Learning Optimal Behavior Through Reasoning and Experiences*

Cosmin Ilut
Duke University & NBER

Rosen Valchev
Boston College & NBER

March 20, 2025

Abstract

We develop a novel framework of bounded rationality under cognitive frictions that studies learning over optimal behavior through *both* deliberative reasoning *and* accumulated experiences. Using both types of information, agents engage in Bayesian non-parametric estimation of the unknown action value function. Reasoning signals are produced internally through mental deliberation, subject to a cognitive cost. Experience signals are based on the observed utility outcomes at previous actions. Agents' subjective estimation uncertainty, which evolves through information accumulation, modulates the two modes of learning in a state- and history-dependent way. We discuss how the model draws on and bridges conceptual, methodological and empirical insights from both economics and the cognitive sciences literature on reinforcement learning.

JEL Codes: D83, D91, E21, E71, C11

*We thank Sam Gershman for extensive and helpful discussions. We also thank seminar and conference participants at AEA Annual Meeting, Duke University and Boston University. Email addresses: Ilut cosmin.ilut@duke.edu, Valchev valchev@bc.edu.
There are no relevant sources of research funding or financial relationships pertaining to this study, all views and opinions are solely those of the authors.

1 Introduction

There is a deep and growing interdisciplinary interest in understanding and modeling cognitive limitations in decision making. Across both economics and cognitive sciences, the literature recognizes two broad ways in which people learn about and deduce what is the best course of action in a given situation. The first is *cognition and reasoning*: through introspective, abstract deliberations, humans can get a better grasp of what is their optimal action in the situation at hand. The second is *accumulated experiences*: by observing realized outcomes of past decisions, agents update their views on the respective benefits of taking those actions in these circumstances. These two sources of information are conceptually distinct and are both limited: experiences are observed only along the realized path of situations the agent actually faced in the past, and while abstract thinking could help the agent deduce outcomes in counterfactual actions and situations, such deliberations are cognitively costly. While the literature recognizes these two distinct methods of learning, they have so far not been studied together in a single, unified framework.

This paper develops a novel framework of constrained-optimal behavior under cognitive frictions that studies *jointly* learning through both reasoning *and* accumulated experiences. To do so, the paper draws on conceptual, methodological and empirical insights from both economics (such as Sims (2003a), Ilut and Valchev (2023)) and the cognitive sciences literature on reinforcement learning (RL) (Kaelbling et al. (1996), Sutton and Barto (2018)).[1]

In our framework, agents are uncertain about optimal behavior in the sense of facing subjective uncertainty over the optimal policy and value functions that characterize their decision problem. In particular, we follow the RL approach of letting agents learn about the *action* value function $Q_\pi(c, s)$, which gives the expected discounted sum of utility when choosing action c in state s and following a given policy $\pi(c|s)$ thereafter.[2] While the standard approach in economics is to endow the agents with perfect knowledge about the optimal policy function $\pi^*(c|s)$ and therefore the action value function $Q^*(c, s)$, in our model agents perceive $Q^*(c, s)$ as uncertain ex-ante, and gradually learn about it over time.

Learning can occur through two different channels. First is cognition – that is thinking through the decision problem at hand and getting a better handle of the best course of action. Thus, cognition helps the agents obtain a more precise estimate of $Q^*(c, s)$, however it is mentally costly. Agents trade off that benefit and cost and decide on the optimal amount

[1] See for example Wiederholt (2010), Gabaix (2019) and Mackowiak et al. (2021) for surveys on imperfect attention in economics, Enke (2024) and Caplin (2025) for recent reviews on cognitive economics, Gershman and Daw (2017) for a survey on RL and Barberis and Jin (2023) for a recent application of RL to economics.

[2] In the language of RL, this "action" value function is closely related and derived from the "state" value function $V(s)$ which is perhaps more familiar to economists.

of cognitive effort to invest in reasoning any given period, giving rise to "resource-rational" use of cognition. Second, agents also update beliefs about optimal behavior based on the experienced flow utility each period. This type of learning is cognitively free and is obtained by default by each agent at the beginning of the period.

We show how both reasoning and experience signals can be incorporated in revising beliefs over the unknown $Q^*(c,s)$ -function using formal non-parametric Bayesian updating formulas, which provide a tractable and recursive structure to the learning dynamics. There are four key specific features of our learning framework. First, we model agents' prior beliefs over the unknown function $Q^*(c,s)$ as a Gaussian Process (GP) distribution.[3] On the one hand, the GP distribution is methodologically appealing as a prior over the space of functions, as it is very flexible, and also tractable in recursively characterizing the conditional moments of the unknown function.[4] On the other hand, cognitive sciences increasingly emphasizes a Bayesian approach and in particular the appeal of GP distributions, both for conceptual and descriptive reasons.[5]

Second, *given* an estimate of the action value function $\widehat{Q}_t(c,s)$, agents take constrained optimal actions trading off *exploitation* and *experimentation* incentives. On the one hand, the agent has incentives to choose the action with the highest estimated value $\widehat{Q}_t(c,s_t)$ at the current state s_t (exploitation). On the other hand, the agent recognizes that $\widehat{Q}_t(c,s_t)$ is just an uncertain estimate so there is an incentive to explore other actions and learn more about the unknown $Q^*(c,s)$. We model this desire for experimentation following the concept of maximum entropy reinforcement learning, which constrains the entropy of the action distribution to be larger than some minimum threshold, thus ensuring randomization.[6] In contrast to the standard approach that sets this minimum threshold as an exogenous time-invariant parameter, we let the entropy lower bound be proportional to the remaining *subjective uncertainty* over $Q^*(c,s_t)$. This captures the intuition that experimentation is valuable only to the extent that the Q-function is uncertain. Putting it all together, the resulting optimal action policy function takes the form of the *softmax* function widely used in statistics and machine learning (see eg. Sutton and Barto (2018)).

Turning to the learning modes and the dynamic evolution of the estimates of the action value function, our third modeling feature is learning from experienced utility. Because it is derived from actual observed outcomes, this learning follows the so-called *"model-free"* or

[3]The GP distribution can be derived from first principles as the limiting case of Bayesian Gaussian kernel regression (eg. Rasmussen and Williams (2006), Liu et al. (2011)).

[4]In economics, these properties have also recently gathered attention to GP, including work like Callander (2011), Dew-Becker and Nathanson (2019), Ilut et al. (2020), Bardhi (2022) and Ilut and Valchev (2023).

[5]See for example Griffiths et al. (2008), Griffiths et al. (2010) and Gershman et al. (2015).

[6]See for example Mnih et al. (2016), Haarnoja et al. (2017), Eysenbach and Levine (2019)).

"Q-learning" dynamic programming solution techniques used in machine learning and cognitive sciences (e.g. Dearden et al. (1998)). In particular, at the beginning of the period, the agent observes the realization of the new state s_t and also the realized utility $u(c_{t-1}, s_{t-1})$ that she experienced based on last period's choice-state pair. Knowing the flow utility yesterday and the realized state today implies an *experience-based* informative signal that updates beliefs about $Q^*(c, s)$. The key intuition behind this update, is that the flow utility $u(c_{t-1}, s_{t-1})$ reveals the "temporal difference" in the Q-function, i.e. $u(c_{t-1}, s_{t-1}) = Q^*(c_{t-1}, s_{t-1}) - \beta \mathbb{E}_{t-1}(Q^*(c_t, s_t))$.[7]

Finally, but importantly, within the same objective function that determines optimal actions, we model the benefit and cost of abstract reasoning. By *reasoning* we mean the internal deliberation process through which the agent produces information and learns about $Q^*(c, s)$. As such, our terminology of reasoning is similar to the notion *model-based learning* in the RL literature (Sutton and Barto (2018)).

The benefit of reasoning is the resulting reduction of uncertainty over $Q^*(c, s_t)$. This benefit emerges through an *interaction* with actions. In particular, since reasoning decreases current conditional uncertainty, it also weakens the desire to experiment, by formally lowering the threshold on the action distribution entropy. Intuitively, the agent values the reduction in uncertainty from reasoning because it allows her to select an action closer to the one with the currently highest estimated, not having to worry about experimentation. In turn, the cost of reasoning is also in utility terms, quantified as the reduction in entropy achieved by acquiring these deliberation signals, as in information theory (e.g. Sims (2003b)), capturing effort as in the cognitive control literature (e.g. Botvinick et al. (2004), Kool et al. (2017)).

Critically, the effective precision of *both* the reasoning and experience-based signals is *endogenous*, and closely related to the agent's beginning of period prior beliefs and uncertainty, which evolve dynamically. In our non-parametric framework, uncertainty reduction is local – in the sense that a signal about the value function at some state of the world s_t updates beliefs about the value function at other states s as well, but the information content of the signal is decreasing in the distance $|s - s_t|$. This local reduction of uncertainty makes the evolution of uncertainty both state and history dependent – it matters at what other states the agent has already reasoned through or experienced his decision problem.

Moreover, the optimal choice of reasoning intensity takes the form of an optimal threshold – the agents only choose to reason as much as needed to lower uncertainty about the value function at the current state s_t down to an optimal level. As a result, if the agent is repeatedly visiting the same states of the world, uncertainty would already be at (or even below) the threshold, and hence the agent will choose to do no further reasoning. Furthermore,

[7] These updates resemble temporal difference solution techniques, used extensively in RL (Sutton (1988)), and in particular their Bayesian, GP-based version, connecting to the approach in Engel et al. (2003, 2005).

while the agent receive the experience signal each period, its precision is endogenous, and in particular, if the agent repeatedly sees the utility outcome of the same combination of actions and states, then the repeated observations of the same experience carry no new information.

The combination of these two forces give rise to "learning traps" – situations in which the agent's beliefs stop evolving, when the state variable is repeatedly cycling in the same part of the state space. We show that these learning traps can be self-fulfilling when the state dynamics are endogenous, and affected by the actions of the agent. Specifically, if the agent has established beliefs $\widehat{Q}(c, s)$, such that it leads to actions that keep the state dynamics s_t oscillating in the same interval of the state space, then these beliefs will not update any further as the agent finds no need to reason and keeps observing the same experiences.[8]

This feedback between beliefs and actions, and the resulting "learning traps', have rich economic implications. As an illustration, we showcase the mechanism in the context of an Aiyagari (1994)-style consumption-savings problem with incomplete markets and a borrowing constraint. A first result is that there are two types of behavior that are consistent with a learning trap in this setting, and thus these are the two types of behavior that emerge prominently in the ergodic distribution of the model.

First, one type of belief that lead to a self-fulfilling cycle is when agents perceive that $\widehat{Q}(c, s)$ is high for high values of consumption c, and hence these agents are systematically overconsuming, and pushing their wealth down towards the borrowing constraint. This type of beliefs create persistently Hand-to-Mouth agents. Second, the other type of beliefs consistent with a learning trap is when agents, on average, consume as much as under the permanent income hypothesis, but locally their implied consumption policy functions are *steep*. This steepness means that agents tend to consume a lot out of positive income shocks, and also cut back on consumption in the face of negative shocks, and thus keep their wealth (the state) stable, giving rise to a learning trap.

We quantify these two characteristic features of the model in Section 4, and find that the model can indeed generate a large fraction of Hand-to-Mouth agents (30%) and also a large MPC for even the wealthiest agents (0.27 MPC in the top quintile of wealth). These two moments were targeted in our calibration, but still constitute significant quantitative success since the full information model (i.e. the standard rational model) generates only 0.4% of Hand-to-Mouth agents and an MPC of just 0.04 for the wealthiest agents.

Moreover, the model also shows great promise in matching important untargeted models. First, in our model there is significant heterogeneity in what level of wealth $\bar{s}$ becomes stable for each agent. Thus, the model features *persistent* wealth inequality, with the particularly

[8]This is a similar phenomenon to the learning traps characterized in Ilut and Valchev (2023). However in that work learning from experiences is not allowed.

interesting feature that the ergodic distribution of $\bar{s}$ features a fat right tail. In other words, there are many rich agents in our model that are rich because they actively saved a lot. This characteristic feature of our model is consistent with the challenging empirical facts of Halvorsen et. al. (2024) who show that the majority of wealthy households are indeed wealthy because of their high savings rates.

Second our model does not only imply high average MPCs for wealthy agents, but in fact implies that there will be rich heterogeneity in the MPCs of different agents. The reason is that the beliefs of agents are subject to idiosyncratic reasoning noise, and hence different agents will have different perceptions of optimal behavior. While on average they all tend to beliefs that imply relatively steep policy functions, there will still be lot of heterogeneity in the actual slopes of these functions. As a result, the model implies significant heterogeneity in MPCs that is not related to any fundamental, observable characteristics of the agents. This is consistent with the empirical evidence of Lewis et al. (2019), who show that a big fraction of the variation in MPCs in the date remains unexplained even after controlling for a rich set of observable household characteristics.

Lastly, we also point out that the welfare losses of the mistakes our boundedly rational agents make is substantial. The reason is that in our framework agents beliefs and actions are not just noisy *around* the true optimal, which implies that any deviations from optimality are transitory. Due to the learning traps, in our framework agents' beliefs systematically deviate from the fully rational value function and behavior even in the long-run, and this has a significant impact on welfare. Taking into account these long-run deviations, we estimate that the welfare loss due to our bounded rationality mechanism is approximately 5% of permanent consumption.

