, where $\pi(w) > 0$ for all $w \in W$. Write $\pi_{\min} = \min_w \pi(w)$.*

Definition 2.2 (Encoding). *Let $X = \{x_1, \ldots, x_M\}$ be a finite set of* percepts*. An* encoding *is any function $p\colon W \to X$ (not necessarily surjective). The encoding induces a partition of W into non-empty fibres (cells) $\mathcal{C} = \{C_x : x \in \operatorname{Im}(p)\}$, where $C_x = p^{-1}(x) = \{w \in W : p(w) = x\}$. We write $m(p) = |\operatorname{Im}(p)| \leq M$ for the number of percept values actually used by p.*

An encoding's partition of W into cells is what matters for performance; the specific labels attached to percepts are arbitrary. The next definition makes this symmetry precise, since uniqueness results throughout the paper hold only "up to relabeling." The formal consequence (risk invariance under relabeling) is stated in Lemma 2.12 after the Bayes risk has been defined.

Definition 2.3 (Percept-label symmetry and quotient)**.** *Let S_M be the permutation group on X. It acts on encodings by relabeling percepts:*

$$(\sigma \cdot p)(w) = \sigma(p(w)), \quad \sigma \in S_M. \tag{1}$$

Define orbit equivalence $p \sim_X p'$ iff $p' = \sigma \cdot p$ for some $\sigma \in S_M$. The orbit space is

$$\bar{\Omega} = \Omega/{\sim_X}, \quad \textit{where } \Omega = X^W. \tag{2}$$

Elements of $\bar{\Omega}$ are encoding classes modulo percept-label names.

With the encoding defined, we can state the central question precisely. An encoding can be more or less faithful to the structure of the world; the following definition distinguishes three grades of fidelity that the theory will characterise.

Definition 2.4 (Veridicality hierarchy)**.**

- ***Full veridicality:*** *p is injective on W (possible iff $M \geq N$).*
- ***Ecological veridicality (relative to μ):*** *p is injective on the quotient $W/{\sim_\mu}$, equivalently $p(w_1) = p(w_2)$ implies $w_1 \sim_\mu w_2$. We abbreviate this to* eco-veridical.
- ***Lossy-optimal encoding at capacity M:*** *any minimiser of the multi-task Bayes risk $R(p)$ (defined in Definition 2.10) over all functions $p\colon W \to X$ with $|X| = M$.*

When μ separates all points, ecological veridicality coincides with full veridicality. When $M < |W/{\sim_\mu}|$, ecological veridicality is infeasible and the relevant notion is lossy optimality.

The remaining ingredient is what the agent must *do* with its percepts. We model the agent's ecological demands as a distribution over tasks.

Definition 2.5 (Task)**.** *A* task *is a bounded function $f\colon W \to [-B, B]$ for some $B > 0$. We identify tasks with vectors $f \in \mathbb{R}^N$. A* task distribution *is a probability measure μ on $\mathbb{R}^N$.*

Defining tasks as scalar-valued functions $f\colon W \to \mathbb{R}$ restricts attention to point estimation of a single quantity. This is the framework used by Hoffman et al. (2015) and Prakash et al. (2021) in formulating the FBT theorem, and we adopt it for direct comparability. Real biological tasks (categorisation, relational judgment, motor planning) are richer. The extension to vector-valued tasks $f\colon W \to \mathbb{R}^D$ is straightforward: the task distance becomes $\sigma^2(w_1, w_2) = \mathbb{E}_\mu[\|f(w_1) - f(w_2)\|^2]$, and the entire theory carries through with D-dimensional variance replacing scalar variance. Tasks with categorical or non-numeric outcomes require replacing squared error with an appropriate proper scoring rule (see the discussion of loss-function dependence after Theorem 4.1). The qualitative structure of the theory (separation, ecological veridicality, evolutionary convergence) does not depend on the scalar restriction.

Note that μ is an arbitrary probability measure, not a uniform one. Tasks that are ecologically more frequent, more fitness-consequential, or both, receive greater weight under μ. This weighting propagates through the entire theory: the task distance $\sigma^2(w_1, w_2) = \mathbb{E}_{f\sim\mu}[(f(w_1)-f(w_2))^2]$ is a μ-weighted average, so pairs of world states that high-weight tasks distinguish contribute more to separation than pairs that only low-weight tasks distinguish. In the lossy case ($M < N$, Theorem 4.2), the optimal partition allocates limited perceptual capacity preferentially to distinctions that high-weight tasks demand. For example, a raptor whose task ecology is dominated by predation (detecting prey against a cluttered background) will have an optimal encoding that finely discriminates motion-related features at the expense of texture distinctions among stationary objects, not because the encoding is "distorted," but because it is ecologically veridical with respect to a μ that heavily weights predation. The apparent distortion (e.g. telephoto acuity at the fovea, poor peripheral texture resolution) is the theory's prediction, not a deviation from it.

Because an organism cannot redesign its sensory transduction for each new task, we require the encoding to be shared across all tasks. This is the *cognitive impenetrability* assumption (Pylyshyn, 1999), which grounds the entire framework: it is what makes multi-task performance, rather than single-task performance, the relevant criterion.

Assumption 2.6 (Cognitive impenetrability)**.** *The encoding p is fixed across all tasks. Only the downstream* readout *(the mapping from percepts to actions) may vary per task.*

2.2 Bayes Risk

Because the encoding is shared, the relevant performance measure is the average over the task ecology, not the risk on any single task. An encoding is "good" to the extent that it allows accurate estimation of task values from percepts alone. The natural loss is squared error, and the natural benchmark is the best possible estimate given the percept: the Bayes-optimal estimate.

Given encoding p and task f, an agent observing percept $x = p(w)$ must estimate $f(w)$. Under squared-error loss, the optimal estimate is the conditional expectation:

$$\hat{f}(x) = \mathbb{E}[f(w) \mid p(w) = x] = \sum_{w\in C_x} \pi(w \mid C_x)\, f(w), \tag{3}$$

where $\pi(w \mid C_x) = \pi(w)/\pi(C_x)$ and $\pi(C_x) = \sum_{w\in C_x} \pi(w)$.

Definition 2.7 (Single-task Bayes risk)**.** *The Bayes risk of encoding p on task f is:*

$$R(p, f) = \mathbb{E}_w\Big[(f(w) - \hat{f}(p(w)))^2\Big] = \sum_x \pi(C_x) \cdot \mathrm{Var}(f \mid C_x), \tag{4}$$

where $\mathrm{Var}(f \mid C_x) = \sum_{w\in C_x} \pi(w \mid C_x)\,[f(w) - \hat{f}(x)]^2$ *is the within-cell conditional variance.*

The Bayes risk has a useful decomposition into "signal" and "noise" that will be needed for the pairwise formula below.

Lemma 2.8 (Variance decomposition)**.** *For any encoding p and task f:*

$$\mathrm{Var}(f) = \sum_x \pi(C_x) \cdot \mathrm{Var}(f \mid C_x) + \mathrm{Var}(\hat{f}(p(w))). \tag{5}$$

That is: total variance = within-cell variance (= Bayes risk) + between-cell variance (= explained variance).

Proof. This is the law of total variance: $\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y \mid Z)] + \mathrm{Var}(\mathbb{E}[Y \mid Z])$ applied with $Y = f(w)$ and $Z = p(w)$. □

Corollary 2.9. *$R(p, f) = 0$ if and only if f is constant on each cell C_x. That is, f is p-measurable.*

Proof. $R(p, f) = \sum_x \pi(C_x)\, \mathrm{Var}(f \mid C_x) = 0$ iff $\mathrm{Var}(f \mid C_x) = 0$ for all x (since $\pi(C_x) > 0$), iff f is constant on each C_x. □

So far we have measured performance on a single task. Since the encoding is fixed across all tasks (cognitive impenetrability), the relevant performance measure is the *average* over the task ecology.

Definition 2.10 (Multi-task Bayes risk)**.** *The expected Bayes risk over the task distribution:*

$$R(p) = \mathbb{E}_{f\sim\mu}[R(p, f)] = \sum_x \pi(C_x) \cdot \mathbb{E}_\mu[\mathrm{Var}(f \mid C_x)]. \tag{6}$$

2.3 The Pairwise Decomposition

The multi-task Bayes risk admits a revealing pairwise decomposition that will be the workhorse of the static optimality theorems. It shows that the risk of an encoding is determined entirely by which pairs of world states it merges and how distinguishable those pairs are across tasks.

Lemma 2.11 (Pairwise Bayes risk formula)**.** *For any encoding p:*

$$R(p) = \frac{1}{2} \sum_x \frac{1}{\pi(C_x)} \sum_{w_1, w_2 \in C_x} \pi(w_1)\, \pi(w_2)\, \sigma^2(w_1, w_2), \tag{7}$$

where $\sigma^2(w_1, w_2) = \mathbb{E}_{f\sim\mu}[(f(w_1) - f(w_2))^2]$.

Proof. We use the identity: for any discrete random variable Z with values $\{z_i\}$ and probabilities $\{p_i\}$:

$$\mathrm{Var}(Z) = \frac{1}{2} \sum_{i,j} p_i\, p_j\, (z_i - z_j)^2. \tag{8}$$

Applied to f restricted to cell C_x with weights $\pi(\cdot \mid C_x)$:

$$\mathrm{Var}(f \mid C_x) = \frac{1}{2} \sum_{w_1, w_2 \in C_x} \pi(w_1 \mid C_x)\, \pi(w_2 \mid C_x)\, (f(w_1) - f(w_2))^2.$$

Taking $\mathbb{E}_\mu$ and using linearity (the exchange of expectation and summation is justified by Fubini's theorem, since all terms are non-negative and the sums are finite):

$$\mathbb{E}_\mu[\mathrm{Var}(f \mid C_x)] = \frac{1}{2} \sum_{w_1, w_2 \in C_x} \pi(w_1 \mid C_x)\, \pi(w_2 \mid C_x)\, \sigma^2(w_1, w_2).$$

Substituting into (6) and using $\pi(w_i \mid C_x) = \pi(w_i)/\pi(C_x)$ gives (7). □

Interpretation. *The Bayes risk is a weighted sum of $\sigma^2(w_1, w_2)$ over all pairs of world states that are* merged *(placed in the same cell). The weight for a merged pair $(w_1, w_2) \in C_x$ is $\pi(w_1)\pi(w_2)/\pi(C_x)$, which is always positive. Every merged pair of task-distinguishable states contributes positively to Bayes risk.*

We can now verify the claim from Definition 2.3 that relabeling percepts does not affect performance.

Lemma 2.12 (Risk invariance under relabeling)**.** *For every $\sigma \in S_M$ and encoding p:*

$$R(\sigma \cdot p) = R(p). \tag{9}$$

Proof. Relabeling only permutes cell names; the underlying partition $\{C_x\}$ of W is unchanged. Both the single-task risk (4) and the multi-task risk (6) depend on p only through this partition and the associated π-masses, hence are invariant under σ. □

Corollary 2.13 (Uniqueness on the quotient)**.** *Any statement of uniqueness "up to percept-label permutation" is equivalent to ordinary uniqueness on $\bar{\Omega}$.*

3 The Separation Condition

The pairwise formula (Lemma 2.11) shows that Bayes risk is controlled by a single object: the expected squared difference $\sigma^2(w_1, w_2)$ between task values at pairs of world states. This quantity plays the role of a distance on W induced by the task ecology. Pairs with $\sigma^2 > 0$ are *separated* by the task distribution, merging them incurs risk, while pairs with $\sigma^2 = 0$ are *equivalent*, merging them is free. The separation condition is the bridge between the task ecology and the structure of optimal encodings: it determines which distinctions the encoding must preserve and which it may discard.

Definition 3.1 (Task distance)**.** *For $w_1, w_2 \in W$, define:*

$$\sigma^2(w_1, w_2) = \mathbb{E}_{f\sim\mu}\Big[(f(w_1) - f(w_2))^2\Big]. \tag{10}$$

This is the expected squared difference in task values between w_1 and w_2.

Lemma 3.2. *The function $d(w_1, w_2) = \sigma(w_1, w_2)$ is a pseudo-metric on W.*

Proof. (i) $d(w, w) = 0$ since $f(w) - f(w) = 0$. (ii) $d(w_1, w_2) = d(w_2, w_1)$ trivially. (iii) Triangle inequality: by Minkowski's inequality in $L^2(\mu)$,

$$\|f(w_1) - f(w_3)\|_{L^2(\mu)} \le \|f(w_1) - f(w_2)\|_{L^2(\mu)} + \|f(w_2) - f(w_3)\|_{L^2(\mu)},$$

i.e. $\sigma(w_1, w_3) \le \sigma(w_1, w_2) + \sigma(w_2, w_3)$. □

That σ is a pseudo-metric means it behaves like a distance except that distinct points can have distance zero. The zero-distance pairs define the equivalence classes that the encoding is free to merge. This is formalised next.

Definition 3.3 (Separation)**.** μ separates w_1 *from* w_2 *if* $\sigma^2(w_1, w_2) > 0$*, equivalently if* $f(w_1) \neq f(w_2)$ *on a set of tasks of positive* μ*-measure. The distribution* μ separates points *(or is* point-separating*) if* $\sigma^2(w_1, w_2) > 0$ *for all* $w_1 \neq w_2$*.*

When μ separates all pairs, the minimum gap quantifies how "hard" it is to distinguish the closest pair of states. This quantity controls the fitness gap and convergence rates in later sections.

Definition 3.4 (Separation margin)**.** *When* μ *separates points:*

$$\delta_\mu = \min_{w_1 \neq w_2} \sigma^2(w_1, w_2) > 0. \tag{11}$$

The zero-sets of the task distance define the "grain" of the task ecology: the level of detail it demands from the encoding.

Definition 3.5 (Task equivalence)**.** *Define* $w_1 \sim_\mu w_2$ *iff* $\sigma^2(w_1, w_2) = 0$*. This is an equivalence relation (reflexivity and symmetry are obvious; transitivity follows from the triangle inequality for* σ*). Equivalently,* $w_1 \sim_\mu w_2$ *iff* $f(w_1) = f(w_2)$ *for* μ*-almost every task* f*. The equivalence classes* $[w]_\mu$ *partition* W *into groups of states indistinguishable up to* μ*-null task sets.*

There is a useful algebraic reformulation. Embed tasks as vectors: define $\Phi \colon W \to L^2(\mu)$ by $\Phi(w) = [f \mapsto f(w)]$. Then

$$\sigma^2(w_1, w_2) = \|\Phi(w_1) - \Phi(w_2)\|^2_{L^2(\mu)}. \tag{12}$$

Separation means Φ is injective. The task distance is the $L^2(\mu)$ distance between the images.

4 Static Optimality Theorems

With the framework in place, we can now state and prove the main results on which encodings minimise multi-task Bayes risk. These are *static* results: they characterise the optimal encoding as a function of the task ecology μ and the capacity M, without reference to any process that might find it. The dynamic question, whether evolution converges to such an optimum, is deferred to Section 7.

The central result (Theorem 4.1) is that zero Bayes risk is achievable if and only if the encoding preserves all task-relevant distinctions (ecological veridicality), and that any encoding that merges a separated pair incurs a quantifiable risk penalty. Theorem 4.2 extends this to the capacity-constrained case. Theorem 4.4 and Proposition 4.6 then analyse how the number of relevant distinctions grows as the task ecology diversifies.

Theorem 4.1 (Multi-Task Veridicality (Equal Complexity))**.** *Let* $k_\mu = |W/{\sim_\mu}|$*. Then:*

(a) $R(p) = 0$ *if and only if for every cell* C_x*, all elements of* C_x *are in the same* μ*-equivalence class* $[w]_\mu$*. Equivalently,* p *is ecologically veridical (injective on* $W/{\sim_\mu}$*).*

(b) Hence $R(p) = 0$ *is achievable if and only if* $M \geq k_\mu$*. In particular:*

- *If μ separates all points ($\delta_\mu > 0$, so $k_\mu = N$): $R(p) = 0$ if and only if p is injective on W (full veridicality).*
- *If μ does not separate all points: zero-risk encodings are exactly those that merge only μ-equivalent states.*

(c) If μ separates points, then for any encoding p that merges at least one pair $w_1 \neq w_2$:

$$R(p) \geq \pi_{\min}^2 \cdot \delta_\mu > 0. \tag{13}$$

Consequently, under full separation and $M \geq N$, the fully veridical encoding is the unique minimiser up to a permutation of percept labels.

Proof. **(a), $\Longrightarrow$:** If every cell C_x contains only μ-equivalent states, then for any pair $w_1, w_2 \in C_x$ we have $\sigma^2(w_1, w_2) = 0$, i.e. $f(w_1) = f(w_2)$ for μ-a.e. f. So $\mathrm{Var}(f \mid C_x) = 0$ for μ-a.e. f, and hence $\mathbb{E}_\mu[\mathrm{Var}(f \mid C_x)] = 0$ for every x. Thus $R(p) = 0$.

(a), $\Longleftarrow$: Suppose $R(p) = 0$. Then $\mathbb{E}_\mu[\mathrm{Var}(f \mid C_x)] = 0$ for all x (since $\pi(C_x) > 0$). This means $\mathrm{Var}(f \mid C_x) = 0$ for μ-a.e. f. For a cell $C_x = \{w_1, \ldots, w_k\}$ with $k \geq 2$, $\mathrm{Var}(f \mid C_x) = 0$ requires $f(w_1) = f(w_2) = \cdots = f(w_k)$ for μ-a.e. f. Taking expectations: $\sigma^2(w_i, w_j) = \mathbb{E}_\mu[(f(w_i) - f(w_j))^2] = 0$ for all pairs i, j within C_x. So every merged pair is μ-equivalent.

