Bank of England

Deep reinforcement learning in a monetary model

Staff Working Paper No. 1,142
September 2025

Mingli Chen, Rama Cont, Andreas Joseph, Michael Kumhof, Xinlei Pan, Wei Xiong and Xuan Zhou

Staff Working Papers describe research in progress by the author(s) and are published to elicit comments and to further debate. Any views expressed are solely those of the author(s) and so cannot be taken to represent those of the Bank of England or any of its committees, or to state Bank of England policy.

Bank of England

Staff Working Paper No. 1,142

Deep reinforcement learning in a monetary model

Mingli Chen,[1] Rama Cont,[2] Andreas Joseph,[3] Michael Kumhof,[4] Xinlei Pan,[5] Wei Xiong[6] and Xuan Zhou[7]

Abstract

We propose deep reinforcement learning (DRL) as a general approach to bounded rationality in dynamic stochastic general equilibrium (DSGE) models. Agents are represented by deep artificial neural networks and learn to maximise their intertemporal objective function by interacting with an a priori unknown environment. Applying this approach to a model from the adaptive learning literature, DRL agents can learn all equilibria irrespective of local stability properties. However, learning is slow and may be unstable without the imposition of early stopping criteria. These findings can have implications for the use and interpretation of DRL agents and of DSGE models more generally.

Key words: Artificial intelligence, deep reinforcement learning, adaptive learning, monetary policy, fiscal policy, multiple equilibria.

JEL classification: C14, C52, D83, E52, E62.

(1) University of Warwick. Email: m.chen.3@warwick.ac.uk
(2) University of Oxford. Email: rama.cont@maths.ox.ac.uk
(3) Bank of England. Email: andreas.joseph@bankofengland.co.uk
(4) Bank of England. Email: michael.kumhof@bankofengland.co.uk
(5) University of California (Berkeley). Email: xinleipan@berkeley.edu
(6) University of Oxford. Email: wei.xiong@maths.ox.ac.uk
(7) Reserve Bank of Australia. Email: zhoux@rba.gov.au

The views expressed in this paper are those of the authors, and not necessarily those of the institutions with which the authors are affiliated, especially not of the Bank of England or the Reserve Bank of Australia. All errors are our own. We thank Angus Armstrong, Marcus Buckmann, Roger Farmer, Jesús-Fernández-Villaverde, Lilia Maliar, Albert Marcet, Chengchun Shi, Ekaterina Svetlova, Cheng Wang, Kai Ye, Yaolang Zhong, Shenghao Zhu and conference/seminar participants at CEBRA, CEF, Chinese University of Hong Kong, De Nederlandsche Bank, Federal Reserve Board, Fudan University, Peking University, Shanghai University of Finance and Economics, Tsinghua University, T2M, University of International Business and Economics, University of Southampton, QuickTalks 2: Macroeconometrics and Applied Macro, and WEHIA for useful comments. Support from the Economic and Social Research Council (ESRC) through the Rebuilding Macroeconomics Network (Grant Ref: ES/R00787X/1) is gratefully acknowledged.

The Bank's working paper series can be found at www.bankofengland.co.uk/working-paper/staff-working-papers

Bank of England, Threadneedle Street, London, EC2R 8AH
Email: enquiries@bankofengland.co.uk

ISSN 1749-9135 (on-line)

Agent expectations are central to macroeconomics. The concept of rational expectations (RE) assumes individual rationality and consistency of expectations for all agents. When applied econometrically, RE models attribute significantly more knowledge to agents than what is typically available to econometricians (Sargent 1993, Evans & Honkapohja 2009). While this framework serves as a logical benchmark, the assumptions underpinning it are strong, and many have argued that they should be relaxed (Woodford 2013, Moll 2024).

The literature on Adaptive Learning (AL) (Sargent 1993, Evans & Honkapohja 2001) is one of the leading paradigms in the learning literature and the one that we take as our starting and reference point. AL retains the assumption of individual rationality, while replacing consistency of expectations with the assumption that agents form their expectations iteratively, and use recursive linear least squares as a forecasting rule. These forecasts are an input into agents' decision rules, and in each period the economy attains a temporary equilibrium. Models populated with AL agents put the agents on an equal footing with an econometrician who is observing data from the model. However, this type of parametric recursive method assumes that agents continue to correctly specify the laws of motion and other relevant functional relationships of the model. By assumption, the predictions of this econometric model need not coincide with the predictions of the true model. This makes it important to specify the reduced form forecasting rule such that the AL agents' expectations converge to those of the RE agents. In this case, an equilibrium is referred to as learnable. Importantly, economic dynamics, e.g. the stability of central bank policies such as Taylor rules or forward guidance, may be different under AL compared to RE (Eusepi & Preston 2018).

In this paper, we combine a classical dynamic stochastic general equilibrium (DSGE) model with flexible expectations formation that employs modern developments in *deep reinforcement learning* (DRL, Mnih et al. (2015), Sutton & Barto (2018)). We populate the model with a household agent who is represented by a set of deep neural networks that encode its knowledge and behavior. Importantly, the agent has no a priori knowledge of the structure of the economy other than its own preferences encoded in a utility function. Instead of actions derived from first-order conditions, this agent uses its utility realizations in response to its actions to learn potentially non-linear decision rules and state values represented by neural networks (Goodfellow et al. 2016).

Our approach enables agents to learn flexibly, in the sense of being unconstrained by a particular functional form, because our learning algorithms are based on nonparametric universal function approximators (Cybenko 1989, Goodfellow et al. 2016). This reduces the risk of misspecification inherent in a parametric approach. Learning that replaces RE agents with DRL agents represented by artificially intelligent neural networks is also reminiscent of the bounded rationality paradigm of Sargent (1993).

While a RE agent is endowed with comprehensive knowledge about the structure of the economy, a DRL agent only knows its reward, which it maximizes through learning from repeated interactions with the model economy environment (Sutton & Barto 2018). Specifically, at each step, the agent observes state variables, takes actions (e.g. consumption), receives a reward from the environment, and observes a new state. Based on this experience, the agent updates the weights of its neural networks to improve future decision-making. The DSGE framework is amenable to the DRL approach, because it can easily be cast in the temporal setting of state, actions, reward, and next state.

The use of deep artificial neural networks in reinforcement learning is at the forefront of advances in artificial intelligence, where agents learn to master complex dynamic environments. DRL has powered landmark AI successes — from Atari and AlphaGo to robotics (Mnih et al. 2013, 2015, Silver et al. 2016) — yet its properties for modelling macroeconomic agents is largely unexplored. We contribute to the literature by showing that DRL can be used by economists to solve complex behavioral problems. We argue that this approach offers a principled computational way to represent bounded rationality.

