Learning to Choose with Categories: Theory and an Application to Algorithmic Prominence*

Philippe Jehiel† Aviman Satpathy‡

November 9, 2026

Click here for the latest version

Abstract

We introduce Coarse Q-learning, in which a decision-maker partitions alternatives into coarse similarity classes, chooses among them by logit, and learns class valuations by averaging experienced payoffs. Even in stationary bandit problems with independently drawn menus, choices change the composition of pooled feedback. This endogenous sampling can generate a unique globally attracting mixed equilibrium with limiting indifference, multiple stable strict rankings, or persistent cycling. Learning can reduce welfare even when the coarse representation admits efficient choice. An application demonstrates how digital platforms may exploit this feedback by promoting high-quality products to sustain purchases of inferior products under a common brand.

Keywords: Coarse Inference, Misspecified Learning, Reinforcement Learning
JEL Codes: C62, C73, D83, D91

*We thank Evan Friedman, Johannes Hörner, Giacomo Lanzani, Margarita Pavlova, Ariel Rubinstein, Jean-Marc Tallon, Olivier Tercieux, and Giacomo Weber for valuable comments. The AI tools, Refine.ink and GPT 6 Astra, were used to review the paper for consistency, accuracy, and clarity. All remaining errors are our own. This work has benefited from the support of the Agence Nationale de la Recherche in France through the program Investissements d'Avenir ANR-17-EURE-0001.

†Paris School of Economics & University College London; jehiel@enpc.fr
‡Paris School of Economics; aviman.satpathy@psemail.eu

1 Introduction

Consider a consumer, Alice, who has recently moved to Paris and dines out regularly. To discover her taste in wine, she orders a bottle to share with friends on each occasion. Facing a vast and unfamiliar selection across restaurants, Alice simplifies her choice by relying on wine categories based on region of origin, such as Bordeaux or Chianti, and price range. She typically favors the category she currently values the most, occasionally explores others, and uses each bottle's realized experience to update her assessment of its category. We ask how she will compare these categories after accumulating sufficient experience.

One might expect Alice to eventually learn each category's average quality, much as a bandit learner discovers an arm's expected payoff. This intuition, however, overlooks how her choices determine the composition of that average. Restaurants differ both in the categories they offer and in the wines representing each category. When Alice comes to favor Bordeaux, she orders it more often in restaurants where Chianti is also available, but her choice is unchanged where Bordeaux is the only available category. Relative to Chianti, her Bordeaux observations therefore place more weight on wines offered alongside Chianti. Even though every wine has a stationary payoff distribution and restaurant menus are drawn independently from a fixed distribution, the mixture of wines she samples within a category changes with her valuations. A similarity class is generally not a bandit arm with a fixed mixture distribution as the choices based on its assessment change the evidence used to assess it.

We study this endogenous sampling feedback mechanism in a stationary bandit problem with independently drawn menus and independent payoff shocks. Alternatives are partitioned into *similarity classes*, and the decision-maker maintains one valuation per class. She chooses among available classes using the logit rule, selects an alternative within the chosen class by a fixed tie-break, and updates the class valuation toward its realized payoff. Harmonic visit-count learning rates make this valuation a sample average of observed payoffs. We call this dynamic process *Coarse Q-learning* (CQL). It combines familiar reinforcement learning (Skinner, 1938; Caplin and Dean, 2008; Glimcher, 2011) with categorization (Rosch and Lloyd, 1978). Memory and sample constraints motivate tracking a few categories rather than many alternatives; limited salience provides a natural basis for organizing those categories around attributes such as region and price (Tversky, 1972; Bordalo et al., 2012, 2013).

With a separate class for every alternative, learning recovers the alternatives' expected payoffs in the long-run, and increasingly sensitive choice selects the best available options. Coarsening changes this conclusion because valuations and the distributions they estimate are jointly determined. The resulting feedback can produce robust indifference, several stable

strict rankings, or persistent fluctuations. These outcomes require neither non-stationary fundamentals nor strategic interaction: they arise from pooling heterogeneous observations whose composition depends on the learner's own choices. Throughout, sensitivity to valuation differences is held fixed during learning; we then study long-run outcomes as sensitivity increases, so that mistakes induced solely by logit exploration vanish in the limit.

Our central result identifies conditions under which experience leads to *indifference*. Return to Bordeaux and Chianti, and suppose, for simplicity, that the wines offered on their own are equally rewarding on average and sufficiently better than those offered when the two categories compete. This may reflect the expertise of restaurants specializing in a particular wine region, whose carefully curated offerings compare favorably with those of establishments catering to a broader range of tastes. A preference for Bordeaux then increases its exposure to less favorable competitive experiences, while Chianti's assessment relies more heavily on favorable experiences where it is the only option. This feedback corrects the initial preference. In the high-sensitivity limit, a strict ranking cannot sustain itself: the category losing every comparison would be assessed through its favorable singleton experiences, undermining its lower valuation. Learning instead sustains indifference and nontrivial mixing between categories, even when their representatives in the competitive menu have different expected payoffs. The result extends beyond this illustration.

With singleton menus for every class and a connected pattern of competition, a sufficiently large common increase in singleton payoffs yields a unique equilibrium to which learning converges from every initial assessment. As sensitivity increases, its limiting policy continues to mix in at least one competitive menu. Indifference thus arises for open sets of primitives, rather than for equal objective payoffs. Nor is the resulting randomization simply residual exploration - the composition of observations must remain balanced to sustain the equal limiting valuations. More generally, we identify a self-correcting feedback condition under which stronger competition from a rival weakly improves a class's sampled payoff, ensuring global convergence to a unique equilibrium, even in the absence of singleton menus.

Reversing the payoff comparison can reverse this feedback. If singleton experiences are sufficiently unfavorable, a class favored in competitive menus receives better observations, while a neglected class is assessed mainly through its disappointing singleton experiences. Under suitably rich menu support, several strict rankings become self-confirming and locally attracting from nearby valuations. Initial assessments and realized experiences can then have persistent path-dependent effects on choice. With three or more classes, and at least one class yielding heterogeneous expected payoffs across competitive menus, corrective and reinforcing feedback can also interact to prevent convergence of learning altogether. We construct an

open set of decision problems with only three classes in which, at every sufficiently high sensitivity, the unique equilibrium is unstable and our stochastic learning almost surely has a non-constant periodic limit set. Persistent cycling can therefore survive stationary fundamentals, sustained exploration, and diminishing learning rates.

To establish these results, we derive the mean-field dynamics of harmonic visit-count coarse Q-learning in continuous time using stochastic approximation. Their steady states, *smooth valuation equilibria* (SVE), exist at every finite sensitivity, and their high-sensitivity limits refine the valuation equilibria (VE) of Jehiel and Samet (2007). Using o-minimal geometry, we show that this refinement is generically finite even when the VE set contains continua. We also characterize precisely which strict VE survive vanishing logit exploration and show that they are approximated by locally stable SVE at all sufficiently high sensitivities.[1]

The welfare analysis separates the cost of coarse representation from the additional cost of coarse learning. Our benchmark is a fully informed oracle using the same categories, within-class choice rule, and common class scores across menus. Its optimal policy is generically unique and can always be implemented by a strict ranking of classes. When each menu has a unique best class, a common ranking selects it in every menu if and only if the required comparisons are acyclic. Even when this efficient policy is representable, learning need not attain it. Singleton payoffs make the distinction transparent as they contribute the same unavoidable payoff to every ranking, yet change the ranking or mixture that learning sustains. In the wine illustrations considered below, Bordeaux is better when both categories are available, but learning can either favor Chianti or mix between them indefinitely in the high-sensitivity limit. Reweighting feedback flexibly across menus and histories cannot generally close this welfare gap. We establish an impossibility result for payoff-assessment learning by constructing a family of decision problems in which any coarse learner who eventually assesses homogeneous classes correctly and uses the same monotone stochastic choice rule across menus suffers a strictly positive welfare loss relative to the oracle in the long-run.

Later in the paper, we develop an application that shows how a digital platform that sells its own products along with third-party ones can exploit this feedback through algorithmic prominence (Lee and Musolff, 2025; Farronato et al., 2023). Consumers pool experiences across products of unknown, heterogeneous quality sold under the platform's common in-house brand label, while a competitor offers constant quality. Prominence removes the competitor from consideration in selected product states. Under complete information, promoting the high-quality product cannot increase sales, since consumers already prefer it to

[1] A strict VE survives the refinement if and only if each class's valuation equals its average payoff across the menus that minimize its valuation disadvantage to competing chosen classes, weighted by their probabilities.

the competitor. Under coarse learning, however, the platform's optimal deterministic policy can promote only the high-quality product, with its favorable feedback sustaining purchases of inferior products under the same brand, even when the competitor remains available. The resulting disappointments offset the favorable promoted experiences, leaving consumers indifferent between brands. In our illustrations, this policy diverts sales from the competitor while preserving consumer welfare at its no-prominence level. Yet consumers could achieve a higher payoff using the same categories by buying the promoted high-quality product and otherwise choosing the competitor, a ranking that payoff-based learning cannot sustain. We also show how randomized prominence can further raise platform profits at consumers' expense, while prominence costs can make deterministic policies optimal.

Section 2 presents the learning model and its mean-field dynamics. Section 3 develops the equilibrium, learning, and welfare results, with illustrations of their mechanisms. Section 4 studies algorithmic prominence. Sections 5 and 6 discuss Bayesian foundations, interpretations, and related literature. Section 7 concludes. All proofs are in the appendix.

2 Model

2.1 Primitives

Let $\mathcal{A}$ be a finite set of alternatives and $\Psi := \{\psi \subseteq \mathcal{A} : \psi \neq \varnothing\}$ the set of menus. Time is discrete, $k \in \mathbb{N}_0$. In each period k, Nature draws a menu ψ_k independently according to an exogenous distribution $f \in \Delta(\Psi)$. Alice chooses an alternative $a \in \psi_k$ and observes a reward $r_{a,k}$. Each alternative a has a payoff distribution $R_a \in \Delta(\mathbb{R})$, which does not depend on the menu in which it appears. For each a, the sequence $\{r_{a,k}\}_{k\geq 0}$ is i.i.d. with law R_a, and these sequences are independent across alternatives and independent of the menu process. We assume uniform square-integrability: $\sup_{a\in\mathcal{A}} \mathbb{E}\left[r_{a,k}^2\right] \leq M_2 < \infty$, and write $\mu_a := \mathbb{E}[r_{a,k}]$ for the expected payoff of alternative a. Alice is risk-neutral and does not know the payoff distributions. She learns about expected payoffs from the rewards she observes.[2]

Alice represents the alternatives through an exogenously fixed partition $\mathcal{S}$ into *similarity classes*. Let $\rho : \mathcal{A} \to \mathcal{S}$, with $\rho(a) = s$ if $a \in s$, be the similarity projection assigning each alternative to its class. Alice represents an alternative only through its class and maintains one payoff estimate per class to guide her choices. This coarse representation economizes on both memory and data. When alternatives are numerous, maintaining a separate payoff estimate for each places substantial demands on memory, while learning about rarely sampled

[2] A standard interpretation is that each alternative a has a deterministic *fundamental* payoff μ_a, and the observed payoff is a noisy signal $r_{a,k} = \mu_a + \varepsilon_{a,k}$, where $\{\varepsilon_{a,k}\}_{k\geq 0}$ are i.i.d. mean-zero shocks. The stochasticity of payoffs may reflect either imperfect monitoring of the fundamentals or idiosyncratic taste shocks.

alternatives rests on few observations. Organizing learning and choice around a smaller set of categories reduces the number of estimates Alice must retain and allows a broader set of experiences within a class to inform the payoff estimate, in line with the principles of categorization studied in psychology (Rosch and Lloyd, 1978; Anderson, 1991).

Attribute salience provides a natural basis for organizing these categories. Alternatives typically differ along many payoff-relevant dimensions, only some of which attract Alice's attention. A consumer, for example, may distinguish wines by region and price while paying little attention to vintage or producer. We assume a fixed attention mechanism: an attribute is either salient or not. Alice groups together alternatives that coincide on the salient dimensions and learns a common expected payoff for each class. She treats the payoffs sampled from a class as i.i.d. draws conditional on a common but unknown stationary distribution. In reality, her classes need not reflect objective payoff homogeneity, because the salient attributes need not coincide with the attributes that are most predictive of payoffs.

2.2 Discrete-time Learning Model

Let $\mathcal{T} = (\mathcal{A}, \Psi, f, r)$ denote the stage decision problem defined by the primitives above. At the beginning of period k, Alice has a valuation $v_k(s) \in \mathbb{R}$ for each similarity class $s \in \mathcal{S}$, representing her current estimate of its expected payoff. Write $\mathbf{v}_k = (v_k(s))_{s\in\mathcal{S}}$ for the valuation vector, with $\mathbf{v}_0$ specifying her initial assessments. Alice perceives the realized menu ψ_k through the set of represented classes, $\omega_k := \rho[\psi_k] = \{\rho(a) : a \in \psi_k\}$. The history of observed class menus, chosen classes, and realized payoffs generate her information filtration $\mathcal{F}_k = \sigma\text{-alg}\Big((\omega_t, s_t, r_t)_{t<k}\Big)$ at the beginning of period k. Her initial assessments $\mathbf{v}_0$ and this history determine her current valuations $\mathbf{v}_k$. Given $\mathbf{v}_k$ and the realized class menu ω_k, she chooses among the available classes according to the multinomial logit (softmax) rule (Luce, 1959), commonly used in reinforcement learning models and logistic fictitious play (Sutton and Barto, 2018; Fudenberg and Levine, 1998; Hofbauer and Sandholm, 2002):

$$\sigma(s \mid \omega_k; \mathbf{v}_k) = \frac{\exp\Big(\beta v_k(s)\Big)}{\sum_{j\in\omega_k} \exp\Big(\beta v_k(j)\Big)}, \qquad s \in \omega_k,$$

with $\sigma(s \mid \omega_k; \mathbf{v}_k) = 0$ for unavailable classes. The parameter $\beta \in [0, \infty)$ measures Alice's sensitivity to valuation differences and remains fixed throughout the learning process. At $\beta = 0$, she chooses uniformly among the available classes. Higher sensitivity places greater weight on classes with higher valuations, although every available class retains positive probability at any finite β. Holding valuations fixed, choice probabilities concentrate on the highest-

valued classes as $\beta \to \infty$. We are particularly interested in analyzing the long-run behavior of the learning process for a decision-maker with sufficiently large sensitivity $\beta < \infty$.