(c): Assume μ separates points. Let p merge at least one distinct pair. Then $\exists$ distinct $w_1, w_2 \in W$ with $p(w_1) = p(w_2) = x_0$. By Lemma 2.11, the contribution of this pair to $R(p)$ is:

$$R(p) \geq \frac{\pi(w_1)\,\pi(w_2)}{\pi(C_{x_0})} \cdot \sigma^2(w_1, w_2). \tag{14}$$

Since $\sigma^2(w_1, w_2) \geq \delta_\mu > 0$, and $\pi(w_1) \geq \pi_{\min}$, $\pi(w_2) \geq \pi_{\min}$, $\pi(C_{x_0}) \leq 1$:

$$R(p) \geq \pi_{\min}^2 \cdot \delta_\mu. \qquad \square$$

Loss-function dependence. The results in this paper divide into two categories with different degrees of generality:

Loss-general results (hold for any strictly proper loss function):

- Theorem 4.1(a): $R(p) = 0$ iff p merges only μ-equivalent states. This is an information-theoretic statement: under any strictly proper loss, the Bayes risk vanishes iff the task value is measurable with respect to the encoding partition.
- Proposition 4.3: Gauge freedom within equivalence classes.
- Theorem 4.4 and Proposition 4.6: The phase transition and graded cascade, which depend on the rank of the task matrix, not on the loss.
- The qualitative direction of selection (Theorem 7.4): merging separated states incurs positive Bayes risk under any proper loss, so selection always pushes against merging.

Squared-error-specific results:

- Theorem 4.1(c): The specific lower bound $R(p) \geq \pi_{\min}^2 \cdot \delta_\mu$.
- The pairwise decomposition (Lemma 2.11) and the spectral analysis of Section 6.
- The Hoeffding concentration bounds of Section 5.
- The quantitative convergence rate in Theorem 7.6(c).

Under a different loss, the task "distance" between world states becomes a different divergence (e.g. total variation, Hellinger, or a loss-specific Bregman divergence), and the pairwise decomposition, spectral structure, and concentration bounds take different forms. The qualitative structure of the theory (separation condition, fitness gap, and evolutionary convergence) is preserved, but the quantitative bounds are not portable. We work throughout with squared error because it is the standard benchmark in the FBT literature (Hoffman et al., 2015; Prakash et al., 2021), because it yields the cleanest algebra, and because it permits the spectral analysis of Section 6. Extending the quantitative theory to specific biologically motivated loss functions (e.g. asymmetric losses reflecting the "smoke detector principle") is a natural direction for future work; the key observation is that asymmetric losses that catastrophically penalise certain errors can only *increase* the cost of merging the relevant states, strengthening the selective pressure toward ecological veridicality for those distinctions.

Theorem 4.1 assumes that the percept space is large enough to represent all equivalence classes. What happens when it is not, that is, when $M < k_\mu$ and the agent *must* merge some separated states? The next result shows that the optimisation problem reduces to a finite weighted partitioning problem.

Theorem 4.2 (Lossy Case (Finite Partition Optimisation))**.** *Let Π_M be the finite set of partitions of W into at most M non-empty cells. For $P = \{C_1, \ldots, C_m\} \in \Pi_M$ define*

$$J(P) := \frac{1}{2} \sum_{C \in P} \frac{1}{\pi(C)} \sum_{w_1, w_2 \in C} \pi(w_1)\, \pi(w_2)\, \sigma^2(w_1, w_2). \tag{15}$$

Then:

(a) For every encoding $p\colon W \to X$ with $|X| = M$, if $P(p)$ is its induced partition, $R(p) = J(P(p))$.

(b) Conversely, for every $P \in \Pi_M$ there exists an encoding p_P with $P(p_P) = P$ and $R(p_P) = J(P)$.

(c) Therefore

$$\min_{p\colon W \to X,\, |X| = M} R(p) = \min_{P \in \Pi_M} J(P), \tag{16}$$

and a minimiser exists (finite search space).

When $M < k_\mu$ (in particular $M < N$ under full separation), every feasible encoding is lossy relative to μ and the optimum is the best M-cell partition under J.

Here p_P denotes any encoding that realises partition P (it is generally not unique).

Proof. (a) is exactly Lemma 2.11 rewritten by cells, so risk depends on p only through $P(p)$. (b) Given P with $m \leq M$ cells, assign each cell a distinct label in X and map each w to its cell label; unused labels are allowed. Then $P(p_P) = P$ and (a) gives $R(p_P) = J(P)$. (c) (a)–(b) establish equivalence between optimisation over encodings and optimisation over Π_M. Since W is finite, Π_M is finite, so a minimiser exists. □

Interpretation. *This is a finite weighted partitioning problem (an analogue of k-means distortion minimisation with fixed cluster count), where distortion is induced by the task distance and prior-weighted within-cell pair penalties.*

The cardinality $M = |X|$ of the percept space is a structural constraint reflecting the organism's metabolic and channel-capacity limits. Maintaining M distinct perceptual states requires neural architecture, receptor diversity, channel bandwidth, cortical representational area, that carries a metabolic cost. Our framework parameterises this cost through fixed capacity M rather than through an explicit penalty term. Theorem 4.1 characterises when zero Bayes risk is feasible ($M \geq k_\mu$); Theorem 4.2 characterises the optimum at any constrained budget, especially $M < k_\mu$. The question "what is the optimal M?" is a separate problem involving metabolic trade-offs that our framework does not address, but neither does Hoffman's FBT, which also fixes perceptual capacity. The lossy k-means structure of Theorem 4.2 is closely related to rate-distortion theory: the optimal M-partition minimises distortion (Bayes risk) at a given rate ($\log M$ bits of channel capacity). An explicit rate-distortion formulation, with a Lagrange multiplier β on the mutual information $I(W; P)$, would yield the same qualitative structure, aggressive compression of task-equivalent states, faithful separation of task-distinguishable states, with β playing the role of metabolic cost per bit.

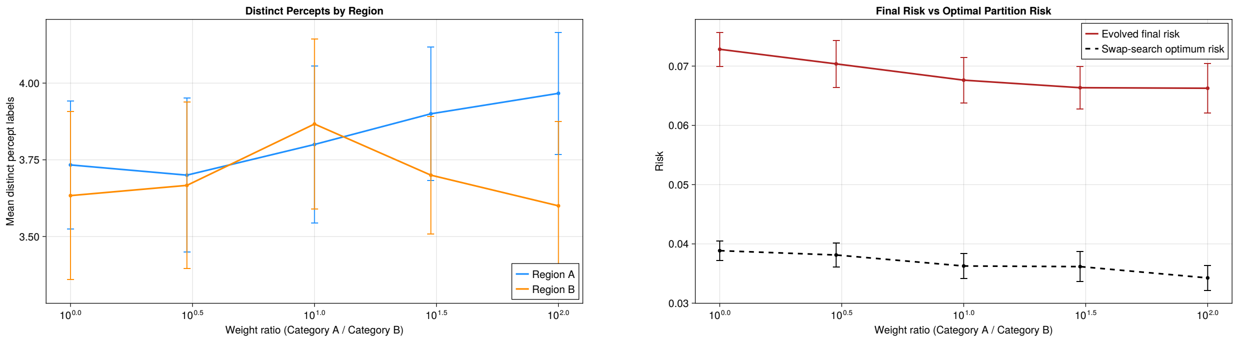
Numerical simulations provide partial support for this weighted-partition prediction. Figure 1(b) shows that evolved final risk tracks the swap-search partition optimum across weight ratios, with a consistent gap reflecting finite-population and finite-generation overhead. Figure 1(a) shows a modest trend in percept allocation: Region A tends to gain distinct labels as its ecological weight increases, though the effect size is small in this parameter regime.

When the task ecology does not separate all points, the optimal encoding is not unique: any assignment of percepts *within* an equivalence class is equally good.

Proposition 4.3 (Gauge degeneracy of ecological optima)**.** *Let $[w]_\mu$ denote a task-equivalence class. Any encoding that is injective on $W/{\sim_\mu}$ has Bayes risk* 0 *(Theorem 4.1), and therefore all such encodings are global minimisers. In particular, when μ does not separate points, there is a nontrivial degenerate set of optimal encodings corresponding to arbitrary assignments* within *each equivalence class, subject only to not merging distinct classes.*

Proof. By Theorem 4.1(a), $R(p) = 0$ iff each cell contains only μ-equivalent states. Therefore every encoding injective on the quotient $W/{\sim_\mu}$ attains the same minimal value 0. If at least one class has cardinality > 1, there are multiple such encodings, giving degeneracy of minimisers. □

Interpretation. *The equivalence classes $[w]_\mu$ define a flat optimum set: any encoding that separates distinct classes but is arbitrary within each class attains the same minimal risk* 0*. This degeneracy disappears when μ separates all points.*



(a) Mean number of distinct percept labels assigned to each world region versus weight ratio. A modest trend toward differential allocation is visible, Region A tends upward and Region B tends downward as the weight ratio increases, though the effect size is small and error bars overlap at most points.

(b) Evolved final risk (red, solid) and swap-search partition optimum (black, dashed) versus weight ratio. Both decrease as the ecology concentrates on one region; the persistent gap reflects finite-population evolutionary overhead.

Figure 1: Effect of ecological weighting in the lossy regime ($M < N$). Both panels share the weight-ratio axis; together they show that evolution shifts representational allocation toward task-heavy regions, as predicted by Theorem 4.2. See Appendix E for full details.

The preceding results concern the population-level task distribution μ. In practice, organisms encounter a finite sample of T tasks. The next result characterises when the empirical analogue of separation holds and when it does not, giving the mechanism behind the phase transitions observed in simulations.

Theorem 4.4 (Empirical Separation and Capacity Thresholds)**.** *Let $f_1, \ldots, f_T$ be T tasks. Form the* task matrix $F \in \mathbb{R}^{T \times N}$ *with $F_{ti} = f_t(w_i)$. Define the empirical task distance:*

$$\sigma_T^2(w_i, w_j) = \frac{1}{T}\sum_{t=1}^{T}(f_t(w_i) - f_t(w_j))^2 = \frac{1}{T}\|F_{\cdot i} - F_{\cdot j}\|^2, \tag{17}$$

where $F_{\cdot i}$ is the i-th column of F. Let $\sim_T$ be the empirical equivalence relation $w_i \sim_T w_j$ iff $\sigma_T^2(w_i, w_j) = 0$, and $k_T = |W/{\sim_T}|$. For each task, write its evaluation vector on W as $v_t := (f_t(w_1), \ldots, f_t(w_N)) \in \mathbb{R}^N$, so $v_t^\top$ is the t-th row of F.

(a) $\sigma_T^2(w_i, w_j) > 0$ for all $i \neq j$ if and only if the columns of F are pairwise distinct (equivalently, $k_T = N$).

(b) $\operatorname{rank}(F) = N$ *is sufficient (not necessary) for pairwise distinct columns.*

(c) Capacity criterion for zero empirical risk:

- *Zero empirical Bayes risk is achievable if and only if $M \geq k_T$.*
- *Full empirical veridicality requires both $k_T = N$ and $M \geq N$.*

(d) Let ν be the sampling law of v_t (the pushforward of the task law under evaluation on W). If $\nu \ll \lambda^N$ (absolutely continuous w.r.t. Lebesgue measure on $\mathbb{R}^N$), then with probability 1 a single task has distinct values on all N states, so $k_1 = N$ and the cascade is trivial. Thus biologically meaningful phase transitions require structured task families.

Proof. (a) Immediate from $\sigma_T^2(w_i, w_j) = 0$ iff $F_{\cdot i} = F_{\cdot j}$.

(b) Linear independence of columns implies pairwise distinctness.

(c) For empirical objective $R_T(p) = (1/T)\sum_t R(p, f_t)$, we have $R_T(p) = 0$ iff $R(p, f_t) = 0$ for every t (nonnegative summands). By Corollary 2.9, this holds iff each task f_t is constant on every cell of p, equivalently iff each cell is contained in an empirical equivalence class of $\sim_T$. Hence one needs at least one percept per $\sim_T$-class, i.e. $M \geq k_T$, and this is also sufficient by assigning one percept per class.

(d) By assumption, $v_1 \sim \nu$ with $\nu \ll \lambda^N$. For each $i \neq j$, the tie event $\{v_{1i} = v_{1j}\}$ is the hyperplane $H_{ij} = \{x \in \mathbb{R}^N : x_i = x_j\}$, which has Lebesgue measure 0. Hence $\nu(H_{ij}) = 0$. A finite union bound over the $\binom{N}{2}$ pairs gives $P(\exists i \neq j : v_{1i} = v_{1j}) = 0$, so the coordinates are pairwise distinct almost surely. □

Corollary 4.5 (Capacity-Aware Transition)**.** *Define empirical ecological complexity at task count T by*

$$k_T := |W/{\sim_T}|.$$

Then the sequence $(k_T)_{T\geq 1}$ satisfies:

- *The relevant transition at fixed capacity M (as T increases) is from* ecologically veridical *($k_T \leq M$) to* lossy-optimal *($k_T > M$).*
- *Full veridicality is a special case requiring $k_T = N$ and $M \geq N$.*
- *There is no universal critical task number T^* independent of task family; instead, $T_{\text{loss}}(M) = \inf\{T : k_T > M\}$, and its value is family-dependent.*

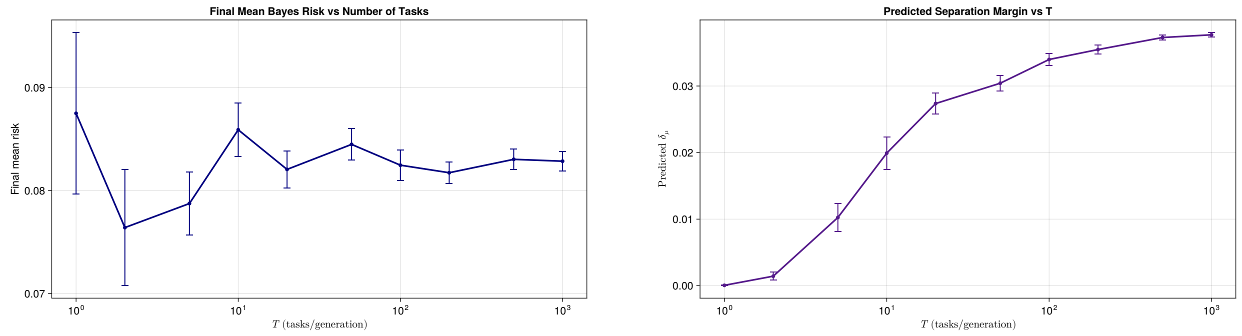
A key qualitative prediction of the theory is that adding tasks can only *increase* the number of distinctions the encoding must represent, never decrease it. This monotonicity gives the "graded cascade" from coarse to fine perception as the task ecology diversifies.

Proposition 4.6 (Graded separation cascade)**.** *Let $f_1, f_2, \ldots$ be a sequence of tasks and define k_T as above. Then:*

(a) k_T is non-decreasing in T: $k_{T+1} \geq k_T$.

(b) $1 \leq k_T \leq N$.

(c) For fixed M, no monotonic law holds in general for the averaged *optimum $R^*(M,T) = \min_p R_T(p)$ when R_T is the sample average over T tasks: adding one task changes both numerator and denominator. (A monotone statement does hold for the cumulative objective $S_T(p) = \sum_{t\leq T} R(p, f_t)$, for which $\min_p S_T$ is non-decreasing.)*



(a) Final mean Bayes risk versus number of sampled tasks per generation ($\log T$ axis). Risk drops sharply from $T = 1$ to $T = 2$, then partially rebounds before settling to a plateau around the $M = 2$ capacity floor. Error bars show run-level standard deviation across replicates.

(b) Estimated empirical separation margin δ_T versus number of tasks. The trend is non-decreasing in expectation, matching Proposition 4.6's graded-cascade prediction.

Figure 2: Two complementary views of the graded cascade (Proposition 4.6) in a Berke-style setting ($N = 11$, $M = 2$, beta-function tasks). See Appendix E for the full recovery analysis.

Proof. (a) If $\sigma_T^2(w_i, w_j) > 0$, then

$$\sigma_{T+1}^2(w_i, w_j) = \frac{T}{T+1}\,\sigma_T^2(w_i, w_j) + \frac{1}{T+1}(f_{T+1}(w_i) - f_{T+1}(w_j))^2 > 0.$$

So distinctions cannot be lost.

(b) Immediate.

(c) For any fixed partition p, $R_{T+1}(p)$ is the average of $T+1$ nonnegative terms and need not be $\leq R_T(p)$. Since the minimisation is over the same feasible set, monotonic decrease is not guaranteed. □

Proposition 4.6 gives only monotonicity and bounds. The *shape* of the staircase $T \mapsto k_T$ is model-dependent: unstructured continuous task families tend to yield $k_1 = N$ almost surely (Theorem 4.4(d), trivial cascade), while structured families can exhibit gradual growth, with transition width governed by geometric anisotropy of the task family.

Condition number and transition width. In structured families, transition sharpness is controlled by anisotropy of task variation. In the Gaussian-linear model (Section 6), this is captured by $\kappa = \lambda_{\max}/\lambda_{\min}$ of Σ_c on the relevant subspace: $\kappa \approx 1$ gives sharp transitions; $\kappa \gg 1$ gives broad transitions. This is a family-specific quantitative prediction, not a universal law across all task families.

Define the *empirical separation margin* $\delta_T = \min_{i \neq j} \sigma_T^2(w_i, w_j)$, the finite-sample analogue of δ_μ. Figure 2 illustrates the graded cascade empirically: panel (a) shows mean Bayes risk dropping sharply at low T and then fluctuating near a capacity-constrained plateau, while panel (b) tracks the monotone growth of δ_T, matching Proposition 4.6's prediction.

5 Concentration and Finite-Sample Guarantees

The static theorems of Section 4 characterise optimality with respect to the population-level risk $R(p) = \mathbb{E}_{f\sim\mu}[R(p,f)]$. An organism facing T sampled tasks per generation sees the empirical risk $\bar{R}_T(p)$, which fluctuates around $R(p)$. How many tasks suffice for the empirical risk to reliably distinguish veridical from non-veridical encodings? The answer is a standard Hoeffding-plus-union-bound concentration argument (Hoeffding, 1963), polynomial rather than exponential in N, that justifies the practical relevance of the separation margin δ_μ.