We apply our approach to a classical model from the learning literature in macroeconomics (Benhabib et al. 2001, Evans & Honkapohja 2005). This model looks at the interaction of monetary and fiscal policies in the presence of a single representative household agent. It studies the dynamics of inflation, debt, and money under a global Taylor rule that generates two steady states of inflation, the inflation target and a low inflation "liquidity trap".

Evans & Honkapohja (2005), Eusepi (2007) and Evans & Honkapohja (2008) have studied the stability properties of this model under AL, and have shown that the learnability of the two steady states depends on the parameterizations of monetary and fiscal policy rules. Specifically, monetary and fiscal policy can both be "active"or "passive"depending on how strongly the central bank or the fiscal authority react to deviations from the inflation target or to the outstanding stock of debt, respectively.

We investigate whether and how our proposed DRL approach enables the household agent to learn the analytically known locally optimal RE solutions, and compare these results with those under AL. When an active fiscal or monetary policy is paired with a passive policy counterpart, the corresponding RE equilibria are determinate. We find that both AL and DRL agents can learn these equilibria. When both monetary and fiscal policies are active (or both are passive), the corresponding RE equilibria are explosive (or indeterminate). We find that, unlike AL agents, DRL agents can learn such equilibria. The main reason is that DRL agents are not constrained by the dynamics of the linearized system, but instead follow the "global map" given by long-term utility maximization. This means that the economy can end up in potentially more states than previously thought.

These findings echo early comparisons between AL and evolutionary algorithms such as genetic learning (Arifovic 1995). Our study also aligns with recent calls to rethink rational expectations in heterogeneous agent models on both empirical and computational grounds (Moll 2024). They underscore the promise of DRL as a flexible, boundedly rational framework in modern macroeconomic environments.

We assess the state of learning or the state of (bounded) rationality of households by measuring how far their actions are from the RE solution. To do so we introduce a set of simple metrics, first-order condition distances (FOC-distances), which quantify an agent's behavior on a well-defined spectrum from random to rational. In a small numerical experiment, we show how the expectations of DRL agents can be extracted from FOC-distances. Together with the proposed measures of bounded rationality, this type of analysis may help to eventually bring the proposed approaches to the data.

The DRL agent is able to find global solutions, so that DRL can be used as a global solution technique. Interestingly, our experiments indicate that, in our model, active monetary policy has a stabilizing effect over large parts of the state space. The household converges to the inflation-target steady state despite this having lower utility than the liquidity trap steady state, suggesting a relatively large basin of attraction for this solution.

DRL also poses several challenges. First, learning can be slow to converge unless one adopts the "social learning" interpretation that we will discuss below. Second, learning can be unstable. While the DRL agent generally learns the RE solution, it does not necessarily remain there as it has no knowledge of its existence or location. This necessitates the introduction of early stopping criteria, which are common in the reinforcement learning literature (Hastie et al. (2009), Liu et al. (2021), Xia et al. (2023), Karwowski et al. (2023)).

Assuming that DRL agents resemble real-world agents who have imperfect knowledge of their economic environment but are fully aware of their own preferences, these findings have wider implications for the use or interpretation of DSGE models. In particular, learning agents may not behave rationally at most times. Linear stability criteria may not be sufficient to describe whether agent behavior is stable or not, and economic mechanisms that keep a learning agent at the RE solution need to be justified.

The remainder of this paper is structured as follows. Section 1 presents the model. Section 2 provides brief introductions to both adaptive learning and deep reinforcement learning. Section 3 presents the main results. Section 4 conducts robustness analysis. We conclude with a general discussion in Section 5. Auxiliary information is provided in the appendix.

1 The Model

The model closely follows Benhabib, Schmitt-Grohe & Uribe (2001) and Evans & Honkapohja (2005). Time is discrete and measured in quarters. Prices are flexible. We focus our presentation on the components of the model that are needed to formulate the learning problem.

Households There is a single representative household who discounts the future at a rate $\beta \in (0,1)$. The agent seeks to maximize its utility, which depends on real consumption c_t, real money balances m_t, and hours worked h_t, subject to an inter-temporal budget constraint. Formally, the household solves the following problem

$$\max_{c_t, m_t, h_t} \quad \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t U(c_t, m_t, h_t) \tag{1}$$

$$\text{s.t.} \qquad m_t + b_t + c_t = \frac{m_{t-1}}{\pi_t} + \frac{b_{t-1}}{\pi_t} R_{t-1} + w_t h_t - \tau_t, \tag{2}$$

where $\pi_t = \frac{P_t}{P_{t-1}}$ is the gross inflation rate with P_t as the price level, R_t is the gross nominal interest rate on government bonds b_t held from period t to $t+1$, w_t is the real wage rate, τ_t is real lump-sum taxation, $m_t = \frac{M_t}{P_t}$, and $b_t = \frac{B_t}{P_t}$.

We follow Evans & Honkapohja (2005) by adopting a utility function of the form

$$U(c_t, m_t, h_t) = \frac{c_t^{1-\sigma}}{1-\sigma} + \chi \frac{m_t^{1-\sigma}}{1-\sigma} - \frac{h_t^{1+\varphi}}{1+\varphi}. \tag{3}$$

Firms The representative firm operates a constant returns to scale production function

$$y_t = \varepsilon_t^y h_t, \tag{4}$$

where ε_t^y is an exogenous and stochastic technology shock with mean one. In each period the firm maximizes profits as the difference between production and the wage bill by setting the real wage rate, i.e.

$$\max_{h_t} y_t - w_t h_t \quad \Rightarrow \quad w_t = \varepsilon_t^y \,. \tag{5}$$

Market Clearing We assume that the goods market clears in every period,

$$c_t = y_t. \tag{6}$$

Government Budget Constraint and Policy Rules The government issues interest-bearing bonds and non-interesting bearing currency (money), and collects taxes. It operates under the real inter-temporal *government budget constraint* (GBC)

$$m_t + b_t + \tau_t = \frac{m_{t-1}}{\pi_t} + R_{t-1}\frac{b_{t-1}}{\pi_t} \,. \tag{7}$$

The above is subject to the transversality condition

$$\lim_{j \to \infty} \prod_{k=0}^{j} \left(\frac{\pi_{t+k}}{R_{t+k-1}} \right) b_{t+j} = 0 \,. \tag{8}$$

Fiscal Policy is represented by the linear tax rule of Leeper (1991),

$$\tau_t = \gamma_0 + \gamma b_{t-1} + \varepsilon_t^\tau \,, \tag{9}$$

where ε_t^τ is an exogenous and stochastic fiscal policy shock with mean zero. We also make the natural assumption that γ is not excessively large, $0 \leq \gamma \leq \beta^{-1}$. We follow the terminology of Leeper (1991) to define fiscal policy as being *active* if $\gamma < \beta^{-1} - 1$ and *passive* if $\gamma > \beta^{-1} - 1$. Active (passive) fiscal policy implies that taxes rise insufficiently (sufficiently) to ensure fiscal solvency through adjustments in the primary surplus.