Our framework connects to several literature strands. In terms of the economics literature, we relate to models of *resource-rationality* that emphasize a *cost-benefit tradeoff of reasoning*. This tradeoff appears in various approaches in economics (including Sims (2003a), Woodford (2020), Gabaix (2014), Alaoui and Penta (2022), Stevens (2020), Ilut and Valchev (2023), and others) and cognitive sciences (e.g. Gershman et al. (2015), Griffiths et al. (2015), Shenhav et al. (2017)). Our key contribution here is to use insights from the RL literature to integrate learning about optimal behavior from experiences *into* such cognitive-based models, while emphasizing the sequential nature of information acquisition.In turn, we connect to a growing interest and evidence in economics on *experience-based learning*, as in Malmendier and Nagel (2016), Malmendier (2021). We differ from this latter approach by studying experiences jointly with reasoning, as well as by modeling experiences as informative not about the state law of motion, but instead about the perceived value of taking specific actions, as in RL.

Our framework innovates within standard RL modeling itself as well, by building a *dynamic* cognitive model with several key features emerging endogenously. These features also differentiate our framework from recent applications of standard RL to economics, like Barberis and Jin (2023). In particular, first, we obtain and characterize an internally consistent, endogenous weighing arbitration between reasoning and experience signals in the form of state- and history-dependent Bayesian updating weights. The property that this weighing is based on uncertainty is shared with cognitive and neuroscience work arguing that the brain puts more weight on the learning mode that it deems more reliable (as in Daw et al. (2005), Lee et al. (2014)). Second, the framework is characterized by an *endogenous decision to engage reasoning*. The implication of adjusting cognition in a state-dependent way is desirable for both empirical and conceptual reasons. Empirically, it is consistent with evidence and a general desired approach (see eg. Kool et al. (2017)) in the neuroscience and RL literatures emphasizing the cognitive cost of mental simulations versus their benefit. Third, the tradeoff between exploitation and experimentation in actions is state- and history-dependent, since it depends on the conditional subjective uncertainty the agent still perceives in her estimate of $Q^*(c, s)$. In contrast, in the RL literature, this tradeoff is governed by an exogenously set parameters.[9]

2 Framework

We model the process of economic agents figuring out optimal behavior using both their abstract thinking abilities, i.e. cognition, and their accumulated experience with past actions.

2.1 Motivation

Consider a generic recursive problem with discrete time indexed by t that can be expressed as a Bellman equation

$$V^*(s_t) = \max_{c_t \in \mathcal{B}(s_t)} u(c_t, s_t) + \beta \mathbb{E}_t V^*(s_{t+1}) \tag{1}$$

At time t an agent makes a choice c_t (potentially a vector of choices) and is perfectly aware of all payoff-relevant details of the environment, including perfectly observing s_t, the vector of state variables, including exogenous and endogenous states. The state follows a Markov law of motion $s_{t+1} = \mathcal{F}(c_t, s_t, \epsilon_{t+1})$, where ϵ_{t+1} is a vector of innovations drawn from a distribution

[9] In that literature there is typically a constant (called temperature) parameter, or a pre-designed/adaptively tuned path (see eg. Kirkpatrick et al. (1983), Ackley et al. (1985), Lin et al. (2018)).

$\mathcal{G}(\epsilon_{t+1})$. The problem is also typically subject to constraints, written as $c_t \in \mathcal{B}(s_t)$, where $\mathcal{B}(s_t)$ defines the feasible budget set.

The primitives of the environment are therefore the per-period utility function $u(.)$, the time discount factor $\beta < 1$, the shock distribution $\mathcal{G}$, law of motion $\mathcal{F}$ and set of constraints $\mathcal{B}$. The value function $V^*(s)$ encodes the continuation value attached to starting in any particular realization of s, when following the optimal policy function $\pi^*(c|s)$. The latter is the optimally selected mapping from any given state s to probabilities over feasible choices $c_t \in \mathcal{B}(s_t)$ from equation (1).[10]

Such dynamic problems are at the core of modern economics. The standard approach to solving them in economics is to outright assume that since the agent knows all the payoff-relevant primitives, the agent also immediately knows the optimal policy function $\pi^*(c|s)$ implied by problem (1), or equivalently the implied value function $V^*(s)$. Alternatively, a large literature on information frictions assumes some form of imperfect observation of the state s_t, but still maintains the assumption that the agent knows perfectly well the implied optimal decision rule π^* *given the agent's information* about the state.

However, since dynamic problems require evaluating infinite paths of actions, or in recursive terms, an action plan with all possible future contingencies, characterizing the value function $V^*(s)$ is generally a challenging functional problem (e.g. Judd et al. (1998), Bertsekas (2019)) - where in fact characterizing the π^* and V^* functions under imperfect information about s_t is often even more complicated. As a result, these complex solutions are often not fully tractable even to highly trained economists, and state-of-the-art approximation techniques take significant effort to implement. We aim to develop a framework which takes seriously that economic agents do not have immediate access to knowledge about $V^*(s)$, and thus $\pi^*(c|s)$ and modeling their gradual learning process.

Learning through experiences and reasoning

Our starting point is to observe that in real life people typically have *two* sources of information about optimal behavior. The first source is *experiences*: by observing the per-period utility outcomes of different actions taken at various states in the past, the agent learns about $V^*(s)$. The second source is *cognition and reasoning*: a unique human characteristic is the ability to think abstractly about the problem at hand. Through such internal deliberations, agents can learn about $V^*(s)$, as the implied value of taking different courses of action – for example, this could take the form of mentally simulating paths forward of possible behavior and comparing expected outcomes. Both learning sources are important, but also both are limited: experiences are realized only along the *actual path* taken by the

[10]Stating policies in terms of probabilities of taking actions is useful in developing the framework later.

agent, while abstract thinking is a scarce cognitive resource.

To model learning from both mental simulation and actual experiences, we connect to the large *reinforcement learning* (RL) literature (Kaelbling et al. (1996), Sutton and Barto (2018)). There an agent interacting with a dynamic and stochastic environment learns an optimal control policy for a sequential decision problem, typically a Markov Decision Process.

Action value function

The key object underscoring learning in RL, and the one we focus on here as well, is the so called "action value function". This is a concept closely related to the familiar Bellman value function in equation (1), and is defined as

$$Q^*(c_t, s_t) = u(c_t, s_t) + \beta \mathbb{E}_t V^*(s_{t+1})$$

This function gives the present discounted value of utility for any given action-state pair (c_t, s_t) at time t, assuming optimal behavior going forward. We can also write the action value function in a recursive form as

$$Q^*(c_t, s_t) = u(c_t, s_t) + \beta \mathbb{E}_t Q^*(\pi^*(c_{t+1}|s_{t+1}), s_{t+1}) \tag{2}$$

where $\pi^*(c|s)$ satisfies the Bellman equation (1). The functions $V^*(s)$ and $Q^*(c, s)$ are closely related, as $V^*(s_t) = \max_{c_t \in \mathcal{B}(s_t)} Q^*(c_t, s_t)$, but methodologically it is useful to work with Q^*.

The key idea we want to analyze is how, starting at the beginning of time, i.e. $t = 0$, the agent is uncertain about $Q^*(c, s)$ and can learn about it over time, utilizing both costly internal deliberation and learning from experienced outcomes of past taken actions c_{t-s}.

2.2 Human learning with Gaussian Processes

We model agent's beliefs over the unknown function $Q^*(c, s)$ as Gaussian Process (GP) distributions. In particular, at the beginning of time the agent has the initial prior

$$Q^*(c, s) \sim GP(\widehat{Q}_0(c, s), \widehat{\Sigma}_0(c, s, c', s')) \tag{3}$$

where the mean function is $\widehat{Q}_0(c, s) = \mathbb{E}(Q^*(c, s))$ and the variance-covariance function is $\widehat{\Sigma}_0(c, s, c', s') = Cov(Q^*(c, s), Q^*(c', s'))$. The defining feature of a GP distribution is that for any two action-state pairs (c, s) and (c', s'), the values $Q^*(c, s)$ and $Q^*(c', s')$ have a joint-Normal distribution with mean and variance-covariance given by $\widehat{Q}_0$ and $\widehat{\Sigma}_0$.

The GP distribution can be derived from first principles as the limiting case of Bayesian Gaussian kernel regression (eg. Rasmussen and Williams (2006), Liu et al. (2011)), and in

this formulation the resulting variance-covariance function $\widehat{\Sigma}_0(c, s, c', s)$ encodes the agent's prior view of how likely correlated are the function's values at different points (here pairs (c, s) and (c', s')), and therefore is inducing a measure of proximity, or similarity, between those points. A functional form that we find particularly tractable is the popular negative exponential, $\widehat{\Sigma}_0(c, s, c', s') = \sigma_0^2 \exp(-\psi_s||s - s'|| - \psi_c||c - c'||)$, where the parameters ψ_s and ψ_c control the prior degree of correlation.

Using GP distributions is appealing for two main reasons. First, due to its flexibility and tractability – GP distributions extend the familiar Kalman filter to learning about functions and the law of motion for the conditional mean and variance can be characterized recursively. As a result, the *conditional* beliefs in *each* period also follow updated GP distributions, as detailed below.

Second, this approach builds on and it is motivated by developments in cognitive sciences. At a broad level, cognitive sciences emphasize that cognition is fundamentally related to forming uncertain conjectures from partial or noisy information, and thus a probabilistic framework is particularly well suited both conceptually and in terms of accounting for data on observed behavior (eg. Chater et al. (2006), Griffiths et al. (2008)). The GP distribution in particular has been increasingly used in this literature, building on connections to statistics and machine learning (on the latter see eg. Barber (2012)). Moreover, recent experimental and neuroscience evidence point to the human learning process as well described by GP (see eg. Gershman et al. (2015) for a survey and Schulz et al. (2018), Wu et al. (2018) for some evidence). Even more specific to RL, the Bayesian approach and the GP distribution have also found applications in this RL literature (see eg. Engel et al. (2005)). As surveyed in Ghavamzadeh et al. (2015), this approach can provide a tractable and coherent way to model the exploitation-exploration tradeoff, fundamental to learning from experiences, as a function of subjective uncertainty over $Q^*(c, s)$ estimates, as well as offering a formal way to incorporate prior beliefs into the action-selection algorithms.

2.3 Learning from experienced outcomes

We start by detailing how we formalize the learning through experienced outcomes and its connection to the concept of "model-free" learning in machine learning. We will consider this type of learning as free of cognitive costs, as it accumulates by the agent simply experiencing utility flows of the specific actions she has taken in the past, and not through abstractly thinking about contingencies and counterfactuals.

Consider a typical period t. The agent enters the period with some prior beliefs, conditional on a given prior information set $\mathcal{I}_{t-1}$. As anticipated earlier, and proven later in

this section, given the time zero prior in equation (3), the conditional $t-1$ beliefs also have a GP distribution:

$$Q^*(c,s)|\mathcal{I}_{t-1} \sim GP(\widehat{Q}_{t-1}(c,s), \widehat{\Sigma}_{t-1}(c,s,c',s')) \tag{4}$$

where $\widehat{Q}_{t-1}(c,s) = \mathbb{E}(Q^*(c,s)|\mathcal{I}_{t-1})$ and $\widehat{\Sigma}_{t-1}(c,s,c',s') = Cov(Q^*(c,s), Q^*(c',s')|\mathcal{I}_{t-1})$ follow a recursive formulation detailed later in section 2.4.

At the beginning of period t, the agent observes the current realized state s_t and the utility flow experienced given the previous period's choices $u(c_{t-1}, s_{t-1})$. This utility flow is in fact informative about the optimal action-value function $Q^*(c,s)$, since rearranging the Bellman equation (2) we can write

$$u(c_{t-1}, s_{t-1}) = Q^*(c_{t-1}, s_{t-1}) - \beta \mathbb{E}_{t-1} Q^*(\pi^*(c_t|s_t), s_t) \tag{5}$$

Thus, agents realize that experienced utility is a linear combination of the unknown $Q^*(c,s)$ and this forms the basis of the information structure of the experience signal we define below.

2.3.1 The experience-based signal

Updating beliefs based on equation (5) involves a potentially complex and cognitively expensive integration step in the computation of the expectation $\mathbb{E}_{t-1} Q^*(\pi^*(c_t|s_t), s_t)$. At the same time, we are interested in modeling the learning from experiences as cognitively free, a type of learning that comes to the agent by default each period. To this end, we simplify the updating step by building on the idea of 'model-free' Q-learning found in RL (eg. Dearden et al. (1998) and Sutton and Barto (2018)) and use the following approximation

$$u(c_{t-1}, s_{t-1}) \approx Q^*(c_{t-1}, s_{t-1}) - \beta Q^*(\pi^*(c_t|s_t), s_t) \tag{6}$$

The key advantage and property of this approximation highlighted by the RL literature is that the temporal difference on the right-hand side does not need to compute the expectation $\mathbb{E}_{t-1}$ and thus also does not need to assume knowledge of the transition probability $\mathcal{F}(c_{t-1}, s_{t-1}, \epsilon_t)$. The agent simply observes the actual realization of s_t and looks back at the old action-state pair (c_{t-1}, s_{t-1}) and the realized utility $u(c_{t-1}, s_{t-1})$. Still, using this approximation to update estimates of the Q-function is *asymptotically consistent* – if agents visit all possible states in the support of s_t with positive probability, then the update based on the approximation will eventually converge to the true $Q^*(c,s)$ as $t \to \infty$ (Dearden et al. (1998)). The way we specify the optimal choice c_t below will indeed satisfy this full spanning condition.