Theorem 5.1 (Exponential Convergence with T Tasks)**.** *Let $f_1, \dots, f_T$ be drawn iid from μ, with $|f(w)| \le B$ for all w, f. For any fixed encoding p with $R(p) > 0$:*

$$P\Big(\bar{R}_T(p) \le \eta\Big) \le \exp\Big(-\frac{2T(R(p)-\eta)^2}{B^4}\Big), \quad \textit{for } \eta < R(p), \tag{18}$$

where $\bar{R}_T(p) = (1/T)\sum_{t=1}^{T} R(p, f_t)$ is the empirical Bayes risk and $\eta > 0$ is a tolerance threshold.

Proof. The random variables $R(p, f_t)$ are iid with $\mathbb{E}[R(p, f_t)] = R(p) > 0$ and $0 \le R(p, f_t) \le B^2$ (since $\mathrm{Var}(f \mid C_x) \le \mathbb{E}[f^2 \mid C_x] \le B^2$). By Hoeffding's inequality for bounded iid variables in $[0, B^2]$ (Hoeffding, 1963):

$$P\Big(\bar{R}_T - R(p) \le -(R(p)-\eta)\Big) \le \exp\Big(-\frac{2T(R(p)-\eta)^2}{(B^2)^2}\Big),$$

which gives the result. □

The per-encoding bound becomes a *uniform* guarantee by a union bound over all possible encodings, using the fitness gap from Theorem 4.1(c).

Corollary 5.2 (Uniform Bound over All Non-Veridical Encodings)**.** *Assume full separation ($\delta_\mu > 0$, so $k_\mu = N$) and $|W| = |X| = N$. Then the number of encodings is N^N, and the probability that any encoding that merges at least one separated pair has empirical Bayes risk below η is:*

$$P\Big(\exists\, p \textit{ that merges a separated pair} : \bar{R}_T(p) \le \eta\Big) \le N^N \cdot \exp\Big(-\frac{2T(\pi_{\min}^2\delta_\mu - \eta)^2}{B^4}\Big). \tag{19}$$

Setting the right side $\le \alpha$ and solving:

$$T \ge \frac{B^4}{2(\pi_{\min}^2\delta_\mu - \eta)^2} \cdot \Big(N \log N + \log(1/\alpha)\Big). \tag{20}$$

Proof. Union bound over at most N^N encodings, each satisfying Theorem 5.1 with $R(p) \ge \pi_{\min}^2\delta_\mu$. □

Under the full-separation assumptions of Corollary 5.2, this bound is deliberately worst-case: it uses the minimum Bayes risk gap $\pi_{\min}^2\delta_\mu$ for every non-veridical encoding, whereas most non-veridical encodings have much larger Bayes risk (merging multiple well-separated pairs). It also uses the crude union bound over all N^N encodings, ignoring correlations between overlapping encodings. The bound is therefore most useful as a qualitative guarantee that exponentially many tasks are *not* required, polynomial in N suffices, rather than as a practical estimate of the transition point for specific task families. Equation (20) should be interpreted as an asymptotic existence guarantee, not as a quantitative predictor of biological phase-transition locations. For the latter, the spectral analysis of Section 6 and the condition-number discussion above provide sharper, family-specific predictions.

6 Gaussian Task Model and Spectral Analysis

The results of Sections 3 to 5 hold for arbitrary task distributions μ. In this section, we develop a specific parametric example, Gaussian-linear tasks, to illustrate how the abstract quantities (σ^2, δ_μ, the separation condition) can be computed in closed form and related to the spectral structure of the task covariance. This is one tractable instantiation, chosen for its analytical transparency; other task families (e.g. the beta-function tasks of Berke et al.) yield the same qualitative structure but require numerical computation of σ^2 and δ_μ rather than closed-form expressions. Nothing in the core theory (Theorems 4.1, 4.2, 7.4 and 7.6) depends on the Gaussian assumption.

6.1 Setup

Assign each world state a feature vector: $w_i \mapsto \varphi_i \in \mathbb{R}^D$. Tasks are random linear functions:

$$f(w_i) = c^\top \varphi_i, \quad c \sim \mathcal{N}(0, \Sigma_c), \tag{21}$$

where $\Sigma_c \in \mathbb{R}^{D\times D}$ is the task covariance matrix (positive semi-definite).

Theorem 6.1 (Task Distance under Gaussian Model)**.**

$$\sigma^2(w_i, w_j) = (\varphi_i - \varphi_j)^\top \Sigma_c \, (\varphi_i - \varphi_j). \tag{22}$$

This is the squared Mahalanobis distance between φ_i and φ_j under Σ_c.

Proof.

$$\begin{aligned} \sigma^2(w_i, w_j) &= \mathbb{E}[(c^\top \varphi_i - c^\top \varphi_j)^2] = \mathbb{E}[(c^\top (\varphi_i - \varphi_j))^2] \\ &= (\varphi_i - \varphi_j)^\top \mathbb{E}[cc^\top] \, (\varphi_i - \varphi_j) = (\varphi_i - \varphi_j)^\top \Sigma_c \, (\varphi_i - \varphi_j). \end{aligned} \qquad \square$$

Theorem 6.2 (Separation Characterisation)**.** *Let $\Delta = \{\varphi_i - \varphi_j : i \neq j\}$ be the set of pairwise difference vectors, and let $V = \operatorname{span}(\Delta)$ be the subspace they span (dimension $r \leq \min(D, N-1)$).*

Exact pairwise condition: *μ separates all points iff*

$$(\varphi_i - \varphi_j)^\top \Sigma_c \, (\varphi_i - \varphi_j) > 0 \quad \textit{for all } i \neq j, \tag{23}$$

equivalently iff no nonzero pairwise difference vector lies in $\ker(\Sigma_c|_V)$*.*

A convenient sufficient condition is that Σ_c is positive definite on V, i.e.:

$$\lambda_{\min}(P_V \Sigma_c P_V) > 0, \tag{24}$$

where P_V is the orthogonal projection onto V.

The separation margin is:

$$\delta_\mu = \min_{i\neq j} (\varphi_i - \varphi_j)^\top \Sigma_c\, (\varphi_i - \varphi_j) \geq \lambda_{\min}(\Sigma_c|_V) \cdot \min_{i\neq j} \|P_V(\varphi_i - \varphi_j)\|^2 = \lambda_{\min}(\Sigma_c|_V) \cdot d^2_{\min}, \tag{25}$$

where $\lambda_{\min}(\Sigma_c|_V)$ is the smallest eigenvalue of Σ_c restricted to V, and $d^2_{\min} = \min_{i\neq j} \|\varphi_i - \varphi_j\|^2$ is the squared minimum distance between world-state features.

Proof. By Theorem 6.1, $\sigma^2(w_i, w_j) = \Delta_{ij}^\top \Sigma_c\, \Delta_{ij}$ with $\Delta_{ij} = \varphi_i - \varphi_j \in V$. Thus μ separates all points iff $\Delta_{ij}^\top \Sigma_c\, \Delta_{ij} > 0$ for all $i \neq j$, proving the exact condition.

If Σ_c is positive definite on V, then $v^\top \Sigma_c\, v > 0$ for every nonzero $v \in V$, hence in particular for every $\Delta_{ij} \neq 0$, so separation follows. For the margin bound, for any $v \in V$ with $\|v\| = 1$:

$$v^\top \Sigma_c\, v \geq \lambda_{\min}(\Sigma_c|_V).$$

So $\sigma^2(w_i, w_j) = \|\Delta_{ij}\|^2 \cdot (\Delta_{ij}/\|\Delta_{ij}\|)^\top \Sigma_c\, (\Delta_{ij}/\|\Delta_{ij}\|) \geq \|\Delta_{ij}\|^2 \cdot \lambda_{\min}(\Sigma_c|_V)$.

Taking the minimum over $i \neq j$ gives $\delta_\mu \geq \lambda_{\min}(\Sigma_c|_V) \cdot d^2_{\min}$. □

Corollary 6.3 (Spectral Control of Convergence)**.** *In the full-separation feasible-veridical regime (μ separates points and $M \geq N$), assume Σ_c is positive definite on V so that (Theorem 6.2)*

$$\delta_\mu \geq \lambda_{\min}(\Sigma_c|_V) \cdot d^2_{\min}.$$

Then for any confidence level $\alpha \in (0, 1)$ and tolerance η with $0 < \eta < \pi^2_{\min}\, \lambda_{\min}(\Sigma_c|_V)\, d^2_{\min}$, a sufficient condition for veridical empirical dominance (via Corollary 5.2) is

$$T \geq \frac{B^4}{2\Big(\pi^2_{\min}\, \lambda_{\min}(\Sigma_c|_V)\, d^2_{\min} - \eta\Big)^2} \cdot \Big(N \log N + \log(1/\alpha)\Big). \tag{26}$$

Hence T scales polynomially in N and inversely with the squared spectral bottleneck $\lambda_{\min}(\Sigma_c|_V)$.

Asymptotic reading. *For fixed α, η and other constants, the leading dependence is*

$$T = O\left(\frac{N \log N}{\pi^4_{\min}\, \lambda^2_{\min}(\Sigma_c|_V)\, d^4_{\min}}\right). \tag{27}$$

The task-distance landscape. The spectral structure of Σ_c directly determines the pairwise task distances $\sigma^2(w_i, w_j)$, and it is these distances, not the condition number κ per se, that control evolutionary outcomes. Figure 3 visualises all $\binom{N}{2} = 55$ pairwise distances for each of the six task families used in simulation 2 ($N = 11$, $T = 10{,}000$ Monte Carlo estimate of the population σ^2 matrix). The isotropic family (`gauss_iso`) has all distances tightly clustered near 1.0, well above the approximate mutation-selection threshold δ^*_μ (derived in Section 7; it depends on the mutation rate ε and fitness offset C via $\delta^*_\mu \approx 2C\varepsilon/\pi^2_{\min} \approx 0.48$ for our simulation parameters). As the condition number increases, the distribution of distances stretches across orders of magnitude: the largest distances remain near 1, but bottleneck pairs fall to 10^{-2} or below. The beta families show a qualitatively similar spread despite different generative mechanisms.

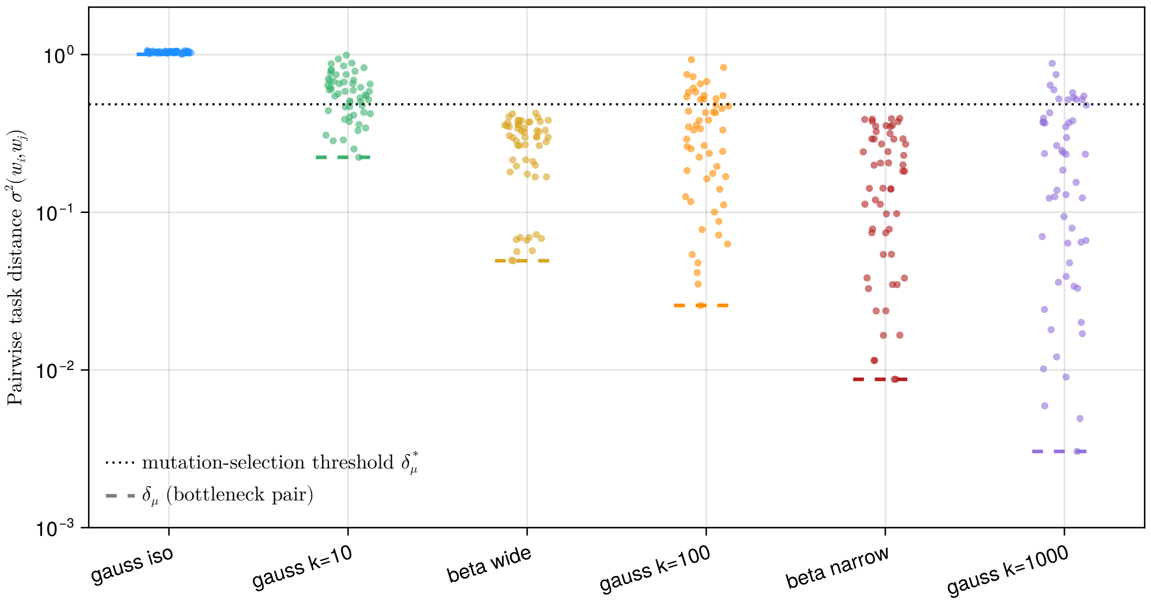


Figure 3: Distribution of all 55 pairwise task distances $\sigma^2(w_i, w_j)$ for each task family (population estimates from 10,000 task samples). Dashed horizontal lines mark δ_μ (the minimum, i.e. the bottleneck pair). The dotted line marks the approximate mutation-selection threshold $\delta^*_\mu \approx 2C\varepsilon/\pi^2_{\text{min}}$: pairs above this line can be maintained by selection against the mutation load; pairs below it cannot. Only `gauss_iso` has *all* pairs above the threshold.

From distance structure to final outcomes. This matters because Theorem 4.1(c) bounds the fitness gap for each merged pair individually: $R(p) \geq \pi^2_{\text{min}} \cdot \sigma^2(w_i, w_j)$. An encoding that fails to separate states w_i, w_j pays a penalty proportional to their task distance $\sigma^2(w_i, w_j)$, regardless of how well it separates other pairs. Full veridicality therefore requires that *every* pairwise distance exceed the mutation-selection threshold, a condition that only well-conditioned ecologies satisfy.

Figure 4 tests this prediction directly: population δ_μ (the bottleneck distance) is a clean monotone predictor of the final veridical fraction across all six families ($N = 11$, $M = 11$, $K = 500$, 30 replicates per condition, all T values pooled). Panel (a) shows that only `gauss_iso` ($\delta_\mu \approx 1.0$) achieves substantial veridicality (~63%), with `gauss_k10` ($\delta_\mu \approx 0.22$) at a marginal ~17% and the remaining families near zero. Panel (b) reveals that mean Bayes risk is nearly flat across these families despite their vastly different veridical fractions; we return to this observation in Section 7 (Figure 8), where the evolutionary trajectory data makes the explanation precise.

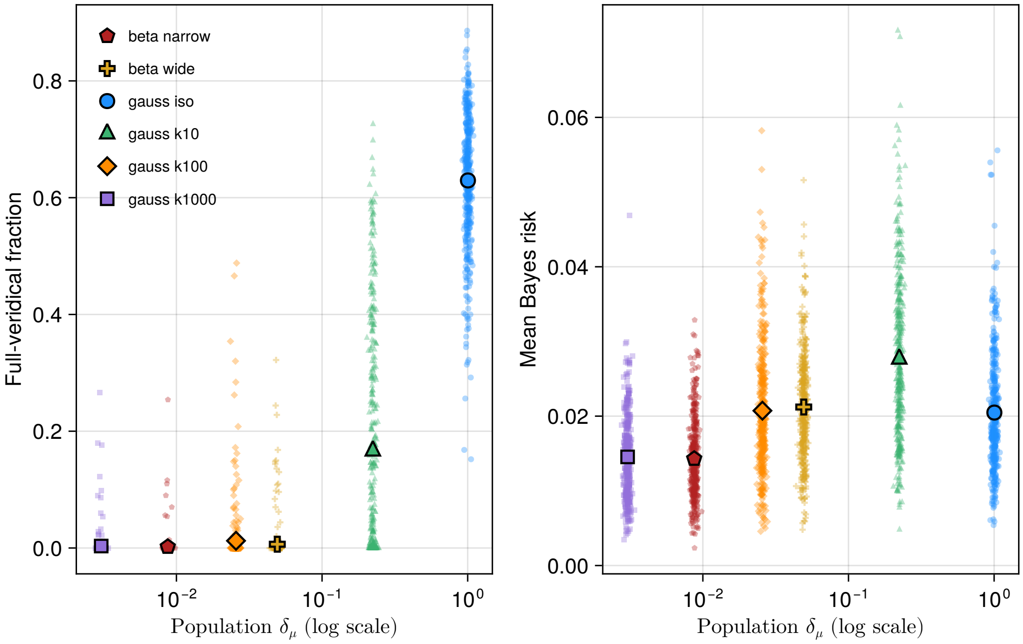
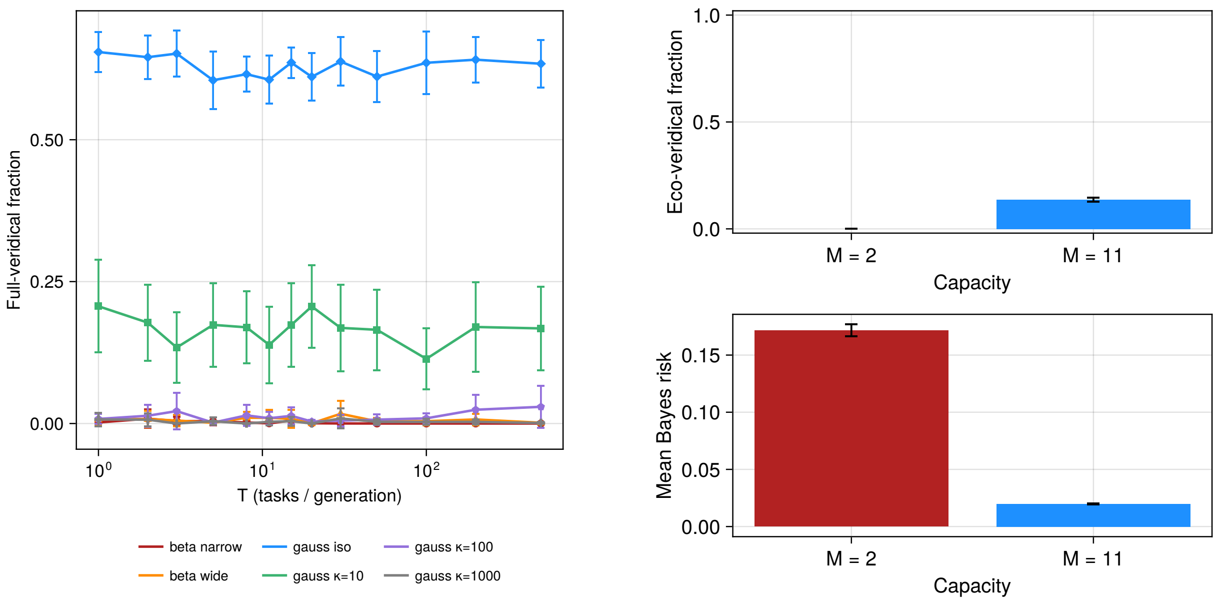


Figure 4: (a) Final full-veridical fraction versus population δ_μ (log scale). Large markers: family means across all T and replicates; small points: individual runs. The monotone trend confirms that the bottleneck distance controls veridicality. (b) Mean Bayes risk versus population δ_μ. Risk is nearly flat across families, and if anything, *higher* for the most veridical family, because ill-conditioned ecologies permit low-risk non-veridical encodings that merge near-indistinguishable states.