Conditional on the chosen class s_k, an alternative a_k is drawn exogenously and uniformly from $s_k \cap \psi_k$.[3] Thus, Alice's valuations determine the probability assigned to each available class, which is then divided equally among its available members.[4] Denoting the cardinality of a set by $|\cdot|$, the resulting probability of choosing alternative $a \in \psi_k$ is

$$\nu_k(a \mid \psi_k, \mathbf{v}_k) := \Pr(a_k = a \mid \psi_k, \mathbf{v}_k) = \frac{\sigma(\rho(a) \mid \omega_k; \mathbf{v}_k)}{|\rho(a) \cap \psi_k|}. \tag{1}$$

After the choice, Alice observes the realized payoff $r_k := r_{a_k,k}$, but receives no counterfactual feedback about unchosen alternatives. She updates the valuation of the chosen class according to the temporal-difference (TD) recursion:

$$v_{k+1}(s) = v_k(s) + \mathbf{1}\{s_k = s\}\, \alpha_k(s)\Big(r_k - v_k(s)\Big), \qquad s \in \mathcal{S}. \tag{2}$$

The reward prediction error $r_k - v_k(s_k)$ measures the difference between the observed payoff and Alice's current assessment of the chosen class. A positive prediction error raises its valuation, while a negative one lowers it. The valuations of unchosen classes remain unchanged. This recursion is the bandit variant of asynchronous Q-learning (Watkins and Dayan, 1992; Sutton and Barto, 2018), with valuations maintained at the coarser similarity class level.[5] We use harmonic learning rates indexed by class visit counts. Let

$$N_k(s) := \sum_{t=0}^{k-1} \mathbf{1}\{s_t = s\}, \qquad \alpha_k(s) := \frac{1}{\kappa_s + N_k(s) + 1}, \qquad \kappa_s > 0, \tag{3}$$

where $N_k(s)$ counts the observations of class s collected before period k. The parameter κ_s determines the weight attached to Alice's initial assessment. The learning rate declines as she accumulates observations of that class: periods in which the class is not chosen leave the weight on its next observation unchanged. Under the maintained assumptions, valuations remain bounded almost surely and every class that appears with positive probability is sampled infinitely often. Consequently, the effective step-sizes satisfy the usual Robbins and

[3] Uniform tie-breaking uniquely obeys label-independence among alternatives within the chosen class. More generally, we allow for any fixed exogenous tie-breaking kernel $\Upsilon(\cdot \mid s, \psi)$ supported on $s \cap \psi$.

[4] At fixed valuations, duplicating an alternative within a class leaves the perceived class menu and class-choice probabilities unchanged. It only redistributes probability within that class, avoiding the usual issue of probability inflation associated with IIA in the "duplicates problem" of Debreu (1960).

[5] In the appendix, we allow for vanishing UCB-style optimism bonuses that encourage finite-sample exploration and show our asymptotic analysis is robust to such modifications of the learning rule. Also, we omit the bootstrapping term usually considered in MDPs since our i.i.d. menu process is action-independent.

Monro (1951) conditions, $\sum_k \mathbf{1}\{s_k = s\}\alpha_k(s) = \infty$ and $\sum_k \mathbf{1}\{s_k = s\}\alpha_k(s)^2 < \infty$, almost surely. The boundedness and sampling arguments are provided in the appendix.

This choice of harmonic learning rates gives valuations a direct interpretation as averages of past observations (see Section 5.1 for a Bayesian foundation). Iterating (2) yields

$$v_k(s) = \frac{\kappa_s v_0(s) + \sum_{t=0}^{k-1} \mathbf{1}\{s_t = s\} r_t}{\kappa_s + N_k(s)}. \tag{4}$$

Every observed payoff from class s receives the same weight in its current valuation, while the initial assessment carries the weight of κ_s observations.[6]

The averaging rule pools observations whose composition is shaped by Alice's choices. Different menus may offer different representatives of the same class, and her current valuations determine how frequently she selects that class in each menu. The conditional distribution of payoffs entering its sample average can therefore change as valuations evolve, even though each objective alternative has a fixed payoff distribution. Valuations influence the evidence used to revise them, creating the feedback at the heart of our coarse learning dynamics.

Together, (1)–(3), the initial valuations $\mathbf{v}_0$, and the pseudo-counts $(\kappa_s)_{s\in\mathcal{S}}$ specify the learning process in the given environment. The next subsection develops a continuous-time approximation that tracks both valuations and class visit frequencies to study its long-run behavior. With harmonic visit-count step sizes, Alice's response to a new observation depends on how often she has sampled the corresponding class. To describe the evolution of her valuations, we therefore keep track of the class sampling frequencies as well. Let $\chi := \sum_{s\in\mathcal{S}} \kappa_s$, and

$$x_{k,s} := \frac{\kappa_s + N_k(s)}{\chi + k}, \ s \in \mathcal{S}.$$

The vector $\mathbf{x}_k$ records the empirical frequencies of class choices, including the initial observation weights $\{\kappa_s\}_{s\in\mathcal{S}}$. Its coordinates are strictly positive and sum to one, and the contribution of the initial weights vanishes as observations accumulate.[7]

2.3 Continuous-time Asymptotic Approximation

To describe the expected motion of this process, fix a valuation vector $\mathbf{v}$. The probability that Alice chooses class s in a period is $\Lambda_s(\mathbf{v}) := \sum_{\psi:\, s\cap\psi\neq\varnothing} f(\psi)\,\sigma(s \mid \rho[\psi]; \mathbf{v})$. Conditional

[6] κ_s can be interpreted as a prior pseudo-count: a larger value makes Alice's initial assessment more resistant to revision, although its weight vanishes as observations accumulate. Under suitable conjugate priors, (4) coincides with the Bayesian posterior mean of a class's expected payoff under Alice's misspecified (subjectively) stationary model in the one-parameter natural exponential family; see Sec. 5.1 for details.

[7] For our analysis, we consider and retain only the classes that are available in at least one supported menu.

on choosing s from menu $\omega = \rho[\psi]$, uniform within-class selection yields expected payoff $\mu_s(\psi) := \dfrac{1}{|s \cap \psi|} \sum_{a \in s \cap \psi} \mu_a$. Averaging over the menus from which Alice chooses s yields

$$g_s(\mathbf{v}) := \frac{\displaystyle\sum_{\psi:\, s \cap \psi \neq \varnothing} f(\psi)\, \sigma(s \mid \rho[\psi]; \mathbf{v})\, \mu_s(\psi)}{\displaystyle\sum_{\psi:\, s \cap \psi \neq \varnothing} f(\psi)\, \sigma(s \mid \rho[\psi]; \mathbf{v})}.$$

$g_s(\mathbf{v})$ is the expected payoff observed when class s is chosen under valuations $\mathbf{v}$. The difference $g_s(\mathbf{v}) - v_s$ is Alice's expected prediction error for that class. A positive expected prediction error raises its valuation, while a negative one lowers it. After rescaling time to account for the diminishing learning rates, the mean-field asynchronous dynamics are

$$\begin{aligned} \dot{v}_s &= \frac{\Lambda_s(\mathbf{v})}{x_s}\Big(g_s(\mathbf{v}) - v_s\Big), \\ \dot{x}_s &= \Lambda_s(\mathbf{v}) - x_s, \end{aligned} \qquad s \in \mathcal{S}. \tag{5}$$

Here, $\mathbf{x}$ is a strictly positive probability vector, and the derivatives are taken with respect to learning time. The two equations reflect the distinction between current behavior and accumulated experience. Class s is currently chosen with probability $\Lambda_s(\mathbf{v})$, while its learning rate depends on its historical sampling frequency x_s. At a sufficiently late date k, a new observation of s receives weight approximately $1/[(\chi + k)x_s]$. Combining this weight with the probability of observing s yields the factor $\Lambda_s(\mathbf{v})/x_s$ in the valuation equation.

For a given expected prediction error, valuations adjust faster when a class is currently sampled more often than its historical frequency would suggest. The frequency equation describes how that history itself changes: x_s rises when the current probability of choosing s exceeds its accumulated share, and falls when the inequality is reversed. Diminishing learning rates also make the effect of payoff and sampling noise small over fixed intervals of learning time. This connection between the stochastic recursion and the ODE can be stated precisely using standard stochastic-approximation results (Kushner and Yin, 2003).

Proposition 1 (Continuous-time approximation)**.** *Under the maintained assumptions, fix $\beta \in [0, \infty)$, positive initial weights $\{\kappa_s\}_{s \in \mathcal{S}}$, and a finite initial valuation vector $\mathbf{v}_0$. Then, $\sup_{k \geq 0} \|\mathbf{v}_k\| < \infty$, and $\liminf_{k \to \infty} \min_{s \in \mathcal{S}} x_{k,s} > 0$ almost surely. Let $Z(t)$ be the piecewise-linear interpolation of $\{(\mathbf{v}_k, \mathbf{x}_k)\}_{k \geq 0}$ at times $\{\tau_k\}_{k \geq 0}$, and let Φ_h denote the flow of (5). Then Z is an asymptotic pseudo-trajectory of this flow: almost surely, for every $T > 0$,*

$$\lim_{t \to \infty} \sup_{0 \leq h \leq T} \Big\| Z(t + h) - \Phi_h(Z(t)) \Big\| = 0.$$

Consequently, the set of subsequential limits of $\{(\mathbf{v}_k, \mathbf{x}_k)\}_{k\geq 0}$ *is almost surely nonempty, compact, connected, and internally chain transitive for the asynchronous joint ODE* (5).[8]

The first conclusion establishes boundedness of the valuation process and continued sampling of every class. The approximation statement then gives a precise sense in which the ODE describes learning: starting sufficiently late, the interpolated process follows the ODE from its current state arbitrarily closely over any fixed interval of learning time.[9] Its long-run outcomes are therefore constrained by the recurrent behavior of the joint ODE system.

Menu variation affects the evidence pooled within a class in two ways. First, different menus may contain different representatives of the same class. Their expected payoffs may generally differ, so the mean payoff from choosing a class can vary across menus, even though each alternative has a fixed payoff distribution.[10] Second, Alice's valuations determine how often she selects that class in each menu through competition with the other available classes. The mixture of menus contributing to its payoff history is therefore selected by her own choices. Current valuations shape the composition of the evidence she observes, and this evidence in turn shapes future valuations. This feedback makes learning *active* and *endogenous*.

Reduction of the tree to class menus: The mean-field dynamics admit a simpler tree representation that groups together menus containing the same similarity classes. Let $\Omega := \{\omega \subseteq \mathcal{S} : \omega \neq \varnothing\}$, and $\Psi_\omega := \{\psi \in \Psi : \rho[\psi] = \omega\}$. The probability of encountering class menu ω is $p(\omega) := \sum_{\psi\in\Psi_\omega} f(\psi)$. For each ω with $p(\omega) > 0$ and each $s \in \omega$, define

$$\pi_s(\omega) \;=\; \frac{\sum\limits_{\psi\in\Psi_\omega} f(\psi)\,\mu_s(\psi)}{\sum\limits_{\psi\in\Psi_\omega} f(\psi)}, \qquad \mu_s(\psi) \;:=\; \frac{1}{|s\cap\psi|}\sum_{a\in s\cap\psi}\mu_a.$$

This is the expected payoff from choosing class s when the perceived menu is ω. To see why the averaging weights depend only on menu probabilities, recall that for a fixed $\mathbf{v}$ Alice uses the same class-choice probabilities in every objective menu inducing ω. Conditional on choosing s, these menus therefore retain their relative occurrence probabilities.

[8] The proof is provided in the appendix. Internal chain transitivity means that any two points in the invariant set can be connected by a finite chain of orbit segments lying in the set, with arbitrarily small jumps between segments and an arbitrarily large lower bound on their durations; see Benaïm (1999).

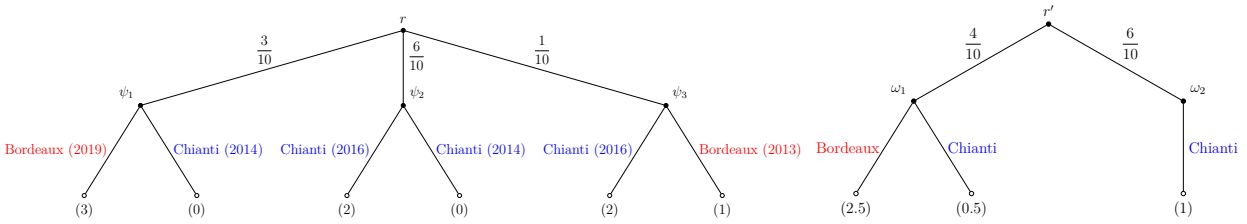
[9] The corresponding clock is $\tau_0 = 0$ and $\tau_k = \sum_{j=0}^{k-1} \frac{1}{\chi + j + 1}$. Thus, $\tau_k \sim \log k$: a fixed interval of learning time contains progressively more observations as k grows.

[10] The same mean-field ODE construction accommodates menu-dependent primitive payoff distributions $R_{a,\psi}$, under the corresponding conditional independence and uniform second-moment assumptions. Writing $\mu_a(\psi)$ for their means, one replaces $\mu_s(\psi)$ by $|s\cap\psi|^{-1}\sum_{a\in s\cap\psi}\mu_a(\psi)$. We maintain menu-independent primitive payoffs to isolate the menu dependence generated by coarse aggregation.

Writing $\sigma_\omega^s(\mathbf{v}) := \sigma(s \mid \omega; \mathbf{v})$, the class-selection probabilities and conditional expected payoffs become $\Lambda_s(\mathbf{v}) = \sum_{\omega \in \Omega:\, s \in \omega} p(\omega)\, \sigma_\omega^s(\mathbf{v})$, and

$$g_s(\mathbf{v}) = \frac{\displaystyle\sum_{\omega \in \Omega:\, s \in \omega} p(\omega)\, \sigma_\omega^s(\mathbf{v})\, \pi_s(\omega)}{\displaystyle\sum_{\omega \in \Omega:\, s \in \omega} p(\omega)\, \sigma_\omega^s(\mathbf{v})}.$$

The reduced decision tree $\mathcal{T}' = (\mathcal{S}, \Omega, p, \pi)$ thus preserves both Λ and g, and hence the joint mean-field dynamics in (5). Figure 1 illustrates the reduction. We use this class-menu representation throughout the analysis.



(a) Alice's original decision tree $\mathcal{T}$ **(b)** Alice's reduced decision tree $\mathcal{T}'$

Figure 1: Example of tree-reduction for the mean-field CQL dynamics

Elementary benchmarks: The following four special benchmarks eliminate the feedback from valuations to expected class payoffs. In each case, $g(\mathbf{v})$ is a constant vector $\bar{\boldsymbol{\pi}}$, and the mean-field dynamics (5) trivially converge to these valuations from every initial condition. The stochastic learning process (2) likewise satisfies $\mathbf{v}_k \to \bar{\boldsymbol{\pi}}$ almost surely.[11]

(i) *Pure exploration:* When $\beta = 0$, Alice chooses uniformly among the available classes, independently of her valuations. The expected payoff observed from class s is therefore a constant $\bar{\pi}_s = \dfrac{\sum_{\omega \ni s} p(\omega) |\omega|^{-1} \pi_s(\omega)}{\sum_{\omega \ni s} p(\omega) |\omega|^{-1}}$. Menu composition affects this average, but Alice's valuations do not affect which menus contribute to it.

(ii) *Finest partition:* When every similarity class contains a single alternative, Alice maintains a separate valuation for each alternative. Every observation entering the valuation of $\rho(a) := a$ then has conditional mean μ_a, so $g_{\rho(a)}(\mathbf{v}) = \mu_a$. Learning recovers each alternative's expected payoff asymptotically as in standard Q-learning over alternatives.

(iii) *Constant class availability:* Suppose every menu contains a representative of every class, so that $p(\mathcal{S}) = 1$. Then $g_s(\mathbf{v}) = (\sigma_{\mathcal{S}}^s(\mathbf{v})\, \pi_s(\mathcal{S}))/(\sigma_{\mathcal{S}}^s(\mathbf{v})) = \pi_s(\mathcal{S})$. Although

[11] For the joint asynchronous dynamics, $\frac{d}{dt}\big[x_s(t)(v_s(t) - \bar{\pi}_s)\big] = -x_s(t)(v_s(t) - \bar{\pi}_s)$. Bounded valuations and positive class availability ensure that sampling frequencies eventually stay away from zero. Hence $\mathbf{v}(t) \to \bar{\boldsymbol{\pi}}$ and $\mathbf{x}(t) \to \Lambda(\bar{\boldsymbol{\pi}})$. The almost-sure convergence of the stochastic process follows from Proposition 1.

different alternatives may appear across physical menus, the set of competing classes is always the same. Alice therefore chooses a given class with the same probability in every menu, leaving the composition of its observed payoffs independent of valuations.