One last computational difficulty is that the last term in equation (6) evaluates $Q^*(c,s)$

at the optimal policy $\pi^*(c|s)$. Since $Q^*(c,s)$ is unknown so is the policy $\pi^*(c|s)$. Although unknown, we can nevertheless use structure on the policy function, by noting that under full information this policy is degenerate, putting all the weight on the 'greedy' action, i.e. $\pi^*(c_t^*|s_t) = 1$ for $c_t^* = \arg\max_{c_t \in B(s_t)} Q^*(c_t, s_t)$, and $\pi^*(c_t|s_t) = 0$ for all other actions $c_t \neq c_t^*$. Since our framework keeps track of beliefs about $Q^*(c,s)$, we can thus approximate $\pi^*(c_t|s_t)$ by the agent's current best guess, namely by the greedy policy conditional on time $t-1$ beliefs about $Q^*(c,s)$:

$$\widehat{\pi}_{t-1}^{greedy}(c_t|s_t) = \begin{cases} 1, & \text{for } c_t = \widetilde{c}_t \equiv \arg\max_{c_t \in B(s_t)} \widehat{Q}_{t-1}(c_t, s_t) \\ 0, & \forall\, c_t \neq \widetilde{c}_t \end{cases} \tag{7}$$

In this way, we use the observed utility flow $u(c_{t-1}, s_{t-1})$, to define the *experience-based* signal η_t^E

$$\eta_t^E \equiv u(c_{t-1}, s_{t-1}) \approx Q^*(c_{t-1}, s_{t-1}) - \beta Q^*(\pi_{t-1}^{greedy}(c_t|s_t), s_t) \tag{8}$$

We can then readily compute a formal Bayesian update based on this experience signal,

$$\widehat{Q}_t^E(c,s) = \widehat{Q}_{t-1}(c,s) + \alpha_t^E(c,s)\left[\eta_t^E - \left(\widehat{Q}_{t-1}(c_{t-1}, s_{t-1}) - \beta \widehat{Q}_{t-1}(\pi_{t-1}^{greedy}(c_t|s_t, s_t))\right)\right] \tag{9}$$

where $\widehat{Q}_t^E(c,s) \equiv \mathbb{E}(Q^*(c,s)|\mathcal{I}_{t-1}, \eta_t^E)$. The signal to noise ratio $\alpha_t^E(s,s)$ reflects the beginning-of-period uncertainty $\widehat{\Sigma}_{t-1}(c,s,c',s')$ and can be derived as

$$\alpha_t^E(c,s) = \frac{Cov(\eta_t^E, Q^*(c,s)|\mathcal{I}_{t-1})}{Var(\eta_t^E|\mathcal{I}_{t-1})} \tag{10}$$

The signal also reduces the conditional uncertainty facing the agent, as

$$\begin{aligned} \Sigma_t^E(c, s_t, c', s_t) &\equiv Cov(Q^*(c,s_t), Q^*(c',s_t)|\mathcal{I}_{t-1}, \eta_t^E) \\ &= \widehat{\Sigma}_{t-1}(c, s_t, c', s_t) - \alpha_t^E(c,s) Cov(\eta_t^E, Q^*(c,s)|\mathcal{I}_{t-1}) \end{aligned}$$

Consistent with the cognitive science literature, we interpret this experiential, "model-free" update, along the realized path of state action pairs (c_t, s_t), as "cognitively free" - in contrast to the cognitively costly source of learning through reasoning modeled in section 2.4. Conditional on the prior information $\mathcal{I}_{t-1}$ and the experience signal η_t^E, the agent's beliefs at time t can so far thus be described as

$$Q^*(c,s)|\mathcal{I}_{t-1}, \eta_t^E \sim GP(\widehat{Q}_t^E(c,s), \widehat{\Sigma}_t^E(c,s,c',s')) \tag{11}$$

2.3.2 Noise in the experience-based signal

The approximation step in equation (6), by avoiding the computational intensity of integrating, induces noise in the update based on the 'model-free', experience-based signal. Specifically, as modelers we know that the implied "surprise" in this experience signal can be written as

$$\eta_t^E - \mathbb{E}(\eta_t^E|\mathcal{I}_{t-1}) = Q^*(c_{t-1}, s_{t-1}) - \beta \mathbb{E}_{t-1} Q^*(\pi^*(c_t|s_t), s_t) - \left(\widehat{Q}_{t-1}(c_{t-1}, s_{t-1}) - \beta \widehat{Q}_{t-1}(\pi_{t-1}^{greedy}(c_t|s_t)), s_t)\right)$$

To gain some intuition, let us zoom in on the extreme case where the prior turns out to equal the truth, i.e. $\widehat{Q}_{t-1}(c, s) = Q^*(c, s)$. Even in this case there is still noise in the signal, which can be written as

$$\eta_t^E - \mathbb{E}(\eta_t^E|\mathcal{I}_{t-1}) = -\beta \left(\mathbb{E}_{t-1} Q^*(\pi^*(c_t|s_t), s_t) - Q^*(\pi^*(c_t|s_t), s_t)\right) \tag{12}$$

Since a 'bad' shock to the state s_t would make $Q^*(\pi^*(c_t|s_t), s_t)$ lower than its conditional expectation $\mathbb{E}_{t-1} Q^*(\pi^*(c_t|s_t), s_t)$, by equation (12) we see that this will show up as effectively 'bad' news in the signal, that is $\eta_t^E - \mathbb{E}(\eta_t^E|\mathcal{I}_{t-1}) < 0$. Here by 'bad shock' we mean an innovation to s_t in the direction in which $Q^*(c, s)$ is decreasing in s. For example, if there is a state like wealth or income, that would be a negative shock to income.

As per equation (6), a 'negative' surprise in the signal η_t^E updates downward the agent's perception of the *difference* between the action value function at yesterday's state of the world s_{t-1} and the action value function at today's state of the world s_t. Intuitively, this would imply that a negative signal η_t^E updates *downwards* beliefs of $Q^*(c_{t-1}, s_{t-1})$, the value of yesterday's action-state pair, and *upwards* the beliefs of $Q^*(c_t, s_t)$.

Intuitively, this means that the agent partly attributes the fact that he is worse off today at time t (which is truly due to the negative exogenous shock to s_t), to the possibility that the action taken yesterday, c_{t-1}, was a 'bad' action that led us down a 'bad path' today.

An illustration: consumption-savings problem

At this stage it is useful to provide an illustration of these forces in a simple economic framework. To do so, we will use a standard consumption-saving dynamic problem with a borrowing constraint as an illustrative application of the framework. In Section 4 we further perform a numerical illustration of the mechanism in this specific environment. This application is a natural laboratory for various reasons. First, this dynamic problem is a core and widely-studied mechanism in macroeconomic models, including the RBC, New Keynesian and Heterogeneous Agent models. Second, the environment is rich but transparent enough to serve as a useful laboratory to understand departures from full rationality. Third, in the micro data, consumption decisions appear to deviate from standard rational models in

important ways, some of which we will speak to in Section 4. For these reasons, recent work on bounded rationality and behavioral macro (e.g. Laibson et al. (Forthcoming), Bianchi et al. (2021), Ilut and Valchev (2023), Lian (2023), Andre et al. (2025)), also typically focus on exploring consumption-savings problems and thus studying the new framework in the same context is further helpful for understanding how key implications differ.

This environment is a specific case of the general notation introduced around equation (1). The agent has concave utility $u(c_t)$ over her consumption c_t, which is the only choice variable.[11] She inelastically supplies at a constant given wage w her stochastic endowment of labor ϵ_t, drawn i.i.d as $\log \epsilon_t \sim N(0, \sigma)$. She can save in an asset that earns a constant return r. In this case, the only relevant state variable for her decision problem is s_t, "cash-on-hand", which here then follows the law of motion $\mathcal{F}(c_{t-1}, s_{t-1}, \epsilon_t)$ given by

$$s_t = (1 + r)(s_{t-1} - c_{t-1}) + w\epsilon_t$$

Imposing a no debt constraint leads in this case to the feasible action set $\mathcal{B}(s_t)$ be $c_t \leq s_t$.

Next, we use Figure I to illustrate the mechanics of the experience signal η_t^E within this straightforward economic framework. For this illustration, we are plotting a situation where the time t income shock is negative, $\epsilon_t < 0$, so that the agent has experienced a "bad" exogenous shock to the state variable, cash on hand. In both panels, we plot conditional beliefs over the action-value function $Q^*(c, s)$ at a fixed value for the state variable s, with the x-axis covering the feasible range for consumption c_t given the relevant value of the state.

The blue line in panel (a) plots the beginning of period conditional expectation of the agent evaluated at the time $t-1$ value of the state variable s_{t-1}, $\widehat{Q}_{t-1}(c, s_{t-1})$, and the blue line in panel (b), instead, plots the beginning of period conditional expectation at the time t value of the state s_t. Keep in mind that in this example $s_t < s_{t-1}$, since we are showcasing a negative income shock at time t, and we can see this from the fact that the x-axis in panel (b) covers a smaller interval of values. So both panels show us the conditional expectation of the optimal action-value function as a function of possible consumption choices, and thus shows the expected payoffs of the agent taking any given consumption action c.

The red lines plot the posterior beliefs $\widehat{Q}_t^E(c, s)$ after updating with experienced utility $u(c_{t-1})$, conditional on a current negative income shock ε_t. So the left panel shows us how the agent updates his beliefs over the value function evaluated at s_{t-1} and the right panel shows us the update of the value function evaluated at s_t. We can clearly see how the conditional belief of the value function at the old state is shifted down, while at the new state it is shifted up, as foreshadowed in the above discussion. That is, overall, a bad exogenous shock to the

[11] Compared to the general setup in equation (1), in this specific case utility does not depend directly on the current state.

state (here a negative income shock), leads the agent to update downwards beliefs of the present discounted value of being at state s_{t-1}, and upwards for being at the state s_t.

More importantly, these shifts affect not only the estimated level of the value functions, but also their *shapes* and thus also change the policy function the agent chooses to follow. We will discuss these effects on the optimal choice of c_t in detail further below.

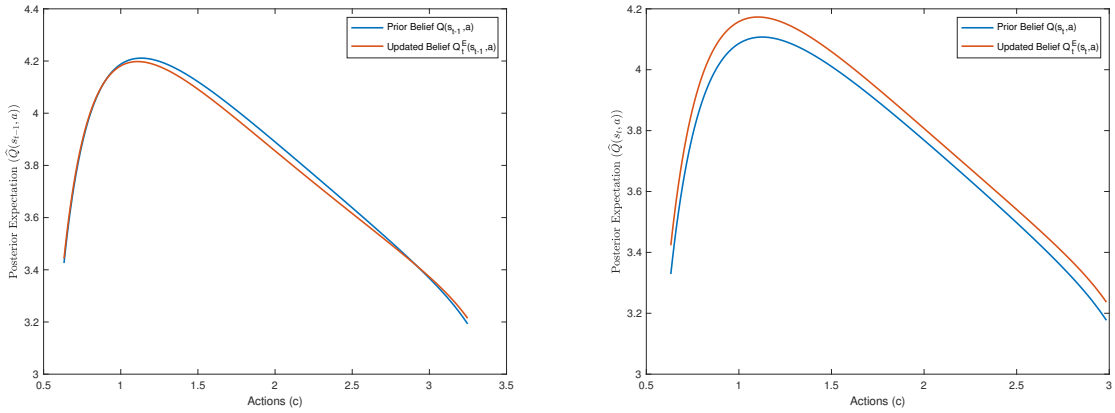


Figure I: Experience signal, negative s_t surprise: updating $Q^*(c, s_{t-1})$ (left) and $Q^*(c, s_t)$ (right)

2.4 Learning through abstract reasoning

After updating with the cognitively-free experience signal at the beginning of period t, the agent's beliefs are described by equation (11). Next, the agent decides how intensely to engage in cognitively costly abstract reasoning, which produces further information about $Q^*(c, s)$. After updating beliefs based on all information obtained at time t, the agent optimally chooses an action policy $\pi_t(c|s_t)$.