Family and capacity decomposition. Figure 5 provides a complementary view. Panel (a) shows, family by family, the full-veridical fraction as a function of task count T at $M = 11$: within each family the fraction is essentially flat in T, confirming that it is the ecology's distance structure, not the number of tasks per generation, that drives the outcome. Panel (b) aggregates across all families and task counts, showing the stark capacity contrast ($M = 2$ versus $M = 11$) in both eco-veridical fraction and mean Bayes risk. Figure 6 summarises the effect sizes across regimes, with a log-scale panel showing all favorable conditions far exceed the random-injective null.



(a) Full-veridical fraction versus T for six task families at $M = 11$. Well-conditioned families (`gauss_iso`, blue) reach ∼0.6; moderately anisotropic ones (`gauss_k10`, green) plateau around 0.15; poorly conditioned families cluster near zero.

(b) Capacity comparison aggregated across all families and task counts. Top: eco-veridical fraction; bottom: mean Bayes risk. $M = 2$ yields near-zero veridicality and high risk; $M = 11$ permits substantially higher veridicality and lower risk.

Figure 5: Geometry and capacity jointly control the veridicality transition, as predicted by Theorem 4.4 and Proposition 4.6. See Appendix E for full details.

Regime-level effect size summary. To make the transition magnitude explicit, Figure 6 compresses the main contrasts into three regimes and includes the random-injective baseline.

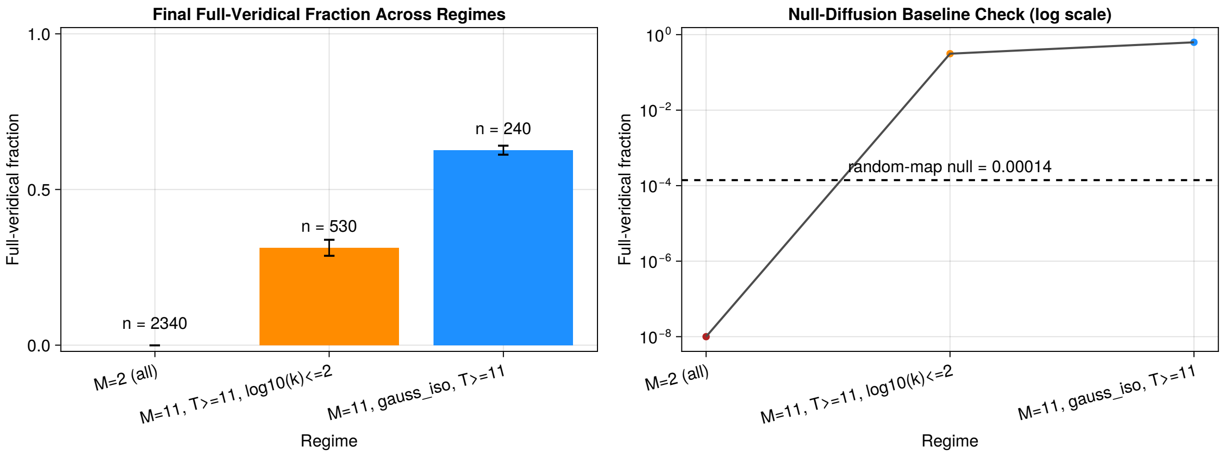


Figure 6: Left panel: full-veridical fraction across three regimes, $M = 2$ (all conditions, $n = 2340$), favorable $M = 11$ ($T \geq 11$, $\log_{10} \kappa \leq 2$, $n = 530$), and `gauss_iso` with $M = 11$, $T \geq 11$ ($n = 240$). Right panel: same regimes on a log scale, with the random-injective null baseline ($\approx 1.4 \times 10^{-4}$, dashed) for reference.

To summarise: the condition number κ of Σ_c is a useful summary because it controls the *ratio* of largest to smallest pairwise distance, but the operative variable is the bottleneck distance δ_μ itself. Figures 3 and 4 make this distinction precise.

7 Evolutionary Dynamics via Price's Equation

We now analyse two coupled levels of evolutionary dynamics: (i) the finite-K stochastic Wright–Fisher process driven by realised fitness w_t; and (ii) its deterministic mean-field limit driven by expected fitness W. This is the dynamic complement to the static theorems of Section 4. The argument uses three tools with distinct roles:

- **Price's equation** (§7.2–7.4) provides a *decomposition* of evolutionary change into selection and transmission components. It identifies the *direction* of selection: mean Bayes risk decreases each generation (Theorem 7.4). However, Price's equation is a statistical identity, namely an exact partition of change, not a dynamical law. It cannot, by itself, predict long-term trajectories or prove convergence to an equilibrium.

- **Quasispecies theory** (§7.5) provides the *dynamical* result. In the frequency-independent setting, the deterministic update is a normalized positive linear map. Perron–Frobenius theory yields convergence to the dominant mutation-accessible asymptotic regime (equilibrium in primitive blocks, periodic limit in irreducible periodic blocks); unique global equilibrium is recovered when quotient mutation is primitive (Theorem 7.6, Remark 7.9).

- **Spectral perturbation theory** (§7.5, Theorem 7.6(c)) provides the *rate*. The convergence speed is governed by a dominant spectral ratio (within the dominant reachable class plus inter-class spectral separation). Under explicit small-mutation assumptions, this ratio is controlled by the class-wise capacity-aware fitness gap. A δ_μ-based bound is recovered as a corollary in the full-separation feasible-veridical primitive regime.

The causal mechanism throughout is ordinary natural selection: organisms with encodings that merge task-distinguishable world states incur positive Bayes risk, hence lower fitness, hence fewer offspring. Price describes one-generation change; the replicator-mutator recursion gives the mean-field dynamical law; Wright–Fisher gives the finite-population stochastic process around that law. This separation also aligns with the selection-for/selection-of distinction: covariance decomposition alone does not establish long-run adaptation, whereas the explicit dynamical model does. In Pigliucci and Kaplan's terms, statistical patterns underdetermine causal history unless they are embedded in explicit, competing causal models tested with additional evidence (Pigliucci and Kaplan, 2006, prelude; ch. 2).

7.1 Population Model

Consider a population of K agents at generation t. Agent i has encoding $p_i \colon W \to X$. Let Ω denote the finite set of all encodings $W \to X$ (so $|\Omega| = M^N$). For each $p \in \Omega$, let $n_t(p)$ be the number of agents with encoding p, and $\pi_t(p) = n_t(p)/K$ the frequency.

Fitness. In generation t, all agents face the same sampled tasks $f_1, \ldots, f_T$. Define realised generational fitness

$$w_t(p) = TC - \sum_{\tau=1}^{T} R(p, f_\tau), \tag{28}$$

with C chosen so $w_t(p) > 0$. Its expectation is

$$W(p) = \mathbb{E}[w_t(p)] = TC - T \cdot R(p), \tag{29}$$

which is a decreasing linear function of $R(p)$. Thus minimising Bayes risk is equivalent to maximising expected fitness.

Reproduction. At the end of each generation, agents reproduce with probability proportional to realised fitness: rate $w_t(p)/\bar{w}_t$, where $\bar{w}_t = \sum_p \pi_t(p)\, w_t(p)$.

Mutation. Offspring encodings mutate according to a kernel Q_ε on Ω. We do **not** assume full support: biologically realistic viability/development constraints may forbid many transitions. A useful concrete family is local per-position mutation plus viability filtering, where only admissible transitions receive positive probability.

Two dynamics and scope. Theorems 7.4 and 7.6 are statements about the deterministic expected-fitness recursion (replace w_t by W). Theorem 7.10 is the stochastic finite-K Wright–Fisher model and links it to that deterministic limit via law of large numbers as $K \to \infty$.

Lemma 7.1 (Label-equivariance of mutation)**.** *Assume the mutation kernel is percept-label equivariant: for every $\sigma \in S_M$ and encodings p, p':*

$$Q_\varepsilon(\sigma \cdot p \to \sigma \cdot p') = Q_\varepsilon(p \to p'). \tag{30}$$

Hence transition probabilities depend only on orbit classes in $\bar{\Omega} = \Omega/{\sim_X}$, so the process admits a well-defined quotient mutation kernel. If equivariance fails, all statements should be read on Ω directly rather than on $\bar{\Omega}$.

Proof. Immediate from the stated equivariance property. For uniform per-position mutation, equivariance holds because transition probabilities depend only on coordinatewise equality/inequality patterns, which relabeling preserves. □

7.2 Price's Equation

We now apply Price's equation (Price, 1970, 1972) to the encoding population. The result is a one-generation accounting identity that decomposes change in any trait into a selection term (covariance with fitness) and a transmission term (mutation bias). It does not, by itself, predict long-term dynamics, but it identifies the *direction* of selection, which Section 7.4 will specialise to Bayes risk.

For a fixed generation, write $w(p) := w_t(p)$ for notational simplicity.

Theorem 7.2 (Price equation for encoding populations)**.** *Let $z\colon \Omega \to \mathbb{R}$ be any real-valued trait of encodings. The change in population mean of z across one generation satisfies the Price identity (Price, 1970, 1972):*

$$\bar{w} \cdot \Delta\bar{z} = \mathrm{Cov}_\pi(w, z) + \mathbb{E}_\pi[w \cdot \delta z], \tag{31}$$

where:

- $\Delta\bar{z} = \bar{z}_{t+1} - \bar{z}_t$ *is the generational change in mean z,*
- $\mathrm{Cov}_\pi(w, z) = \sum_p \pi_t(p)(w(p) - \bar{w})(z(p) - \bar{z})$ *is the selection covariance,*
- δz *is the expected change in z due to mutation within a lineage,*
- *the second term $\mathbb{E}_\pi[w \cdot \delta z]$ is the transmission bias.*

Proof. This is the standard Price equation derivation (Price, 1970, 1972). Write $\pi'(p)$ for the frequency after selection but before mutation:

$$\pi'(p) = \pi_t(p) \cdot w(p)/\bar{w}.$$

Then $\bar{z}' = \sum_p \pi'(p)\, z(p) = (1/\bar{w}) \sum_p \pi_t(p)\, w(p)\, z(p)$. So:

$$\bar{w} \cdot \bar{z}' = \sum_p \pi_t(p)\, w(p)\, z(p) = \mathrm{Cov}_\pi(w, z) + \bar{w}\,\bar{z}.$$

Thus $\bar{w}(\bar{z}' - \bar{z}) = \mathrm{Cov}_\pi(w, z)$, which is the selection-only Price equation.

Including mutation: $\bar{z}_{t+1} = \sum_p \pi'(p)\,[z(p) + \delta z(p)] = \bar{z}' + \sum_p \pi'(p)\,\delta z(p)$. So:

$$\bar{w}(\bar{z}_{t+1} - \bar{z}_t) = \mathrm{Cov}_\pi(w, z) + \sum_p \pi_t(p)\, w(p)\, \delta z(p) = \mathrm{Cov}_\pi(w, z) + \mathbb{E}_\pi[w \cdot \delta z]. \qquad \square$$

7.3 Fisher's Fundamental Theorem

Setting the trait z equal to fitness itself recovers a version of Fisher's fundamental theorem: selection always increases mean fitness (in the absence of mutation).

Theorem 7.3 (Fitness increase under selection)**.** *Apply Price's equation with $z = w$ (the fitness trait itself):*

$$\bar{w} \cdot \Delta\bar{w} = \mathrm{Var}_\pi(w) + \mathbb{E}_\pi[w \cdot \delta w]. \tag{32}$$

Selection only *($\varepsilon = 0$): $\Delta\bar{w} = \mathrm{Var}_\pi(w)/\bar{w} \geq 0$, with equality iff the population is monomorphic.*

With mutation *($\varepsilon > 0$): $\Delta\bar{w} = \mathrm{Var}_\pi(w)/\bar{w} + \mathbb{E}_\pi[w \cdot \delta w]/\bar{w}$, where the mutation term $\mathbb{E}_\pi[w \cdot \delta w]$ can be negative (mutations are typically deleterious).*

Proof. Set $z = w$ in (31). Then $\mathrm{Cov}_\pi(w, w) = \mathrm{Var}_\pi(w)$. The result follows. □

For Berke's asexual haploid population (no recombination), the additive genetic variance equals the total phenotypic variance in fitness, so $\mathrm{Var}_\pi(w)$ is the full variance. In sexual diploid populations, Fisher's theorem involves only the additive component; our asexual setting avoids this complication.

7.4 Price Equation for Bayes Risk

The payoff of the Price decomposition comes from setting $z = R$ (Bayes risk). Under our expected-fitness specification (29), fitness is a decreasing linear function of risk, so the selection covariance becomes an exact negative-variance term. Thus, the selection component pushes mean risk down, while mutation/transmission can offset part of that decrease.

Theorem 7.4 (Bayes risk decrease in the deterministic mean-field dynamics)**.** *In the deterministic expected-fitness dynamics (replace w by W from* (29)*), applying Price with $z = R$ gives*

$$\bar{W} \cdot \Delta\bar{R} = \mathrm{Cov}_\pi(W, R) + \mathbb{E}_\pi[W \cdot \delta R], \tag{33}$$

and since $W(p) = TC - T\,R(p)$,

$$\Delta\bar{R} = -\frac{T \cdot \mathrm{Var}_\pi(R)}{\bar{W}} + \frac{\mathbb{E}_\pi[W \cdot \delta R]}{\bar{W}}. \tag{34}$$

Selection only (under the linear map (29)): $\Delta\bar{R} = -T \cdot \mathrm{Var}_\pi(R)/\bar{W} \leq 0$. *Mean Bayes risk strictly decreases unless all types in support have equal Bayes risk.*

Proof. Apply Theorem 7.2 with trait $z = R$ and fitness w replaced by expected fitness W. Then

$$\bar{W} \cdot \Delta\bar{R} = \mathrm{Cov}_\pi(W, R) + \mathbb{E}_\pi[W \cdot \delta R].$$

Using $W(p) = TC - TR(p)$:

$$\mathrm{Cov}_\pi(W, R) = \mathrm{Cov}_\pi(TC - TR, R) = -T \cdot \mathrm{Var}_\pi(R),$$

which yields (34). Under selection only ($\delta R \equiv 0$), the right-hand side is $-T \cdot \mathrm{Var}_\pi(R) \leq 0$, with equality iff R is constant on the support. □

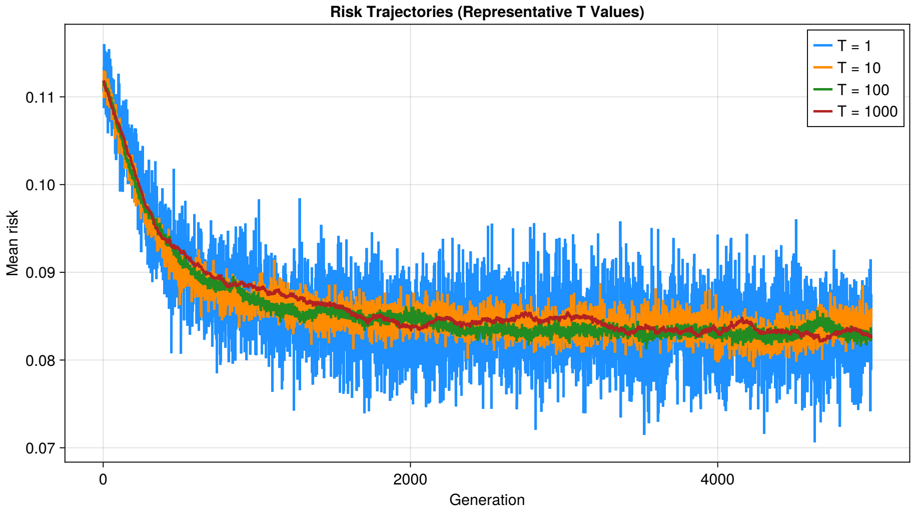


Figure 7: Mean Bayes risk over 5000 generations for representative task counts ($T = 1, 10, 100, 1000$) in the Berke-style setting ($N = 11$, $M = 2$). All trajectories decline, confirming Theorem 7.4. They converge to approximately the same asymptotic value, the $M = 2$ capacity floor (Theorem 4.2), since the optimal 2-cell partition is the same regardless of T. Higher T reduces per-generation sampling noise but does not change the asymptotic level in this model.

For finite K with realised fitness, (34) holds with sampling noise terms; monotonic decrease is then in expectation or in the mean-field limit, not pathwise for every finite population trajectory.