(iv) *Homogeneous class means:* The finest-partition benchmark extends to any partition satisfying $\mu_a = \bar{\pi}_s$ for every $a \in s$. In that case, $\mu_s(\psi) = \bar{\pi}_s$ whenever class s is available, and hence $g_s(\mathbf{v}) = \bar{\pi}_s$. Changes in the composition of sampled alternatives leave expected class payoffs unchanged as they are constant across menus.

These benchmarks identify special circumstances in which the composition of experience has no feedback effect on expected class payoffs. With heterogeneous alternatives and varying sets of competing classes, Alice's valuations determine her choices which can in turn change the distribution of expected rewards entering her class estimates. This feedback raises two questions: which valuations are consistent with the choices they induce, and when does learning converge to such valuations? We call these self-consistent steady states *smooth valuation equilibria (SVE)* and turn next to their structure and stability.

3 Results

We begin by analyzing the structure of SVE, their high-sensitivity limits LSVE, and the local stability of strict limiting equilibria. We then examine global convergence, the effects of singleton-payoff shifts on indifference and multiplicity, and persistent cycling. Finally, we distinguish the welfare costs of coarse representation and learning. Examples throughout illustrate the feedback between choice and experience driving these outcomes.

3.1 Equilibrium Structure

For equilibrium and local stability analysis, it is convenient to introduce the *normalized mean-field dynamics*

$$\dot{v}_s = g_s(\mathbf{v}) - v_s, \qquad s \in \mathcal{S}. \tag{6}$$

Under these simpler dynamics, each valuation adjusts at unit rate toward the expected payoff generated by the current choice policy. The endogenous composition of observed payoffs is captured by the same map g, while adjustment rates are synchronized across classes. Its steady states correspond one-to-one to the steady states of the asynchronous joint dynamics (5): a valuation vector $\mathbf{v}^*$ is stationary if and only if $\mathbf{v}^* = g(\mathbf{v}^*)$, and the corresponding visit frequencies are uniquely determined by $\mathbf{x}^* = \Lambda(\mathbf{v}^*)$. The two systems also agree on the local asymptotic stability of corresponding hyperbolic steady states. At corresponding steady states, the Jacobian of the asynchronous joint dynamics (5) has the

eigenvalues of $Dg(\mathbf{v}^*) - I$, together with $|\mathcal{S}| - 1$ additional eigenvalues equal to -1. Hence hyperbolicity and, for hyperbolic steady states, local asymptotic stability coincide across the two systems.[12] The appendix formally establishes both the stated correspondences.

Thus, at a steady state, Alice's valuation of each class equals the expected payoff she receives conditional on choosing it. Her choices determine which observations enter this assessment, while the assessment itself guides her choices. A smooth valuation equilibrium is a fixed point that requires these two elements to be mutually consistent.

Definition 1 (Smooth valuation equilibrium)**.** *Fix a reduced decision tree $\mathcal{T}' = (\mathcal{S}, \Omega, p, \pi)$ and sensitivity $\beta \in [0, \infty)$. A pair $(\mathbf{v}, \sigma)$ is a* ***smooth valuation equilibrium*** *(SVE) if:*

$$\sigma_\omega^s = \mathbf{1}\{s \in \omega\} \frac{\exp(\beta v_s)}{\sum_{r \in \omega} \exp(\beta v_r)}, \qquad v_s = \frac{\sum_{\omega \ni s} p(\omega)\, \sigma_\omega^s\, \pi_s(\omega)}{\sum_{\omega \ni s} p(\omega)\, \sigma_\omega^s}.$$

The first equality applies to every supported menu ω, and the second to every $s \in \mathcal{S}$.

The first condition specifies logit-optimal choices given valuations. The second, consistency, requires each class valuation to equal the expected payoff in the sample generated by those choices. Thus, Alice correctly assesses the average payoff of each class under her own sampling behavior. The equilibrium condition allows the alternatives within a class to have different expected payoffs: their contributions to its valuation depend endogenously on how frequently they are encountered and chosen. Equivalently, equilibrium valuations satisfy $\mathbf{v}^* = g(\mathbf{v}^*; \beta)$, with the associated choice profile uniquely determined by the logit rule. At every finite sensitivity, each available class is chosen with positive probability. Consequently, its equilibrium valuation is an average of its expected payoffs across all supported menus in which it appears, with strictly positive weight on each such menu. Writing

$$\Omega_s^+ := \{\omega \in \operatorname{supp}(p) : s \in \omega\}, \qquad K := \prod_{s \in \mathcal{S}} \left[\min_{\omega \in \Omega_s^+} \pi_s(\omega), \max_{\omega \in \Omega_s^+} \pi_s(\omega) \right],$$

every SVE valuation lies in the relative interior of K. We denote the set of equilibrium valuations by $\mathcal{V}(\beta) := \{\mathbf{v} \in K : \mathbf{v} = g(\mathbf{v}; \beta)\}$.

High-sensitivity limits of SVE: To understand equilibrium behavior when choices respond strongly to valuation differences, we study SVE as $\beta \to \infty$, holding the decision problem fixed. A vector $\mathbf{v}^\infty$ is a *limiting smooth valuation equilibrium* (LSVE) if there exist

[12] A steady state is hyperbolic (hence, regular) if its Jacobian has no eigenvalue with zero real part. Additionally, it is locally asymptotically stable if every such eigenvalue has a negative real part.

$\beta_n \to \infty$ and $\mathbf{v}_n \in \mathcal{V}(\beta_n)$ such that $\mathbf{v}_n \to \mathbf{v}^\infty$. We write $\mathcal{V}(\infty)$ for the set of all such limits.[13] The natural benchmark is the arg max (or greedy) counterpart of SVE, called a *valuation equilibrium* (VE), following the analogous solution concept introduced by Jehiel and Samet (2007) for extensive-form games. A VE consists of a valuation vector and a choice profile such that Alice chooses only among valuation-maximizing classes in each menu, allowing menu-specific mixing over ties. Each class chosen with positive aggregate probability is valued consistently at its expected payoff conditional on being chosen. For classes that are never chosen, payoff consistency imposes no additional restriction, although their valuations remain subject to the menu-wise optimality conditions (see the appendix for a formal definition). Let $\mathcal{VE}$ denote the set of valuation vectors that can be supported in this way.

Theorem 1 proves the existence of SVE, describes how their structure varies with sensitivity, and relates their high-sensitivity limits to VE. In particular, while every accumulation point of the SVE correspondence as $\beta \to \infty$ is a VE; generically, the converse need not hold.

Theorem 1. *Fix a decision tree $(\mathcal{S}, \Omega, p, \pi)$. For $\beta \geq 0$, let $\mathcal{V}(\beta) := \{\mathbf{v} \in K : \mathbf{v} = g(\mathbf{v}; \beta)\}$ denote the SVE set, and let $\mathcal{V}(\infty)$ denote the set of accumulation points of $\{\mathcal{V}(\beta)\}_{\beta \uparrow \infty}$. Then:*

(i) For every $\beta < \infty$, $\mathcal{V}(\beta)$ is nonempty and compact, and the correspondence $\beta \mapsto \mathcal{V}(\beta)$ is upper hemicontinuous and compact-valued on $[0, \infty)$. The SVE graph $\{(\mathbf{v}, \beta) : \mathbf{v} \in \mathcal{V}(\beta)\}$ has finitely many path-connected components. For generic parameters (p, π) and all but finitely many $\beta \in [0, \infty)$, $\mathcal{V}(\beta)$ is finite and consists of an odd number of hyperbolic equilibria, each lying on a locally unique real-analytic branch in β.[14]

(ii) Every accumulation point of SVE as $\beta \to \infty$ is a VE: $\mathcal{V}(\infty) \subseteq \mathcal{VE}$. Equivalently, for every $\varepsilon > 0$, $\exists\ \widehat{\beta} < \infty$ such that $\forall\ \beta \geq \widehat{\beta}$, every $\mathbf{v} \in \mathcal{V}(\beta)$ lies in an ε-neighborhood of some $\mathbf{v}^\infty \in \mathcal{VE}$. For generic (p, π), $\mathcal{V}(\infty)$ is a nonempty finite subset of $\mathcal{VE}$.[15]

Theorem 1 establishes that smooth valuation equilibrium is well-defined and places substantial restrictions on its multiplicity. Equilibria exist at every nonnegative sensitivity and, for generic decision trees, are hyperbolic and finite in number at all but finitely many sensitivities. More strikingly, their high-sensitivity limits form a nonempty finite subset of VE.

[13] This limit concerns sensitivity across learning environments. Within each environment, β remains fixed along each learning path for all forward time.

[14] Throughout, genericity is understood with respect to the numerical parameters of a fixed reduced tree structure: we hold $\mathcal{S}$, Ω, and $\text{supp}(p)$ fixed and vary the positive menu probabilities and the class-menu expected payoffs. The resulting parameter space carries the relative Euclidean topology, with probabilities constrained to sum to one, and Lebesgue measure on its affine hull. A property holds *generically* if it holds on a comeager subset of this space that also has full measure. Valuation vectors and sensitivities carry their usual Euclidean topologies, and the equilibrium graph carries the corresponding product topology.

[15] The appendix also shows that the SVE connected component containing the unique equilibrium at $\beta = 0$ contains an equilibrium at every finite sensitivity and, generically, has a unique high-sensitivity limit.

The finiteness of the limiting set is substantive because the underlying set of VE need not be finite, even for generic primitives. Optimal choice leaves the valuation of a never-chosen class unrestricted by payoff consistency. An underdetermined VE can also arise when every class is chosen: ties permit menu-specific mixing, and these mixing probabilities can adjust together with valuations while preserving both indifference and payoff consistency. Such continua can persist under small perturbations of the primitive payoffs and probabilities.[16]

The proof sidesteps this difficulty by studying the entire SVE correspondence as sensitivity varies. We show that the SVE graph in $(\mathbf{v}, \beta)$-space is definable in the o-minimal structure $\mathbb{R}_{\exp}$ (real ordered field with the exponential function) since its equilibrium equations involve only algebraic operations and exponentials. Generic regularity via standard transversality arguments makes this graph one-dimensional; o-minimality then ensures finitely many path-connected components and, after compactifying sensitivity, finitely many boundary points corresponding to $\beta = \infty$. This yields a generically finite refinement even when VE form a continuum, retaining precisely those equilibrium valuations that persist as limits of SVE under vanishing logit exploration. This refinement also has an axiomatic foundation.[17]

We next study the stability and convergence of the CQL dynamics.[18] We assess local stability using the eigenvalues of the Jacobian of (5), and establish global convergence through strict Lyapunov functions or contraction arguments in the sup norm.

3.2 Strict Equilibria: Refinement and Local Stability

Theorem 1 establishes that every LSVE (high-sensitivity limit of SVE) is a VE, but the converse need not hold. We now characterize which strict VE survive this refinement. The additional restriction concerns classes that Alice never chooses under the limiting optimal policy. Although VE leaves their valuations unrestricted, these classes continue to be sampled at every finite sensitivity. Their valuations must therefore remain consistent with the observations generated by this increasingly rare exploration as $\beta \to \infty$.

Definition 2 (Strict VE)**.** *A VE valuation $\mathbf{v}^*$ is strict if: (i) every supported menu ω has a unique valuation-maximizing class chosen in equilibrium, denoted by $s^*(\omega) := \arg\max_{j \in \omega} v_j^*$;*

[16] For example, consider three equally likely menus $\{A, B\}$, $\{B, C\}$, and $\{C, A\}$, with payoffs $(1, 0)$ to the classes listed in each menu. For every $t \in (0, 1)$, choosing the first-listed class with probability t in each menu generates conditional mean payoff t for every class. Thus (t, t, t) is a VE valuation vector. The consistency equations are regular in the three mixing probabilities, so a local continuum of such valuation vectors persists under small perturbations of the primitives. The unique LSVE here is $(1/2, 1/2, 1/2)$.

[17] On a rich domain of menus, positivity, continuity, label- and context-independence, IIA, translation invariance, and monotonicity characterize the logit family up to nonnegative sensitivity (Breitmoser, 2021).

[18] Local asymptotic stability requires trajectories starting sufficiently close to remain close and converge to the equilibrium. Global asymptotic stability additionally requires convergence from every initial condition.

and (ii) the payoff-consistent valuations induce a strict total order on the set of classes chosen in at least one supported menu, $\mathcal{W}(\mathbf{v}^*) := \{s^*(\omega) : \omega \in \text{supp}(p)\}$.

Thus, distinct classes in $\mathcal{W}(\mathbf{v}^*)$ have distinct valuations, while no ordering among never-chosen classes is required. We refer to condition (i) alone as *menu-wise strictness*. For generic decision trees, menu-wise strictness already implies condition (ii).[19]

To understand how the remaining classes are assessed, consider the valuation loss associated with choosing class s in menu ω: $\Delta_s(\omega; \mathbf{v}^*) := v^*_{s^*(\omega)} - v^*_s,\ \omega \in \Omega^+_s$. Under logit choice, a larger loss makes selection exponentially less likely as sensitivity increases. Consequently, conditional on choosing a class that is sub-optimal in every menu, Alice's observations exponentially concentrate on the menus in which its valuation disadvantage is the smallest. For the class's valuation to survive vanishing logit exploration, it must equal the p-weighted average payoff across these *closest-loss menus* denoted by $\Omega^\star_s(\mathbf{v}^*) := \arg\min_{\omega \in \Omega^+_s} \Delta_s(\omega; \mathbf{v}^*)$. For a class chosen in some menu, the minimum loss is zero, and its closest-loss menus are precisely those in which it is optimal. The next result shows that closest-loss consistency exactly characterizes which strict VE are LSVE, and that each such LSVE is approached by a locally unique, locally stable family of SVE defined at every sufficiently large sensitivity.

Theorem 2 (Characterization and local stability)**.** *Let* $\mathbf{v}^*$ *be a strict VE valuation. Then:*

(i) The valuation vector $\mathbf{v}^*$ *is an LSVE if and only if it satisfies closest-loss consistency:*

$$v^*_s = \frac{\sum_{\omega \in \Omega^\star_s(\mathbf{v}^*)} p(\omega)\pi_s(\omega)}{\sum_{\omega \in \Omega^\star_s(\mathbf{v}^*)} p(\omega)}, \qquad s \in \mathcal{S}. \tag{7}$$

(ii) Whenever (7) holds, there exist a neighborhood U *of* $\mathbf{v}^*$ *and* $\widehat{\beta} < \infty$ *such that,* $\forall\ \beta \geq \widehat{\beta}$*,* U *contains a unique SVE, denoted by* $\mathbf{v}^\beta$*, and* $\mathbf{v}^\beta \to \mathbf{v}^*$ *as* $\beta \to \infty$*. This SVE is locally exponentially stable in the normalized dynamics (6) and the corresponding equilibrium* $\left(\mathbf{v}^\beta, \Lambda(\mathbf{v}^\beta; \beta)\right)$ *is locally exponentially stable under the asynchronous dynamics (5).*

Thus, closest-loss consistency identifies which strict patterns of choice can be supported in equilibrium by valuations learned from the observations that those choices generate. This sampling structure also explains local stability of nearby SVE for sufficiently large sensitivity. Near a strict equilibrium, small valuation perturbations preserve the optimal class in each

[19] For any fixed pure choice profile, VE consistency determines each chosen class's valuation as a conditional payoff average. Equality between two such valuations is a nongeneric restriction on the payoffs. Since there are finitely many pure choice profiles, the conclusion holds simultaneously across all menu-wise strict VE.

menu and thus the set of closest-loss menus for each class. Hence, at high sensitivity, they have little effect on the composition of an optimal class's observations, with nearly fixed relative weights. Even for a rarely chosen class, whose overall sampling frequency can respond sharply to valuation shocks, its conditional expected payoff remains almost unchanged as the relative weights of its closest-loss menus remain nearly invariant. This allows the mean-reverting update rule (6) to correct valuation errors by pulling assessments back toward these nearly fixed payoff targets. This yields local attraction under the normalized dynamics and, by the local-stability correspondence, under the harmonic visit-count dynamics (5).