2.4.1 Reasoning signal

To understand these choices, consider first the reasoning process. By *abstract reasoning* we mean the *internal* deliberation process through which the agent deduces information about $Q^*(c, s)$ by doing mental calculations and analysis, not through directly experiencing outcome of actions. This notion of reasoning is similar to the notion of *planning* or *model-based learning* in the RL literature (Sutton and Barto (2018)). Reasoning can in practice work through various specific processes. For example, the agent might perform (mentally) value function or policy function iteration to varying degrees of precision, or simulate in their mind a number of future paths of actions and compute resulting discounted utility flow (a common

solution technique in RL, e.g. Moerland et al. (2023)).[12]

We do not take a stance on the specific methodology used by the agent to produce information about $Q^*(c,s)$, but model this process in a flexible reduced form way that aims to capture a key tradeoff that many different solution methods share – taking more computational steps in any given method (e.g. simulate more and further paths forward), leads to (i) higher accuracy of the resulting solution, but (ii) is cognitively costlier.

To do so, first we assume that through reasoning the agent can generate a vector of signals η_t^R that are unrestricted linear functions of $Q^*(c,s_t)$

$$\eta_t^R = \Omega_t' Q^*(c,s_t) + \varepsilon_{\eta,t} \quad ; \quad \varepsilon_{\eta,t} \sim N(0,\Sigma_{\eta,t}) \tag{13}$$

The agent optimally chooses the structure of the loadings matrix Ω_t and the noise variance matrix $\Sigma_{\eta,t}$, subject to a cost-benefit tradeoff introduced further below.

The reasoning signal η_t^R enhance the current information set, which becomes $\mathcal{I}_t = (\mathcal{I}_{t-1}, \eta_t^E, \eta_t^R)$, and thus updates the experienced-based beliefs. The resulting posterior end-of-period belief over the vector of Q-values at all (c,s) is:

$$\widehat{Q}_t(c,s) = \widehat{Q}_t^E(c,s) + \alpha_t^R(c,s)' \left(\eta_t^R - \Omega_t' \widehat{Q}_t^E(c,s_t)\right) \tag{14}$$

where the signal-to-noise ratio $\alpha_t^R(c,s)$ is given by

$$\alpha_t^R(c,s) = Cov\left(Q^*(c,s), \Omega_t' Q^*(c,s_t) | \mathcal{I}_{t-1}, \eta_t^E\right) \left(Var(\Omega_t' Q^*(c,s_t)|\mathcal{I}_{t-1}, \eta_t^E) + \Sigma_{\eta,t}\right)^{-1} \tag{15}$$

which can be tractably computed from $\widehat{\Sigma}_{t-1}(c,s,c',s')$ and the update based on η_t^E described above. Importantly, the reasoning signals also further reduce uncertainty, so that

$$\widehat{\Sigma}_t(c,s_t,c',s_t) \equiv Cov(Q^*(c,s_t), Q^*(c',s_t)|\mathcal{I}_t) = \Sigma_t^E(c,s_t,c',s_t) - \alpha_t^R(c,s)\Sigma_t^E(c,s_t,\Omega_t c;s_t)' \tag{16}$$

denotes the posterior variance function over the vector $Q^*(c,s_t)$. Like with the mean, the signal updates beliefs over the co-variance function at all states.

So far we have showed how both reasoning and experience signals can be incorporated in revising beliefs over the unknown Q-function using formal non-parametric Bayesian updating formulas. This joint integration of both types of learning is present in both some recent treatments of RL (see Sutton and Barto (2018)) and also in classical RL algorithms such as Dyna-Q (Sutton (1990, 1991)).[13] Nevertheless, the RL literature usually resorts to ad-

[12] By producing new information internally to the agent, without being observed by an outsider, reasoning here also resembles that of 'fact-free learning' in Aragones et al. (2005) and Alaoui and Penta (2022).

[13] These algorithms include many extensions, such as Dyna-Q+ (Sutton (1990)), which like in the current

hoc assumptions on how the two types of learning are mixed together, while our framework proposes an endogenous arbitration between the two types of signals based on formal Bayesian updating notions. Furthermore, the specific formulation of belief updating for both signals in equations (14) and (9) resembles temporal difference solution techniques, used extensively in RL (Sutton (1988)), and in particular their Bayesian, GP-based version, connecting to the approach in Engel et al. (2003, 2005).

Put together, these two sources of information update agents' conditional mean and uncertainty over the whole function $Q^*(c,s)$, providing a recursive structure to the GP distribution entering next period. In particular, conditional on the entire information set $\mathcal{I}_t$, agent's beliefs at time t are thus given by

$$Q^*(c,s)|\mathcal{I}_t \sim GP(\widehat{Q}_t(c,s), \widehat{\Sigma}_t(c,s,c',s')) \tag{17}$$

Equations (14) and (16) describe the updated beliefs, taking as given the structure for the reasoning signal in equation (13). We now explain how given the current state s_t the agent chooses her optimal reasoning structure and action policy function.

We break the choice over reasoning and actions in two sequential, but connected steps. First, the agent chooses her optimal reasoning (i.e. Ω_t and $\Sigma_{\eta,t}$) given her expectation of how different reasoning choices, and associated realizations of the reasoning signals, will affect her optimal action at the end of the period. This choice is naturally conditional on her beliefs before obtaining the reasoning signals, i.e. conditional on the information set $\mathcal{I}_t^E$. Second, given the optimally chosen reasoning structure, the agent implements an action policy function that is conditional on the ex-post realization of the reasoning signal η_t^R. This latter policy governs the actual actions taken at the end of the current period t.

2.4.2 Action distribution conditional on reasoning signals

It is most instructive to consider these choices in reverse order, starting with the optimally chosen action distribution given a reasoning choice and realized reasoning signals η_t^R. Given the end of period beliefs, equation (17), the agent chooses the optimal action distribution that maximizes the following objective

$$\max_{\pi_t(c|s_t)} \mathbb{E}\left(\sum_c Q^*(c,s_t)\pi_t(c|s_t)\middle|\mathcal{I}_t\right)$$

subject to two types of constraints on the action distribution. First, is that of feasibility: $\pi_t(c|s_t) = 0$ for actions that do not satisfy the budget constraint $c \notin \mathcal{B}(s_t)$, and that the

framework further stimulates exploration.

action distribution probabilities sum to one $\sum_c \pi_t(c|s_t) = 1$. Second, the action distribution is subject to minimum entropy constraint:

$$-\sum_c \ln(\pi_t(c|s_t))\pi_t(c|s_t) \geq h \sum_c \sigma_t^2(c, s_t) \tag{18}$$

where the LHS is the entropy of the actions distribution $\pi_t(c|s_t)$, and $\sigma_t^2(c, s_t)$ denotes the diagonal entries of the end-of-period posterior variance $\Sigma_t(c, s_t, c', s_t)$.

Let us describe and motivate the tradeoffs characterizing this constrained optimization. In terms of actions, our objective captures, typical to the RL literature, the *exploitation-experimentation tradeoff* in taking actions in the presence of uncertainty over the function $Q^*(c, s)$. In particular, the first term in the objective (20) reflects the agent's *benefit of exploitation* (or "greediness" in the language of RL). This term incentivizes the agent to choose the action with the highest estimated value $\widehat{Q}_t(c, s_t)$.

In turn, the entropy constraint (18) captures this *desire for experimentation*. In particular, leveraging insights in the RL literature, we model the desire to experiment in a parsimonious reduced form way with the idea of "*entropy regularization*", or maximum entropy RL, (eg. Mnih et al. (2016), Haarnoja et al. (2017), Eysenbach and Levine (2019)), which requires the entropy of the action distribution to be bigger than some minimum threshold. This regularization ensures randomization in actions as long as there still remains uncertainty about the optimal action value function $Q^*(c, s_t)$ – the chosen action distribution $\pi_t(c|s_t)$ puts some probability of exploring any feasible action unless $\sigma_t^2(c, s_t) = 0$ for all c). This tractably captures the exploration incentives inherent in dynamic learning problems – as long as agents are uncertain about the true payoffs of any given action, they should not only select the action with the current highest estimated payoff but also put some weight on further exploration.

The solution to this constrained optimization yields the following action distribution

$$\widehat{\pi}_t(c|s_t) = \begin{cases} \frac{\exp\left(\frac{\widehat{Q}_t(c,s_t)}{\delta_t}\right)}{\sum_c \exp\left(\frac{\widehat{Q}_t(c,s_t)}{\delta_t}\right)} & \text{if } \delta_t > 0 \\ \arg\max_c \widehat{Q}_t(c, s_t) & \text{if } \delta_t = 0 \end{cases} \tag{19}$$

where δ_t is the Lagrange multiplier on the minimum entropy constraint (18).

The policy in (19) takes the form of the *softmax* function widely used in statistics and machine learning. This function takes the expected rewards of following any given action and transforms them into action probabilities (eg. Sutton and Barto (2018)). The lower is δ_t, the more the expected rewards affect the probability of taking actions. When the entropy

constraint is not binding (that is when there is no remaining uncertainty about $Q^*(c, s_t)$), $\delta_t = 0$ and the policy becomes the degenerate 'greedy' one, that puts all the weight on the highest-expected reward inducing action. Thus, the agent deviates away from selecting the single action with the highest expected payoff and experiments with other actions, only to the extent to which there is remaining uncertainty about the optimal action-value function and $\delta_t > 0$.

Moreover, the magnitude of δ_t effectively reflects the immediate opportunity cost of deviating from the action with the current highest expected payoff (that is experimenting, rather than exploiting the current expected value function). Thus, the multiplier tends to be higher when (i) the beliefs $\widehat{Q}_t(c, s_t)$ are more peaked, so that deviating from the highest-perceived rewards is perceived to be more costly in terms of foregone exploitation, and when (ii) the uncertainty over actions, the RHS of equation (18), is larger, since that pushes up the lower bound of the constraint inducing more experimentation. This fundamental link with uncertainty will influence the optimal reasoning choice, as we discuss below.

Graphical illustration

Figure II illustrates both how beliefs get updated through reasoning as per the updating equation (14) and how the beliefs affect the chosen action distribution $\pi_t(c|s_t)$ as per equation (19). Panel (a) illustrates the update in the end of period beliefs $\widehat{Q}_t(c, s_t)$, where the value function is evaluated at the current state s_t and along the x-axis we have the feasible interval of c_t choices. The blue line plots the pre-reasoning estimate $\widehat{Q}^E_t(c, s_t)$, and the red and yellow lines show the updated beliefs $\widehat{Q}_t(c, s_t)$ for two different realizations of the reasoning signal η^R_t, resulting in two different shapes of the end-of-period beliefs.

Consider first the posterior described by the red line. In this case, the agent has updated his beliefs about the payoff of high consumption actions downward, and updated the beliefs about lower consumption actions upward, relative to the prior (blue line). We see the resulting optimal policy $\pi_t(c|s_t)$ in the red line of panel (b). It is useful to compare this action distribution to the action distribution, $\pi^E_t(c|s_t)$, that would have resulted if the reasoning signals had not updated the agent's beliefs at all (i.e. the signals corresponded identically to the prior) – this is plotted with the blue line. Thus, we see that updating the beliefs of $\widehat{Q}_t(c, s_t)$ to the red line has the following two main effects: (i) it shifts the action distributions towards lower consumption choices, and (ii) it makes the distribution more peaked around those lower consumption actions. This is very intuitive given panel (a), since the updated beliefs $\widehat{Q}_t(c, s_t)$ are both (i) lower for higher values of c and (ii) more peaked, thus resulting in a higher perceived loss of deviating from the action with the highest expected payoff.

The yellow lines in Figure II indicate an update based on a different signal realization.

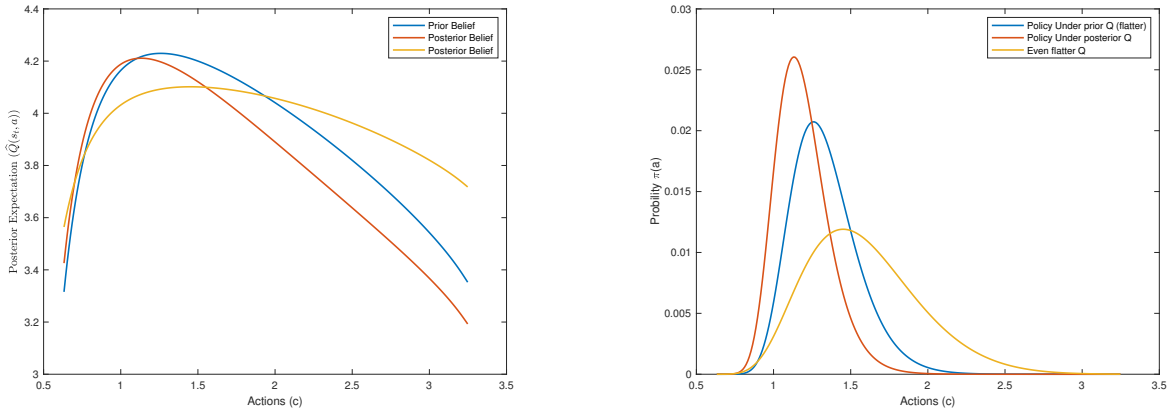


Figure II: Beliefs updated by the reasoning signal (panel a) and corresponding policy distributions (panel b).