Monetary Policy follows Benhabib, Schmitt-Grohe & Uribe (2001) and Evans &

Honkapohja (2005), with a nonlinear global interest rate rule

$$R_t - 1 = f(\pi_t)\varepsilon_t^r. \tag{10}$$

The function $f(\pi)$ is assumed to be non-negative and non-decreasing, while ε_t^r is an exogenous and stochastic monetary policy shock with mean one. We use the functional form

$$f(\pi_t) = (R^* - 1)\left(\frac{\pi_t}{\pi^*}\right)^{\frac{AR^*}{R^*-1}}, \tag{11}$$

where $A > \beta$, π^* is the inflation target of the monetary authority, and $R^* = \beta^{-1}\pi^*$ is the steady state nominal interest rate that is consistent with this inflation target. We adopt the notation

$$\alpha := f'(\pi) = \frac{A}{\beta}\left(\frac{\pi}{\pi^*}\right)^{\frac{R^*(A-1)+1}{R^*-1}}, \tag{12}$$

where α measures how strongly monetary policy responds to inflation around a steady state value of inflation given by π.

This specification of monetary policy implies that the nominal interest rate is strictly positive and strictly increasing in the inflation rate. We refer to monetary policy as *active (passive)* if the monetary authority raises the nominal interest rate by *more (less) than one-for-one* in response to an increase in the inflation rate around the steady state π, that is, if $\alpha > (<)1$.

Optimality Conditions While our learning approach will not rely on optimality conditions, some of the measures that we will define will reference them. The first-order conditions are given by the Euler equation

$$1 = \beta E_t \left(\frac{c_{t+1}}{c_t}\right)^{-\sigma} \frac{R_t}{\pi_{t+1}}, \tag{13}$$

the real money demand

$$m_t = c_t \left(\frac{R_t - 1}{\chi R_t}\right)^{-1/\sigma}, \tag{14}$$

and the labor supply equation

$$w_t = c_t^{\sigma} h_t^{\varphi}. \tag{15}$$

Under optimality, the labor market clears. Combining (4), (5), (6) and (15) we find that consumption, output and labor depend on the technology shock,

$$c_t = y_t = \varepsilon_t^y h_t = \varepsilon_t^{y\,\frac{1+\varphi}{\sigma+\varphi}}. \tag{16}$$

The Two Steady States Both deterministic steady states are characterized by the following set of equations:

$$\text{Euler / Fisher Equation:} \quad R = \frac{\pi}{\beta} \tag{17}$$

$$\text{Money Demand:} \quad m = y\left(\frac{\pi-\beta}{\chi\pi}\right)^{-1/\sigma} \tag{18}$$

$$\text{Monetary Policy:} \quad R = 1 + (R^* - 1)\left(\frac{\pi}{\pi^*}\right)^{\frac{AR^*}{R^*-1}} \tag{19}$$

$$\text{Fiscal Policy \& GBC:} \quad b = \left(\frac{1}{\beta} - 1 - \gamma\right)^{-1}\left(\gamma_0 + (1-\frac{1}{\pi})m\right) \tag{20}$$

$$\text{Output:} \quad y = 1 \tag{21}$$

Equation (17) and (19) together determine the steady state of inflation:

$$\frac{\pi}{\beta} = 1 + (R^* - 1)\left(\frac{\pi}{\pi^*}\right)^{\frac{AR^*}{R^*-1}}. \tag{22}$$

If $f(\cdot)$ is continuous and differentiable as in (11), and has a steady state π^* with $f'(\pi^*) > 1$ in accordance with the Taylor principle, there exists a second low inflation steady state π_L with $f'(\pi_L) < 1$. Figure 1 illustrates this multiplicity of steady-state inflation via the intersection of the Fisher equation and the monetary policy rule.

These results are formalized by

Proposition 1 *[Benhabib, Schmitt-Grohe & Uribe (2001)] There exist two steady states of inflation. The first steady state is characterized by an inflation rate $\pi^* \geq 1$ such that the steady state Fisher equation is satisfied and the feedback rule is active, $R^* = \frac{1}{\beta}\pi^*$ and $f'(\pi^*) = \frac{A}{\beta} > \beta^{-1}$. The second steady state is characterized by an inflation rate $\pi_L < \pi^*$*

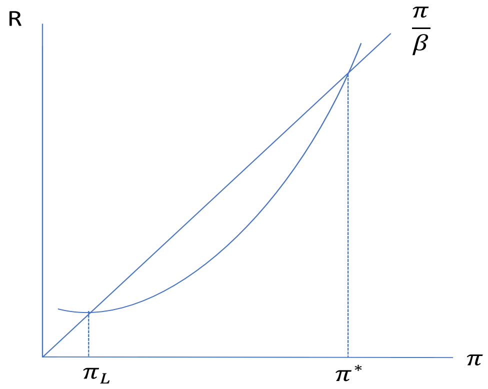


Figure 1 The Two Steady States of Inflation

such that the steady state Fisher equation is satisfied and the interest rate rule is passive, $R_L = \frac{1}{\beta}\pi_L$ *and* $f'(\pi_L) = \frac{A}{\beta}(\frac{\pi_L}{\pi^*})^{\frac{(A-1)R^*+1}{R^*-1}} < \beta^{-1}$.

Once inflation is determined, real money balances are determined by (18) and real debt by (20). The linearized dynamic system of equations used in the RE and AL settings is presented in the Appendix. In the neighborhood of either steady state, our model can be described by a linear approximation in two variables, π_t and b_t.

2 Learning Approaches

In this section we first review the AL approach and then give a general introduction to the main concepts of (deep) reinforcement learning and to the specific DRL algorithm used in this paper. We apply this algorithm to our model and derive state transition and learning protocols tailored to the model setting. Finally, we put both learning approaches into context using the concept of generalized policy iteration, which offers a unifying framework.

2.1 Adaptive Learning

Learning in macroeconomics represents deviations from the RE hypothesis while still adhering to the general equilibrium principle (Sargent 1993, Evans & Honkapohja 2001, 2009, Eusepi & Preston 2018). In this sense, all learning approaches contribute to the study of the general notion of bounded rationality, which may include RE as a special case.