Therefore, if a decision tree admits a strict VE that satisfies closest-loss consistency (7), the set of asymptotically stable SVE for sufficiently high sensitivity in the CQL dynamics is nonempty. In the appendix, we show that every locally asymptotically stable SVE of the joint asynchronous dynamics (5) has strictly positive probability of being the long-run outcome of the stochastic CQL process (2), from any finite initial valuations and positive prior counts. Thus, asymptotic stability serves as a dynamic equilibrium selection criterion.

As a special case that shall be useful later, suppose $\mathbf{v}^*$ is a menu-wise strict VE at which every class wins in at least one supported menu. Each class's closest-loss menus are precisely those it wins, so ordinary VE consistency implies closest-loss consistency. The appendix shows that even if classes that win in different menus have equal valuations, $\mathbf{v}^*$ is the limit of a locally unique branch of locally exponentially stable SVE defined for all sufficiently large β, as in Theorem 2(ii). This applies in particular when every singleton menu is supported.

3.3 Convergence with Constant Competitive-Menu Payoffs

We now establish the global convergence of the asynchronous CQL dynamics (5) to SVE in all decision trees with at most two similarity classes. This insight follows from the more general result stated below, which allows arbitrarily many classes provided each class's expected payoff is constant across the set of competitive menus in which it appears. Formally, suppose that, for each class s, there exists a constant $b_s \in \mathbb{R}$ such that

$$\pi_s(\omega) = b_s \qquad \text{for every } \omega \in \Omega_s^+ \text{ with } |\omega| \geq 2. \tag{8}$$

Thus, a class s yields the same expected payoff whenever it competes with another class. Its expected payoff in the singleton menu $\{s\}$, when supported, is unrestricted - allowing its valuation to eventually lie between $\pi_s\{s\}$ and b_s.[20] While condition (8) holds trivially with two classes, it also holds for any number of classes whenever every competitive menu

[20] We call a menu *competitive* if it contains at least two similarity classes. A singleton menu may nevertheless contain several alternatives: it is the available class, rather than the available alternative, that is unique.

features at least one alternative from each class, so that $\mathcal{S}$ is the only non-singleton menu in the reduced tree. This condition preserves endogenous sampling bias. As a class becomes more attractive, Alice chooses it more frequently in competitive menus, increasing the weight of those observations relative to its singleton observations. We show this restriction ensures global convergence of learning while allowing for both multiplicity of equilibria and indifference with endogenous mixing in the high-sensitivity limit.

Theorem 3. *Suppose* (8) *holds. For any sensitivity* $\beta \in [0, \infty)$*, every solution of the joint asynchronous dynamics* (5)*, starting from* $\mathbf{v}(0) \in \mathbb{R}^{\mathcal{S}}$ *and* $\mathbf{x}(0) \in \operatorname{ri} \Delta(\mathcal{S})$*, approaches the equilibrium set* $\mathcal{E}(\beta) := \{(\mathbf{v}, \Lambda(\mathbf{v}; \beta)) : \mathbf{v} \in \mathcal{V}(\beta)\}$*. Generically, since* $\mathcal{V}(\beta)$ *is finite for a.e.* β*, each trajectory converges to an individual hyperbolic equilibrium.*[21]

Also, for fixed sensitivity and menu probabilities, convergence generically survives all sufficiently small perturbations of the constant-competitive-payoff benchmark (8). This restriction makes competitive-menu observations of a class interchangeable in expectation. Alice's choices therefore affect its assessed payoff $g_s(\mathbf{v})$ only through the balance between its singleton and competitive observations. Since singleton observations arrive at a fixed rate, the class's total sampling frequency determines their weight in its assessment. This dependence on aggregate sampling frequencies aligns choice adjustments across menus with a common potential whose payoff component has conditional mean payoffs as marginal returns and whose entropy component captures logit randomization. At a given logit policy, a marginal shift toward a class with a larger expected prediction error raises the potential. The proof combines this alignment between choice and payoff feedback with a nontrivial smoothing argument based on harmonic averaging to establish convergence toward the SVE set.

The potential need not be concave in general, allowing for multiple stable SVE and therefore path dependence. However, if each class's singleton payoff weakly exceeds its competitive payoff on average, competitive observations dilute favorable singleton feedback, creating diminishing returns to sampling. The potential is then strictly concave in choice probabilities, yielding a unique, globally attracting SVE, as illustrated below.

The following two-class illustrations based on Fig. 2 demonstrate how convergence can coexist with path dependence or produce indifference in the high-sensitivity limit. Consider two classes, Bordeaux (B) and Chianti (C). The menus $\{B, C\}$, $\{B\}$, and $\{C\}$ occur with equal probability. The wines offered in the competitive menu yield expected payoffs 2 and 1, respectively, while those offered in either singleton menu yield a common $z \in \mathbb{R}$. Changing z reverses the feedback between choices and the evidence used to evaluate each class.[22]

[21] With two classes, the SVE set is finite at every sensitivity, yielding convergence to an isolated equilibrium.
[22] At an SVE with competitive-menu choice probability q, the normalized Jacobian has eigenvalues -1 and

Writing $q = (1+e^{-\beta(v_B - v_C)})^{-1}$ for the probability of choosing Bordeaux in the binary menu, the conditional payoff means are $g_B(q) = \dfrac{2q+z}{1+q}$, and $g_C(q) = \dfrac{1-q+z}{2-q}$.

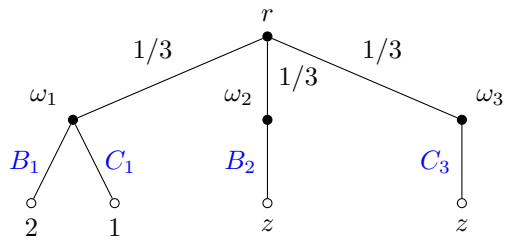


Figure 2: Example of a Decision Tree with Two Similarity Classes

Path dependence: Let $z = 0$. Choosing a class more frequently in the competitive menu increases the share of favorable observations in its payoff history, reinforcing its valuation advantage. For sufficiently large β, there are two locally attracting SVEs approaching the strict VEs $(v_B, v_C) = (1, 0)$ with Bordeaux chosen in the binary menu and $(0, \frac{1}{2})$ with Chianti chosen in the binary menu; both sets of eigenvalues converge to -1. A third mixed SVE is a saddle, with one positive eigenvalue. Thus initial assessments can lead to different long-run rankings, even though Bordeaux yields the higher payoff whenever Alice chooses between the two classes. Figure 3(a) illustrates these distinct basins of attraction.

Limiting indifference: Now let $z = 11/3$. More frequent selection in the competitive menu exposes a class disproportionately to unfavorable payoffs below its singleton benchmark, correcting and shrinking its valuation advantage. The SVE is unique for every $\beta \geq 0$, and all eigenvalues are negative. Theorem 3 therefore gives global convergence under both dynamics, as illustrated in Figure 3(b). As $\beta \to \infty$, $(v_B, v_C) \to (3, 3)$, and $q \to 2/3$. Alice consequently becomes indifferent between Bordeaux and Chianti despite their unequal competitive-menu payoffs. This limiting indifference is sustained by endogenous mixing with Bordeaux chosen twice as often as Chianti when both are available.

3.4 Self-Correcting Feedback and Global Convergence

We now identify a condition under which endogenous sampling leads to a unique long-run outcome, even with heterogeneous competitive-menu payoffs. This mechanism concerns which observations used in the assessment of a class get displaced when a competing class becomes more attractive. If these observations have relatively low payoffs on average, the residual observations improve the assessment of the class losing choice share. This response counteracts the initial change in relative attractiveness and stabilizes valuations and choices.

$-1 + \beta q(1-q)\left[\dfrac{2-z}{(1+q)^2} + \dfrac{1-z}{(2-q)^2}\right]$. The joint asynchronous Jacobian has one additional eigenvalue -1.

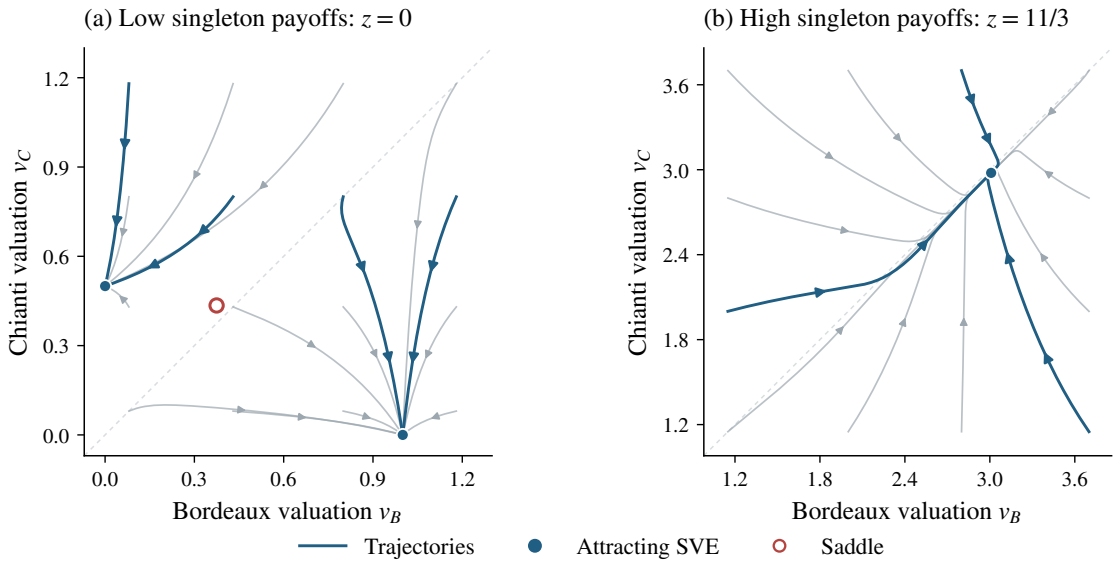


Figure 3: Valuation projections of the asynchronous joint dynamics at $\beta = 20$, initialized with $x_B = x_C = 1/2$. (a) Low singleton payoffs, $z = 0$: two attracting SVE and a saddle. (b) High singleton payoffs, $z = 11/3$: a unique globally attracting SVE.

To express this condition, fix a sensitivity $\beta \geq 0$ and let $w_s^{\mathbf{v}}(\omega) := p(\omega)\sigma_\omega^s(\mathbf{v})/\Lambda_s(\mathbf{v})$, $\omega \in \Omega_s^+$, denote the distribution of menus conditional on choosing class s. These are the weights with which observations from different menus enter its assessed expected payoff $g_s(\mathbf{v})$. For another class $r \neq s$, the probability $\sigma_\omega^r(\mathbf{v})$ measures the exposure of an s-observation from menu ω to crowding by r, with $\sigma_\omega^r(\mathbf{v}) = 0$ when $r \notin \omega$. An increase in v_r relative to s proportionally reduces the probability of choosing s most acutely in menus where r already has a high choice probability. Simply differentiating the conditional expected payoff map w.r.t. v_r gives

$$\frac{\partial g_s(\mathbf{v})}{\partial v_r} = -\beta \sum_{\omega \in \Omega_s^+} w_s^{\mathbf{v}}(\omega)\Big[\pi_s(\omega) - g_s(\mathbf{v})\Big]\sigma_\omega^r(\mathbf{v}) = -\beta \operatorname{Cov}_{w_s^{\mathbf{v}}}\left(\pi_s(\omega), \sigma_\omega^r(\mathbf{v})\right), \quad r \neq s. \tag{9}$$

Nonpositive payoff-crowding covariance implies a rise in v_r marginally displaces observations of class s that are no better, on average, than those currently used to assess it. Their removal weakly raises $g_s(\mathbf{v})$, providing a corrective response to the initial drop in the valuation of s relative to r. If this condition holds for each pair of classes over all $\mathbf{v} \in \operatorname{ri} K$, we obtain the following result on global convergence to a unique SVE with arbitrarily many classes.

Theorem 4 (Global convergence under non-positive payoff-crowding covariance)**.** *Fix a tree $\mathcal{T}'$ and $\beta \in [0, \infty)$. Suppose that, for every $\mathbf{v} \in K$ and every pair of distinct classes $s, r \in \mathcal{S}$,*

$$\operatorname{Cov}_{w_s^{\mathbf{v}}}\left(\pi_s(\omega), \sigma_\omega^r(\mathbf{v})\right) \leq 0. \tag{10}$$

Then there is a unique SVE $\mathbf{v}^\beta$*, and the corresponding equilibrium* $\left(\mathbf{v}^\beta, \Lambda(\mathbf{v}^\beta)\right)$ *is globally asymptotically stable under the joint asynchronous dynamics* (5)*, for arbitrary finite initial valuations and positive initial sampling frequencies.*

The covariance condition makes selection self-correcting: raising a rival's valuation weakly improves a class's assessed payoff, while raising its own valuation weakly lowers it.[23] Since a common shift in valuations leaves choice probabilities unchanged, $g(\mathbf{v} + c\mathbf{1}) = g(\mathbf{v})$ and hence $Dg(\mathbf{v})\mathbf{1} = 0$. The normalized drift $F_s(\mathbf{v}) := g_s(\mathbf{v}) - v_s$ therefore satisfies

$$\frac{\partial F_s}{\partial v_s} = -1 - \sum_{r \neq s} \left| \frac{\partial g_s}{\partial v_r} \right| = -1 + \beta \sum_{r \neq s} \mathrm{Cov}_{w_s^{\mathbf{v}}}\left(\pi_s(\omega), \sigma_\omega^r(\mathbf{v})\right) \leq -1, \qquad \mathbf{v} \in K.$$

The restoring own-valuation effect thus strictly dominates the combined cross effects, making the largest absolute valuation error shrink uniformly on K.[24] This gives contraction for the normalized dynamics. The same error bound yields convergence for the asynchronous joint dynamics because their adjustment factors $\Lambda_s(\mathbf{v})/x_s$ remain positive and bounded away from zero along each trajectory. Valuation convergence then implies $x \to \Lambda(\mathbf{v}^*)$ through $\dot{x}_s = \Lambda_s(\mathbf{v}) - x_s$. The argument also covers initial valuations outside the attracting, positively invariant box K. Proposition 1 therefore gives $(\mathbf{v}_k, \mathbf{x}_k) \to \left(\mathbf{v}^\beta, \Lambda(\mathbf{v}^\beta)\right)$ almost surely under stochastic CQL (2). We next give a simple condition on mean payoffs that ensures (10).