Compared to pre-reasoning beliefs, this signal decreases the estimate of $Q^*(c, s_t)$ for lower consumption actions, while increasing it for higher actions. As a result, the update is less peaked, and rather flatter, than the prior belief $\widehat{Q}_t^E(c, s_t)$. The effect of this update in beliefs on the policy distribution in panel (b) is intuitive. Compared to the policy $\pi_t^E(c|s_t)$, this alternative reasoning signal (i) shifts the action distributions towards higher consumption choices, and (ii) makes the distribution overall flatter. Thus, in contrast to the distribution induced by the update in the red line, the yellow line updated policy is more dispersed, leading to more effective randomization, as the reasoning signal informs the agent that a wider set of consumption choices is estimated to have similar values.

2.4.3 Optimal reasoning

Given this optimal action strategy, given a realization of the reasoning signal η_t^R, the agent solves for the optimal reasoning strategy maximizing

$$\max_{\Sigma_{\eta,t},\Omega_t} \underbrace{\mathbb{E}\left(\sum_c \widehat{Q}_t(c, s_t)\pi_t(c|s_t)\middle|\mathcal{I}_t^E\right)}_{\text{exploitation benefit}} - \underbrace{\kappa \ln\left[\frac{|\Sigma_t^E(c, s_t, c', s_t)|}{|\Sigma_t(c, s_t, c', s_t)|}\right]}_{\text{reasoning cost}} \tag{20}$$

where we model the cognitive cost of reasoning as proportional to the information flow, measured in entropy units, in the reasoning signals η_t^R.

Thus, when considering how much ($\Sigma_{\eta,t}$) and how (Ω_t) to reason, the agent is facing a *cost-benefit tradeoff*. The *cost* of reasoning is cognitive, given by the second term in the objective (20). This cost is increasing in the information content of the chosen reasoning signals η_t^R, capturing that increasing the intensity of reasoning (e.g. mentally simulating more paths forward) requires higher cognitive effort. It is in particular quantified as the

reduction in entropy from pre to post-reasoning achieved by acquiring the signals η_t^R (as in information theory, e.g., Sims (2003b)).

Like in the cognitive control literature (e.g. Kool et al. (2017), Botvinick et al. (2004)), the constant marginal cost $\kappa > 0$ can be interpreted as the opportunity cost of cognitive capacity, which can be otherwise employed on other, outside-the-model tasks. Thus, κ will be higher if individuals have a higher opportunity cost of cognitive capacity or if their particular deliberation process behind η_t^R takes longer to produce a given amount of information. In addition, κ will also be higher if the environment is more complex and it is thus objectively harder to come up with insights on the unknown optimal Q-function (see eg. Oprea (2020) for measurement of complexity). The latter interpretation connects to existing models and evidence on how computational complexity increases the effective cognitive noise (eg. Enke and Graeber (2023)).

The *benefit* of reasoning is that the eventual information it provides reduces uncertainty, which relaxes the minimum entropy constraint (18), and allow the agent to select actions with higher expected payoffs as he becomes more certain about the values of $Q^*(c, s_t)$. This minimizes the deviations from the action with the highest expected payoff, and thus increased the expected gain in the current period. Again, this minimum entropy constraint on the actions captures the fact that, as in typical bandit problems, the more uncertain is the action value function $Q^*(c, s)$, the more there is a desire to explore the function in other parts of the state space. In its extreme, if $Q^*(c, s_t)$ was known, then there is no point to experimenting as there is nothing further to learn. So the more the agent reasons, and the further he learns about $Q^*(c, s_t)$, the less he needs to worry about the remaining uncertainty and experimenting away from the action with the highest estimated payoff.

Consider now the optimal reasoning choice, given this objective. Intuitively, the marginal cognitive cost will be set to equate the marginal benefit of reduction in uncertainty. The latter is given in the objective by the value of relaxing the minimum entropy constraint in the choice of actions, i.e. the ex-post Lagrange Multiplier δ_t. To simplify things, we will make a certainty equivalence assumption, and assume that the agent chooses her optimal reasoning under the mean realization of the reasoning signals, so that we avoid an extra expectations step and set $\delta_t = \delta_t^E$ (the Lagrange Multiplier of the optimal actions choice, given that the conditional expectation remains equal to $\widehat{Q}_t^E(c, s_t)$).

Formally, the optimal reasoning choice consists of choosing the loading matrix Ω_t and the signal noise variance matrix $\Sigma_{\eta,t}$. To solve this problem, we use insights from rate-distortion theory. In particular, let $\lambda_t^{(i)}$ denote the eigenvalues of posterior variance $\Sigma_t(c, s_t, c', s_t)$. Using

the standard properties for eigenvalues that

$$|\ln(\Sigma_t(c, s_t, c', s_t)|) = \sum_i \ln(\lambda_t^{(i)}), \quad \text{and} \quad \sum_c \sigma_t^2(c, s_t) = \sum_i \lambda_t^{(i)} \tag{21}$$

optimal reasoning can be expressed as choosing the eigenvalues $\lambda_t^{(i)}$ optimally.

We sort the eigenvalues in descending order so $\lambda_t^{(1)} \geq \lambda_t^{(2)} \geq ... \geq \lambda_t^{(N)}$. By standard properties of variance matrices $\Sigma_{\eta,t}$ is positive definite, and thus the (sorted) eigenvalues of the posterior variance $\Sigma_t(c, s_t, c', s_t)$ must be weakly smaller than those of the (pre-reasoning) one $\Sigma_t^E(c, s_t, c', s_t)$, denoted by $\lambda_{E,t}^{(i)}$. Hence, $\lambda_t^{(i)} \leq \lambda_{E,t}^{(i)}$. Optimizing the objective function in (20) over $\lambda_t^{(i)}$ implies

$$\lambda_t^{(i)} = \min\{\frac{\kappa}{h\delta_t}, \lambda_{E,t}^{(i)}\} \tag{22}$$

Hence, the agent has an optimal target level for uncertainty and wants to only reduce eigenvalues that before reasoning are bigger than $\frac{\kappa}{h\delta_t^E}$ – a property known as reverse 'water filling' in rate-distortion theory (see Cover and Thomas (1999), chapter 13). On the one hand, this means that there could be situations in which the agent chooses to not reason and acquire any further information – those would be situations (i.e. states of the world s_t) where uncertainty is already low enough to not justify any further cognition cost (i.e. $\lambda_{E,t}(i)$ is already lower than $\frac{\kappa}{h\delta_t}$).

On the other hand, equations (22) and (19), uncover a tight intuitive *interaction* between reasoning intensity and actions distribution. For example, take states s_t where in the absence of reasoning, the multiplier δ_t would be high, reflecting the high opportunity cost of experimentation. The agent has a lever to reduce that cost, namely to reason more intense. Indeed, she knows that more intense reasoning leads to information that deterministically (through Bayesian updating) lowers the conditional uncertainty, which in turn, holding everything else constant, lowers the incentive to randomize and thus the expected loss from experimentation. Indeed, by (22), when this loss is larger, the agent optimally employs more cognition to further reduce uncertainty. She values that reduction in uncertainty precisely because it allows her to select an action closer to the perceived greedy policy, not having to worry about experimentation.

Lastly, having characterized the reasoning solution in terms of the optimal choice of eigenvalues of the resulting posterior variance matrix, we can immediately derive that the optimal signal structure is for Ω_t to be the matrix of eigenvectors of $\widehat{\Sigma}_t^E(c, s_t, c', s_t)$, and $\Sigma_{\eta,t}$

be a diagonal matrix with entries[14]

$$\sigma_{i,t}^2 = \begin{cases} \frac{\kappa \lambda_{E,t}^{(i)}}{\lambda_{E,t}^{(i)} h \delta_t^E - \kappa} & \text{when} \quad \lambda_{E,t}^{(i)} > \frac{\kappa}{h\delta_t} \\ \infty & \text{when} \quad \lambda_{E,t}^{(i)} \leq \frac{\kappa}{h\delta_t} \end{cases} \tag{23}$$

Moreover the update from $\lambda_{E,t}^{(i)}$ to $\lambda_t^{(i)}$ is straightforward because it is an update over orthogonal normal variables with a vector of orthogonal signals. Finally, we note that the reasoning intensity and multiplier δ_t are jointly determined, since the latter reflects the intensity of the entropy constraint at the new reasoning choice.

Graphical illustrations

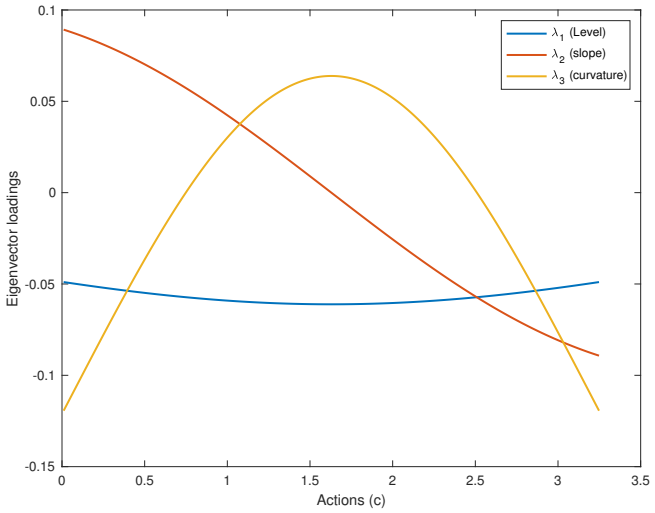


Figure III: Reasoning choice: Eigenvectors of Variance matrix

Expressing the reasoning problem in terms of the eigenvalues and eigenvectors of the variance matrix $\widehat{\Sigma}_t^E$ is very convenient analytically, but it also turns out to be quite intuitive since it is related to familiar properties of Principal Component Analysis (PCA). To illustrate this intuition, in Figure III we plot the first three columns of the optimally chosen Ω_t in our running example of a consumption savings problem.

These loadings are of course just the eigenvectors of the variance-covariance matrix $\widehat{\Sigma}_t^E$, and as such they are intuitively related to the shape, slope and curvature of the underlying

[14]In practice, the eigenvalues are likely to decline in value quickly, so only a few of these signals will typically have non-infinite noise variance making the solution very tractable and easy to implement.

unknown function $Q^*(c, s)$. As is familiar from PCA, these three components account for the bulk of the uncertainty over the non-parametric shape of $Q^*(c, s_t)$, and in this particular illustrative example we find that the agent indeed optimally chooses to obtain three reasoning signals, with structure given by these three vectors. So intuitively, in the next few figures of this illustration, as reasoning signals realize the agent is updating beliefs of the level, slope and curvature of Q^*.

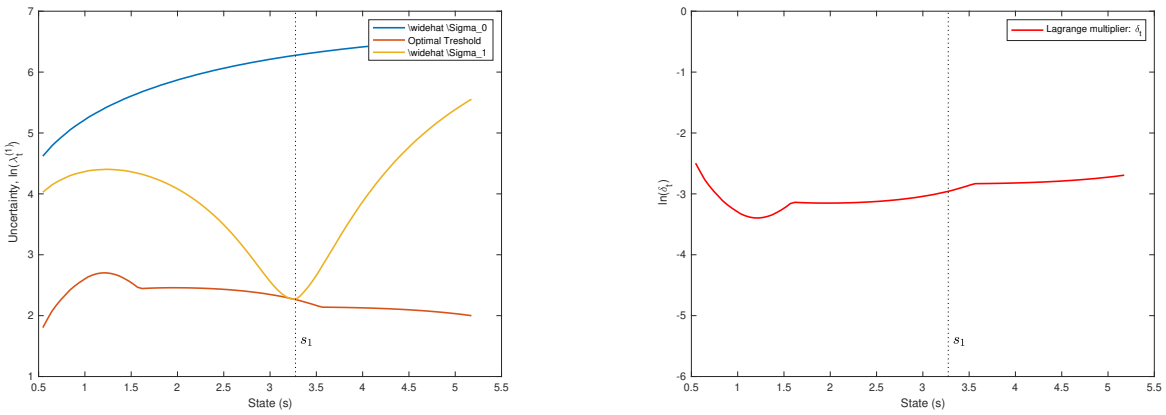


Figure IV: Subjective Uncertainty, pre and post reasoning (panel a) and Lagrange Multiplier δ_t^E (panel b)

.

First, in Figures IV we illustrate the optimal reasoning strategy of equation (22). Panel (a) plots in the blue line the pre-reasoning uncertainty over $Q^*(c, s)$ at various states s_t, as measured here by the (log of) highest eigenvalue $\lambda_{E,t}^{(1)}$. The first interesting thing to observe is that this measure of subjective uncertainty is increasing in the value of the state, cash on hand s_t (x-axis). The reason is that in this consumption-savings environment, at low values of the state the set of feasible consumption actions is smaller, being restricted by the constraint of no debt we have $c_t \leq s_t$. Thus, at low state values there are fewer consumption actions available for the agent to consider (and they are all more similar to each other), and this translates in the overall uncertainty the agent faces over the function $Q^*(c, s_t)$ as being lower.