Remark 7.5 (Non-linear fitness maps)**.** *The exact identity* $\mathrm{Cov}_\pi(W, R) = -T\,\mathrm{Var}_\pi(R)$ *is specific to the linear expected-fitness map* (29)*. If instead $W(p) = g(R(p))$ with g strictly decreasing, then* $\mathrm{Cov}_\pi(g(R), R) \leq 0$ *(strict unless R is almost surely constant), so the* direction *of selection remains toward lower risk, but the covariance no longer reduces to a variance term. Quantitatively, higher moments of the risk distribution enter, so* (34) *should then be read as a sign statement, not an exact identity.*

Figure 7 confirms the Price decomposition computationally. Recall that Bayes risk quantifies how poorly an encoding serves the task ecology: $R(p) = 0$ means the encoding preserves all task-relevant distinctions (full ecological veridicality), while $R(p) > 0$ means ecologically meaningful information is lost. Figure 7 shows that mean Bayes risk declines over generations for all task-diversity levels T, consistent with a dominant negative selection term in (34) and with the mean-field tendency of Theorem 7.4. All four trajectories converge to approximately the same asymptotic risk (≈ 0.08), which is the capacity floor imposed by $M = 2 \ll N = 11$: with only two percepts, the best achievable encoding is the optimal 2-cell partition (Theorem 4.2), regardless of how many tasks are sampled. What T controls is not the asymptotic level under this update rule but the *noise*: higher T gives a more precise per-generation estimate of the population risk, reducing the stochastic variance visible in the trajectories ($T = 1$, blue, fluctuates widely; $T = 1000$, red, is nearly deterministic).

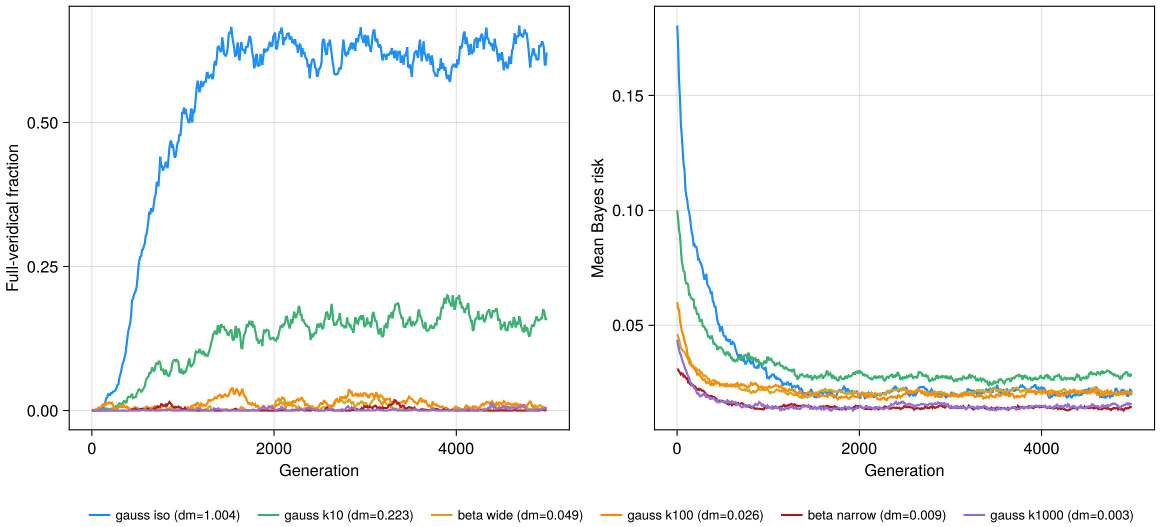


Figure 8: Evolutionary trajectories at $M = N = 11$ ($T = 50$, $K = 500$, 30 replicates averaged). (a) Full-veridical fraction over 5000 generations. Only families with δ_μ above the mutation-selection threshold (`gauss_iso`, `gauss_k10`) achieve substantial veridicality; others stall near zero. (b) Mean Bayes risk over generations. All families converge to low risk (0.01–0.05), but the most veridical family (`gauss_iso`, blue) has the *highest* asymptotic risk. See text for the explanation via Theorem 4.1(c).

Figure 8 extends this analysis to the $M = N = 11$ regime across all six task families. When capacity is sufficient for full veridicality, the task-distance structure σ^2 becomes the binding constraint. Panel (a) shows the full-veridical fraction rising over generations: `gauss_iso` ($\delta_\mu \approx 1.0$) converges to ~65% veridical within ~1000 generations; `gauss_k10` ($\delta_\mu \approx 0.22$) reaches ~15%; the remaining families, whose bottleneck distances fall below the mutation-selection threshold, stall near zero. Panel (b) reveals a striking and theoretically important pattern: **mean Bayes risk is inversely related to veridicality**. The most veridical family (`gauss_iso`) has the *highest* asymptotic risk, while the least veridical families (`gauss_k1000`, `beta_narrow`) have the lowest. This apparent paradox is explained by Theorem 4.1(c): in ill-conditioned ecologies, the non-veridical encodings merge pairs with tiny σ^2, incurring negligible risk per merge. A population of such encodings has low average risk despite low veridicality. In well-conditioned ecologies, by contrast, even a small fraction of non-veridical individuals (maintained by mutation pressure) carry encodings that merge well-separated pairs, contributing substantial risk. This finding cautions against using mean Bayes risk as a summary diagnostic: a population can have uniformly low risk while being far from ecologically veridical, if the ecology's distance structure permits low-cost merges.

7.5 Quotient-Space Convergence Dynamics

Price's equation tells us the direction of selection but cannot, by itself, predict where the population ends up. For that we need a dynamical theory. A key structural feature of our model is that expected fitness $W(p)$ depends only on the encoding p, not on the population composition. This is the frequency-independent setting of quasispecies theory (Eigen, 1971; Park et al., 2010).

7.5.1 The Quasispecies Recursion

For the deterministic generation-to-generation dynamics on a canonical representative set $\Omega_c \subset \Omega$ (one encoding per percept-label orbit; equivalently $\bar{\Omega} = \Omega/{\sim_X}$), define $D = \mathrm{diag}(W(p_1), \ldots, W(p_{|\Omega_c|}))$ and $A_\varepsilon := Q_\varepsilon^\top D$. The replicator-mutator update is

$$x_{g+1} = \frac{A_\varepsilon\, x_g}{\mathbf{1}^\top A_\varepsilon\, x_g}, \tag{35}$$

where $x_g \in \Delta^{|\Omega_c|}$ is the encoding-frequency vector at generation g.

By induction,

$$x_g = \frac{A_\varepsilon^g\, x_0}{\mathbf{1}^\top A_\varepsilon^g\, x_0}. \tag{36}$$

Hence fixed points satisfy

$$A_\varepsilon\, x^* = \lambda\, x^*, \tag{37}$$

for $\lambda = \mathbf{1}^\top A_\varepsilon\, x^*$. Thus fixed points are Perron eigenvectors of A_ε after simplex normalisation.

7.5.2 The Perron–Frobenius Argument

The recursion (35) is a normalised iteration of a nonnegative matrix, exactly the setting of Perron–Frobenius theory for nonnegative matrices (Seneta, 2006). The challenge is that the mutation graph need not be strongly connected (primitive): biological viability constraints can partition encoding space into isolated communicating classes. We therefore work with the full reducible structure, identifying which classes are reachable from the initial population and which among those dominate asymptotically.

By Lemma 7.1, recursion (35) descends to the quotient $\bar{\Omega} = \Omega/{\sim_X}$ (Definition 2.3). To avoid label-symmetry multiplicity, we work on a canonical representative set Ω_c (one encoding per orbit), equivalently on $\bar{\Omega}$. In this subsection, all vectors/matrices and equations (35)–(37) are interpreted on Ω_c; for readability, we drop bars.

Define the directed mutation graph G_ε on Ω_c by edges $p \to q$ iff $Q_\varepsilon(p \to q) > 0$. Denote communicating classes by $\mathcal{K}$ (reserving C for the fitness offset). Let $\mathcal{K}_1, \ldots, \mathcal{K}_L$ be the closed communicating classes. For each closed class $\mathcal{K}$, define the class block

$$A_{\varepsilon,\mathcal{K}} := Q_{\varepsilon,\mathcal{K}}^\top D_{\mathcal{K}}, \tag{38}$$

where $Q_{\varepsilon,\mathcal{K}}$ is the mutation kernel restricted to $\mathcal{K}$ and $D_{\mathcal{K}}$ is the fitness diagonal restricted to $\mathcal{K}$.

For a given initial condition x_0, let $\mathcal{R}(x_0)$ be the set of closed classes reachable from $\mathrm{supp}(x_0)$ in G_ε. For each $\mathcal{K} \in \mathcal{R}(x_0)$, define $\lambda_{\mathcal{K}} := \rho(A_{\varepsilon,\mathcal{K}})$ (spectral radius).

Fix a dominant reachable class $\mathcal{K}^*$ satisfying

$$\lambda_* := \lambda_{\mathcal{K}^*} > \max_{\mathcal{K} \in \mathcal{R}(x_0) \setminus \{\mathcal{K}^*\}} \lambda_{\mathcal{K}}. \tag{39}$$

Inside $\mathcal{K}^*$, define the class-wise capacity-aware optimum and gap:

$$R^*_{\mathcal{K}} = \min_{p \in \mathcal{K}^*} R(p), \qquad P^*_{\mathcal{K}} = \operatorname*{argmin}_{p \in \mathcal{K}^*} R(p).$$

If $P^*_{\mathcal{K}} \neq \mathcal{K}^*$, set

$$\Delta R_{\mathrm{gap},\mathcal{K}} = \min_{p \in \mathcal{K}^* \setminus P^*_{\mathcal{K}}} [R(p) - R^*_{\mathcal{K}}] > 0;$$

if $P^*_{\mathcal{K}} = \mathcal{K}^*$, define $\Delta R_{\mathrm{gap},\mathcal{K}} := 0$ (degenerate no-gap case). Choose $p^*_{\mathcal{K}} \in P^*_{\mathcal{K}}$ and define

$$W^*_{\mathcal{K}} = W(p^*_{\mathcal{K}}), \quad W_{2,\mathcal{K}} = \max_{p \in \mathcal{K}^* \setminus P^*_{\mathcal{K}}} W(p), \quad \Delta W_{\mathcal{K}} = W^*_{\mathcal{K}} - W_{2,\mathcal{K}} = T \cdot \Delta R_{\mathrm{gap},\mathcal{K}},$$

and $A_{0,\mathcal{K}} := D_{\mathcal{K}}$.

Theorem 7.6 (Constrained-mutation quasispecies convergence)**.** *Assume:*

(i) Frequency-independent expected fitness $W(p) > 0$ for all $p \in \Omega_c$;

(ii) For each closed class $\mathcal{K} \in \mathcal{R}(x_0)$, the block $A_{\varepsilon,\mathcal{K}}$ is primitive;

(iii) The dominant-class condition (39) *holds.*

Then:

(a) ***Class-conditional convergence.*** *The recursion* (35) *converges to the Perron equilibrium of the dominant reachable class:*

$$x_g \to x^*_{\mathcal{K}^*},$$

where $x^*_{\mathcal{K}^*}$ is the normalized Perron right eigenvector of $A_{\varepsilon,\mathcal{K}^*}$, embedded in Ω_c (zero mass outside $\mathcal{K}^*$).

(b) ***Small-mutation concentration within the dominant class.*** *If $P^*_{\mathcal{K}}$ is a singleton $\{p^*_{\mathcal{K}}\}$ and the mutation kernel admits a first-order expansion*

$$Q_{\varepsilon,\mathcal{K}^*} = I + \varepsilon L_{\mathcal{K}^*} + O(\varepsilon^2),$$

where $L_{\mathcal{K}^*}$ is the first-order mutation generator on $\mathcal{K}^*$ (nonnegative off-diagonals, row sums zero), then $x^*_{\mathcal{K}^*} \to \delta_{p^*_{\mathcal{K}}}$ as $\varepsilon \to 0$ and, for $p \in \mathcal{K}^*$, $p \neq p^*_{\mathcal{K}}$:

$$x^*_{\mathcal{K}^*}(p) = \frac{\varepsilon \cdot W(p^*_{\mathcal{K}})\, L_{\mathcal{K}^*}(p^*_{\mathcal{K}} \to p)}{W(p^*_{\mathcal{K}}) - W(p)} + O(\varepsilon^2). \tag{40}$$

For the uniform per-position mutation model, $L_{\mathcal{K}^*} = N(1 - 1/M)\,(Q_{1,\mathcal{K}^*} - I)$, which recovers the previous effective-rate form $\tilde{\varepsilon} = N\varepsilon(1 - 1/M)$ as a special case.

(c) ***Rate via dominant class spectral ratio.*** *Let* $\lambda_{1,\mathcal{K}} > |\lambda_{2,\mathcal{K}}| \geq \cdots$ *be eigenvalues of* $A_{\varepsilon,\mathcal{K}^*}$ *by modulus and define*

$$\rho_{\mathcal{K}^*} = |\lambda_{2,\mathcal{K}}|/\lambda_{1,\mathcal{K}}, \quad \eta_{\varepsilon,\mathcal{K}} = \|A_{\varepsilon,\mathcal{K}^*} - A_{0,\mathcal{K}}\|_2.$$

If $\eta_{\varepsilon,\mathcal{K}} < \Delta W_{\mathcal{K}}/2$*, then*

$$\rho_{\mathcal{K}^*} \leq \frac{W_{2,\mathcal{K}} + \eta_{\varepsilon,\mathcal{K}}}{W^*_{\mathcal{K}} - \eta_{\varepsilon,\mathcal{K}}} < 1. \tag{41}$$

Define

$$\theta := \max\Big\{\rho_{\mathcal{K}^*}, \max_{\mathcal{K}\in\mathcal{R}(x_0)\setminus\{\mathcal{K}^*\}}(\lambda_{\mathcal{K}}/\lambda_*)\Big\} < 1, \tag{42}$$

with the convention max *over an empty set equals* 0*. Then there exist* $C_0 > 0$ *and integer* $s \geq 1$ *such that*

$$\|x_g - x^*_{\mathcal{K}^*}\| \leq C_0 \cdot g^{s-1} \cdot \theta^g. \tag{43}$$

If $A_{\varepsilon,\mathcal{K}^*}$ *is diagonalizable,* $s = 1$ *for the within-class contribution.*

If $\eta_{\varepsilon,\mathcal{K}} \leq \Delta W_{\mathcal{K}}/4$*, then*

$$1 - \rho_{\mathcal{K}^*} \geq \frac{\Delta W_{\mathcal{K}}}{2W^*_{\mathcal{K}}} \geq \frac{\Delta W_{\mathcal{K}}}{2TC} = \frac{\Delta R_{\text{gap},\mathcal{K}}}{2C}. \tag{44}$$

Proof sketch. The full proof is in Appendix D; we outline the architecture here.

Write the replicator-mutator update as the normalised iteration $x_g = A^g x_0/\mathbf{1}^\top A^g x_0$ with $A = A_\varepsilon = Q_\varepsilon^\top D$, a nonnegative matrix. Because mutation can be constrained, A may be reducible.

Step 1 (Frobenius decomposition). Permute A into lower block-triangular form, separating a transient block T from closed communicating-class blocks $B_1, \ldots, B_L$ (Seneta, 2006). The closed-part dynamics decomposes into independent primitive blocks plus transient feeding terms that decay exponentially (Lemma D.1).

Step 2 (Perron expansion on each block). Each primitive block B_i has a simple dominant eigenvalue λ_i with positive left and right eigenvectors. The Perron–Frobenius theorem gives $B_i^g = \lambda_i^g v_i u_i^\top + O(r_i^g)$ with $r_i < \lambda_i$ (Seneta, 2006) (Lemma D.2).

Step 3 (Dominant class wins). Among all reachable closed classes, the one with the largest λ_i eventually dominates the unnormalised iterate $A^g x_0$: all other contributions decay exponentially relative to λ_*^g (Seneta, 2006). After normalisation, the population concentrates on the Perron profile of $\mathcal{K}^*$, proving (a).

Step 4 (Small-mutation perturbation). At $\varepsilon = 0$, the dominant block is the diagonal fitness matrix $D_{\mathcal{K}}$, with eigenvector δ_{p^*}. Analytic perturbation of this simple eigenpair (Kato, 1995; Greenbaum et al., 2020) yields (b).

Step 5 (Spectral ratio via Bauer–Fike). Since $D_{\mathcal{K}}$ is diagonal (hence normal), the Bauer–Fike theorem localises every eigenvalue of $A_{\varepsilon,\mathcal{K}^*}$ within $\eta_{\varepsilon,\mathcal{K}}$ of a diagonal entry (Greenbaum et al., 2020). When the mutation perturbation is smaller than half the fitness gap ($\eta_{\varepsilon,\mathcal{K}} < \Delta W_{\mathcal{K}}/2$), the resulting eigenvalue disks are disjoint, giving the explicit ratio bound (41) and hence the convergence rate (43), proving (c). □

Tie case (outside assumption (iii)). If the unique-dominant condition (39) is dropped and multiple reachable closed classes tie for maximal $\lambda_{\mathcal{K}}$, the limit need not be unique; asymptotically the process concentrates on the top reachable classes, with weights determined by initial condition and transient connectivity. See Proposition D.4 in Appendix D (§D.4) for the full statement and proof.