3.5 Singleton Payoffs: Indifference and Multiplicity

Singleton menus, though not necessary for the preceding result, provide a natural source of this corrective feedback. We now examine how their payoffs shape this feedback and, separately, whether limiting choices are strict or involve indifference. Singleton menus supply observations whose frequency is unaffected by competition. Whenever $\{s\}$ occurs, Alice samples class s regardless of her valuations. Yet the weight of these observations in her assessment of s depends on how frequently she samples it elsewhere. Observations obtained without competition between classes can thus shape subsequent choices in competitive menus.

This role motivates interpreting singleton menus as *anchor menus*: situations in which one class is effectively the only viable option. Indeed, richer menus containing a *focal class* alongside strictly obviously dominated classes provide a similar source of observations at high sensitivity, when the focal class is selected with probability close to one.[25] For the

[23] This mechanism resembles the *self-defeating externalities* of stable games (Hofbauer and Sandholm, 2009). Their condition imposes monotonicity on the population payoff map; ours restricts how valuations change the composition of payoff observations. Neither condition implies the other.

[24] This argument immediately yields local stability and uniqueness of SVE (by Poincaré-Hopf index theorem).

[25] A sufficient separation condition is that the highest expected payoff of each accompanying class, across all

remainder of this subsection, we maintain the following assumption.

Assumption 3.1. *There are at least two similarity classes, every singleton menu has positive probability, and the co-occurrence graph is connected. This graph has vertex set $\mathcal{S}$ and an edge between distinct classes s and r whenever they appear together in a supported menu.*

Starting from a fixed decision tree $\mathcal{T}'$, introduce a common singleton-payoff shift $z \in \mathbb{R}$:

$$\pi_s^z(\omega) := \pi_s(\omega) + z\,\mathbf{1}\{\omega = \{s\}\}. \tag{11}$$

Positive shifts make singleton experiences more favorable, while negative shifts make them less favorable. Menu probabilities and competitive-menu expected payoffs remain fixed, as do differences between singleton expected payoffs across classes. We examine how this change in the evidence available to Alice affects learning and its long-run outcomes.

3.5.1 High Singleton Payoffs and Limiting Indifference

Sufficiently favorable singleton experiences have two consequences. They generate the corrective feedback identified in the preceding subsection, ensuring convergence to a unique SVE. A sufficiently large common shift also makes choices that are strict in every menu inconsistent with the assessments they generate, forcing indifference as $\beta \to \infty$.

Theorem 5. *Under Assumption 3.1, there is a finite threshold $\bar{z} \geq 0$, depending on the underlying decision tree, such that, for every fixed $z > \bar{z}$:*

(i) For every $\beta < \infty$, there is a unique SVE $\mathbf{v}^\beta(z)$. The corresponding equilibrium $\left(\mathbf{v}^\beta(z), \Lambda(\mathbf{v}^\beta(z))\right)$ is globally asymptotically stable under the joint dynamics (5).

(ii) As $\beta \to \infty$, the valuations $\mathbf{v}^\beta(z)$ and their associated logit choice profiles converge to a VE $\left(\sigma^\infty(z), \mathbf{v}^\infty(z)\right)$. This limit exhibits indifference in at least one competitive menu: $|\operatorname{argmax}_{s\in\omega} v_s^\infty(z)| \geq 2$ for some $\omega \in \operatorname{supp}(p)$ with $|\omega| \geq 2$. Moreover, $\sigma^\infty(z)$ assigns positive probability to at least two classes in some supported menu.

The convergence conclusion follows from non-positive payoff-crowding covariance. Each class is selected whenever its singleton menu occurs, so singleton observations retain a share of its conditional payoff distribution bounded away from zero, uniformly over valuations and sensitivity. For sufficiently large z, these favorable observations raise its assessed payoff above every payoff it receives in competitive menus. A competing class therefore displaces

its supported menus, lies strictly below the focal class's lowest expected payoff across its supported menus. Equilibrium valuations inherit this gap, so the probability of selecting an accompanying class vanishes exponentially with sensitivity. The formal results retain singleton class menus.

relatively poor observations, improving the assessment of the class it crowds out. Theorem 4 then gives uniqueness and convergence. The appendix provides the bounds establishing a threshold $\bar{z}$ that works for every finite sensitivity.

Limiting indifference requires a further high-sensitivity limit argument. Suppose every supported menu had a unique valuation maximizer. A lowest-valued class would then be selected only in its singleton menu, so its valuation would incorporate the entire shift z. A class beating it in a competitive menu would also receive observations from that menu, incorporating only a fraction of the singleton shift into its valuation. For sufficiently large z, the supposedly lower-valued class would consequently have the higher assessment. This contradicts the proposed strict choices, ruling out valuation equilibria with a unique maximizer in every menu. Thus, the sufficiently large common shift overturns every pattern of menu-wise strict choices. Heterogeneous primitive expected payoffs can therefore generate persistent indifference through the endogenous composition of experience (see Fig. 3(b) for example). Part (ii) shows that the unique SVE branch selects a unique limiting VE with nontrivial mixing in at least one competitive menu, even though infinitely many VE may coexist. A stronger full-support restriction on menu probabilities yields indifference across all classes.

Assumption 3.2. *All singleton menus are equally likely, and* $\omega \subseteq \omega' \implies p(\omega) \leq p(\omega')$.

For instance, uniform probabilities over all nonempty menus satisfy this condition. Under Assumption 3.2, sufficiently high singleton payoffs force indifference across all classes.

Proposition 2 (Total indifference)**.** *Under Assumption 3.2, there is a finite threshold* $\bar{z}_1$ *such that, for every* $z > \bar{z}_1$*, every VE satisfies* $v_s = v_r$ *for all* $s, r \in \mathcal{S}$.

Any strict valuation gap would give the higher-valued class a larger sampling frequency and a smaller singleton share. A sufficiently large positive shift would then raise the lower-valued class's assessment more strongly, contradicting the gap. The LSVE selected by the unique branch of globally stable SVE is unique, fully-mixed and exhibits total indifference. Retaining inclusion monotonicity, this result persists with nearly equal singleton probabilities.

However, high singleton payoffs are not necessary for (10). The appendix provides additional primitive sufficient conditions, including examples without singleton menus and one with intermediate singleton payoffs that also sustains robust limiting indifference and mixing.

3.5.2 Low Singleton Payoffs and Multiplicity of Strict VE

When singleton experiences are sufficiently unfavorable, the composition of observations can reinforce the valuation differences that guide choice. A class selected infrequently in

competitive menus is assessed largely through its poor singleton experiences, sustaining its low valuation. More frequent competitive choices dilute these unfavorable observations. Different patterns of selection can consequently support different self-confirming rankings. Under Assumption 3.1, every class is selected in its singleton menu. Theorem 2 implies that every strict VE is an LSVE, approached by a branch of locally stable SVE as $\beta \to \infty$.

Theorem 6. *Fix a reduced decision tree $\mathcal{T}'$ satisfying Assumption 3.1. There is a finite threshold $\underline{z}$ such that, for every $z < \underline{z}$, the following statements hold:*

(i) If the decision tree is generic, there exists at least one strict VE.

(ii) If every binary menu is supported, there exist at least $|\mathcal{S}|$ distinct strict VE, one with each class uniquely lowest-ranked. This conclusion does not require genericity.

For each fixed such z, there is a finite sensitivity threshold $\widehat{\beta}(z)$ such that, for every $\beta \geq \widehat{\beta}(z)$, a unique SVE lies in a sufficiently small neighborhood of each strict VE. These SVE converge to their respective VE as $\beta \to \infty$ and are locally exponentially stable under (5).

The mechanism is transparent when choices are held fixed. For a pure (deterministic) choice rule φ selecting one class in each menu, let $D_s^\varphi := \sum_{\omega:\, \varphi(\omega)=s} p(\omega)$ be the aggregate probability of selecting class s across all menus featuring it. Its consistent valuation is

$$v_s^\varphi(z) = a_s^\varphi + \lambda_s^\varphi z, \qquad \lambda_s^\varphi := \frac{p(\{s\})}{D_s^\varphi}, \tag{12}$$

where a_s^φ is its conditional expected payoff before the shift. The coefficient λ_s^φ is the share of singleton observations in its payoff history. As z becomes increasingly negative, differences in these shares dominate the fixed differences in baseline payoffs. Classes whose observations are concentrated in singleton menus receive the largest effective penalty.

The existence argument constructs a priority order whose choices generate the valuation comparisons needed to sustain it. With every binary menu supported, this construction can place any designated class s last. Every other class r then receives competitive observations from $\{s, r\}$, so its singleton share is strictly below one, whereas s bears the entire singleton penalty. The construction also orders the remaining classes strictly: assigning a binary menu to an earlier-priority class removes a positive source of competitive observations from the later-priority class, increasing the latter's eventual singleton share. For sufficiently negative shifts, the resulting valuations reproduce the entire priority order. The appendix provides the construction and the associated thresholds. Genericity in part (i) provides a strict total order on valuations. Theorem 2 then supplies the nearby locally exponentially stable SVE for sufficiently large sensitivity.

Under the stronger Assumption 3.2, sufficiently low singleton payoffs make every strict ranking self-confirming. For any proposed strict ranking, the menu probabilities ensure that higher-ranked classes receive more competitive observations and place lesser weight on the singleton penalty, so lower-ranked assessments are depressed more strongly. The appendix proves, for each sufficiently negative z, all sufficiently large sensitivities admit at least $|\mathcal{S}|!$ locally exponentially stable SVE, one associated with each strict ranking (see Fig. 3(a)). Initial valuations then affect which valuation ranking is sustained in the long-run, as the choices they induce shape the evidence subsequently used to evaluate each class.

3.6 Instability and Persistent Cycling

The preceding results establish stability and convergence in several important classes of decision problems. However, they leave open the possibility that learning can fail to converge in a decision tree with three or more similarity classes. We now answer this question in the affirmative by showing that stationary decision problems can indeed generate persistent oscillations in the learning dynamics. In fact, with just three similarity classes, we show there are open sets of decision problems for which sufficiently sensitive choice admits no asymptotically stable SVE and learning persistently oscillates.

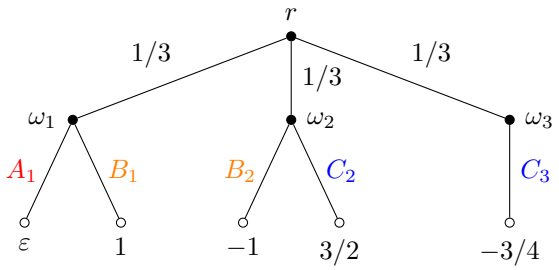


Figure 4: A stationary decision tree generating persistent cycling. Colors identify similarity classes; terminal labels give the expected payoffs of individual alternatives with $\varepsilon > 0$.

Consider the decision tree in Figure 4. Five alternatives form three similarity classes: $A = \{A_1\}$, $B = \{B_1, B_2\}$, and $C = \{C_2, C_3\}$. The induced class menus are $\{A, B\}$, $\{B, C\}$, and $\{C\}$, each occurring with equal probability 1/3. The feedback operates through the observations entering class B which links the two competitive menus.

B pays well when competing with A and poorly when competing with C. Suppose Alice begins by choosing A more often. She then observes fewer of B's favorable payoffs from $\{A, B\}$, so her assessment of B deteriorates. This reinforces A's position and also encourages her to choose C over B. Choosing C has the opposite effect on B's selected sample, however, as it removes unfavorable observations from $\{B, C\}$. As C gains ground, its favorable competitive

payoffs also raise its own valuation, further displacing B from the menu $\{B, C\}$ where B performs poorly. The observations that remain for the rarely chosen B are consequently more favorable, allowing its valuation to recover. That recovery sets up the next reversal. As Alice returns to B, its unfavorable payoffs in $\{B, C\}$ re-enter her sample. At the same time, choosing C less often in that menu leaves its valuation increasingly shaped by its poor singleton payoff, reinforcing Alice's return to B. The renewed exposure to B's unfavorable payoffs pulls its valuation down again, allowing A to regain ground. The cycle is sustained by these changes in the composition of experience: every alternative's mean payoff is fixed, but Alice's choices continually change the evidence used to evaluate its class.[26]

At $\varepsilon = 0$, the construction is particularly transparent: the unique SVE is $\mathbf{v}^* = (0, 0, 0)$ at every finite sensitivity, and it is linearly unstable whenever $\beta > 8$. A small positive value of ε yields the following robust conclusion (see Figure 5 for an illustration).

Theorem 7 (Persistent cycling)**.** *Fix the menu support $\Omega_+ = \{\{A, B\}, \{B, C\}, \{C\}\}$. There exist $\varepsilon > 0$, an open neighborhood $\mathcal{U}$ of the decision tree in Figure 4, and a finite $\bar{\beta}$ such that, for every tree in $\mathcal{U}$ and every finite $\beta \geq \bar{\beta}$:*

(i) There is a unique SVE, which is linearly unstable under both the normalized (6) and joint asynchronous dynamics (5).

(ii) Every trajectory of both the normalized and the joint asynchronous dynamics outside the equilibrium's stable manifold has a non-trivial periodic orbit as its omega-limit set.

(iii) From any finite initial valuations and positive prior counts, the limit set of $(\mathbf{v}_k, \mathbf{x}_k)$ under the stochastic CQL process (2) is almost surely a non-constant periodic orbit of the asynchronous dynamics (5). Neither valuations nor choice probabilities converge.

Here openness is relative to the space of positive menu probabilities and expected payoffs of alternatives on the specified support. The proof uses a fixed $\varepsilon > 0$, allowing all three menu probabilities and all five payoffs to vary in a sufficiently small neighborhood. Thus the absence of a stable equilibrium and the persistence of oscillation survive perturbations of the decision problem. Harmonic averaging makes the argument tractable. In the sufficient-statistic coordinates $(\mathbf{x}, \mathbf{y})$, where $y_s = x_s v_s$, three affine discrepancies decay exponentially, leaving an invariant attracting plane. The unique equilibrium is a source on this plane. Boundedness and the Poincaré–Bendixson theorem then imply that every nonstationary

[26] The interaction between B and C creates a negative feedback loop: a higher valuation of C improves the composition of B's observations, whereas a higher valuation of B worsens the composition of C's observations. This interacts with the self-reinforcing feedback in $\{A, B\}$, where choosing A more often removes favorable observations of B. These opposing feedbacks sustain oscillation at high sensitivity.

planar trajectory has a periodic omega-limit set.[27] The appendix extends the conclusion to the full joint dynamics and combines the planar limit-set results of Benaïm (1999) with a verification that the stochastic process almost surely avoids the unstable equilibrium.