The yellow line plots the optimally chosen posterior uncertainty, measured as the (log of) $\lambda_t^{(1)}$. The red line depicts the optimal threshold $\frac{\kappa}{h\delta_t^E}$ the agent wants to achieve through reasoning. As per the logical of this optimal uncertainty target, the agent's posterior uncertainty after reasoning (yellow line) is equal to the threshold at the current state of the world s_t.

Importantly, the reasoning signal obtained at that state is also informative about the function $Q^*(c,s)$ at other states, but this information is decreasing in the distance $|s - s_t|$ (as per the key property of the underlying Gaussian Process distribution). This creates a characteristic *local reduction in uncertainty*, where the posterior uncertainty increases (locally) away from s_t where the agent's reasoning this period was centered at. The resulting of the posterior uncertainty (yellow line) combines the local informativeness of the reasoning signal and the shape of the prior uncertainty, and thus it is locally increasing ass it moves away from s_t, but is also low for low values of cash-on-hand.

The reasoning threshold in the red line of Figure IV is simply inversely related to the endogenously-determined multiplier δ_t, plotted in panel (b). A stark visual property of δ_t is that it is decreasing in the state for low values of the state s_t. This reflects a significantly larger utility loss from experimentation at low states. The main reason has to do with the typical shape of beliefs $\widehat{Q}_t(c,s)$, which are usually strongly peaked around low values of wealth as they inherit that property from Q^*(on average). In such situations, any deviations from the optimal are quite costly and thus at low s_t the opportunity cost of experimentation and randomization across actions are large.

2.4.4 Summary

Put together, the recursive nature of the agent's dynamic evolution of beliefs and actions can be summarized as

1. Priors: the agent enters period t with a GP prior over the function $Q^*(c,s)$ summarized in equation (4) by the prior mean function $\widehat{Q}_{t-1}(c,s)$ and co-variance function $\widehat{\Sigma}_{t-1}(c,s,c',s')$.

2. State evolution: the current state s_t is realized, following the law of motion $\mathcal{F}(c_{t-1}, s_{t-1}, \epsilon_t)$.

3. Agent's observables: the agent observes the current s_t, the previous action-state (c_{t-1}, s_{t-1}) pair and the utility realized at that pair.

4. Experience: the utility $u(c_{t-1}, s_{t-1})$ serves as her experience-based signal η_t^E of equation (8), which updates prior beliefs in equation (11).

5. Reasoning: Conditional on these updates, she reasons to reduce uncertainty by following the strategy in equation (22).

6. Posteriors: The particular realized reasoning signal η_t^R of equation (13) updates the mean $\widehat{Q}_t(c,s)$ and co-variance function $\widehat{\Sigma}_t(c,s,c',s')$ that form the end-of-period t beliefs in equation (17). This posterior becomes the prior for period $t+1$.

7. Actions and utility: The time-t action c_t is drawn from the soft-max policy distribution in equation (19) which uses posterior beliefs. Utility $u(c_t, s_t)$ is determined.

3 Learning dynamics

Next, we consider how learning dynamics over time. First, in Figure V we show how the uncertainty facing the agent evolves as the agent moves into the second period, $t = 2$. For this illustration, we assume there was a negative income shock $\epsilon_2 < 0$, and hence we are in a situation where $s_2 < s_1$. With the blue line, we plot first the posterior uncertainty at the end of period $t = 1$, $\widehat{\Sigma}_1$. As the agent enters period $t = 2$, she observes the realized experience signal η_2^E and this updates her beliefs and lowers her uncertainty down to $\widehat{\Sigma}_2^E$, which is plotted with the yellow line. The optimal reasoning choice follows the now familiar threshold strategy characterized by equation (22), so to illustrate that with the red line we plot the resulting optimal threshold $\frac{\kappa}{h\delta_2}$.

There is a new result illustrated in this Figure, in that the optimal threshold is in fact now above $\widehat{\Sigma}_2^E$, the agent's beginning of period uncertainty, for a range of values around the current state s_2. Thus, there is not a neighborhood in the state space (including the current state s_2), where the agent's uncertainty is already lower than the target level, and hence she optimally chooses to not reason any further but condition her actions on existing beliefs. Intuitively, this situation arises naturally as the agent accumulates information. The accumulation of both experience and reasoning signals will eventually lower the agent's uncertainty below the optimal threshold in any parts of the state space the agent visits often, and hence the agent will stop reasoning when faced with these "familiar" situations.

3.1 Learning Traps

Moreover, if the agent is repeatedly visiting the same part of the state space, not only will the agent stop reason, but the experience signal η_t^E will also provide limited new information. The structure of the experience signal, equation (2.3.1), is such that it provides new information only to the extent to which the agent keep sampling new combinations of action-state pairs (c, s). If the agent's state starts fluctuating in a given interval of the state space, as the uncertainty of the agent naturally falls in this familiar part of the state space, the optimal policy function will also imply actions c that become less randomized. If this persists, eventually the agent will be revisiting the same action-state pairs (c, s) and the experience signal will not contain any new information, and formally we will have $\alpha_t^E(c, s) \to 0$.

In such situations the evolution of beliefs $\widehat{Q}_t(c, s)$ will stop (or at least significantly slow

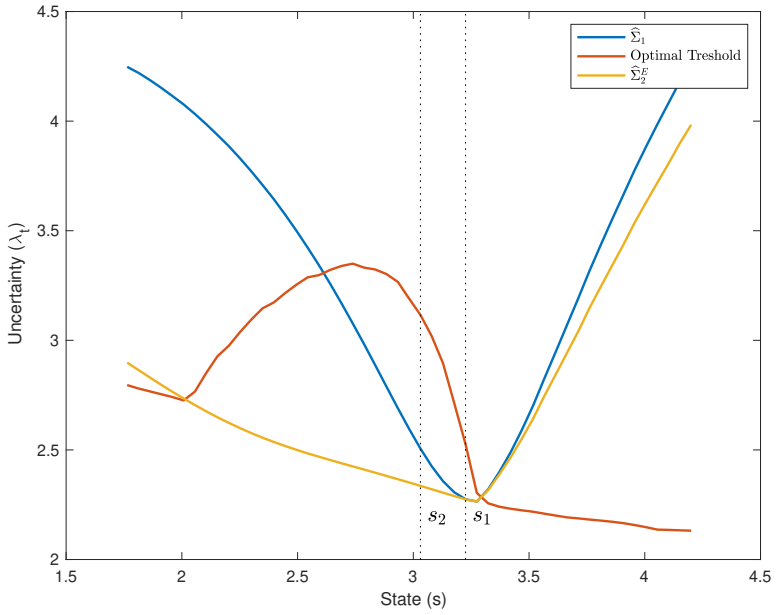


Figure V: A no reasoning situation

down), as the agent is not receiving new information from either reasoning or experiences. If the state evolution is exogenous, then such situations are not going to be systematically related to the agent's beliefs $\widehat{Q}_t(c,s)$, and thus while learning stops and the agent's beliefs never fully converge to $Q^*(c,s)$, the resulting beliefs will be on average correct.

However, when the state is *endogenous*, then there is a feedback between the noise in the estimated $\widehat{Q}_t(c,s)$ and action policy $\pi_t(c|s_t)$, and the dynamics of the state variable. For example, in the consumption-savings illustration if the agent mistakenly beliefs that high consumption actions are best, then this will tend to push cash on hand s_t down over time, and vice versa.

This feedback between beliefs and the state evolution, could lead to situations that we term "learning traps". This are situations where the established beliefs of the agent $\widehat{Q}_t(c,s)$ imply actions $\pi_t(c|s)$ that keep the dynamics of the state s_t stable around some endogenous point $\bar{s}$. As the state repeatedly cycles in the same region of the state space, the agent's beliefs stabilizes and stop updating as per the logic explained above. Only certain types of beliefs, however, are consistent with stable s_t dynamics, and thus the model predicts that the agent's learning dynamics will be asymptotically drawn to such points. The characteristics of these stable long-run beliefs depend on the broader economic framework the agent acts in. A similar phenomenon also arises in the framework of Ilut and Valchev (2023) where there is no learning from experience, but here we show that the learning traps in fact persist even when the agent receives a new experience signals each period.

For example, in the consumption-savings illustration we observe that for any given choice of consumption today c_t, the state evolution can be expressed as follows

$$\mathbb{E}_t(s_{t+1}) = s_t + (1+r)(c^{PIH}(s_t) - c_t), \tag{24}$$

where $c^{PIH}(s_t) = \frac{r}{1+r}s_t + \frac{w}{1+r}$ is the policy function under the permanent income hypothesis. This policy function has the special feature that acting under $c^{PIH}(s)$ makes the state a random walk – i.e. the expected future state is the same as today's state when consuming $\frac{r}{1+r}s$ fraction of the state each period.

Thus, we can observe that two types of action policies will keep the state s_t stable. First, if the agent's implied behavior under $\pi_t(c|s)$ is such that there is a point $\bar{s}$ around which the agent (on average, given the noisy nature of $\pi_t(c|s)$) consumes more than under the PIH policy for $s_t > \bar{s}$, and less than the PIH policy for $s_t < \bar{s}$, then the state dynamics become locally stable around $\bar{s}$. We can see this directly from the state evolution equation (24) above – when $c_t > c^{PIH}(s_t)$ for $s_t > \bar{s}$, then the state is on average trending down from s_t. And vice versa for $s_t < \bar{s}$. Hence, in a neighborhood around this point $\bar{s}$ the state dynamics will be mean reverting towards $\bar{s}$, thus the agent is likely to repeatedly experience the same situations and stop learning. We observe that the defining feature of this learning trap is that around the stable wealth point $\bar{s}$, the action policy of the agent is *steeper* than the PIH consumption policy. That is, for s close to $\bar{s}$

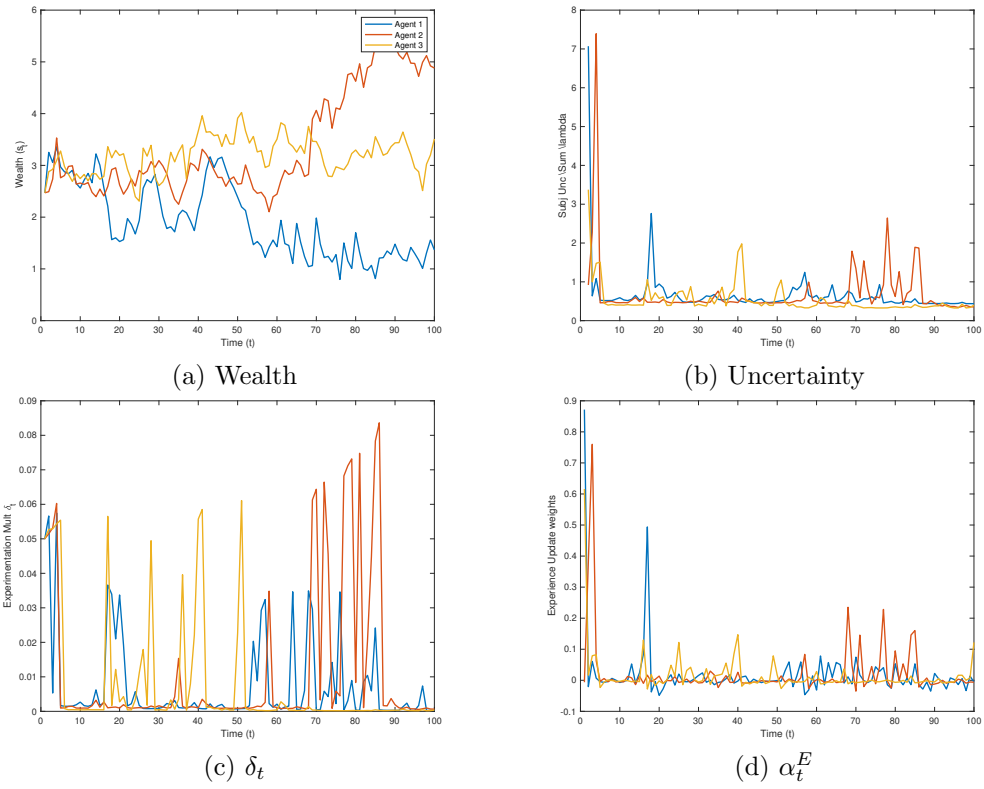
$$\frac{\partial \mathbb{E}(c_t)}{\partial s} > \frac{\partial c^{PIH}(s)}{\partial s}$$

Lastly, the stable points $\bar{s}$ are endogenous to the beliefs of the agent, and due to the reasoning noise $\varepsilon_{\eta,t}$ they will be heterogeneous and disperses across agents. However, because there is a borrowing constraint and thus a lower bound on wealth, the borrowing constraint itself could become a mass point in the distribution of stable $\bar{s}$. That it, there could be a large fraction of agents "stuck" with beliefs such that the borrowing constraint itself is the stable wealth point for all of those agents, giving rise to endogenous "Hannd-to-Mouth" agents.