One of the main approaches in the economic learning literature is *AL*. Private agents, the household agent in our case, make forecasts using a reduced form econometric model of the relevant variables, and estimate the parameters of this model in a self-referential system based on past data. In each period, the economy uses the agent forecast as an input and attains a temporary equilibrium. This provides a new data point for the next period's forecast. This sequence of temporary equilibria may generate parameter estimates that converge to a fixed point corresponding to a RE equilibrium. In this case, the RE equilibrium is stable under learning, or learnable.

Evans & Honkapohja (2001) show that there is a close connection between the convergence of least squares learning to a RE equilibrium and a stability condition, known as *E-stability*, based on a mapping from a *perceived law of motion* (that private agents are estimating) to an *implied actual law of motion* that generates the data under these perceptions. E-stability represents the local stability properties at a rational expectations equilibrium of a system of equations based on this map.

If there are multiple RE equilibria, the econometric model that agents use for forecasting can determine which of these equilibria are learnable. This may then serve as an equilibrium selection criterion.

We translate this to our model setting. In doing so we focus on the case in which the exogenous shocks are i.i.d. processes, which simplifies our analysis without affecting the theoretical results. The linearized model for π_t and b_t is shown in equation A.1 in the appendix. The rational expectation solutions of π_t and b_t are then i.i.d. processes around their deterministic steady states. In AL, the agents treat the system describing π_t and b_t as i.i.d. processes with unknown means that they try to estimate by least squares (perceived law of motion). The agents forecast π^e_{t+1} and b^e_{t+1} by estimating the mean values of π_t and b_t, which is called *steady state learning*, and where the expectations operator no longer signifies rational expectations. We can then identify the expectations of the variables with the estimates of their means. This can be written as a simple recursive algorithm

$$x^e_{t+1} = x^e_t + \phi_t(x_t - x^e_t)\,, \tag{23}$$

with $x \in \{\pi, b\}$. The superscript e refers to the agent's expected quantity, and ϕ_t is a gain sequence. Under least-squares learning it is usually taken to be $\phi_t = \frac{1}{t}$, often termed a "decreasing-gain"sequence, where the influence of new observations decreases over time.

E-stability can be evaluated using a linearized system matrix $\mathbf{B}$ evaluated at the steady state, a derivation of which is presented in Section A.2 in the appendix. For a system comprised of two variables, the E-stability condition is that one eigenvalue of $|\mathbf{B} - I|$ have a real part less than zero and the other an eigenvalue with a real part greater than zero.

The formal learning results are summarized by

Proposition 2 *Under steady state learning, if the supports of shocks are sufficiently small, we have: (i) If fiscal policy is passive, $|\gamma - \beta^{-1}| < 1$, the steady state π^* is locally stable under learning and the steady state π_L is not locally stable under learning.*
(ii) If fiscal policy is active, $|\gamma - \beta^{-1}| > 1$, the steady state π^* is not locally stable under learning and the steady state π_L is locally stable under learning.

Proof: The eigenvalues of $|\mathbf{B} - I|$ are $ev_1(\pi) = \frac{1}{\beta f'(\pi)} - 1$ and $ev_2 = \frac{1}{1/\beta - \gamma} - 1$. Since $f'(\pi^*) > \frac{1}{\beta}$ and $f'(\pi_L) < 1$, we have $ev_1(\pi^*) < 0$ and $ev_1(\pi_L) > 0$. When fiscal policy is passive, $ev_2 > 0$ and when fiscal policy is active, $ev_2 < 0$. ■

Proposition 2 states that the stance of fiscal policy determines which of the two steady states is learnable via AL. Only the combinations active-passive and passive-active of the monetary-fiscal policy mix are learnable via AL.

2.2 (Deep) Reinforcement Learning

Reinforcement learning studies the problem of maximizing the long-run reward of an agent within a modeling environment that is unknown to the agent. The objective is to find behavioral rules or policies that, as a function of state observations, lead to agent actions that maximize the expected discounted reward. Instead of relying on extensive knowledge of the model on the part of the agent, reinforcement learning imposes minimal requirements on agents' knowledge and behavior. Specifically, the agent observes state variables and responds to reward signals without explicit transition dynamics. Here we give a brief introduction to reinforcement learning and relate it to the model of Section 1. This introduces features that can be readily transferred to other model settings. A comprehensive introduction to reinforcement learning is given in Sutton & Barto (2018).

An agent in reinforcement learning aims to maximize its expected cumulative lifetime reward, or the *expected return*,

$$\max_{\mathcal{P}} \mathbb{E}_t[G_t] \quad \text{with} \quad G_t \equiv \sum_{k=0}^{\infty} \beta^k r(s_{t+k}, a_{t+k}), \tag{24}$$

where $\beta \in (0, 1]$ is a discount factor, $r(s_t, a_t) \in \mathbb{R}$ is the reward given state $s_t \in \mathcal{S} \subset \mathbb{R}^{n_s}$ and action $a_t \in \mathcal{A} \subset \mathbb{R}^{n_a}$, where n_s and n_a are the dimensions of the state and action spaces. The agent achieves maximization of (24) by optimizing its behavioral rule, or

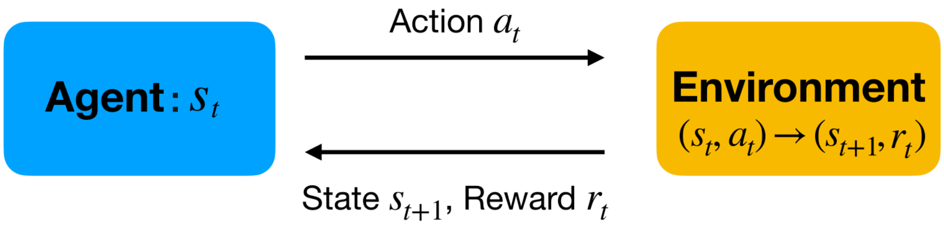

Action a_t
Agent: s_t
Environment
$(s_t, a_t) \rightarrow (s_{t+1}, r_t)$
State s_{t+1}, Reward r_t

Figure 2 Agent-Environment Interaction in Reinforcement Learning

policy, $\mathcal{P} : s_t \rightarrow a_t \in \mathcal{A} \subset \mathbb{R}^{n_a}$ based on observed state transitions. These actions interact with the environment in which the agent lives, leading to the next state as well as returning the current reward, i.e. $\mathcal{E} : (s_t, a_t) \rightarrow (s_{t+1}, r_t)$. This process is schematically shown in Figure 2. This can be formulated as a Markov decision process defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, r)$. The transition probability $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ describes the probability of the next state $Pr(s_{t+1}|s_t, a_t) = F(s_{t+1}|s_t, a_t)$ given the current state s_t and action a_t, where $F(\cdot)$ describes the model environment. This transition function fulfills the Markov property in that it only depends on the current state and action, but not on the history of state transitions.