Corollary 7.7 (Primitive special case: global convergence). *If the induced mutation kernel on Ω_c is primitive, there is a single closed class $\mathcal{K}^* = \Omega_c$. Theorem 7.6 reduces to a unique interior fixed point and global convergence from every interior initial condition, with the same spectral-rate bound.*

Corollary 7.8 (δ_μ specialization). *In the full-separation feasible-veridical primitive regime (μ separates points and $M \geq N$), define the global quantities*

$$R^* = \min_{p\in\Omega_c} R(p), \quad P^* = \operatorname*{argmin}_{p\in\Omega_c} R(p), \quad \Delta R_{\text{gap}} = \min_{p\notin P^*}[R(p) - R^*],$$

$$A_0 = D, \quad \eta_\varepsilon := \|A_\varepsilon - A_0\|_2, \quad \Delta W := T \cdot \Delta R_{\text{gap}}.$$

Write $\rho_\varepsilon := \rho_{\mathcal{K}^*}$ for the spectral ratio of the unique closed block (Theorem 7.6(c) with $\mathcal{K}^* = \Omega_c$). Then $R^* = 0$ and $\Delta R_{\text{gap}} \geq \pi_{\min}^2 \delta_\mu$ by Theorem 4.1(c). Therefore, under $\eta_\varepsilon \leq \Delta W/4$:

$$1 - \rho_\varepsilon \geq \frac{\pi_{\min}^2 \delta_\mu}{2C}. \tag{45}$$

Remark 7.9 (General irreducible-but-periodic extension). *Theorem 7.6 assumes primitivity of each relevant closed block to obtain pointwise convergence and a strict spectral gap ratio < 1. If a dominant closed block is only irreducible with period $d > 1$, normalized iterates need not converge to a single fixed vector; instead, they approach a d-cycle (one limit point per residue class mod d), while Cesàro averages converge to the Perron profile. The accessible-class conclusion remains unchanged (selection still concentrates on dominant reachable classes), but the asymptotic object is periodic rather than stationary. Appendix D (§D.8) records this extension. Figure 9 illustrates these dynamics numerically.*

7.5.3 Finite-Population Convergence

Theorem 7.6 is a statement about the deterministic mean-field limit. Real populations are finite and stochastic (Park et al., 2010). The next result closes the gap: the Wright–Fisher process tracks the deterministic limit on any fixed time horizon as the population size K grows, with $O(1/\sqrt{K})$ fluctuations.

Theorem 7.10 (Wright–Fisher under constrained mutation). *For the finite-population Wright–Fisher process with K agents, selection proportional to realised fitness $w_t(p)$, and mutation kernel Q_ε:*

(a) The composition chain is finite-state and Markov. Its communication structure is inherited from the mutation viability graph on encodings: the chain decomposes into transient sets and one or more closed communicating classes. Thus a stationary distribution always exists, but uniqueness holds only class-by-class (not globally unless the chain is irreducible).

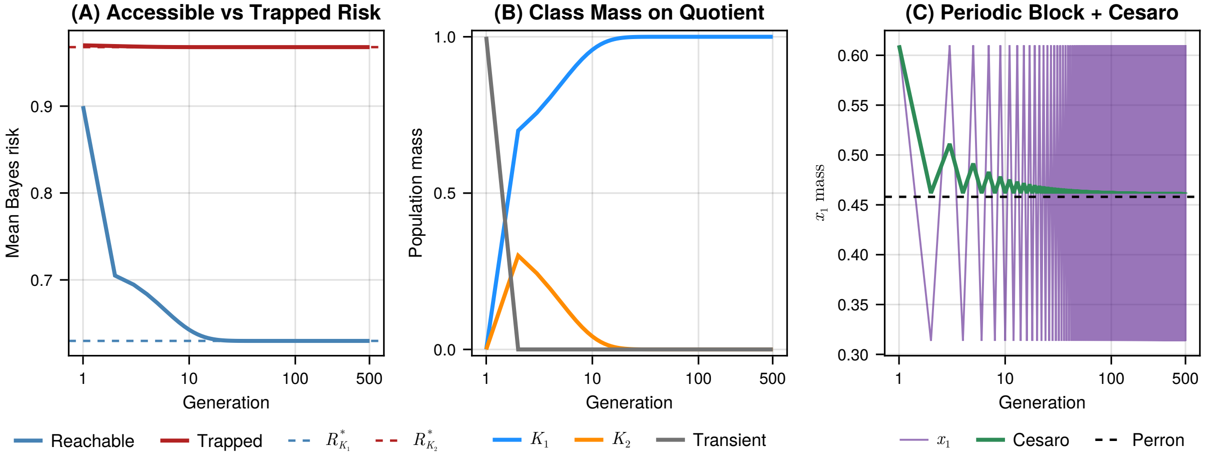


Figure 9: Numerical illustration of Theorem 7.6 and Remark 7.9. **(A)** Constrained mutation with two closed classes ($\mathcal{K}_1$ dominant, $\mathcal{K}_2$ subdominant) and one transient state. From the transient state (reachable to both classes), mean risk drops to the dominant-class asymptote $R^*_{\mathcal{K}_1}$; from initial support restricted to $\mathcal{K}_2$, the population is trapped at $R^*_{\mathcal{K}_2}$. Dashed lines: theoretical asymptotic risk from Perron eigenvectors of each block. **(B)** Quotient dynamics for the accessible case: mass on $\mathcal{K}_1$ (dominant) grows monotonically while $\mathcal{K}_2$ and the transient set are driven to zero, consistent with the spectral-radius ordering $\lambda_{\mathcal{K}_1} > \lambda_{\mathcal{K}_2}$. **(C)** Irreducible period-2 mutation (Remark 7.9): the raw trajectory oscillates indefinitely, but the Cesàro mean converges to the Perron profile (dashed).

(b) *Restricted to any closed class that is aperiodic and irreducible, the chain has a unique stationary distribution and finite mixing time within that class. If a closed class is irreducible but periodic, the stationary distribution is still unique but trajectory convergence is periodic unless one time-averages.*

(c) *As $K \to \infty$ (with iid task sampling each generation), the composition process converges on each fixed finite time horizon in probability to the deterministic expected-fitness replicator-mutator recursion* (35) *on Ω_c (Ethier and Kurtz, 1986, Ch. 11). For initial conditions satisfying Theorem 7.6 assumptions and sufficiently large K, finite-horizon trajectories track the deterministic class-conditional limit: single equilibrium under a primitive dominant class, periodic orbit under an irreducible periodic dominant class, or multi-modal behaviour when several classes are co-dominant.*

Proof. (a)–(b) are standard finite Markov-chain facts once reducibility is allowed explicitly. For (c), apply the standard density-dependent population-process limit (Ethier and Kurtz, 1986, Ch. 11): conditional one-step drift is the expected-fitness map (35), sampling fluctuations are $O(1/\sqrt{K})$, and class-conditional asymptotics follow from Theorem 7.6 (primitive blocks) together with Remark 7.9 and Proposition D.8 (irreducible periodic dominant block). (Morandotti and Orlando 2025 gives a recent rigorous derivation for Moran-type processes in weak-selection scaling.) □

The deterministic timescale is governed by the dominant-class ratio in Theorem 7.6(c). Under primitive mutation this reduces to the global ratio ρ_ε. Corollary 7.8 provides the δ_μ specialization in the full-separation feasible-veridical primitive regime. Finite-population Wright–Fisher dynamics add $O(1/\sqrt{K})$ fluctuations; with constrained mutation, stationarity and mixing should be interpreted within reachable communicating classes.

7.6 Combining: Complete Evolutionary Picture

We can now assemble the static optimality results (Section 4), the Price decomposition (Section 7.4), the quasispecies convergence (Theorem 7.6), and the finite-population link (Theorem 7.10) into a single statement.

Theorem 7.11 (Main synthesis)**.** *Under multi-task selection with mutation rate $\varepsilon > 0$ and positive expected fitness:*

(a) ***Static optimality*** *(Theorem 4.1): in the full-separation, sufficient-capacity regime, the fully veridical encoding is optimal (unique up to percept-label permutation); in general, ecological optima are capacity-aware minimisers of R on Ω_c.*

(b) ***Directional selection decomposition*** *(Theorem 7.4):*

$$\Delta \bar{R} = -T \cdot \mathrm{Var}_\pi(R)/\bar{W} + \mathbb{E}_\pi[W \cdot \delta R]/\bar{W}.$$

Selection decreases mean risk; mutation contributes a transmission term.

(c) ***Deterministic convergence on the quotient*** *(Theorem 7.6): with constrained mutation, the replicator-mutator recursion converges to the quasispecies equilibrium of the dominant reachable communicating class $\mathcal{K}^*$ (when dominance is unique). The limiting optimum is therefore the best* accessible *optimum, not necessarily the global optimum on Ω_c.*

(d) ***Rate statement (explicit form):*** *convergence is exponential/geometric with ratio controlled by the dominant class (Theorem 7.6(c)). If $\eta_{\varepsilon,\mathcal{K}} := \|A_{\varepsilon,\mathcal{K}^*} - A_{0,\mathcal{K}}\|_2 \leq \Delta W_{\mathcal{K}}/4$, then*

$$1 - \rho_{\mathcal{K}^*} \geq \frac{\Delta R_{\mathrm{gap},\mathcal{K}}}{2C}.$$

Under primitive mutation, this reduces to the global bound (with the corresponding global gap ΔR_{gap}):

$$1 - \rho_\varepsilon \geq \frac{\Delta R_{\mathrm{gap}}}{2C}.$$

In the full-separation feasible-veridical primitive regime, Corollary 7.8 yields $1 - \rho_\varepsilon \geq \pi_{\min}^2 \delta_\mu/(2C)$.

(e) ***Finite populations*** *(Theorem 7.10): with iid task sampling, Wright–Fisher dynamics converge on each fixed finite horizon in probability to the deterministic limit as $K \to \infty$, with $O(1/\sqrt{K})$ fluctuations; ergodic behaviour is class-conditional unless mutation renders the quotient chain irreducible, and periodic classes require time-averaged interpretation.*

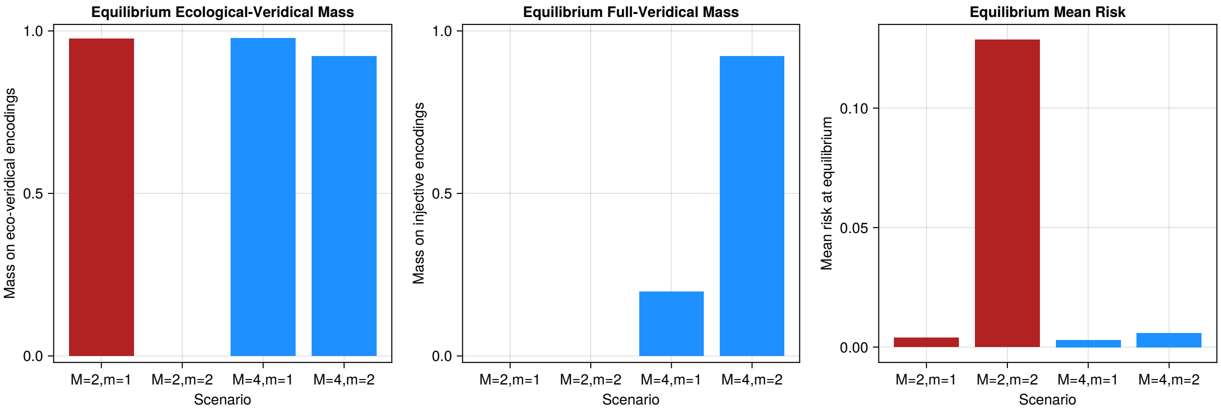


Figure 10: Exact deterministic replicator-mutator equilibrium for $N = 4$, four scenarios varying capacity $M \in \{2, 4\}$ and active task dimensions $m \in \{1, 2\}$. Left: mass on eco-veridical encodings (high when $M \geq k_\mu$). Centre: mass on fully injective encodings (high only when $M \geq N$ and separation is rich). Right: mean Bayes risk at equilibrium (near zero when eco-veridical, elevated in the lossy $M = 2$, $m = 2$ case). This validates Theorems 4.1, 4.2 and 7.6 without stochastic confounds.

Figure 10 provides a direct computational test of the synthesis in a small exact system ($N = 4$ world states, varying M and number of active task dimensions m). The deterministic replicator-mutator recursion is iterated to equilibrium with no sampling noise. When $M \geq k_\mu$ (capacity sufficient for the task-ecology's equivalence classes), equilibrium mass concentrates overwhelmingly on ecologically veridical encodings ($\geq$0.93) and risk is near zero; when $M < k_\mu$ (e.g. $M = 2$, $m = 2$), the population converges to lossy-optimal encodings with higher risk, exactly as Theorems 4.2 and 7.6 predict. Full injectivity (middle panel) is achieved only when $M \geq N$ and m is large enough to separate all points.

8 Recovery of Prior Results

A theory that claims to unify prior work must demonstrate that it recovers the prior results as special cases. Our framework does so in two directions:

- **Hoffman's FBT as the $T = 1$ special case.** For a single non-injective task f, the interface encoding p_f (which merges all states with equal f-value) attains zero Bayes risk, but so does any veridical encoding. In this regime there is no selective pressure for full veridicality: Hoffman's insight is exactly the case where μ does not separate points. See Appendix E for the formal statement (Proposition E.1).
- **Berke et al.'s simulations as a numerical instance.** With $N = 11$, $M = 2$, and beta-function tasks, our capacity-aware framework predicts a broad crossover from interface-like to ecologically optimal coarse encodings as T grows (Figure 2), matching the empirical pattern Berke et al. report. Full veridicality is impossible at $M = 2$; the relevant target is the best 2-cell ecological partition (Theorem 4.2). The detailed recovery is in Appendix E.

The capacity-dependent and geometry-dependent predictions of Sections 4 and 6 are further confirmed numerically: task families with well-conditioned geometry show much stronger veridicality transitions (Figures 5 and 6), the bottleneck distance δ_μ is a clean monotone predictor of veridicality across all families (Figures 3 and 4), evolutionary trajectories confirm that family structure, not task count, drives outcomes (Figure 8), and weighted ecologies produce lossy optima that match Theorem 4.2 (Figure 1). Detailed numerical analysis is in Appendix E. The heuristic correspondence to Frank (2025)'s force–metric–bias framework is discussed in Appendix C.

9 Discussion

9.1 Summary of Contributions

1. **The separation theorem and fitness gap** (Theorem 4.1): Multi-task Bayes risk vanishes if and only if the encoding is injective across task-equivalence classes, and is feasible if and only if $M \geq k_\mu$. Any encoding that merges a μ-separated pair incurs risk at least $\pi_{\min}^2 \cdot \delta_\mu$, providing strict selective pressure against ecologically invalid merges.

2. **Evolutionary convergence** (Theorems 7.4 and 7.6 and Remark 7.9): Price's equation identifies the direction of selection; quasispecies analysis via Perron–Frobenius proves class-conditional convergence of deterministic selection-mutation dynamics to the dominant reachable asymptotic regime on quotient encoding space. Rate is governed by the dominant spectral ratio, with an explicit lower bound in terms of the fitness gap.

3. **Graded separation cascade** (Proposition 4.6, Theorem 4.4): the resolved ecological complexity k_T is monotone non-decreasing as tasks are added. For structured task families (the biologically relevant case), this produces gradual transitions whose width is controlled by the condition number of the task family, a quantitatively testable prediction. Full veridicality is recovered only when both separation is complete and capacity is sufficient.

4. **Spectral control** (Theorem 6.2): In the Gaussian model, everything reduces to the spectrum of the task covariance Σ_c. The bottleneck direction, the task dimension with least diversity, controls both the fitness gap and the convergence rate. Simulations confirm that the operative variable is the bottleneck distance δ_μ itself: it predicts veridicality monotonically across all task families (Figure 4), while the condition number κ is a useful proxy but not the fundamental driver (Figure 3).

5. **Risk is a poor proxy for veridicality** (Figures 4 and 8): Mean Bayes risk is nearly flat across families with wildly different veridical fractions (0% to 63%), because ill-conditioned ecologies permit low-risk non-veridical encodings that merge pairs with negligible σ^2. The fitness gap for such merges is too small to overcome mutation pressure. This finding cautions against interpreting low risk as evidence of veridicality.

6. **Ecological Umwelt** (§9.4): The equivalence classes $[w]_\mu$ provide a mathematical framework that gives formal content to the concept of a species-specific perceptual world. Veridicality is always relative to the task ecology, resolving the Hoffman–Berke debate.

7. **Unification**: Hoffman's FBT and Berke's simulations are recovered as special cases / limiting regimes, with a heuristic bridge to Frank's FMB in Appendix C.

9.2 The Role of Cognitive Impenetrability

The entire argument depends on the encoding being *fixed* across tasks. If the organism could switch encodings per task, it could maintain a different (non-veridical) interface for each, and Hoffman's FBT would apply separately to each. Cognitive impenetrability is therefore the mechanism that creates selective pressure for richer shared representations; under mutation-development constraints this pressure can terminate at accessible local optima rather than a global optimum.

This connects to a deep point in information theory. A *sufficient statistic* $S(X)$ for a family of parameters $\{\theta\}$ is one that preserves all information about every θ. A fixed encoding that must serve all tasks is being pressed toward sufficiency for the entire task family. Under full separation ($k_\mu = N$) and sufficient capacity ($M \geq N$), the minimum-dimension sufficient encoding is the identity up to relabeling (veridical encoding). Outside that regime, sufficiency is only relative to $W/{\sim_\mu}$ (ecological veridicality).

9.3 Levels of Analysis: Static Optimality, Evolution, and Development

The results in this paper operate at two distinct levels, and conflating them leads to errors. In both levels, the mathematics is non-teleological: asymptotic states are consequences of local update rules and constraints, with no backward causal pull from future states.

Static optimality (Sections 3 to 6) characterises the minimisers of multi-task Bayes risk, and the associated objective gaps. These results are properties of the objective function $R(p)$, not of any particular optimisation process. They hold regardless of whether p is found by natural selection, gradient descent, reinforcement learning, exhaustive search, or divine fiat. If a process approximately minimises $R(p)$, the resulting encoding is approximately ecologically veridical. This generality is the reason the framework applies beyond biology.

Evolutionary dynamics (Section 7) proves that one *specific* optimisation process, natural selection modelled at mean-field level as a replicator-mutator recursion on a population of organisms, converges to the dominant mutation-accessible asymptotic regime (equilibrium under primitive mutation blocks; periodic cycle in irreducible periodic blocks; global optimum only under primitive global connectivity). The finite-population Wright–Fisher process tracks this deterministic flow on fixed finite horizons as K grows. The entities undergoing evolution are whole organisms, each carrying an encoding as a heritable trait. Selection operates on whole-organism fitness across all tasks. The Price equation, quasispecies theory, and spectral analysis are tools for analysing this particular population process. They do not apply to within-organism processes such as neural development, learning, or synaptic plasticity.