Graphical illustrations

Let us illustrate these forces in our running consumption-savings example. In Figure VI we plot the simulated paths of four variables of interest for three different agents. In panel (a), we plot the evolution of the state (cash-on-hand). As a reference, we note that for the particular parameterization we use, the steady state wealth in the full information model is $\bar{s}^{FI} = 3.25$. In the case of our three bounded rational agents, we see that the paths of their

wealth fluctuate around in this simulation, but eventually stabilize around three different levels $\bar{s}$. The blue path showcases an example of an agent whose wealth converges to and stabilizes around the borrowing constraint. At the borrowing constraint the agent brings no assets into the period, hence cash-on-hand is equal to the income flow for that period $s_t = w\epsilon_t$. The average labor income is 1.15 and this is what s_t stabilizes around for this agent. On the other hand, the yellow line showcases an agent whose wealth stabilizes close to the full information steady state $\bar{s}^{FI}$, and the red line showcases an agent whose wealth stabilizes around a significantly higher point $\bar{s} \approx 5$. Thus, this panel showcases how we can get persistent differences in the wealth levels of agents, and thus generate persistent wealth inequality.



(a) Wealth (b) Uncertainty

(c) δ_t (d) α_t^E

Figure VI: Learning Dynamics over Time

In panel (b), we plot the evolution of the subjective uncertainty over $Q^*(c, s)$ as perceived by these three different agents. As a sufficient statistic for $\widehat{\Sigma}_t$, we plot the sum of its diagonal entries which matches the notion of uncertainty that enters the RHS of the entropy regularization constraint (18). As we can see, this subjective uncertainty is high for all three agents in the first few periods and then on average drops off significantly. By comparing panels (a) and (b), we can see that the spikes in uncertainty are associated with shifts in the state variable – the agents become more uncertain if they move into a new, unfamiliar part

of the state space. We can see this for example with the spike up in the blue line in panel (b) around period 18, which corresponds with the first big drop in wealth in period 18 for the blue agent in panel (a). Similarly, the red line in panel (b) shows an elevated level of uncertainty between periods 70 to 90, which matches the time when the wealth of the red agent starts experiences a significant trend up before stabilizing in panel (a).

Naturally, these shifts in uncertainty also affect the incentive for the agent to act more stochastically and experiment away from the action with the current highest estimated payoff. We can see this illustrated in the time paths of δ_t, which we show in panel (c). The periods of elevated uncertainty in panel (b) correspond to periods of high δ_t in panel (c), which implies agents act under more dispersed action distributions $\pi_t(c|s)$. And the key observation here is that in the last 20 or so periods of this simulation, for all agents δ_t drops off to approximately zero, as uncertainty in panel (b) also stabilizes at a low level, and the agents settle around their stable $\bar{s}$ points in panel (a).

Since in these last few periods $\delta_t \approx 0$, we can intuitively see that the optimal reasoning threshold $\frac{\kappa}{h\delta_t}$ is high and reasoning optimally stops. Still, the agents receive experience signals each period and this could potentially change their stable beliefs. However, as conjectured above and as illustrated in panel (d), once agents start repeating the same action-state pairs (c_t, s_t), the experience signals contain no new information and hence learning from experience stops as well – this can be seen from $\alpha_t^E(c, s_t) \approx 0$ in the last 20 periods in panel (d).

It is instructive to also consider the consumption policy functions that each agent perceives as optimal at the end of this simulation. Since $\delta_t \approx 0$, the softmax policy function essentially converges to the greedy policy $\widehat{c}_t(s_t) = \arg\max_c \widehat{Q}(c, s_t)$, so for simplicity we just plot these univariate policy functions in Figure VII. We use the same colors for the lines as in the simulated paths in Figure VI. The blue line shows the long-run policy for the agent that stabilized wealth around the borrowing constraint, and the long-run policy can help us understand why – as we can see this agent's estimated policy is significantly above the $c^{PIH}(s)$ policy (green dash-dot-line) and thus pushes up against the borrowing constraint (the dashed line).

The stable wealth points implied by these long-run policy functions are given by their intersections with $c^{PIH}(s)$, and we can indeed see that the positions of these intersections correspond to the stable wealth levels in panel (a) of Figure VI – at the borrowing constraint for the blue line, close to the full information stable point for the yellow line, and higher than that for the agent with the red line. Also, all of these policies have the characteristic feature that they are *steeper* than the $c^{PIH}(s)$. This is particularly interesting for the two agents whose wealth is significantly away from the borrowing constraint – the yellow and red agents. These agents both have significant accumulation of liquid wealth assets, and at the

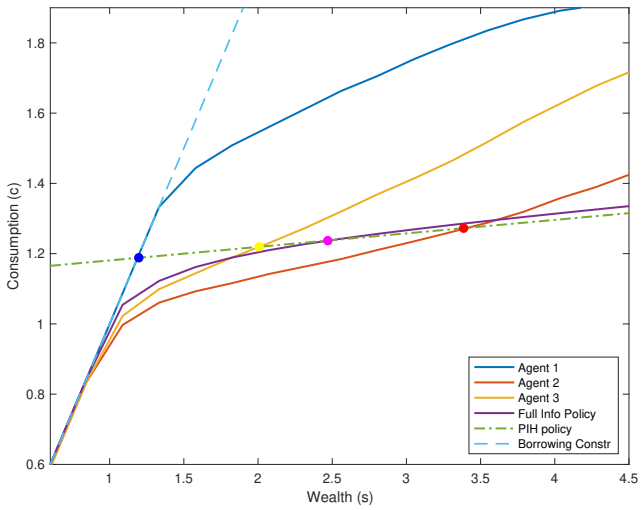


Figure VII: Long run policy functions

same time display MPCs that are much higher than what would be implied under the full information policy function. We explore this further in the next section.

4 Numerical Results

In this section, we quantify some of the characteristic implications of the bounded rational framework, as it pertains to a consumption-savings problem, that we have touched upon qualitatively earlier. For this section, we calibrate the key preference parameters of the model $(\kappa, h, \psi_s = \psi_c)$ to target two important moments in the resulting ergodic distribution – (i) the fraction of Hand-to-Mouth agents and (ii) the average MPC of wealthy agents – and also to target a first-period $\delta_1 = 0.05$ in accordance with the typical calibration of the temperature parameter in RL studies. This is a preliminary calibration, meant to mainly illustrate the model at this stage. Still, the target moments were chosen in a way that disciplines the framework. For example, the fraction of Hand-to-Mouth agents and the avg MPC of wealth agents both speak to the characteristic learning traps that emerge, and we calibrate the model to make sure we are consistent with the empirical evidence on those moments.

We calibrate the other parameters to standard parameters – i.e. $\beta = 0.96, r = 0.04$ (so one period is a year) and we use the income process parameters from Aiyagari (1994). To compute the ergodic distribution, we simulate a panel of 1000 agents for 1000 periods. We also assume that agents die with iid probability 0.02 (expected life of 50 years). Upon death, agents transfer all of their remaining wealth to a newly born agent who starts the learning problem from scratch – there is no transfer of information and/or experience across generations. We then discard the first 500 periods, and compute moments on the second part

of the sample.

Table I: Numerical Results

Moments	Data	BR Model	Full Info
Fraction of Hand-to-Mouth	20-40%	30%	0.4%
Average (aggregate) MPC	0.2-0.6	0.29	0.05
Average MPC | top 20% of assets	0.1-0.3	0.27	0.04
Average MPC | *HtM*		0.45	0.22

Some moments of interest of the resulting ergodic distribution are reported in Table I below. In column (1) we list the empirical value of these moments (when available), in column (2) we report the moments of our bounded rationality model and in column (3) we compare with the full information version (i.e. fully ration) of the model.

In the first line, we see that the model successfully generates a large fraction of Hand-to-Mouth agents. We define Hand-to-Mouth agents as those who have assets less than 2 months worth of income (in our yearly calibration, this means assets less than 1/6 of the mean per period income). In our calibration, 30% of the agents are designated as Hand-to-Mouth which is an interesting juxtaposition with the full information model, where for the same income process only 0.4% of the agents are Hand-to-Mouth.

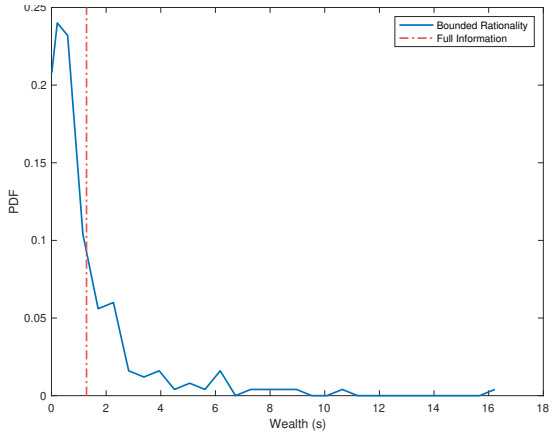
The reason is that under full information agents have strong incentives to avoid the borrowing constraint and engage in significant amount of precautionary savings. As a result, only as mall fraction of particularly unlucky agents end up with such low level of assets. In our model, however, there is a significant number of agents whose beliefs over $Q^*(c, s)$ result in a stable $\bar{s} = w$, which implies a steady state position of zero assets $\bar{a} = 0$. Thus, the model is able to generate a significant mass of *persistently* Hand-to-Mouth agents, due to its characteristic interaction between the errors in the estimation of $Q^*(c, s)$ and the endogenous dynamics of the state s_t.[15]

The second moment targeted by our calibration is the MPC of the unconstrained agents. In particular, we calibrate the model to imply a 27% MPC among the agents at the top of the wealth distribution – those in the top quartile of assets. This moment is reported in the third row of Table I. It is again interesting to contrast our bounded rationality model, which can obtain large MPCs for eve these rich agents (in line with the empirical evidence), while

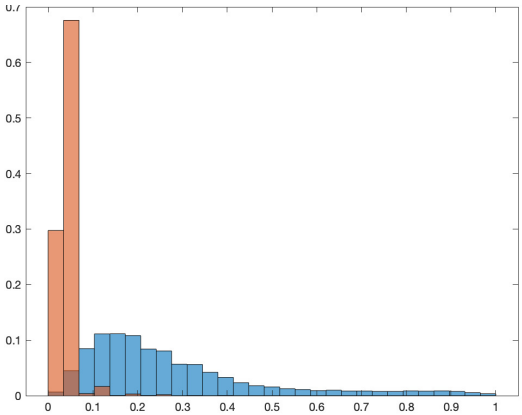
[15]A similar phenomenon also occurs in Ilut and Valchev (2023), but in that case the only source of learning are the reasoning signals. That paper does not consider the possibility of learning from experiences.

Figure VIII: Wealth and MPC Heterogeneity

(a) $\bar{s}$ dist



(b) MPC dist



the full information model (Column (3)) implies an average MPC of only 4% for these agents. The reason is again the characteristic feature of the long-run policy function that emerges in our model – it is steeper than $c^{PIH}(s)$, as beliefs only stop evolving once s_t stabilizes in some region.

As a result of both of these fundamental features of the model – a large fraction of Hand-to-Mouth agents and large MPC for agents in the right tail of the asset distribution – the model implies a sizeable aggregate MPC of 0.29. This is six times bigger than the full information model, which can only deliver a tiny aggregate MPC of 0.05.

4.1 Wealth and MPC Heterogeneity

Now let us discuss some untargeted moments. As mentioned a few times already, the agents in our model establish different stable wealth points $\bar{s}$, mainly due to the idiosyncratic noise in their reasoning signals. We alluded above that part of the reason the model can generate a robust number of Hand-to-Mouth agents is that many agents establish stable wealth points $\bar{s}$ that are at or close to the borrowing constraint.

In panel (a) of Figure VIII we quantify this point directly, by plotting the distribution of stable points $\bar{s}$ in the cross-section of agents in our simulation. To make the discussion a little bit more intuitive, we transform the $\bar{s}$ into the implied stable level of assets $\bar{a}$, by using the steady state budget relation

$$\bar{s} = \bar{a} + w\mathbb{E}()\epsilon)$$

First, we observe that there is indeed a large amount of heterogeneity in these stable wealth points. There is an obvious large mass close to zero – these are the persistent Hand-to-Mouth agents. But in addition to that, there is also a significant fat right tail in these stable wealth points $\bar{a}$, signifying that there is also a significant amount of *persistent wealth inequality*. This fat tail of persistently wealthy agents are essentially agents that are wealthy and remain wealthy because of actively saving sufficient amount to maintain their high wealth levels. This characteristic of the wealth distribution of our model is consistent with the empirical evidence of Halvorsen et. al. (2024) who show that the majority of wealthy households are wealthy indeed because of their savings rate. This is a puzzle for standard models, which would imply that wealthy agents should dissave, not maintain their high levels of wealth.

Second, in panel (b) we also plot the distribution of MPCs in the cross-section of agents. With this graph we want to showcase that it is not the case that all unconstrained agents in our model have the same MPC. Again due to the noise in the reasoning and experience signals different agents establish different beliefs, and that could lead to significant heterogeneity in the implies MPCs as the Figure showcases. This MPC heterogeneity is not be correlated with fundamental features of the agents such as income or wealth, but is rather driven by the noise in beliefs of $Q^*(c, s)$. This is consistent with the empirical evidence in Lewis et. al. (2023), which shows that a big fraction of the large variation in MPCs in the data remains unexplained even after controlling for a rich set of observable household characteristics.