Finding the optimal policy $\mathcal{P}^*$ can be approached from the *state value function*

$$\begin{aligned} V^*(s) &= \max_{a \in \mathcal{A}} \mathbb{E}_{\mathcal{P}^*} \left[G_t | s_t = s, a_t = a \right] \\ &= \max_{a \in \mathcal{A}} Q^*(s, a) \,, \end{aligned} \tag{25}$$

where the last expression defines the *action-value function* Q, i.e. the expected return from following a behavioral rule $\mathcal{P}$ given a state and an action. The optimal policy $\mathcal{P}^*$ optimizes both state and action values, which also maximizes expected return - our final goal.

The action-value function fulfills the recursive *Bellman equation*

$$Q^*(s_t, a_t) = r(s_t, a_t) + \beta \mathbb{E}_{\mathcal{P}} \big[\max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1}) | s_t, a_t \big]. \tag{26}$$

The current state-action value is the current reward plus the expected discounted value of the next period's state-action value. These components form the backbone of many different reinforcement learning algorithms, including value-based (Q-learning), policy-based (Lillicrap et al. (2015), Schulman et al. (2017)), and hybrid approaches. The quality of learning may differ between problems, algorithms, and their implementation.

We will primarily use the Soft Actor-Critic (SAC) approach (Haarnoja et al. 2018, Raffin et al. 2021). This is based on the iterative improvement of a policy $a_t = \mathcal{P}(s_t)$ (actor) that is evaluated by the action-value function $Q(s_t, a_t)$ (critic). That is, we will estimate two separate objects, $\mathcal{P}$ and Q, where one can be used to evaluate the other. These can be parameterized using general function approximators in the form of deep artificial neural networks with internal weights ϕ and θ, denoted by $\mathcal{P}_\phi$ and Q_θ.[1] This combination of traditional reinforcement learning and deep artificial neural networks is referred to as *deep* reinforcement learning.[2] General function approximators dramatically increase the capabilities of reinforcement learning and have facilitated many recent advances.

Interactions of the agent and the environment produce state transitions that the agent samples as observations. Using standard optimization techniques like stochastic gradient descent, the policy and action-value function networks can be trained by iteratively minimizing the Bellman residuum

$$L(\phi, \theta) \;=\; \mathbb{E}_{s_t, a_t, r_t}\left[\frac{1}{2}\big(Q_\theta(s_t, a_t) - \hat{Q}_\theta(s_t, a_t)\big)^2\right], \tag{27}$$

with the entropy-augmented target given by

$$\hat{Q}_\theta(s_t, a_t) \;=\; r(s_t, a_t) + \beta\, \mathbb{E}_{s_{t+1} \sim P(\cdot|s_t, a_t), \mathcal{P}}\left[Q_{\bar{\theta}}\big(s_{t+1}, \mathcal{P}_\phi(s_{t+1}) + \alpha^{\mathrm{SAC}} \mathcal{H}(\mathcal{P}_\phi(\cdot|s_{t+1}))\big)\right], \tag{28}$$

where $\bar{\theta}$ is a target network updated via Polyak averaging, $\mathcal{H}(\mathcal{P}_\phi(\cdot|s_t)) = -\log \mathcal{P}_\phi(a_t|s_t)$ is the policy entropy, $\alpha^{\mathrm{SAC}} > 0$ is the temperature parameter adjusted to match a target entropy $\bar{\mathcal{H}}$. The entropy term encourages exploration while optimizing returns.

Our learning process is summarized in Algorithm 1. Learning takes place in episodes that are initiated with a random state drawn uniformly from a region of interest in the state space. Subsequent iterations between agent actions and the environment result in state transitions, and optimization routines are applied to update the neural network weights in $\mathcal{P}_\phi$ and Q_θ.[3] A learning episode ends when one of two termination criteria is reached, either a maximal number of steps N_{epi}^{max} or a utility change that remains below a fixed small threshold d_u^{min}.[4] This approach allows the agent to exploit knowledge to improve its

[1] We use standard feed-forward networks with two or three hidden layers and 32 nodes in each layer. See Table A.1 for details.

[2] See Goodfellow et al. (2016) on the use of deep artificial neural networks.

[3] The learning rate ζ_{learn} controls the size of parameter updates in $\mathcal{P}_\phi$ and Q_θ. Larger values imply faster learning, but at the risk of destabilized learning with explosive or stagnant behavior.

[4] A smaller value of d_u^{min} means that the agent spends more time within a narrow region of the state

utility while experiencing a sufficiently large part of the state space.

Learning happens via a combination of updates of neural network weights between steps and explorative actions. Explorative actions are not optimal according to the currently learned behavioral policy. Instead they are either purely random during a "burn-in" period ($\mathcal{P}_{\phi}^{random}(s_t)$),[5] or they follow the currently learned policy $\mathcal{P}_{\phi}$ but with an added noise component ($\mathcal{P}_{\phi}^{expl}(s_t)$). This is a crucial part in DRL, as it allows the agent to discover new and ultimately better actions. The magnitude of the random component in actions characterizes the exploration-exploitation trade-off. Set too small or too large, the agent will fail to learn efficiently, or will not learn at all. Specifically, our action space is continuous, and exploration after "burn-in" is achieved by drawing from a normal distribution generated from the currently learned policy function $\mathcal{P}_{\phi}(s_t)$. That is, most actions will be close to the mean of this action distribution, which represents the currently learned best action, while deviations from this mean explore the action space. If such actions yield higher utility at the current state, the weights of the action network $\mathcal{P}_{\phi}$ are updated accordingly.

For updates of the parameters in $\mathcal{P}_{\phi}$ and Q_{θ} the agent draws randomly from a fixed-size memory of experience consisting of N_{mem} past state transitions, and on this basis performs stochastic batch gradient descent. The oldest transition drops out if the memory is full. The overall learning phase is set to last for a maximal number of N_{learn} steps, where N_{learn} is the number of parameter updates by which we expect the agent to solve its optimization problem. We consider the agent's problem solved if it finds action values that correspond to one of the RE steady states of the model.

Testing takes place at intervals of $N_{interval} \ll N_{learn}$ learning steps. During testing, parameter updates are halted temporarily, and learning goals such as distances to one of the steady states are evaluated. Testing consists of a fixed number of test episodes N_{test}, each of which is initiated with a random state drawn uniformly from the state space. Like in learning, testing iterations between agent actions and the environment result in state transitions, but unlike in learning, neural network weights remain fixed at their previously optimized values. The termination criteria for testing episodes are identical to the criteria for learning episodes. Testing is likely to happen within an unfinished learning episode due to the stochastic nature of the length of both learning and testing episodes. The

space. This gives the agent less experience in terms of the number of states it has seen, but it may lead to increased learning precision.