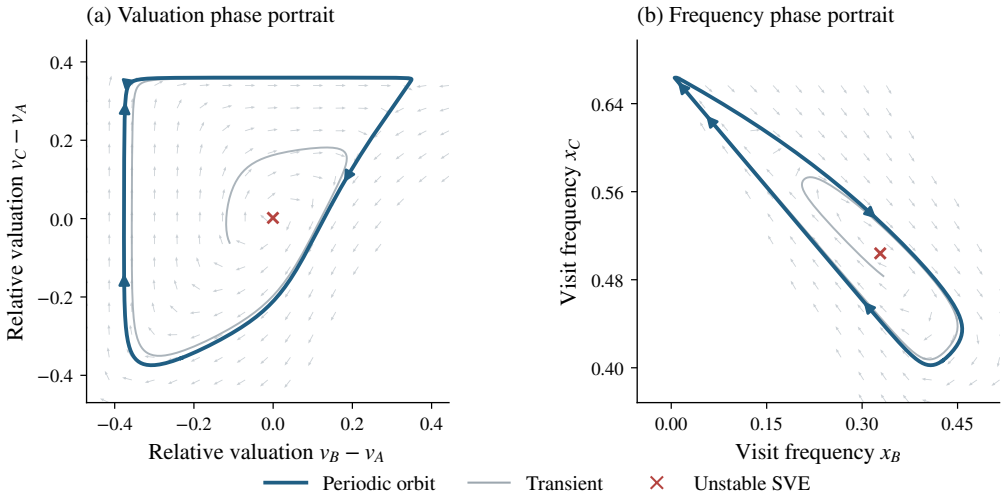


Figure 5: Phase portraits of the joint CQL dynamics for tree in Fig. 4 with $\varepsilon = 0.01$ and $\beta = 20$

The construction is minimal as cycling requires at least 3 classes (by Theorem 3) and 5 alternatives.[28] Its three-menu support is also irreducible. The self-correcting feedback condition (10) fails here as competition removes favorable observations: a higher valuation of A crowds B out of its favorable menu, lowering g_B; a higher valuation of B similarly crowds C out of its favorable menu, lowering g_C. Building on this observation, our next result characterizes the *absence of asymptotically stable SVE* in decision problems with *three classes* under the joint dynamics (5). At an SVE $\mathbf{v}^*$, define the trace and the determinant of the linearized normalized dynamics (6) in the two independent valuation differences by

$$T(\mathbf{v}^*) := -2 + \beta \sum_{s \neq r} \mathrm{Cov}_{w_s^{\mathbf{v}^*}}\Big(\pi_s(\omega), \sigma_\omega^r(\mathbf{v}^*)\Big), \quad d(\mathbf{v}^*) := \det\Big(I - Dg(\mathbf{v}^*)\Big).$$

Lemma 3.1 (Absence of stable SVE with three classes)**.** *Fix $\beta \in [0, \infty)$ and a decision tree with three classes and suppose all SVE are hyperbolic.*[29] *At least one SVE has $d(\mathbf{v}^*) > 0$. No asymptotically stable SVE exists if and only if $T(\mathbf{v}^*) > 0$ at every SVE with $d(\mathbf{v}^*) > 0$.*

The two eigenvalues governing relative valuations have sum T and product d, with -2

[27] The full joint state space is five-dimensional. Its equilibrium has a three-dimensional stable manifold, whose intersection with the invariant plane consists only of the equilibrium.

[28] Instability does not require this particular menu arrangement. The appendix gives further examples with three binary menus whose payoff advantages form a cycle, supplemented either by a ternary menu, without singletons, or by all three singleton menus at intermediate payoff levels.

[29] By Theorem 1, this assumption holds for generic (p, π) at all but finitely many sensitivities.

capturing the unit restoring force in each independent valuation difference. Under the payoff-crowding covariance condition (10), every covariance term is nonpositive, so $T \leq -2$. The same quantities also characterize instability with three classes: at equilibria with $d > 0$, positive covariances must outweigh negative ones sufficiently to overcome these restoring forces, giving $T > 0$. SVE with $d < 0$ are already unstable. Thus, no stable SVE exist precisely when this aggregate reinforcing feedback dominates at every SVE with $d > 0$.

3.7 Coarse representation and the welfare cost of learning

The preceding result shows that with three or more classes, Alice's valuations need not converge. Even when her valuations converge, her choices need not attain the best expected payoff her coarse representation permits. To separate the costs of coarse representation and coarse learning, consider a fully informed oracle who knows the menu distribution and expected payoffs. He uses Alice's similarity partition, within-class implementation, and logit choice rule, with one common valuation per class across menus.[30] We allow high-sensitivity limits of logit policies, making deterministic choices admissible.[31] Unlike Alice, however, he chooses these valuations freely, without requiring consistency with experienced payoffs.

The proposition below characterizes the oracle's optimal choice policy and gives necessary and sufficient conditions for its efficiency. For a fixed strict ranking $\succ$, write $\max_{\succ} \omega$ for the highest-ranked class in menu ω. When each supported menu has a unique payoff-maximizing class, define the *efficient-comparison graph* G^{eff} by drawing an arrow $s \to r$ whenever s is the (objective) payoff-maximizer in a supported menu containing $r \neq s$.

Proposition 3 (The informed coarse benchmark)**.** *An optimal oracle policy is represented by a strict class ranking, and its expected payoff is*

$$W^{\text{coarse}} = \max_{\succ} \sum_{\omega \in \Omega} p(\omega)\, \pi_{\max_{\succ} \omega}(\omega).$$

Generically, every optimal class-choice policy is pure. If each supported menu has a unique payoff-maximizing class, then $W^{\text{coarse}} = \sum_{\omega \in \Omega} p(\omega) \max_{s \in \omega} \pi_s(\omega) \iff G^{\text{eff}}$ *is acyclic.*

Logit randomization cannot improve this benchmark. Its random-utility representation amounts to drawing a strict ranking of classes independently of the menu and choosing the highest-ranked available class. Expected welfare is therefore a weighted average of the welfare generated by strict rankings, and cannot exceed the best of them. Conversely, any strict ranking can be implemented in the limit by increasing the gaps between appropriately

[30] At any $\beta > 0$, logit depends only on $\beta \mathbf{v}$, so unrestricted valuations allow us to normalize $\beta = 1$.
[31] Over finite valuation vectors, the benchmark above is generally only a supremum.

ordered valuations. The graph records the comparisons required for efficiency: an acyclic graph can be completed into a common ranking that selects the best class in every menu, whereas a directed cycle makes these requirements inconsistent.[32]

The welfare gap between unconstrained informed choice and this benchmark measures the cost of coarse representation; any further shortfall under coarse learning reflects the welfare cost of learning. We examine high-sensitivity limits of long-run outcomes to isolate losses that persist as logit noise vanishes. Singleton payoffs illustrate the distinction: they contribute the same payoff to every ranking, yet can change the outcome selected by learning. In the two-class illustrations in Figure 3, the oracle is efficient and always chooses Bordeaux when both classes are available. The limiting Chianti-favoring strict outcome with low singleton payoffs loses 1/3 per period relative to this benchmark, while the limiting mixed outcome with high singleton payoffs loses 1/9. Thus even globally convergent learning can sustain a welfare loss while the optimal policy is representable by the same class scores.

Our harmonic benchmark gives equal weight to observations from the same class, regardless of the menus in which they arise. Alice observes those menus, so in principle she could instead weight their feedback differently while retaining one valuation per class. Such reweighting might correct the choice distortions in the stylized two-class illustrations above. What follows shows why it cannot provide a general remedy for closing the welfare gap.

Indeed, a persistent welfare loss can arise even when learning converges globally to a unique SVE whose high-sensitivity limit is strict. Consider two equally likely binary menus, $\{A, C\}$ and $\{B, C\}$, with deterministic payoffs $(0.6, 1)$ and $(0.4, 0)$, respectively. The oracle ranking $B \succ C \succ A$ selects the best class in each menu and yields $W^{\text{coarse}} = 0.7$. Thus, coarse representation permits efficient choice. Coarse Q-learning, however, cannot sustain these comparisons. Every observation of A yields 0.6, while every observation of B yields 0.4, so their learned ordering contradicts the efficient ranking. For every finite sensitivity, this example has a unique, globally stable SVE, to which harmonic visit-count learning converges almost surely. As sensitivity increases, its valuations approach $(0.6, 0.4, 0)$, a strict VE that selects A and B in their respective menus and yields welfare 0.5. The limiting loss is thus 0.2 per period, despite convergence to a strict outcome and the absence of singleton menus.

The obstruction extends beyond CQL. Call a learning rule *constant-payoff calibrated* if a class whose observations always yield c is assessed at c asymptotically, almost surely. Constant-payoff calibration is a minimal consistency requirement for any reasonable payoff-assessment

[32] Efficiency here holds the within-class implementation fixed. Unrestricted informed choice among objective alternatives may yield a higher payoff.

learning rule. Suppose Alice uses common class valuations and the same monotone choice rule across menus.[33] Calibration eventually ranks A above B, making C weakly less likely to be chosen against A than against B, although efficiency requires the opposite. Changing C's assessment cannot resolve this conflict. The following result formalizes this observation.

Proposition 4 (A limit to payoff-assessment learning)**.** *Under the given choice restrictions, no learning rule satisfying constant-payoff calibration can guarantee long-run average welfare at least* W^{coarse} *in every decision tree. In the preceding example, every such rule yields*

$$\limsup_{K\to\infty} \frac{1}{K}\sum_{k=0}^{K-1} r_k \leq \frac{1}{2} < \frac{7}{10} = W^{\text{coarse}} \qquad \textit{almost surely.}$$

The appendix establishes a positive welfare gap on an open set of payoff and menu-probability specifications with this support. The bound requires only calibration of A and B; neither C's valuation nor the choice rule need converge. Constant-payoff calibrated learning covers arbitrary nonnegative weighted averages of own-payoff observations and initial scores whenever the averages are eventually well-defined and initial-score shares vanish. The weights may depend arbitrarily on past menus and rewards, even with full knowledge of the primitives. Efficient common scores encode priorities across menus; learning them as estimates of a class's expected payoff need not recover those priorities. The impossibility identifies this limitation, even when the efficient policy is representable by a common ranking of categories.

4 Algorithmic Prominence & Consumer Learning

Online marketplaces influence which products consumers consider through search, recommendations, and algorithmic prominence (Lee and Musolff, 2025), with evidence that Amazon gives preferential search rankings to its own brands (Farronato et al., 2023). Moreover, by emphasizing a common in-house brand label and presenting its underlying products homogeneously with a uniform price, a hybrid platform may encourage consumers to attend to brand identity rather than less salient product-specific attributes. Consumers who organize learning around these shared cues pool experiences across the brand, as in our limited-salience interpretation of similarity classes. When consumers learn through such categories, prominence also shapes the experiences that inform subsequent choices. In this section, we demonstrate as a natural application of our coarse Q-learning framework, how a digital platform can use favorable purchases of its own brand to sustain demand for inferior products sold under the same umbrella label, including when competitors remain available.

[33] For binary menus, $\Pr_k(s \mid \{s,r\};\mathbf{v}) = Q_k(v_s, v_r)$, with Q_k nondecreasing in its first argument and nonincreasing in its second, and $Q_k(x,y) + Q_k(y,x) = 1$. Scores are determined before the menu is drawn.

Consider a consumer who uses three similarity classes: the platform's in-house brand H, a competitor C, and an outside option O (shopping elsewhere). In each period, a product state $i \in \{1, \ldots, J\}$ is drawn independently and uniformly. The platform's product has expected payoff μ_i and margin $m_i \geq 0$, while the competitor and the outside option have constant expected payoffs $\mu_C > \mu_O$. Prices are fixed, and we normalize the platform's profit to zero when its own product is not purchased. The consumer maintains one valuation per class and follows our harmonic visit-count coarse Q-learning rule (2). Throughout, suppose

$$\bar{\mu} := \frac{1}{J} \sum_{i=1}^{J} \mu_i < \mu_O < \mu_C, \qquad \max_i \mu_i > \mu_O. \tag{13}$$

The in-house brand's expected payoff is below the outside option on average, although some of its products are better. The platform commits to a stationary prominence rule $\tau = (\tau_i)_{i=1}^J \in [0,1]^J$. After observing state i, it promotes its product with probability τ_i, using an independent draw each period. Prominence removes the competitor from consumers' consideration set, reducing the class menu from $\{H, C, O\}$ to $\{H, O\}$. We initially assume that prominence is costless. For each fixed prominence rule and finite sensitivity, the appendix establishes (by extending Theorem 3) that the induced stationary decision problem has a unique SVE. The corresponding joint equilibrium is globally asymptotically stable under (5), and harmonic visit-count learning (2) converges to it almost surely. We evaluate platform profit and consumer welfare by taking the high-sensitivity limit of these long-run outcomes. Let $d_i(\tau)$ denote the resulting probability of an in-house product sale in state i, whether or not the product is promoted. The platform seeks to maximize expected profit per period, $\Pi^\infty(\tau) := J^{-1} \sum_{i=1}^J m_i d_i(\tau)$, breaking ties in favor of less frequent prominence.

4.1 Optimal Deterministic Prominence

To begin, restrict the platform's strategies to deterministic prominence $\tau = (\tau_i)_{i=1}^J \in \{0,1\}^J$. We start with two states, h and ℓ, satisfying $\mu_h > \mu_C > \mu_O > \mu_\ell$ and $m_h, m_\ell > 0$. Under complete information (the consumer observes the state and knows the expected payoffs), prominence cannot increase sales: the high-quality product is preferred to the competitor, while the low-quality product is worse than the outside option.[34] Our tie-breaking convention therefore selects no prominence. Under coarse learning, however, no prominence leaves the overall brand valued at $\bar{\mu} < \mu_O$, so the consumer chooses the competitor in the limit.

Promoting only the high-quality product changes this outcome. Its favorable payoffs raise

[34] The same benchmark arises under learning with the finest partition: valuations converge to alternatives' expected payoffs, and the subsequent high-sensitivity limit yields informed choice in every menu.

the common brand valuation and encourage purchases of the low-quality product when the competitor is available. Those disappointing payoffs pull the valuation back down. In the limit, the two sources of evidence endogenously balance at indifference $v_H = \mu_C$, with purchase probabilities 1 and $q_2 = (\mu_h - \mu_C)/(\mu_C - \mu_\ell) \in (0,1)$ in states h and ℓ, respectively. Limiting platform profit is $(m_h + q_2 m_\ell)/2$. The other three deterministic rules yield zero limiting brand sales, so *high-only prominence is uniquely optimal* among them. The same corrective feedback behind our limiting-indifference result in Theorem 5 thus sustains inferior purchases despite continued learning from their payoffs.[35]

Adding a medium-quality product makes the platform's trade-off more substantive. Suppose now there are three states, h, m, ℓ, with

$$\mu_h > \mu_C > \mu_m > \mu_O > \mu_\ell, \qquad \frac{\mu_h + \mu_m}{2} \leq \mu_C, \qquad m_h, m_m, m_\ell > 0, \tag{14}$$

and $\bar{\mu} < \mu_O$ as before. Under complete information, prominence increases sales only in state m; our tie-breaking convention therefore selects the medium-only prominence policy. Under coarse learning, that policy also sells only the medium-quality product: its payoff calibrates the brand valuation to μ_m, leaving even the high-quality product unpurchased. Promoting both the high- and medium-quality products improves on this policy. Promoting only the high-quality product can be more profitable still. Under high-only prominence, the in-house brand's limiting purchase probabilities in states (h, m, ℓ) are $(1, q_3, q_3)$, where

$$q_3 := \frac{\mu_h - \mu_C}{2\mu_C - \mu_m - \mu_\ell} \in (0,1).$$

High-and-medium prominence instead induces purchases $(1, 1, 0)$. The platform compares the additional low-quality sales under the first rule with the additional medium-quality sales under the second. If $m_\ell \gg m_m$, i.e. if the margins from the former are sufficiently higher (as under a uniform umbrella-brand retail price P where unit production costs increase with quality, so that net consumer payoffs $\mu_i = u_i - P$ rise with quality while margins $m_i = P - c_i$ fall), the platform optimally uses deterministic prominence in the high state only.