The distinction matters because a different optimisation process, such as stochastic gradient descent training a neural network, may also converge to an encoding that satisfies the static optimality conditions, but the *convergence proof* would require entirely different mathematics (e.g. the theory of implicit regularisation in gradient descent). The static theorems still characterise the minimiser set; only the dynamical proof of convergence changes. In evolutionary biology, this parallels the standard distinction between the *optimality model* (what should evolution produce?) and the *population-genetic model* (does the evolutionary process actually get there?). Our Sections 3 to 6 answer the first question; Section 7 answers the second, for the specific case of frequency-independent selection with mutation.

This two-level structure reflects a fundamental point about evolutionary theory that Sober (2008, pp. 362–363) articulates with particular clarity:

> *The distinction between laws and initial conditions also is important in evolutionary biology. The "laws of motion of populations" are general statements that are conditional in form. They say that if a population has a given set of properties at time t_1 and is subject to this or that evolutionary process then it has various probabilities of exhibiting different properties at time t_2. These laws make no predictions until initial conditions are specified. Duhem's thesis applies to evolutionary biology no less than it applies to physics, though it, of course, needs to be understood probabilistically.*

Our theorem layout follows this distinction exactly: Sections 3 to 6 state conditional laws (if the task ecology has property X, then the optimal encoding has property Y); Section 7 adds the initial conditions (mutation structure, starting distribution, population size) that convert these conditional laws into dynamical predictions. The same adaptationist methodology is stressed in Sober (2024): optimality models are methodologically informative when paired with explicit process models and explicit constraint statements, objective landscape first, process on that landscape second, scope limits third (§9.5).

This separation of concerns also clarifies the scope of the Baldwin effect (Baldwin, 1896; Hinton and Nowlan, 1987): if organisms can *learn* a good encoding within their lifetime (developmental optimisation), this learning does not replace evolutionary convergence but can accelerate it, because learned encodings that improve fitness increase the organism's reproductive success and hence the heritability of the architectural traits that enable such learning. The static optimality results characterise the objective landscape for both development and evolution; the dynamic results establish mean-field convergence and its spectral rate, plus finite-population approximation on fixed horizons.

9.4 Relative Veridicality and the Ecological Umwelt

A natural objection arises: no species is ever exposed to the full universe of possible tasks. A lineage of primates has never had to navigate the echolocation tasks of bats, the magnetoreception tasks of migratory birds, or the infrared detection tasks of pit vipers. Even over evolutionary time, the task distribution μ that a lineage encounters is a proper subset of all conceivable tasks. Does our theorem then require an unrealistic assumption of universal task exposure?

It does not. The theorem's conclusion is always *relative to the task distribution μ actually encountered.* This is not a weakness but the central insight.

What the mathematics actually says. The task distribution μ induces an equivalence relation $\sim_\mu$ on W (Definition 3.5): world states w_1 and w_2 are equivalent iff they are equal on μ-almost every task (equivalently $\sigma^2(w_1, w_2) = 0$). The equivalence classes $[w]_\mu$ partition the world into groups of states that are, from the organism's ecological standpoint, *functionally identical.* Theorem 4.1 and its generalisation (Theorem 4.2) then say:

- **Across equivalence classes:** The risk-minimising encoding is injective. Distinct classes must receive distinct percepts, because some task distinguishes them and merging incurs positive Bayes risk.
- **Within equivalence classes:** The encoding is *free.* No task distinguishes states within a class, so any assignment is equally fit. This is the gauge freedom of Proposition 4.3.

The result is not absolute veridicality but *ecological veridicality*: the encoding is injective up to the resolution of the task ecology. It preserves every distinction that matters for any task the lineage faces, and nothing more.

The graded separation cascade. As the task ecology expands, for example when a species enters a new niche, develops new behaviours, or faces new selective pressures, previously equivalent states may become distinguished. This is a monotone process: once a pair is separated it remains separated (adding tasks can only increase σ^2). The number of effective percept categories $k_\mu = |W/{\sim_\mu}|$ grows as a staircase function of ecological complexity. Each step in the staircase represents the emergence of a new perceptual distinction driven by a new task demand. The encoding tracks this cascade: it remains optimally adapted to the current task ecology, becoming progressively more veridical as that ecology diversifies.

Full veridicality as a limiting case. The condition $\delta_\mu > 0$ for all pairs, which gives full injectivity, is the limit of maximal task diversity. It is an idealisation, just as an ideal gas or a frictionless surface is an idealisation. In practice, every organism inhabits a niche with finite task diversity, and its perception is veridical only up to the resolution of that niche. The theory predicts exactly which distinctions are represented and which are not: the equivalence classes $[w]_\mu$ are a formal characterisation of the organism's *perceptual grain.*

Connection to Umwelt theory. This framework offers a mathematical formalisation of aspects of von Uexküll's (2010) concept of the *Umwelt*, the species-specific perceptual world. In our model, each species' Umwelt is determined by its task ecology μ. A bat and a primate sharing a forest canopy have different μ, hence different equivalence classes, hence different optimal encodings. Each is ecologically veridical, that is, veridical with respect to its own tasks, without either having privileged access to mind-independent reality. The gauge freedom within equivalence classes provides a formal analogue of what it means for a species to be "blind" to certain aspects of the world: not that it misrepresents them, but that no task exerts selective pressure to represent them at all. We note that this is a formalisation within a specific mathematical model, not a philosophical derivation of the Umwelt concept in its full richness.

Resolving the philosophical tension. This framing clarifies the apparent conflict between Hoffman and Berke. Hoffman is correct that fitness-maximisation does not require ontological truth, and indeed, within the unresolved equivalence classes, the encoding is unconstrained by selection (any assignment of percepts to states within a class yields equal fitness). Berke is correct that multi-task selection with cognitive impenetrability creates directional pressure toward preserving task-relevant world structure, and indeed, across resolved classes, the encoding must respect the separation structure induced by σ^2. Both are right, about different parts of the partition. Our contribution is to show that the partition itself, which classes are resolved and which are not, is determined by a single object: the task distribution μ. The mathematical freedom within equivalence classes is a statement about the fitness landscape (multiple encodings are equally fit), not a metaphysical claim about the organism's experience.

Anderson (2015, p. 1509) arrived at a closely related conclusion by philosophical argument. He proposed that veridicality should not be understood as correspondence with an inaccessible world-in-itself, but rather as congruence between different sets of observables:

> *If the fundamental "elements" to be compared in assessing veridicality are observables, the issue of veridicality becomes determining whether the equivalence classes generated by perceptual experience map onto (or into) the set of observables generated by some other given procedure of measurement. In this context, veridicality is a measure of the congruence between different sets of observables; it is not a measure of discordancy between perception and truth, because the latter has no meaning apart from an inaccessible God's eye view.*

Our equivalence classes $[w]_\mu$ and Theorem 4.1 give this argument precise mathematical form: ecological veridicality is exactly the condition that the perceptual partition (the encoding's equivalence classes) is at least as fine as the task-induced partition (the equivalence relation $\sim_\mu$). The "other procedure of measurement" in Anderson's formulation corresponds, in our framework, to the task ecology itself: the set of fitness-relevant observables that the organism's lineage has encountered.

9.5 Limitations and Shared Assumptions

Our convergence theorems (Section 7) separate two claims that are often conflated. Under constrained mutation, selection provably pushes populations toward the best *accessible* optimum (dominant reachable communicating class). Reaching the global optimum additionally requires strong connectivity of the mutation graph on quotient encoding space (primitive mutation), which is a modelling assumption that may fail under evo-devo viability constraints, canalisation, or strong genotype-phenotype restrictions. The space of possible encodings is combinatorially vast ($|\Omega| = M^N$ for unrestricted maps), and realistic mutational neighbourhoods are sparse relative to this space. This is the same constraint-sensitive lesson stressed in philosophy-of-biology critiques of simple adaptationism: treat adaptive hypotheses as one component in a plural causal analysis including developmental and historical alternatives (Pigliucci and Kaplan, 2006, ch. 5), and avoid reducing evolutionary explanation to a single cause when selection may be "main but not exclusive" (Sober, 2008, p. 361). More broadly, our simulation results (decreasing risk trajectories, increasing veridicality fractions) are what Pigliucci and Kaplan (2006, pp. 4–5) call “statistical shadows,” namely patterns compatible with the causal model we propose but not, by themselves, proof of that model over alternatives. As they emphasise, “the statistical shadows cannot be used as direct supporting evidence for any particular causal model”; what *can* be done is to derive alternative causal hypotheses, project their expected statistical patterns, and compare these against the observed data. Our analytic strategy follows this counsel: we derive predictions under distinct causal regimes (primitive vs constrained mutation, full-separation vs lossy capacity) and test each against simulation, rather than inferring mechanism from a single goodness-of-fit.

This limitation is shared with every optimality argument in the literature, including Hoffman’s FBT. Hoffman’s theorem proves that interface encodings are fitter than veridical ones for single tasks, but any dynamic conclusion still depends on a mutation-accessibility model (what transitions are viable) and population process assumptions. If developmental bottlenecks prevent access to some veridical optima, they can equally prevent access to some interface optima. The question “does evolution reach the theoretical optimum?” remains distinct from “what *is* the theoretical optimum?”; our static theorems (Section 4) answer the latter, while Section 7 answers the former conditionally on explicit mutation-accessibility assumptions.

As noted in Section 9.2, cognitive impenetrability may itself be an idealisation. If organisms can partially adapt their encoding to the current task (attentional modulation, context-dependent gain control), the effective task diversity is reduced. Our framework accommodates this as a reduction in the effective separation margin: a partially adaptive encoding faces a smaller effective task set, yielding ecological veridicality at a coarser grain. Full cognitive penetrability (instantaneous, costless re-encoding per task) recovers Hoffman’s FBT as the limiting case.

Finally, one might object that the separation condition is circular, namely that the theory predicts veridicality only by assuming, via the task distribution, that world-state distinctions matter. This misreads the logical structure. The separation condition $\sigma^2(w_1, w_2) > 0$ is an *empirical property* of the task ecology, not a theoretical assumption chosen for convenience. The theorem is a conditional: IF the task ecology separates a pair of states, THEN the optimal encoding distinguishes them. The direction of explanation runs from observable facts about fitness landscapes to predictions about encoding structure, not the other way around. The theory does not tell you which pairs are separated; that is an empirical input, just as Newtonian mechanics does not tell you where the masses are. What the theory provides is the machinery to convert a measured task ecology into precise predictions about which perceptual distinctions are maintained and which are not. Moreover, complete non-separation, the condition under which the theory would have nothing to say, is biologically implausible for any organism with more than one behavioural goal. Any two tasks that rank world states differently will separate at least some pairs. The only question is how many pairs and by how much, which is precisely what δ_μ quantifies.

An additional limitation concerns ecological endogeneity. We treat the task distribution μ as externally given, but in many lineages the task ecology is partly constructed by the organisms themselves (niche construction, environmental engineering). In that regime μ can be history-dependent (μ_t), and mutational accessibility can co-evolve with developmental architecture ($Q_{\varepsilon,t}$). Recent work emphasizing organism-environment feedback (Godfrey-Smith,

2024) makes this extension biologically important. Our current theorems are then interpreted piecewise-conditionally: at each ecological-developmental regime, they characterize the direction and accessible asymptotics of selection for that regime.

References

Barton L. Anderson. Where does fitness fit in theories of perception? *Psychonomic Bulletin & Review*, 22:1507–1511, 2015.

James Mark Baldwin. A new factor in evolution. *American Naturalist*, 30:441–451, 1896.

Jonathan Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

Marlene D. Berke, Robert Walter-Terrill, Julian Jara-Ettinger, and Brian J. Scholl. Flexible goals require that inflexible perceptual systems produce veridical representations. *Cognitive Science*, 46(10):e13195, 2022.

Manfred Eigen. Self-organization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58:465–523, 1971.

Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence.* Wiley, 1986.

Steven A. Frank. The Price equation reveals a universal force–metric–bias law of algorithmic learning and natural selection. *Entropy*, 27:1129, 2025.

Peter Godfrey-Smith. *Darwinian Populations and Natural Selection.* Oxford University Press, 2009.

Peter Godfrey-Smith. *Living on Earth: Forests, Corals, Consciousness, and the Making of the World.* Farrar, Straus and Giroux, 2024.

Anne Greenbaum, Ren-Cang Li, and Michael L. Overton. First-order perturbation theory for eigenvalues and eigenvectors. *SIAM Review*, 62(2):463–482, 2020. arXiv:1903.00785.

Geoffrey E. Hinton and Steven J. Nowlan. How learning can guide evolution. *Complex Systems*, 1:495–502, 1987.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Donald D. Hoffman. *The Case Against Reality.* W. W. Norton, 2019.

Donald D. Hoffman and Manish Singh. Computational evolutionary perception. *Perception*, 41:1073–1091, 2012.

Donald D. Hoffman, Manish Singh, and Chetan Prakash. The interface theory of perception. *Psychonomic Bulletin & Review*, 22:1480–1506, 2015.

Tosio Kato. *Perturbation Theory for Linear Operators.* Springer, 1995.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *JMLR*, 17:1–32, 2016.

Marco Morandotti and Gianluca Orlando. Replicator dynamics as the large population limit of a discrete Moran process in the weak selection regime. 2025. arXiv:2501.12688.

Jeong-Man Park, Enrique Muñoz, and Michael W. Deem. Quasispecies theory for finite populations. *Physical Review E*, 81(1):011902, 2010. doi: 10.1103/PhysRevE.81.011902. URL `https://link.aps.org/doi/10.1103/PhysRevE.81.011902`.

Massimo Pigliucci and Jonathan Kaplan. *Making Sense of Evolution: The Conceptual Foundations of Evolutionary Biology.* University of Chicago Press, 2006.

Chetan Prakash, Chris Fields, Donald D. Hoffman, Robert Prentner, and Manish Singh. Fact, fiction, and fitness. *Entropy*, 22(5):514, 2020.

Chetan Prakash, Kyle D. Stephens, Donald D. Hoffman, Manish Singh, and Chris Fields. Fitness beats truth in the evolution of perception. *Acta Biotheoretica*, 69:319–341, 2021.

George R. Price. Selection and covariance. *Nature*, 227:520–521, 1970.

George R. Price. Extension of covariance selection mathematics. *Annals of Human Genetics*, 35:485–490, 1972.

Zenon W. Pylyshyn. Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, 22(3):341–365, 1999.

Eugene Seneta. *Non-negative Matrices and Markov Chains.* Springer, 2006.

Elliott Sober. *Evidence and Evolution: The Logic Behind the Science.* Cambridge University Press, 2008.

Elliott Sober. *The Philosophy of Evolutionary Theory.* Cambridge University Press, 2024.

Jakob von Uexüll. *A Foray into the Worlds of Animals and Humans.* University of Minnesota Press, 2010. Trans. J. D. O'Neil; orig. pub. 1934.

A Proof Details for the Pairwise Variance Identity

Lemma A.1. *For a discrete random variable Z with $P(Z = z_i) = p_i$:*

$$\mathrm{Var}(Z) = \frac{1}{2} \sum_{i,j} p_i \, p_j \, (z_i - z_j)^2. \tag{46}$$

Proof.

$$\begin{aligned}
\operatorname{Var}(Z) &= \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 \\
&= \sum_i p_i z_i^2 - \Big(\sum_i p_i z_i\Big)^2 \\
&= \sum_i p_i z_i^2 \cdot \Big(\sum_j p_j\Big) - \Big(\sum_i p_i z_i\Big)\Big(\sum_j p_j z_j\Big) \\
&= \sum_{i,j} p_i p_j z_i^2 - \sum_{i,j} p_i p_j z_i z_j \\
&= \frac{1}{2}\sum_{i,j} p_i p_j (z_i^2 - 2 z_i z_j + z_j^2) \\
&= \frac{1}{2}\sum_{i,j} p_i p_j (z_i - z_j)^2.
\end{aligned}$$

(The last line uses the symmetry of the double sum to combine z_i^2 and z_j^2 terms.) □

B Hoeffding's Inequality

Theorem B.1 (Hoeffding 1963)**.** *Let $Y_1, \dots, Y_n$ be independent with $a_i \le Y_i \le b_i$. Then:*

$$P\Big(\bar{Y} - \mathbb{E}[\bar{Y}] \le -t\Big) \le \exp\left(-\frac{2n^2t^2}{\sum_i (b_i - a_i)^2}\right). \tag{47}$$

For iid variables in $[a, b]$: $P(\bar{Y} - \mathbb{E}[\bar{Y}] \le -t) \le \exp(-2nt^2/(b-a)^2)$.

C Heuristic Connection to Frank's FMB Law

This appendix is interpretive rather than theorem-proving. It maps the dynamics of Section 7 onto Frank's force–metric–bias (FMB) template:

$$\Delta\theta = M\, f + b + \xi, \tag{48}$$

with the following correspondences:

- f: local gradient-like selection pressure ($-\nabla R$ under smooth parameterisations of encodings);
- M: geometry/metric term (natural-gradient-style preconditioning, e.g. Fisher metric in softmax coordinates);
- b: systematic non-gradient drift terms (e.g. momentum/transport terms in algorithmic analogues);
- ξ: mutation or sampling noise.

Why heuristic only:

- The main model in Sections 2 to 7 is discrete over finite encoding classes; FMB is a differential parameter-space description.
- Softmax/Fisher parameterisations introduce gauge/degeneracy subtleties unless explicitly fixed.
- A full equivalence proof would require a separate manifold-level setup and regularity assumptions not needed for the core results.