[5] To overcome randomness in initialization at the beginning of each learning episode, the algorithm starts by generating N_{burn} state transitions based on random actions. These are put into memory to allow learning to begin.

Algorithm 1 Learning and Testing Protocol of the Household Agent

```
Initialize environment $\mathcal{E}$ (parameterized model) and agent (parameterized by $\mathcal{P}_\phi$, $Q_\theta$)
  for steps = 1 to $N_{learn}$ do
    Initialize new learning episode with random state $s_t$
    while new learning episode is not done do
      if steps $\leq N_{burn}$ then
        Take allowed random action $a_t = \mathcal{P}_\phi^{random}(s_t)$
      else
        Draw exploration action $a_t = \mathcal{P}_\phi^{expl}(s_t)$
      end if
      Environment returns $(r_t, s_{t+1}) = \mathcal{E}(s_t, a_t)$
      Add state transition $(s_t, a_t, r_t, s_{t+1})$ to memory
      Update the weight in $\mathcal{P}_\phi$ and $Q_\theta$ using batch gradient descent with gradient $\nabla(\mathcal{P}_\phi, Q_\theta)$ from memory of $N_{mem}$ state transitions:
      $(\phi, \theta) \leftarrow (\phi, \theta) - \zeta_{learn} \nabla_{(\phi,\theta)} L(\phi, \theta)$ .
      if $mod(\text{steps}, N_{interval}) = 0$ then
        $n_{train} = steps/N_{interval}$
        for test = 1 to $N_{test}$ do
          Initialize new testing episode with random state $s_t$
          Run testing episode without updates of $\mathcal{P}_\phi$ and $Q_\theta$
          Testing episode termination criteria ($N_{epi}^{max}$, $d_u^{min}$)
          Save state transitions (*)
        end for
        Save current agent ($\mathcal{P}_\phi^{steps}$, $Q_\theta^{steps}$)
      end if
      State update $s_t \leftarrow s_{t+1}$
      Learning episode termination criteria ($N_{epi}^{max}$, $d_u^{min}$)
    end while
  end for
  Save final agent ($\mathcal{P}_\phi^{final}$, $Q_\theta^{final}$) =0
```

Values used in main experiments (Figure 5 top, Figures 6,7,8,9,11): $N_{learn} = 1,500,000$, $N_{mem} = 25,000$, $N_{burn} = 10,000$, $N_{interval} = 10,000$, $N_{test} = 10$, $N_{epi}^{max} = 25,000$

Values used in alternative experiments (Figure 5 bottom, Figure 10): $N_{learn} = 10,000,000$, $N_{mem} = 25,000$, $N_{burn} = 10,000$, $N_{interval} = 10,000$, $N_{test} = 10$, $N_{epi}^{max} = 25,000$

motivation for using a fixed number of learning steps between sequences of test episodes is that this allows us to measure the agent's learning progress uniformly based on the number of parameter updates or learning steps.

For each testing episode, and therefore at different stages of learning, we save the agent parameters $(\mathcal{P}_\phi, Q_\theta)$ as well as all state transitions. This allows ex-post experimentation, the reproduction of test results, or the flexible adjustment of the learning setting.

We have now defined the general DRL framework and algorithm. Next, we describe how this setting can be applied to our model.

2.3 DRL Applied to Our Model

The household's problem (1) is analogous to the learning agent's problem (24) when replacing the general reward $r(s_t, a_t)$ with the household's period utility (3). The environment $\mathcal{E}$, about which the agent is ignorant, is given by the production process (4), wage setting (5), market clearing (6), the government budget constraint (7), fiscal policy (9) and monetary policy (10). It does not include the first-order conditions (13)–(15).

The state at time t, s_t, is given by last period's real money balances, bond holdings, inflation, consumption, and hours worked:

$$s_t = (m_{t-1}, b_{t-1}, \pi_{t-1}, c_{t-1}, h_{t-1}) \tag{29}$$

The state representation is not unique,[6] but it does need to fulfill the Markov property when combined with the agent's actions and the environment.

The household's actions a_t at each time step t are a tuple of consumption, bond saving, and hours worked denoted by

$$a_t = \left(c_t^{act}, b_t^{act}, h_t\right) \tag{30}$$

where x_t^{act}, $x \in \{c, b\}$, represents actions with reference to last period's price level, i.e. X_t/P_{t-1} with X_t being nominal consumption and nominal bond holdings. The action specification is again not unique. It must only allow for valid state transitions that satisfy the Markov property.

The state at time $t+1$, s_{t+1}, is determined by the interactions of the household's actions (30) and the model environment. These actions set the levels of inflation, real consumption

[6] Inflation can be replaced by the gross nominal interest rate according to (10).

and real bond holdings as

$$\begin{aligned} \pi_t &= c_t^{act}/y_t\,, && (31)\\ c_t &= c_t^{act}/\pi_t\,, && (32)\\ b_t &= b_t^{act}/\pi_t\,. && (33)\end{aligned}$$

The first relation states that prices ensure market clearing according to (6), where the relationship between the agent's action (here, choosing consumption with reference to last period's price level) and the price adjustment process between different periods has been made explicit. This mechanism respects the information flow in the model.

We next define the state transition (*) of a single testing step in Algorithm 1.

Step sequence for a single transition: $s_t \rightarrow s_{t+1}$

1. Observe state s_t
2. Take actions $\mathcal{P}_\phi(s_t) = a_t = (b_t^{act}, c_t^{act}, h_t)$
3. Shock realizations $(\epsilon_t^\tau, \epsilon_t^R, \epsilon_t^y)$
4. Production y_t takes place according to (4) and the firm sets wages according to (5)
5. Markets clear: Inflation π_t is set by (31)
6. This determines real consumption c_t and real bond holdings b_t according to (32)-(33)
7. Policy realizations:
    - Government sets taxes τ_t based on the taxation rule (9)
    - Monetary authority sets interest rate R_t based on the Taylor rule (10)
8. Money holdings m_t are realized from the government budget constraint (7)
9. Agent obtains reward $r_t = U(c_t, m_t, h_t)$
10. State is updated $s_t \leftarrow s_{t+1} = (m_t, b_t, \pi_t, c_t, h_t)$