4.2 Welfare Losses due to suboptimal behavior

It is interesting to consider the implied welfare loss of the mistakes implied by our bounded rationality mechanism. A classic argument by Cochrane (1989) is that bounded rationality mistakes only carry second order losses. This argument relies on approximating welfare losses as one-period deviations from optimal behavior, around the rational steady state.

Mapping this into our notation, this would give us the following formula for welfare loss, which we call ΔW^1, as it will be the first measure of welfare loss we consider

$$\Delta W_t^{(1)} = \frac{Q^*(c_t, s_t) - Q^*(c_t^*, s_t)}{Q^*(c_t^*, s_t)}$$

We compute this object for each period and each agent, as it gives us the implied value function loss in each period due to selecting the optimal action $c_t \neq c_t^*$. This gives us a time series for each agent, that we can then average across both agents and time. This gives us an average welfare loss of 0.25% of permanent consumption units.

This is arguably not an overly large welfare loss, hence one might think that the bounded

rationality friction is not having significant real effects. However, this is only a *partial* measure of the welfare loss in our framework precisely because the *steady state* behavior of our agents is different than that of rational agents. Intuitively, the above notion of welfare loss assumes that agents will deviate for one period and then be optimal from the next period onward.

We propose then the following welfare loss measure, in order to also take into account the fact that the long-run behavior of the agents will potentially be different from optimal due to the learning traps feature of our framework

$$\Delta W^{(2)} = \frac{\widehat{V}(s_t) - V^*(s_t)}{V^*(s_t)}$$

We compute this object at the end of the simulation ($t = 1000$) for each agent, and it effectively compares the full information Value Function $V^*(s_t)$ at the relevant state of the world today, to the present discounted value of following the long-run policy function that has emerged at the end of the simulation. The key difference between these two measures of welfare loss is that $\Delta W^{(2)}$ takes into account the fact that, the boundedly rational agent does not deviate from optimality in just one period, but across the whole path of his actions going forward, resulting in a different ending steady state.

Computing the average $\Delta W^{(2)}$ for all agents in our simulation we obtain a welfare loss of approximately 5% of permanent consumption. Thus, taking into account that our bounded rationality friction changes behavior not only in the short-run, but also in the long-run, results in welfare losses that are an order of magnitude bigger than otherwise.

References

Ackley, David H, Geoffrey E Hinton, and Terrence J Sejnowski, "A learning algorithm for Boltzmann machines," *Cognitive science*, 1985, *9* (1), 147–169.

Aiyagari, S Rao, "Uninsured idiosyncratic risk and aggregate saving," *The Quarterly Journal of Economics*, 1994, *109* (3), 659–684.

Alaoui, Larbi and Antonio Penta, "Cost-benefit analysis in reasoning," *Journal of Political Economy*, 2022, *130* (4), 881–925.

Andre, Peter, Joel P Flynn, Georgios Nikolakoudis, and Karthik Sastry, "Quick-fixing: Near-rationality in consumption and savings behavior," Technical Report, NBER Working Paper 33464 2025.

Aragones, Enriqueta, Itzhak Gilboa, Andrew Postlewaite, and David Schmeidler, "Fact-Free Learning," *The American Economic Review*, 2005, *95*, pp–1355.

Barber, David, *Bayesian reasoning and machine learning*, Cambridge University Press, 2012.

Barberis, Nicholas C and Lawrence J Jin, "Model-free and model-based learning as joint drivers of investor behavior," 2023. NBER Working Paper 31081.

Bardhi, Arjada, "Optimal discovery and influence through selective sampling," 2022. Duke University, mimeo.

Bertsekas, Dimitri P, *Reinforcement learning and optimal control*, Athena Scientific Belmont, MA, 2019.

Bianchi, Francesco, Cosmin L Ilut, and Hikaru Saijo, "Diagnostic Business Cycles," 2021. NBER WP 28604.

Botvinick, Matthew M, Jonathan D Cohen, and Cameron S Carter, "Conflict monitoring and anterior cingulate cortex: an update," *Trends in cognitive sciences*, 2004, *8* (12), 539–546.

Callander, Steven, "Searching and learning by trial and error," *American Economic Review*, 2011, *101* (6), 2277–2308.

Caplin, Andrew, *An introduction to cognitive economics: The science of mistakes*, Springer Nature, 2025.

Chater, Nick, Joshua B Tenenbaum, and Alan Yuille, "Probabilistic models of cognition: Conceptual foundations," *Trends in cognitive sciences*, 2006, *10* (7), 287–291.

Cochrane, John H, "The Sensitivity of Tests of the Intertemporal Allocation of Consumption to Near-rational Alternatives," *The American Economic Review*, 1989, *79* (3), 319.

Cover, Thomas M and Joy Thomas, *Elements of information theory*, John Wiley & Sons, 1999.

Daw, Nathaniel D, Yael Niv, and Peter Dayan, "Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control," *Nature neuroscience*, 2005, *8* (12), 1704–1711.

Dearden, Richard, Nir Friedman, and Stuart Russell, "Bayesian Q-learning," *Aaai/iaai*, 1998, *1998*, 761–768.

Dew-Becker, Ian and Charles G Nathanson, "Directed attention and nonparametric learning," *Journal of Economic Theory*, 2019, *181*, 461–496.

Engel, Yaakov, Shie Mannor, and Ron Meir, "Bayes meets Bellman: The Gaussian process approach to temporal difference learning," in "Proceedings of the 20th International Conference on Machine Learning (ICML-03)" 2003, pp. 154–161.

_ , _ , and _ , "Reinforcement learning with Gaussian processes," in "Proceedings of the 22nd international conference on Machine learning" 2005, pp. 201–208.

Enke, Benjamin, "The cognitive turn in behavioral economics," Technical Report, Working Paper 2024.

_ and Thomas Graeber, "Cognitive uncertainty," *The Quarterly Journal of Economics*, 2023, *138* (4), 2021–2067.

Eysenbach, Benjamin and Sergey Levine, "If MaxEnt RL is the answer, what is the question?," *arXiv preprint arXiv:1910.01913*, 2019.

Gabaix, Xavier, "A sparsity-based model of bounded rationality," *The Quarterly Journal of Economics*, 2014, *129* (4), 1661–1710.

_ , "Behavioral inattention," in "Handbook of Behavioral Economics: Applications and Foundations 1," Vol. 2, Elsevier, 2019, pp. 261–343.

Gershman, Samuel J and Nathaniel D Daw, "Reinforcement learning and episodic memory in humans and animals: an integrative framework," *Annual review of psychology*, 2017, *68*, 101–128.

_ , Eric J Horvitz, and Joshua B Tenenbaum, "Computational rationality: A converging paradigm for intelligence in brains, minds, and machines," *Science*, 2015, *349* (6245), 273–278.

Ghavamzadeh, Mohammad, Shie Mannor, Joelle Pineau, Aviv Tamar et al., "Bayesian reinforcement learning: A survey," *Foundations and Trends® in Machine Learning*, 2015, *8* (5-6), 359–483.

Griffiths, Thomas L, Charles Kemp, and Joshua B Tenenbaum, "Bayesian models of cognition.," in "Annual Meeting of the Cognitive Science Society, 2004; This chapter is based in part on tutorials given by the authors at the aforementioned conference as well as the one held in 2006." Cambridge University Press 2008.

_ , Falk Lieder, and Noah D Goodman, "Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic," *Topics in cognitive science*, 2015, *7* (2), 217–229.

_ , Nick Chater, Charles Kemp, Amy Perfors, and Joshua B Tenenbaum, "Probabilistic models of cognition: Exploring representations and inductive biases," *Trends in cognitive sciences*, 2010, *14* (8), 357–364.

Haarnoja, Tuomas, Haoran Tang, Pieter Abbeel, and Sergey Levine, "Reinforcement learning with deep energy-based policies," in "International conference on machine learning" PMLR 2017, pp. 1352–1361.

Ilut, Cosmin and Rosen Valchev, "Economic agents as imperfect problem solvers," *The Quarterly Journal of Economics*, 2023, *138* (1), 313–362.

_ , _ , and Nicolas Vincent, "Paralyzed by Fear: Rigid and Discrete Pricing under Demand Uncertainty," *Econometrica*, 2020, *88* (5), 1899–1938.

Judd, Kenneth L et al., "Numerical Methods in Economics," *MIT Press Books*, 1998, *1*.

Kaelbling, Leslie Pack, Michael L Littman, and Andrew W Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, 1996, *4*, 237–285.

Kirkpatrick, Scott, C Daniel Gelatt Jr, and Mario P Vecchi, "Optimization by simulated annealing," *science*, 1983, *220* (4598), 671–680.

Kool, Wouter, Samuel J Gershman, and Fiery A Cushman, "Cost-benefit arbitration between multiple reinforcement-learning systems," *Psychological science*, 2017, *28* (9), 1321–1333.

Laibson, D, P Maxted, and B Moll, "Present Bias Amplifies the Household Balance-Sheet Channels of Macroeconomic Policy," *Quarterly Journal of Economics*, Forthcoming.

Lee, Sang Wan, Shinsuke Shimojo, and John P O'Doherty, "Neural computations underlying arbitration between model-based and model-free learning," *Neuron*, 2014, *81* (3), 687–699.

Lewis, Daniel J, Davide Melcangi, and Laura Pilossoph, "Latent heterogeneity in the marginal propensity to consume," 2019. FRB of New York Staff Report No. 902.

Lian, Chen, "Mistakes in future consumption, high MPCs now," *American Economic Review: Insights*, 2023, *5* (4), 563–581.

Lin, Junyang, Xu Sun, Xuancheng Ren, Muyu Li, and Qi Su, "Learning when to concentrate or divert attention: Self-adaptive attention temperature for neural machine translation," *arXiv preprint arXiv:1808.07374*, 2018.

Liu, Weifeng, Jose C Principe, and Simon Haykin, *Kernel adaptive filtering: a comprehensive introduction*, Vol. 57, John Wiley & Sons, 2011.

Mackowiak, Bartosz, Filip Matejka, and Mirko Wiederholt, "Rational inattention: A review," *Journal of Economic Literature*, 2021. forthcoming.

Malmendier, Ulrike, "Experience effects in finance: Foundations, applications, and future directions," *Review of Finance*, 2021, *25* (5), 1339–1363.

_ **and Stefan Nagel**, "Learning from inflation experiences," *The Quarterly Journal of Economics*, 2016, *131* (1), 53–87.

Mnih, Volodymyr, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in "International conference on machine learning" PMLR 2016, pp. 1928–1937.

Moerland, Thomas M, Joost Broekens, Aske Plaat, Catholijn M Jonker et al., "Model-based reinforcement learning: A survey," *Foundations and Trends® in Machine Learning*, 2023, *16* (1), 1–118.

Oprea, Ryan, "What makes a rule complex?," *American Economic Review*, 2020, *110* (12), 3913–3951.

Rasmussen, Carl Edward and Christopher KI Williams, *Gaussian processes for machine learning*, Vol. 1, MIT press Cambridge, 2006.

Schulz, Eric, Emmanouil Konstantinidis, and Maarten Speekenbrink, "Putting bandits into context: How function learning supports decision making.," *Journal of experimental psychology: learning, memory, and cognition*, 2018, *44* (6), 927.

Shenhav, Amitai, Sebastian Musslick, Falk Lieder, Wouter Kool, Thomas L Griffiths, Jonathan D Cohen, and Matthew M Botvinick, "Toward a rational and mechanistic account of mental effort," *Annual review of neuroscience*, 2017, *40*, 99–124.

Sims, Christopher A, "Implications of rational inattention," *Journal of Monetary Economics*, 2003, *50* (3), 665–690.

_ , "Implications of rational inattention," *Journal of Monetary Economics*, 2003, *50* (3), 665–690.

Stevens, Luminita, "Coarse pricing policies," *The Review of Economic Studies*, 2020, *87* (1), 420–453.

Sutton, Richard S, "Learning to predict by the methods of temporal differences," *Machine learning*, 1988, *3*, 9–44.

_ , "Integrated architectures for learning, planning, and reacting based on approximating dynamic programming," in "Machine learning proceedings 1990," Elsevier, 1990, pp. 216–224.

_ , "Dyna, an integrated architecture for learning, planning, and reacting," *ACM Sigart Bulletin*, 1991, *2* (4), 160–163.

_ **and Andrew G Barto**, *Reinforcement learning: An introduction*, MIT press, 2018.

Wiederholt, Mirko, "Rational Inattention," in Steven N. Durlauf and Lawrence E. Blume, eds., *The New Palgrave Dictionary of Economics*, Vol. 4, Palgrave Macmillan, 2010.

Woodford, Michael, "Modeling imprecision in perception, valuation, and choice," *Annual Review of Economics*, 2020, *12*, 579–601.

Wu, Charley M, Eric Schulz, Maarten Speekenbrink, Jonathan D Nelson, and Björn Meder, "Generalization guides human exploration in vast decision spaces," *Nature human behaviour*, 2018, *2* (12), 915–924.