What is rigorous in the main text is unaffected: static optimality (Theorems 4.1 and 4.2), concentration (Section 5), and quotient-space evolutionary convergence with spectral rates (Theorems 7.4, 7.6, 7.10 and 7.11). This appendix is a conceptual bridge to adjacent learning-dynamics frameworks.

D Full Proof of Theorem 7.4 (Reducible, Primitive-Block Case)

We provide the full proof of Theorem 7.6 in the constrained-mutation setting, where A_ε can be reducible.

D.1 Setup and Frobenius Form

Fix $\varepsilon > 0$ and write $A := A_\varepsilon = Q_\varepsilon^\top D$ on Ω_c (dimension n). Since $Q_\varepsilon \geq 0$ and D is positive diagonal, A is nonnegative and has the same directed support graph as Q_ε.

Lemma D.1 (Frobenius decomposition and closed-part representation)**.** *There exists a permutation matrix P (ordering transient classes first) such that*

$$PAP^\top = \begin{bmatrix} T & 0 \\ U & B \end{bmatrix}, \tag{49}$$

where T is the transient block (possibly empty) and $B = \operatorname{diag}(B_1, \ldots, B_L)$ collects closed communicating-class blocks $\mathcal{K}_1, \ldots, \mathcal{K}_L$.

For $y_0 := Px_0 = (y_0^{\mathrm{tr}}, y_0^{\mathrm{cl}})$, define $y_g := PA^g x_0$ and $z_g := (y_g)_{\mathrm{cl}}$. Then

$$z_g = B^g\, y_0^{\mathrm{cl}} + \sum_{t=0}^{g-1} B^{g-1-t}\, U\, T^t\, y_0^{\mathrm{tr}}. \tag{50}$$

Proof. Frobenius normal form for nonnegative matrices gives (49) after a suitable permutation of classes. Expanding powers of a block lower triangular matrix yields

$$(PAP^\top)^g = \begin{bmatrix} T^g & 0 \\ \sum_{t=0}^{g-1} B^{g-1-t}\, U\, T^t & B^g \end{bmatrix}.$$

Left-multiplying by y_0 and taking the closed component gives (50) after reindexing. □

D.2 Perron Expansion on Closed Blocks

Lemma D.2 (Block Perron expansions). *For each closed block B_i, define $\lambda_i := \rho(B_i)$. Under Theorem 7.6(ii), B_i is primitive, so there exist positive vectors v_i, u_i ($u_i^\top v_i = 1$), constants $C_i > 0$, and integer $s_i \geq 1$ such that*

$$B_i^g = \lambda_i^g \, v_i \, u_i^\top + R_i(g), \quad \|R_i(g)\| \leq C_i \, g^{s_i - 1} \, r_i^g, \tag{51}$$

where $r_i := \max_{k \geq 2} |\lambda_{k,i}| < \lambda_i$.

Define

$$I_{\max} := \{i : \mathcal{K}_i \textit{ reachable from } \operatorname{supp}(x_0),\ \lambda_i = \lambda_{\max}\}, \quad \lambda_{\max} := \max_{i \textit{ reachable}} \lambda_i. \tag{52}$$

Proof. Primitive Perron–Frobenius gives a simple dominant eigenvalue and positive eigenvectors; Jordan decomposition gives the remainder bound with polynomial factor $g^{s_i - 1}$. □

D.3 Unique Dominant Reachable Class (Theorem 7.6(a))

Proposition D.3 (Unique dominant reachable class convergence). *Assume Theorem 7.6(iii): there is a unique dominant reachable class $\mathcal{K}^*$ with $\lambda_* > \lambda_i$ for all other reachable i. Then there exist $c_* > 0$, $C > 0$, $s \geq 1$, $\theta \in (0, 1)$, and embedded Perron vector $\bar{v}_*$ (v_* on $\mathcal{K}^*$, zero elsewhere) such that*

$$A^g x_0 = \lambda_*^g (c_* \, \bar{v}_* + e_g), \quad \|e_g\| \leq C \, g^{s-1} \, \theta^g. \tag{53}$$

Consequently,

$$x_g := \frac{A^g x_0}{\mathbf{1}^\top A^g x_0} \to \frac{\bar{v}_*}{\mathbf{1}^\top \bar{v}_*} = x^*_{\mathcal{K}^*}. \tag{54}$$

Proof. Combine Lemma D.1 with Lemma D.2 on each reachable closed block. Every contribution has exponential rate λ_i^g times at most polynomial factors. Uniqueness of λ_* makes all non-$\mathcal{K}^*$ terms exponentially smaller; their maximum relative rate is absorbed into $\theta < 1$. Reachability of $\mathcal{K}^*$ from $\operatorname{supp}(x_0)$ implies positive mass transfer into $\mathcal{K}^*$ via (50), so the Perron projection coefficient $c_* > 0$. Normalize (53) to obtain the result. □

D.4 Tie Case (Outside Theorem 7.6 Assumption (iii))

Proposition D.4 (Tie-case limit set). *If $|I_{\max}| > 1$, then*

$$A^g x_0 = \lambda_{\max}^g \Big(\sum_{i \in I_{\max}} c_i \, \bar{v}_i + r_g \Big), \tag{55}$$

with $c_i \geq 0$ (not all zero) and $\|r_g\| / \lambda_{\max}^g \to 0$. Therefore every subsequential limit of normalized iterates lies in

$$\operatorname{conv}\Big\{ \bar{v}_i / (\mathbf{1}^\top \bar{v}_i) : i \in I_{\max} \Big\}. \tag{56}$$

In general, the limit is not unique.

Proof. Apply Lemmas D.1 and D.2 with equal top rates $\lambda_i = \lambda_{\max}$ for $i \in I_{\max}$. Lower-rate reachable blocks vanish relatively. Normalization converts nonnegative leading combinations into convex combinations of normalized Perron rays. Coefficients depend on initial projections/feeding in (50). □

D.5 Small-Mutation Expansion in Dominant Class (Theorem 7.6(b))

Lemma D.5 (Small-mutation dominant-block perturbation). *In dominant block $\mathcal{K}^*$, assume $P^*_{\mathcal{K}} = \{p^*_{\mathcal{K}}\}$ and*

$$Q_{\varepsilon,\mathcal{K}^*} = I + \varepsilon L_{\mathcal{K}^*} + O(\varepsilon^2),$$

Then $A_{\varepsilon,\mathcal{K}^*} = Q^\top_{\varepsilon,\mathcal{K}^*} D_{\mathcal{K}^*}$ has a simple Perron eigenpair near $\varepsilon = 0$, and the normalized Perron vector satisfies

$$x^*_{\mathcal{K}^*}(p) = \frac{\varepsilon \cdot W(p^*_{\mathcal{K}})\, L_{\mathcal{K}^*}(p^*_{\mathcal{K}} \to p)}{W(p^*_{\mathcal{K}}) - W(p)} + O(\varepsilon^2), \quad p \neq p^*_{\mathcal{K}}, \tag{57}$$

with $x^*_{\mathcal{K}^*} \to \delta_{p^*_{\mathcal{K}}}$ as $\varepsilon \to 0$.

In the uniform per-position mutation model, $L_{\mathcal{K}^*} = N(1-1/M)\,(Q_{1,\mathcal{K}^*} - I)$, yielding the equivalent effective-rate notation $\tilde{\varepsilon} = N\varepsilon(1-1/M)$ used in earlier drafts.

Proof. At $\varepsilon = 0$, $A_{0,\mathcal{K}} = D_{\mathcal{K}}$ is diagonal with simple top eigenvalue $W(p^*_{\mathcal{K}})$. Analytic perturbation of simple eigenpairs (Kato, 1995; Greenbaum et al., 2020) yields analyticity and the first-order correction. □

D.6 Spectral-Ratio Bound in Dominant Class (Theorem 7.6(c))

Proposition D.6 (Dominant-class spectral ratio and convergence rate). *Set $A_{0,\mathcal{K}} = D_{\mathcal{K}}$ and $\eta_{\varepsilon,\mathcal{K}} := \|A_{\varepsilon,\mathcal{K}^*} - A_{0,\mathcal{K}}\|_2$. If $\eta_{\varepsilon,\mathcal{K}} < \Delta W_{\mathcal{K}}/2$, then*

$$\rho_{\mathcal{K}^*} := \frac{|\lambda_{2,\mathcal{K}}|}{\lambda_{1,\mathcal{K}}} \leq \frac{W_{2,\mathcal{K}} + \eta_{\varepsilon,\mathcal{K}}}{W^*_{\mathcal{K}} - \eta_{\varepsilon,\mathcal{K}}} < 1. \tag{58}$$

With

$$\theta := \max\Big\{\rho_{\mathcal{K}^*},\ \max_{i \text{ reachable}, i \neq *} (\lambda_i/\lambda_*)\Big\} < 1, \tag{59}$$

(convention: max *over an empty set equals* 0*), there exist $C_0 > 0$ and $s \geq 1$ such that*

$$\|x_g - x^*_{\mathcal{K}^*}\| \leq C_0\, g^{s-1}\, \theta^g.$$

If $A_{\varepsilon,\mathcal{K}^*}$ is diagonalizable, the within-class polynomial factor is absent.

If $\eta_{\varepsilon,\mathcal{K}} \leq \Delta W_{\mathcal{K}}/4$, then

$$1 - \rho_{\mathcal{K}^*} \geq \frac{\Delta W_{\mathcal{K}}}{2W^*_{\mathcal{K}}} \geq \frac{\Delta W_{\mathcal{K}}}{2TC} = \frac{\Delta R_{\text{gap},\mathcal{K}}}{2C}. \tag{60}$$

Proof. Since $A_{0,\mathcal{K}}$ is diagonal/normal, Bauer–Fike gives

$$\min_j |\lambda - W(p_j)| \leq \eta_{\varepsilon,\mathcal{K}}$$

for every eigenvalue λ of $A_{\varepsilon,\mathcal{K}^*}$. If $\eta_{\varepsilon,\mathcal{K}} < \Delta W_{\mathcal{K}}/2$, disks around $W^*_{\mathcal{K}}$ and the rest are disjoint, implying

$$\lambda_{1,\mathcal{K}} \geq W^*_{\mathcal{K}} - \eta_{\varepsilon,\mathcal{K}}, \quad |\lambda_{2,\mathcal{K}}| \leq W_{2,\mathcal{K}} + \eta_{\varepsilon,\mathcal{K}}.$$

Proposition D.3 gives dominant-class convergence; replacing its abstract relative rate by the explicit max yields the rate bound. The inequality $1-\rho_{\mathcal{K}^*} \geq \Delta W_{\mathcal{K}}/(2W_{\mathcal{K}}^*)$ follows by direct algebra:

$$1-\rho_{\mathcal{K}^*} \geq 1-\frac{W_{\mathcal{K}}^*-\Delta W_{\mathcal{K}}+\eta_{\varepsilon,\mathcal{K}}}{W_{\mathcal{K}}^*-\eta_{\varepsilon,\mathcal{K}}} = \frac{\Delta W_{\mathcal{K}}-2\eta_{\varepsilon,\mathcal{K}}}{W_{\mathcal{K}}^*-\eta_{\varepsilon,\mathcal{K}}} \geq \frac{\Delta W_{\mathcal{K}}/2}{W_{\mathcal{K}}^*},$$

then $W_{\mathcal{K}}^* \leq TC$. □

D.7 Primitive Special Case (Corollaries 7.7 and 7.8)

Corollary D.7. *If quotient mutation is primitive on Ω_c, there is one closed class $\mathcal{K}^* = \Omega_c$. Then Theorem 7.6 reduces to standard primitive Perron–Frobenius global convergence. Corollary 7.8 follows by substituting Theorem 4.1(c): $\Delta R_{\text{gap}} \geq \pi_{\min}^2\delta_\mu$.*

D.8 Irreducible but Periodic Dominant Class (Extension of Remark 7.9)

Proposition D.8 (Periodic dominant block)**.** *Keep the setup of Theorem 7.6 except allow the dominant reachable block B_* to be irreducible with period $d > 1$ (not primitive). Then:*

(a) Normalized iterates need not converge to a single fixed point.

(b) There exist d limit points $x^{(0)}, \ldots, x^{(d-1)}$ such that along each residue class, $x_{nd+r} \to x^{(r)}$.

(c) The Cesàro average converges:

$$\frac{1}{G}\sum_{g=0}^{G-1} x_g \to x_*^{\text{PF}}, \tag{61}$$

where x_*^{PF} is the Perron profile on $\mathcal{K}^*$ (embedded in Ω_c).

Proof. Frobenius cyclic decomposition gives a permutation of B_* into d cyclic blocks; B_*^d is block diagonal with primitive diagonal blocks. Applying Proposition D.3 to B_*^d on each residue class yields subsequence limits $x^{(r)}$. Averaging over residues cancels periodic phases and converges to the Perron profile (standard imprimitive Perron–Frobenius asymptotics; Seneta 2006, Ch. 1). □

E Detailed Recovery of Prior Results

This appendix provides the full formal statements and numerical details summarised in Section 8.

E.1 Hoffman's FBT as the $T = 1$ Special Case

Proposition E.1. *For a single task f, the interface encoding p_f defined by:*

$$p_f(w_1) = p_f(w_2) \iff f(w_1) = f(w_2)$$

achieves $R(p_f, f) = 0$ while being non-injective whenever f is non-injective (i.e., whenever two world states yield the same fitness). Any veridical encoding also achieves $R(p, f) = 0$ but "wastes" percepts distinguishing states with equal fitness.

In the competition, the interface encoding wins in Hoffman's regime not because it has lower Bayes risk (both can attain zero for the single task) but through mutational/neutral-network effects among equal-fitness encodings. Which interface family dominates is then controlled by accessibility and mutation structure, not by a Bayes-risk advantage over veridical encodings as such.

This is exactly Hoffman's insight: for a single non-injective task, there is no selective pressure for full veridicality. Our framework recovers this as the case where task-equivalence classes are nontrivial (and typically δ_μ is not point-separating).

E.2 Berke et al.'s Simulations as Numerical Instance

Setting. Berke et al. use $N = 11$, $M = 2$, with tasks drawn as discretised beta functions.

Our capacity-aware framework predicts a transition in ecological fit as k_T grows with T. Full veridicality is impossible ($M < N$); the relevant target is the best 2-cell ecological partition. In structured beta families, k_T grows gradually rather than collapsing to N at $T = 1$, producing a broad crossover rather than a universal sharp threshold.

Berke et al. observe:

- $T = 1$: 92% interface (matches Proposition E.1)
- $T = 5$: early movement away from single-task interface-like encodings
- $T = 200$: majority near the ecologically optimal coarse partition
- $T = 2000$: strong concentration near that coarse optimum

These observations are consistent with Theorem 4.4/Corollary 4.5 (family-dependent, capacity-aware transition) and Theorem 7.11 (selection toward the capacity-aware optimum in the mutation-accessible class when separation pressure is present).

E.3 Task-Family Geometry and Capacity Threshold

The main transition claim in Theorem 4.4 is family-dependent rather than universal. Figure 5(a) shows that families with better-conditioned task geometry attain much higher full-veridical fractions at the same T; for example, in our runs with $M = 11$, `gauss_iso` reaches mean full-veridical fraction 0.6265 for $T \geq 11$, while `beta_narrow` remains near 0.009. Figure 5(b) isolates the capacity effect: with $M = 2$, full veridicality is identically zero across all conditions; with $M = 11$, it emerges broadly once separation is sufficiently rich.

A compact regime summary makes the effect size explicit (Figure 6): $M = 2$ gives mean full-veridical fraction 0.000 ($n = 2340$), while in favorable $M = 11$ regimes ($T \geq 11$, $\log_{10}(\kappa) \leq 2$) the mean rises to 0.3128 ($n = 530$), and for `gauss_iso` with $T \geq 11$ it reaches 0.6265 ($n = 240$). For reference, the null random-injective baseline at $N = 11$ is $11!/11^{11} \approx 1.399 \times 10^{-4}$, far below the observed favorable-regime fractions.

E.4 Ecological Weighting in the Lossy Regime

Theorem 4.2 predicts that when full injectivity is impossible, optimal encodings allocate resolution toward task-heavy regions. Figure 1(a) shows this reallocation directly: as the weight ratio (Category A / Category B) increases from 1 to 100, Region A gains distinct percept labels while Region B loses them. Figure 1(b) shows evolved final risk and the swap-search partition optimum as parallel curves over the same weight-ratio axis; the gap (mean approximately 0.032 in our sweep) reflects finite-population evolutionary overhead rather than a failure to track the optimum's direction.

F Supplementary Computational Figures

The main text carries only the figures needed for the argument arc. Additional corroboration, diagnostics, and finite-size checks are listed here for reproducibility and robustness inspection.

All code and data are available at `https://github.com/gvdr/evo_interface_veridicality`. Numerical summaries for these figures are recorded in `data/README.md` and the figure-specific notes in `figures/*/README.md`.

Supplementary figure	Main purpose	Generated by
`sim2_family_risk_vs_T`	Risk-view companion to the full-veridical transition by family	`analyze_sim2_summary.jl`
`sim2_kappa_scatter_M11`	Scatter-level check of condition-number vs veridicality trend	`analyze_sim2_summary.jl`
`sim2_risk_trajectories_M11`	Risk trajectories at $M = N = 11$ across task families	`analyze_sim2_summary.jl`
`sim3_risk_trajectories_by_weight_ratio`	Convergence diagnostics for non-uniform weighting scenarios	`analyze_and_plot_results.jl`
`quick_claim_check_panels`	Fast stochastic sanity check of the main transition claim	`quick_claim_check.jl`
`deterministic_claim_check`	Deterministic replicator-mutator bridge to theorem-level dynamics	`deterministic_claim_check.jl`
`quick_claim_bridge`	Parameter-sensitivity bridge (mutation/population/time horizon)	`quick_claim_bridge.jl`

Table 1: Supplementary computational figures. All scripts are run as `julia -project=. scripts/<script>`.