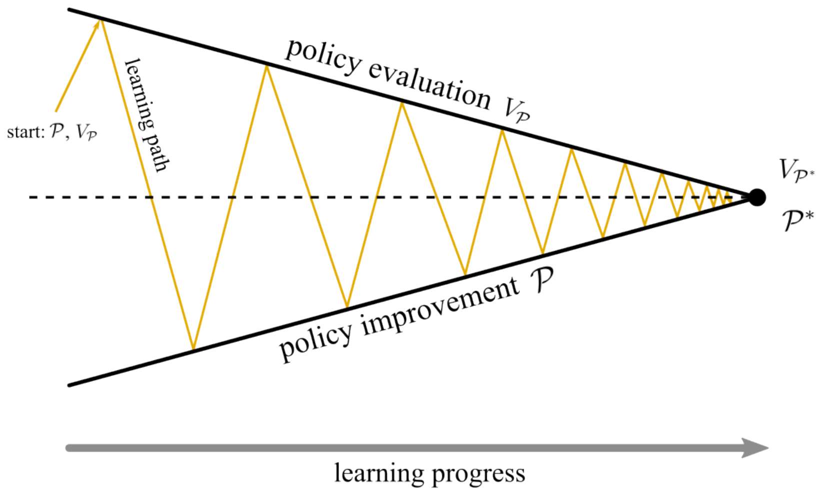

policy evaluation $V_{\mathcal{P}}$
learning path
start: $\mathcal{P}$, $V_{\mathcal{P}}$
$V_{\mathcal{P}^*}$
$\mathcal{P}^*$
policy improvement $\mathcal{P}$
learning progress

Figure 3 Schematic Representation of Generalized Policy Iteration

2.4 Generalized Policy Iteration

AL and DRL can be conceptually compared from the point of view of *generalized policy iteration* (GPI). In GPI, policy evaluation delivers a state value $V_{\mathcal{P}}$, policy improvement $\mathcal{P}$ delivers a change of agent behavior to attain higher returns, and both interact iteratively (Sutton & Barto 2018) - see Figure 3. Under the (non-trivial) assumption that learning converges, this results in a fixed point of an optimal policy $\mathcal{P}^*(s)$ and maximal-return state values $V_{\mathcal{P}^*}(s)$ that represent the RE equilibrium in our model. As long as the agent has not converged to this point, it is called *boundedly rational*.

In our model, optimal agent *behavior* is given by the household first-order conditions, which are listed in the appendix as equations A.12–A.13. The (locally) optimal *state values* are the two steady states corresponding to π^* and π_L in Fig. 1. AL fixes behavior by using the agent's first-order conditions from the outset – the agent does not have to learn behavior. This can be thought of as a horizontal rather than an upward-sloping policy or agent behavior line, as indicated by the dashed line in Figure 3. State values are described by the state equations (23), and they converge to their optimum according to Proposition 2 – the agent does have to learn state values. DRL does not fix behavior, instead both the policy and the corresponding state-action values are learned simultaneously. GPI can also be used to describe *Euler learning* in macroeconomics, where the behavioral rules differ from the first-order conditions (Eusepi & Preston 2018). GPI convergence cannot happen if the behavioral rules are not flexible enough to converge to the first-order conditions. DRL addresses this problem through the flexibility of deep neural networks.

2.5 Early Stopping

In the machine learning and reinforcement learning literatures, *early stopping* is a commonly used technique to mitigate overfitting and to enhance model performance (Hastie et al. 2009, Liu et al. 2021, Xia et al. 2023, Karwowski et al. 2023). That is, when one or several criteria are met by the agent's actions or by the corresponding state transitions, learning stops, either partially or completely.[7] Here, we impose early stopping by making the steady state *absorbing*. Specifically, we introduce small catchment areas around the steady state values of agent actions, which we set to 0.01% of the corresponding steady state values. Let $x_t \in a_t$ and x denote its steady state value, and let $\delta_{ES} = 1e - 4$. Then

$$x_{t'} \quad \to \quad x \quad \forall t' \geq t, \quad \text{if} \quad |(x_t - x)/x| < \delta_{ES}\,. \tag{34}$$

Eq. (34) fixes the three household actions at their steady state values once these have been reached. However, the first order conditions (13) – (15) also make use of the household budget constraint (2), which necessitates the imposition of an additional criterion to guarantee that the steady state is absorbing. We do this by making use of the government budget constraint (7) once (b_t^{act}, c_t^{act}) have reached their steady state values. The reason is the state transition protocol, where the money demand m_t is the residual of all previous events at time t. We make a one-time adjustment to the household's saving decision b_t before the absorbing state is reached based on the model's steady state values (b, m, π, R) and money holdings m_t of the form

$$b_t \quad \to \quad b_{ES} = \frac{\pi\,(m + b + \gamma_0 - m_t/\pi)}{R - \gamma\,\pi}\,. \tag{35}$$

This adjustment is needed because the presence of intertemporal assets (money and bonds) creates a path dependence in the model structure such that reaching optimal steady state action values does not automatically guarantee reaching the steady state of the system across all variables. This can be seen in the household or government budget constraints, which connect previous period money and bond holdings to the current period.[8]

[7] Early stopping is natural both from a behavioral and a computational perspective. It is a simple implementation of the exploration-exploitation trade off when agents can be thought of as having a finite "search budget".

[8] Eq. 35 is derived from the government budget constraint at time $t+1$ to derive a relation $b_{ES}(m_t)$ which is consistent with $m_{t+1} = m$ and $b_{t+1} = b$ assuming all other state variables have reached their steady state values.

3 Quantitative Evaluation

We apply the DRL framework of Section 2 to the model of Section 1.

3.1 Calibration

The model parameterization is given in Table 1. For preferences, we set $\beta = 0.99$, which implies a steady state real interest rate of about 4 percent, $\varphi = 1$, implying a unitary Frisch elasticity of labor supply, $\sigma = 3$, which is within the range of 1 to 3.5 found in the literature, and $\chi = 0.1$, following Evans & Honkapohja (2005). The fiscal policy rule coefficient that corresponds to passive fiscal policy is $\gamma_P = 0.02$, while $\gamma_A = 0$ corresponds to active fiscal policy. The monetary policy rule coefficient $A = 1.3$ gives two steady states of inflation, one at $\pi^* = 1.01$ (4% net per annum) and the other at $\pi_L = 1.0014$ (0.56% net per annum). The shock series ϵ_t^τ, ϵ_t^R, ϵ_t^y follow log-normal, normal and normal distributions, with means of zero, one, and one, respectively. The standard deviations of the shocks either take standard values from the range found in the empirical literature (monetary policy and technology), or are calibrated to the US economy (taxation).[9] Steady state values for the high and low inflation steady states, as well as for passive and active monetary and fiscal policies, are presented in Table 2. For better comparability across regimes, the fiscal policy intercept γ_0 is recalibrated for each policy regime such that bond holdings equal annualized output.[10]

Steady state money holdings equal between 40-50% of annual output, which for the US equals approximately double/half of M1/M3. While strictly speaking our model only represents narrow money, extensions with a financial sector could represent broader aggregates. Note that money and therefore household utility is generally higher in the low-inflation steady state π_L, with passive monetary policy.