Proposition 5 (Optimal deterministic prominence)**.** *Under* (13) *and* (14)*, the optimal deterministic rule, with ties resolved in favor of less frequent prominence, is*

$$\tau^* = \begin{cases} (1,0,0), & \text{if } \; q_3 m_\ell \geq (1-q_3) m_m, \\ (1,1,0), & \text{if } \; q_3 m_\ell < (1-q_3) m_m, \end{cases} \tag{15}$$

[35] While there are no literal singleton menus in this illustration, under high-state prominence only, $\{H, O\}$ serves as a focal menu for H with Alice choosing H almost surely in the long-run, high-sensitivity limit.

with coordinates ordered as (h, m, ℓ)*. The profit-maximizing rule is unique whenever* $q_3 m_\ell \neq (1 - q_3)m_m$*. The maximal limiting profit is* $\Pi^*_{\text{det}} = (m_h + \max\{q_3(m_m + m_\ell),\, m_m\})/3$.

Promoting the medium-quality product secures its sales but dilutes the favorable evidence that sustains $v_H = \mu_C$ under high-only prominence. Relative to promoting both products, high-only prominence sacrifices a fraction $1 - q_3$ of medium-product sales while generating a fraction q_3 of low-product sales. Condition (15) compares these gains and losses. Note that with positive margins, while the complete-information optimal prominence rule is independent of their magnitudes, the coarse-learning optimum depends on them.

Under high-only prominence, consumer welfare equals μ_C, its no-prominence level, in both illustrations. Thus, the platform gains from prominence, not at the cost of the consumer, but of its rival. Nonetheless, from the consumer's standpoint, the same coarse representation permits a better allocation. Holding prominence fixed, the ranking $C \succ H \succ O$ selects the high-quality product under prominence and the competitor elsewhere, yielding an additional payoff of $(\mu_h - \mu_C)/J$ for $J = 2, 3$. CQL cannot sustain this ranking: purchasing the brand only in state h would raise its valuation to $\mu_h > \mu_C$, inducing purchases elsewhere. The gap isolates the welfare cost of learning beyond the restriction to a common class ranking.

4.2 Optimal randomized prominence

Beyond deterministic prominence, randomization lets the platform adjust the composition of purchases more precisely, expanding inferior product sales while accepting a lower learned brand valuation. Return to the two-state illustration and consider the prominence strategy $\tau = (1, t)$: always promote the high-quality product and promote the low-quality product with probability t. Initially, low-state prominence only replaces purchases that already occurred in competitive menus. Once $t > q_2$, however, it increases low-quality sales. For

$$q_2 < t < t^*_2, \qquad \text{where } t^*_2 := \frac{\mu_h - \mu_O}{\mu_O - \mu_\ell} \in (q_2, 1),$$

the promoted mean $(\mu_h + t\mu_\ell)/(1 + t)$ lies strictly between μ_O and μ_C. Consumers therefore buy the brand under prominence and the competitor otherwise. Low-quality sales increase with t, raising profit even as the learned brand valuation falls. The platform approaches its profit supremum by taking $t \uparrow t^*_2$, pushing the brand valuation toward the outside option from above. The three-state illustration follows the same logic. Consider $\tau = (1, 1, t)$: always promote the high- and medium-quality products, and randomize prominence in the low state. For

$$0 < t < t^*_3, \qquad \text{where } t^*_3 := \frac{\mu_h + \mu_m - 2\mu_O}{\mu_O - \mu_\ell} \in (q_3, 1),$$

consumers again buy the brand precisely when it is promoted. Randomization thus secures medium-quality sales while retaining profitable low-quality sales, resolving the tradeoff between the two deterministic candidates. The profit suprema in the two illustrations are

$$\Pi_2^{\sup} = \frac{m_h + t_2^* m_\ell}{2}, \qquad \Pi_3^{\sup} = \frac{m_h + m_m + t_3^* m_\ell}{3},$$

each strictly exceeding every deterministic rule's profit. These gains, however, have a welfare cost. In both illustrations, every sequence approaching the respective profit supremum yields consumer welfare and competitor sales share converging to

$$W_j^* = \mu_C - \frac{\mu_C - \mu_O}{j}(j - 1 + t_j^*) < \mu_C, \qquad c_j^* = \frac{1 - t_j^*}{j} > 0, \qquad j = 2, 3.$$

The competitor retains a positive share, but consumer welfare falls below its no-prominence level. Along the stated policies, the platform's randomization generates these sales shares: consumers choose the brand under prominence and the competitor otherwise.

The same logic characterizes optimal randomized performance with any number of states. Let $d_i \in [0, 1]$ denote the limiting probability of a brand purchase in state i. Valuation consistency equates the learned brand payoff with the purchase-weighted average of product expected payoffs. Positive limiting sales therefore require

$$\sum_i (\mu_i - \mu_O) d_i \geq 0 : \tag{16}$$

favorable purchases must, on average, offset the unfavorable feedback from products below the outside option. The supremum of limiting profit is exactly the maximum of $J^{-1} \sum_i m_i d_i$ over sales allocations satisfying this constraint. With nonnegative margins, a maximizing allocation can fully include every product with $\mu_i \geq \mu_O$. Products below the outside option are then included in decreasing order of $m_i/(\mu_O - \mu_i)$, until the constraint binds, with at most one product partially included. The platform thus prioritizes margin per unit of unfavorable feedback. Slightly reducing prominence in the partially included below-outside-option state, or in a fully included such state if none is partially included, keeps the promoted mean just above μ_O and approaches the bound. Hence approximately optimal stationary rules need randomize in at most one state.[36] The appendix gives the characterization and proof.

[36] With strictly positive margins, no fixed stationary rule attains the supremum. At a rule whose promoted mean equals μ_O, the limiting purchase probability under prominence is $1/2$; approaching that mean from above instead sustains purchase with probability 1. Each rule is evaluated after learning and the high-sensitivity limit, before taking the policy limit. At any fixed finite sensitivity, an exact profit-maximizing rule exists and can also be chosen to randomize in at most one state.

Heterogeneous consumers: The randomized-policy characterization (16) extends to heterogeneous partitions. Suppose a positive fraction of consumers use the coarse brand category, the remainder distinguish individual products, and the platform applies a common prominence rule. The same family of policies approaches the profit supremum for every positive coarse share. Promoting every product at least as good as the outside option already maximizes profit from finest-partition consumers, while adjusting prominence below that threshold affects brand sales only to coarse consumers. In our illustrations, finest-partition consumers also suffer from low-quality prominence: they reject the promoted product but lose access to the competitor and take the outside option. Coarse learning can therefore affect welfare even for consumers who distinguish every product.[37]

Regulation with prominence caps: In the all-coarse case, a cap on the frequency of prominence at the optimal deterministic frequency may seem promising but need not preserve consumer welfare. The platform can reallocate prominence from favorable products toward inferior ones with higher margins, changing the composition of purchases without increasing its prominence frequency. The appendix characterizes the profit supremum under any cap and shows it can be approached by stationary prominence rules that randomize in at most two states, and two-state randomization can be necessary even with a positive share of finest-partition consumers.[38] It identifies when policies approaching the profit supremum preserve no-prominence welfare, yet gives an example where approximately optimal policies lower welfare under every positive cap, though the welfare loss vanishes as the cap falls to 0.

4.3 Optimal costly stationary prominence

In practice, prominence may require advertising and promotional expenditure, or displace other valuable uses of consumers' attention. We show such costs can make deterministic prominence optimal even when randomization is available, while preserving consumer welfare at its no-prominence level. Suppose each use incurs a cost $c \geq 0$, regardless of whether a purchase follows. The platform maximizes limiting net profit $\Pi^{\infty}(\tau) - cT(\tau)$, where $T(\tau) := J^{-1}\sum_i \tau_i$ is its prominence frequency. Consumer payoffs and learning remain unchanged. Costs make purchases of the in-house brand in competitive menus particularly valuable to the platform since they generate revenue without requiring prominence.

[37] Consumer composition can change the optimal deterministic policy. In the three-state illustration (15), a larger finest-partition share favors high-and-medium prominence over high-only prominence. The appendix characterizes this comparison and proves the heterogeneous-consumer extension.

[38] More generally, for any fixed consumer mix, every stationary rule admits a replacement with the same prominence frequency and weakly higher limiting population profit that randomizes in at most two states.

To illustrate, let's specialize the two-state environment to

$$(\mu_h, \mu_\ell) = (4, 0), \qquad \mu_O = \frac{5}{2}, \qquad \mu_C = 3, \qquad m_h = m_\ell = 1.$$

Without costs, the platform approaches its profit supremum by always promoting the high-quality product and promoting the low-quality product with probability tending to 3/5. At $c = 2/3$, however, high-only prominence uniquely maximizes net profit among all stationary rules.[39] It induces purchases $(1, 1/3)$: favorable high-quality experiences sustain low-quality purchases in the competitive menu, keeping the brand valuation at $\mu_C = 3$. Gross profit is 2/3 and prominence frequency is 1/2, giving net profit 1/3.

The distinction between sales and prominence is essential. Promoting the low-quality product with probability 1/3 generates exactly the same sales, but requires prominence in 2/3 of periods and reduces net profit to 2/9. The sequence approaching the costless profit supremum generates more sales, yet its net profit approaches only 4/15 at this cost. Competitive mixing economizes on prominence, while the additional sales generated by more frequent promotion no longer justify their cost. Consumer welfare consequently equals its no-prominence level 3, compared with the lower costless-optimal limit 13/5.

Costs can also change which products receive prominence. Return to the three-state illustration and suppose all margins equal $m > 0$. Since $q_3 < 1/2$, high-and-medium prominence is optimal among deterministic rules when prominence is costless. At $c = m$, high-only prominence instead becomes uniquely optimal among all stationary rules, including randomized ones. The high-state sale just covers its prominence cost; its favorable feedback supports profitable purchases in both unpromoted states. Net profit and consumer welfare are

$$\Pi_h^\infty - cT_h = \frac{m(1 + 2q_3) - m}{3} = \frac{2mq_3}{3} > 0, \qquad W_h^\infty = \mu_C.$$

Thus costs can favor concentrating prominence on the best product, with competitive mixing sustaining sales of the others. For arbitrary J, we optimize over the closure of feasible sales allocations. Positive sales require a purchase-weighted brand mean payoff between μ_O and μ_C, by valuation consistency. For each allocation, we minimize the prominence needed to implement it, allowing approximation at the outside-option boundary. When the brand valuation lies strictly between these thresholds, every brand purchase requires prominence. At valuation μ_C, competitive purchases instead provide a baseline: $d_i = q + (1 - q)\tau_i$, where d_i is the purchase probability in state i and q is the purchase probability in a competitive

[39] High-only prominence is uniquely optimal precisely for $c \in [4/9, 8/9]$, as established in the appendix.

menu. Holding sales fixed, increasing q reduces the prominence required. The least costly implementation therefore uses $q = \min_i d_i$. If an optimal allocation takes only the values $q > 0$ and 1, deterministic prominence implements it: promote the states with purchase probability one and let competitive mixing supply purchases elsewhere.

The net-profit supremum equals the maximum gross profit net of this exact or limiting prominence cost. An optimal policy, when one exists, or a sequence approaching the supremum can be chosen to randomize in at most one state.[40] The appendix gives the formal optimization problem and proof. Whenever the optimum sustains brand valuation μ_C, consumer welfare also equals μ_C: brand purchases yield that payoff on average, and remaining purchases go to the competitor. Costs can therefore preserve welfare through competitive mixing, although they need not select this outcome in every decision problem.

5 Discussion

5.1 A misspecified Bayesian learning interpretation

Limited salience provides a natural source of misspecification. Alice treats payoffs from alternatives within the same similarity class as independent draws from a common stationary distribution, so experience with any one alternative informs her assessment of the entire class. Yet alternatives that coincide on salient dimensions may differ systematically in expected payoffs because of attributes she overlooks. Alice therefore interprets variation arising from these differences as noise around a common class mean.

We show that while coarse Q-learning requires neither a subjective likelihood for payoffs nor a prior, its harmonic updating rule admits an exact Bayesian interpretation under a misspecified model of class-level payoffs. Suppose Alice treats the payoffs observed from each class s, conditional on an unknown mean μ_s, as independent draws from $N(\mu_s, \sigma_s^2)$, where $\sigma_s^2 > 0$ is known in her subjective model. She regards this payoff distribution as independent of the menu and the particular alternative implementing the class. Suppose further that her class means have independent priors $\mu_s \sim N(v_0(s), \sigma_s^2/\kappa_s)$, $\kappa_s > 0$. These are subjective assumptions: the objective payoff distributions need not be Gaussian. Since the normal likelihood is strictly positive at every real-valued payoff, Bayesian updating remains well-defined after every finite realized history. Let $\mathcal{H}_k$ denote Alice's observed history strictly before date k, and let $N_k(s) := \sum_{t=0}^{k-1} \mathbf{1}\{s_t = s\}$. Normal conjugacy yields a Gaussian

[40] A positive supremum is attained if and only if some maximizing sales allocation has a purchase-weighted mean strictly above μ_O. Otherwise it is approached at the outside-option boundary, as in the costless case.

posterior for μ_s with mean and variance

$$m_k(s) := \mathbb{E}_A[\mu_s \mid \mathcal{H}_k] = \frac{\kappa_s v_0(s) + \sum_{t=0}^{k-1} \mathbf{1}\{s_t = s\} r_t}{\kappa_s + N_k(s)},$$
$$V_k(s) := \mathrm{Var}_A(\mu_s \mid \mathcal{H}_k) = \frac{\sigma_s^2}{\kappa_s + N_k(s)},$$

where the subscript A indicates that posterior moments are computed under Alice's subjective model. The posterior mean satisfies

$$m_{k+1}(s) = m_k(s) + \mathbf{1}\{s_k = s\} \frac{r_k - m_k(s)}{\kappa_s + N_k(s) + 1}.$$

This is exactly the CQL update in (2) with the harmonic learning rates in (3). Since $m_0(s) = v_0(s)$, it follows that $m_k(s) = v_k(s)$ at every date along every realized path. The parameter κ_s has the usual Bayesian interpretation as a prior pseudo-count. The representation also describes how Alice's subjective uncertainty evolves. Her posterior variance of the class mean falls whenever that class is sampled and remains unchanged otherwise. For every fixed $\beta \geq 0$, Prop. 1 implies every class is sampled with positive lower limiting frequency. Consequently, $V_k(s) \to 0$ for every $s \in \mathcal{S}$, almost surely under the objective law. Thus, Alice becomes arbitrarily confident about each class's expected payoff in her subjective model.

Vanishing posterior uncertainty does not, however, require convergence of posterior means. Under the conditions of Theorem 7, the stochastic joint process almost surely has a non-trivial periodic orbit as its limit set, and valuations and choice probabilities do not converge. In these environments, Alice's posteriors become increasingly concentrated around means whose vector continues to move. She interprets her observations as evidence about fixed class-level payoff distributions, while her changing valuations alter the composition of the menus and alternatives generating those observations. Endogenous selection can therefore sustain non-stationary behavior even as subjective uncertainty vanishes.[41]

The Bayesian interpretation of valuation updating has a complementary choice-theoretic representation. Given current posterior means $\mathbf{v}_k$ and menu ω, suppose Alice chooses $\sigma \in$

[41] The posterior-mean correspondence extends to other regular one-parameter natural exponential families with scalar sufficient statistic $T(r)$ and Diaconis-Ylvisaker conjugate priors (Diaconis and Ylvisaker, 1979). With a positive prior pseudo-count and a prior mean in the interior of the convex support of $T(r)$, the posterior expectation of the mean statistic is the corresponding pseudo-count-weighted average of the observed $T(r)$ values. Objectively realized observations must be compatible with the subjective likelihood, and the posterior must remain proper after every finite history. When $T(r) = r$, this reproduces the original CQL update; otherwise, the correspondence concerns the recursion using $T(r)$ as its update target. Examples include Bernoulli-Beta, Binomial-Beta, and Poisson-Gamma specifications.

$\Delta(\omega)$ to maximize current subjective expected payoff net of an entropy cost:

$$\max_{\sigma\in\Delta(\omega)} \left\{ \sum_{s\in\omega} \sigma_s v_k(s) - \frac{1}{\beta} \sum_{s\in\omega} \sigma_s \log\left(\frac{\sigma_s}{1/|\omega|}\right) \right\}, \qquad \beta > 0.$$

The cost is proportional to the relative entropy from uniform choice, penalizing concentration on particular classes and thereby encouraging exploration. The unique optimizer is

$$\sigma_s = \frac{\exp\{\beta v_k(s)\}}{\sum_{r\in\omega} \exp\{\beta v_k(r)\}}, \qquad s \in \omega,$$

precisely our logit rule. Larger β lowers the cost relative to payoff differences; uniform choice is recovered as $\beta \downarrow 0$. Thus the benchmark combines Bayesian updating under Alice's subjective payoff model with entropy-regularized choice based on current posterior means.

5.2 Cognitive Complexity & Choice Overload

Beyond limited salience, CQL also admits a tractability interpretation, taking the similarity partition as given. Let $n := |\mathcal{S}|$. Separately evaluating each class conditional on the menu would remove the choice distortions seen above as Alice would asymptotically learn the expected payoff of each class in every menu correctly, and her welfare would approach the efficient benchmark $\sum_{\omega\in\Omega} p(\omega) \max_{s\in\omega} \pi_s(\omega)$ as $\beta \to \infty$. However, it would require one valuation $v(s,\omega)$ for every supported pair with $s \in \omega$, giving $\sum_{\omega\in\text{supp}(p)} |\omega| \leq n \cdot 2^{n-1}$ coordinates, with equality when all nonempty class menus are supported. Alternatively, directly evaluating performance under every strict class ranking guarantees the oracle welfare W^{coarse} asymptotically, yet entails $n!$ candidates, exceeding $n \cdot 2^{n-1}$ for $n \geq 5$. On the contrary, CQL maintains only n valuations and n visit counts, independently of the number of menus.

The burden of conditioning on menus is also statistical: each $v(s,\omega)$ is updated only when that particular menu occurs and class s is chosen. Rarely encountered pairs therefore provide little feedback. CQL pools observations across menus, reducing the number of quantities to estimate. This economy comes at a cost when class payoffs vary across menus. Current valuations determine which menus contribute observations to each class, so the pooled assessment reflects a composition of experiences shaped by Alice's own choices. The feedback between valuations and this endogenous selection drives our results.

The finite-class representation also accommodates infinitely many objective alternatives. Within-class choice can be specified by a fixed measurable probability kernel supported on the available alternatives. Whenever the induced reduced decision tree is finite and satisfies our maintained sampling and moment assumptions, our results apply to that tree.

We also note that menu variation need not reflect changes in physical availability. As the prominence application illustrates, it can instead arise from changes in which alternatives enter the consumer's consideration set. More generally, our analysis applies to such effective menus whenever their evolution satisfies the maintained independence assumptions. The similarity partition describes how Alice groups alternatives, while the considered menu determines which alternatives, and hence which classes, enter her current choice.

Finally, singleton-payoff shifts admit an interpretation as reduced-form choice aversion or affinity across classes (Iyengar and Lepper, 2000). If $u_a(\psi) = \mu_a(\psi) + \xi(\psi)$, the component $\xi(\psi)$ is common to all alternatives in a menu and leaves within-menu payoff comparisons unchanged, but affects assessments pooled across menus. Setting $\xi(\psi) = z\,\mathbf{1}\{|\rho[\psi]| = 1\}$ assigns a bonus to the absence of between-class choice when $z > 0$, and a penalty when $z < 0$. Under the conditions of Theorems 5 and 6, sufficiently large positive shifts produce limiting indifference, whereas sufficiently negative shifts yield multiple locally stable SVE at high sensitivity when all binary menus are supported.

6 Related Literature

Our closest point of contact is Jehiel and Samet (2007) who introduce valuation equilibrium (VE) for extensive-form games with similarity classes. They motivate VE through learning but leave the convergence of learning over similarity classes open. We analyze such learning explicitly in stationary single-agent bandit problems. Its steady states are smooth valuation equilibria (SVE), whose high-sensitivity limits refine VE. Using o-minimal geometry, we show that the set of limiting SVE is generically finite even when VE form continua. For strict VE, closest-loss consistency characterizes this refinement: valuations of classes left unchosen by the limiting policy must reflect the mean payoff in menus in which their valuation disadvantage is smallest. Every strict VE satisfying this condition is approached by locally stable SVE at all sufficiently high sensitivities. Additionally, we show how the endogenous composition of observations can produce either limiting indifference or multiple self-confirming strict rankings or oscillating rankings when no equilibrium is stable.

The endogenous composition of experience also connects CQL to Bayesian learning under misspecification. Berk-Nash equilibrium (Esponda and Pouzo, 2016) requires beliefs to best fit the observations generated by behavior, while Esponda et al. (2021) characterize learning through the dynamics of empirical action frequencies. Our harmonic update has an exact posterior-mean interpretation under a subjective Gaussian model, although objective payoff laws need not be Gaussian. With finite reward outcomes, product-Dirichlet priors also give

an exact Bayesian interpretation. Under *singleton-menu support*, the argmax counterpart of CQL (optimal choices as opposed to logit-optimal) embeds in the framework of Fudenberg et al. (2021): VE correspond to strongly uniform Berk–Nash equilibria, and strict VE inherit local stochastic stability from sufficiently concentrated priors. Applying their strict-stability theorem to finite menu-contingent plans also gives every mixed VE a stable purification through arbitrarily small payoff shocks. The appendix develops these connections formally. They do not by themselves determine the refinement induced by logit exploration, where we fix sensitivity during learning before taking its high-sensitivity limit. We also identify general conditions for global convergence of the CQL dynamics: constant competitive-menu payoffs ensure convergence toward the SVE set, while nonpositive payoff-crowding covariance ensures convergence to a unique SVE, even when singleton menus are not in support.

Our learning rule builds on payoff-assessment models that update values directly from realized rewards (Sarin and Vahid, 1999; Jehiel and Samet, 2005; Cominetti et al., 2010). Our departure is to pool rewards across alternatives in a class, so the distribution being assessed depends on the learner's choices across menus. This connects the paper to analogy-based expectation equilibrium (Jehiel, 2005) in games, where players aggregate observations when forming expectations, and case-based decision theory (Gilboa and Schmeidler, 1995), where evaluations depend on outcomes in similar past cases.[42] We take the similarity partition as given and study the feedback it creates between choice and experience. The resulting distortions can persist even when every class is sampled indefinitely. Coarsening therefore changes long-run behavior, beyond its familiar benefits for memory and sample complexity.

Our cycling result contributes to the study of persistent oscillation under misspecification (Nyarko, 1991; Fudenberg et al., 2017; Esponda et al., 2021). It isolates endogenous selection of observations under a fixed similarity partition as a source of non-convergence. We show the existence of an open set of decision trees with three classes and three independently drawn menus, in which sufficiently sensitive choice yields a unique unstable SVE, and the stochastic learning process almost surely has a nontrivial periodic limit set. Coarse inference can thus sustain oscillation in a stationary bandit problem even as learning rates vanish and, under the Bayesian interpretation, subjective uncertainty disappears.

The application adds a learning channel to the economic analysis of the effects of algorithmic prominence and self-preferencing by online marketplaces on consumer choice (Lee and Musolff, 2025; Hagiu et al., 2022) and derives novel conclusions: prominence changes the experiences pooled under a brand and thereby demand for the platform's own brand when

[42] VE is also conceptually related to Personal Equilibrium (Spiegler, 2016): Alice's choices generate the data used to fit a misspecified subjective model, and the fitted beliefs/valuations in turn justify those choices.

the competitor remains available. We characterize optimal stationary prominence and show how its costs can favor deterministic policies that sustain competitive mixing, while improving consumer welfare relative to costless promotion by preserving its no-prominence level. Our use of similarity classes connects CQL to the theory of coarse thinking. Mullainathan et al. (2008) show how categorical inference transfers the meaning of information across situations, creating scope for persuasion. Our application develops a related dynamic channel of category transference: prominence changes actual purchases, and favorable experiences with promoted products sustain demand for other inferior products under the same brand. Brand assessments are thus jointly determined with the purchases they induce.

The endogenous feedback mechanism also connects CQL to selection neglect. Jehiel (2018) studies investors who assess projects using the returns of previously implemented projects without correcting for the selection governing the observed sample.[43] In our setting, a single agent generates her own sample through repeated choices across menus and pools the experienced payoffs within classes. We characterize how the neglected dependence of this sample on her valuations shapes learning. We show selection can produce corrective feedback and globally convergent learning whose high-sensitivity limit sustains indifference and mixing, or reinforcing feedback supporting distinct stable strict rankings. These effects arise with stationary payoff distributions and independently drawn menus, and persist despite continued sampling and vanishing learning rates. Our welfare analysis separates the restriction imposed by common class valuations from the additional losses induced by learning those valuations as payoff assessments. An informed oracle can optimize within the same coarse representation, yet we show mere constant-payoff calibration and a monotone choice rule preclude attaining its optimal performance in some decision problems, even with arbitrary history-dependent reweighting of observations.

7 Conclusion

Coarse Q-learning makes the composition of experience endogenous even in stationary bandit problems with independently drawn menus. Although each alternative has a fixed payoff distribution with independent draws, the mixture sampled within a class depends on the valuations guiding choice. This feedback can produce globally attracting equilibria with limiting indifference and endogenous mixing, multiple stable strict rankings, or persistent oscillations. These outcomes arise from the interaction between coarse representation and the selection

[43] A related mechanism appears in Esponda (2008): naive buyers learn average quality from purchased goods but neglect how changing their bids alters the composition of purchases. In CQL, choices similarly determine the composition of payoff observations pooled within each class, although Alice interprets each class as a stationary mixture bandit arm.

of observations, without non-stationary fundamentals or strategic interaction. The welfare cost of coarse learning extends beyond the restrictions imposed by coarse representation. Even when a better policy is available under the same categories and common class scores, learning those scores as averages of experienced payoffs need not recover it. Reweighting observations across menus does not provide a general remedy: efficient scores encode priorities across menus that need not agree with the ordering of payoff estimates.

The application shows why this distinction matters for digital marketplaces. Prominence shapes the experiences pooled under a brand, so promoting good products can sustain purchases of inferior ones even when competitors remain available. With costless prominence, randomized policies can yield higher profit than every deterministic rule while reducing consumer welfare below its no-prominence level. Prominence costs can instead make deterministic policies sustaining competitive mixing optimal and preserve that welfare benchmark. These findings highlight the composition of promoted experiences, alongside the frequency of prominence, as a determinant of market outcomes. A natural next step for future research is to study when decision-makers revise their categories, and whether they can learn better choice priorities while retaining a small set of valuations.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological review*, 98(3):409.

Benaïm, M. (1999). Dynamics of stochastic approximation algorithms. *Séminaire de probabilités de Strasbourg*, 33:1–68.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2012). Salience Theory of Choice Under Risk. *The Quarterly Journal of Economics*, 127(3):1243–1285.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2013). Salience and consumer choice. *Journal of Political Economy*, 121(5):803–843.

Breitmoser, Y. (2021). An axiomatic foundation of conditional logit. *Economic Theory*, 72(1):245–261.

Caplin, A. and Dean, M. (2008). Dopamine, reward prediction error, and economics. *The Quarterly Journal of Economics*, 123(2):663–701.

Cominetti, R., Melo, E., and Sorin, S. (2010). A payoff-based learning procedure and its application to traffic games. *Games and Economic Behavior*, 70(1):71–83.

Debreu, G. (1960). Review of individual choice behavior. *The American Economic Review*, 50(1):186–188.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *The Annals of Statistics*, 7(2):269–281.

Esponda, I. (2008). Behavioral equilibrium in economies with adverse selection. *American Economic Review*, 98(4):1269–91.

Esponda, I. and Pouzo, D. (2016). Berk–nash equilibrium: A framework for modeling agents with misspecified models. *Econometrica*, 84(3):1093–1130.

Esponda, I., Pouzo, D., and Yamamoto, Y. (2021). Asymptotic behavior of bayesian learners with misspecified models. *Journal of Economic Theory*, 195:105260.

Farronato, C., Fradkin, A., and MacKay, A. (2023). Self-preferencing at amazon: Evidence from search rankings. *AEA Papers and Proceedings*, 113:239–43.

Fudenberg, D., Lanzani, G., and Strack, P. (2021). Limit points of endogenous misspecified learning. *Econometrica*, 89(3):1065–1098.

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

Fudenberg, D., Romanyuk, G., and Strack, P. (2017). Active learning with a misspecified prior. *Theoretical Economics*, 12(3):1155–1189.

Gilboa, I. and Schmeidler, D. (1995). Case-based decision theory. *The Quarterly Journal of Economics*, 110(3):605–639.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *PNAS*, 108(3):15647–15654.

Hagiu, A., Teh, T.-H., and Wright, J. (2022). Should platforms be allowed to sell on their own marketplaces? *The RAND Journal of Economics*, 53(2):297–327.

Hofbauer, J. and Sandholm, W. H. (2002). On the global convergence of stochastic fictitious play. *Econometrica*, 70(6):2265–2294.

Hofbauer, J. and Sandholm, W. H. (2009). Stable games and their dynamics. *Journal of Economic Theory*, 144(4):1665–1693.e4.

Iyengar, S. S. and Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of personality and social psychology*, 79(6):995.

Jehiel, P. (2005). Analogy-based expectation equilibrium. *Journal of Economic theory*, 123(2):81–104.

Jehiel, P. (2018). Investment strategy and selection bias: An equilibrium perspective on overoptimism. *The American Economic Review*, 108(6):1582–1597.

Jehiel, P. and Samet, D. (2005). Learning to play games in extensive form by valuation. *Journal of Economic Theory*, 124(2):129–148.

Jehiel, P. and Samet, D. (2007). Valuation equilibrium. *Theoretical Economics*, 2(2):163–185.

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

Lee, K. H. and Musolff, L. (2025). Two-sided markets shaped by platform-guided search.

Luce, R. D. (1959). *Individual choice behavior*, volume 4. Wiley New York.

Mullainathan, S., Schwartzstein, J., and Shleifer, A. (2008). Coarse thinking and persuasion. *The Quarterly Journal of Economics*, 123(2):577–619.

Nyarko, Y. (1991). Learning in mis-specified models and the possibility of cycles. *Journal of Economic Theory*, 55(2):416–427.

Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407.

Rosch, E. and Lloyd, B. B. (1978). *Principles of categorization*. MIT press.

Sarin, R. and Vahid, F. (1999). Payoff assessments without probabilities: A simple dynamic model of choice. *Games and Economic Behavior*, 28(2):294–309.

Skinner, B. (1938). *The Behavior of Organisms: An Experimental Analysis*.

Spiegler, R. (2016). Bayesian networks and boundedly rational expectations. *The Quarterly Journal of Economics*, 131(3):1243–1290.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. The MIT Press, second edition.

Tversky, A. (1972). Elimination by aspects: a theory of choice. *Psychological Review*, 79.

Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3):279–292.