3.2 AL Learnability

The learnability and local stability characteristics of the policy regimes in Table 2 are determined by the parameterization of the fiscal and monetary policy rules as described in Proposition 2. This is summarized graphically in Figure 4 for the parameterization in Table 1. The horizontal axis shows the fiscal response parameter γ and the vertical axis

[9] We compute the standard deviation of the ratio of the nominal primary surplus to detrended nominal GDP using either the Hodrick–Prescott or the Hamilton filter. We take the mean of these two measures and adjust the standard deviation of ϵ_t^τ to match their model equivalent.

[10] This does not affect the local stability properties of the model or the learning dynamics of the agent.

parameter	value	description
β	0.9900	discount factor
σ	3.0000	inverse of intertemporal elasticity of consumption and money holdings
φ	1.0000	inverse of Frisch elasticity of labor supply
χ	0.1000	relative preference weight of money holdings
γ_P	0.0200	passive fiscal policy (PFP) coefficient
γ_A	0.0000	active fiscal policy (AFP) coefficient
A	1.3000	Taylor rule coefficient
π^*	1.0100	target gross high-inflation rate (4% net per annum)
π_L	1.0014	implied gross low-inflation steady state (0.56% net per annum)
$sd(\epsilon_t^\tau)$	0.0080	fiscal policy shock
$sd(\epsilon_t^R)$	0.0010	monetary policy shock
$sd(\epsilon_t^y)$	0.0100	technology shock

Table 1 Baseline Model Parameterization

	AMP		PMP	
	PFP	AFP	PFP	AFP
π_{ss}	1.0100	1.0100	1.0014	1.0014
m_{ss}	1.7157	1.7157	2.0614	2.0614
$c_{ss}/n_{ss}/y_{ss}$	1	1	1	1
b_{ss}	4	4	4	4
u_{ss}	-1.0170	-1.0170	-1.0118	-1.0118
γ_0	-0.0566	0.0234	-0.0426	0.0375

Table 2 Steady State Values under Different Policy Regimes

shows the inflation rate π, where each π corresponds to a value of α through (12). The two steady-state inflation rates are marked by the dashed horizontal lines at $\pi^* = 1.01$ (active monetary policy; AMP) and $\pi_L = 1.0014$ (passive monetary policy; PMP). The two fiscal policy rule parameters are marked by vertical red lines at $\gamma_P = 0.02$ (passive fiscal policy; PFP) and $\gamma_A = 0$ (active fiscal policy; AFP). Note that this graph combines a model parameter γ with a state variable π on the two axes, because for any given fiscal stance (γ) there are two possible steady states. This results in the four regimes marked by their local stability characteristics. These are separated by the vertical and horizontal solid gray boundary lines in Figure 4, which are located at $\gamma^{boundary} = 1/\beta - 1 = 0.0101$ and $\pi^{boundary} = 1.0059$, where the latter is implicitly determined by $f'(\pi_L)\beta = 1$. As shown in the AL literature (Evans & Honkapohja 2001), the determinate regimes (AMP-PFP and PMP-AFP) are learnable while the explosive (AMP-AFP) and indeterminate (PMP-PFP) regimes are not. We next investigate which of these policy regimes are learnable by DRL.

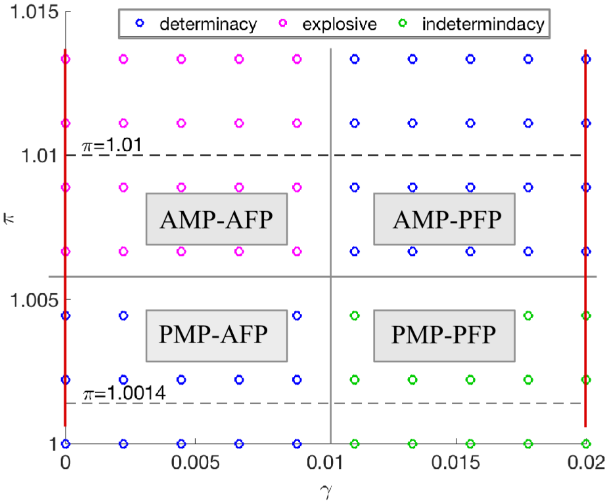


Figure 4 Local Dynamic Stability Properties by Policy Regime

3.3 DRL Learnability

While there exists a correspondence between the dynamic stability properties of non-linear maps and of local linear approximations, such a correspondence does not exist for DRL. Instead, learnability is an empirical question, which we address via numerical simulations. In each of our experiments, we follow the learning protocol described in Algorithm 1 and, depending on the policy regime, we define symmetric and local regions in the action and state spaces, around either the low or high inflation steady state. Details of the settings of the learning algorithm are listed in Table A.1 in the appendix. Exogenous shocks remain switched off in Sections 3.3 and 3.4, and will instead be investigated separately in Section 4.3. Most of our experiments run for $N_{learn} = 1,500,000$ learning steps.[11] We conduct $N_{test} = 10$ test episodes for each $N_{interval} = 10,000$ learning steps. A state or regime is described as DRL learnable if the household's action values converge to the corresponding steady state values during testing. All results presented below are taken from test episodes between learning intervals. We focus on the final action and state values of each test episode to assess the state of convergence of household behavior. We initially focus on the AMP-PFP regime, the classical policy regime of monetary dominance around the target inflation level π^*.

[11] All learning experiments are performed once with a fixed random seed. The effects of different random initializations are investigated in Sextion 4.4.

3.3.1 Unconstrained DRL

Learning is unconstrained when we allow the household agent to explore the state and action spaces for long periods without imposing stopping criteria. The agent is assumed to have learned the RE solution for a given number of learning steps if all of its action values are at the policy regime's steady state values at the ends of the test episodes. Normalized learning outcomes for inflation, bond holdings, and hours worked after 1,500,000 learning steps are shown in the upper part of Figure 5. Normalization divides all values by their corresponding steady state values and subtracts one.

In this and all following figures, the horizontal line is the RE benchmark and the hatched areas (if shown) mark different learning phases. All learning curves are constructed using the averages of the last 50 state transitions within the ten test episodes that occur at each learning interval. Solid lines are the means across test episodes, after taking moving averages of 10 learning-test cycles, or 100,000 learning steps, to focus on learning trends.

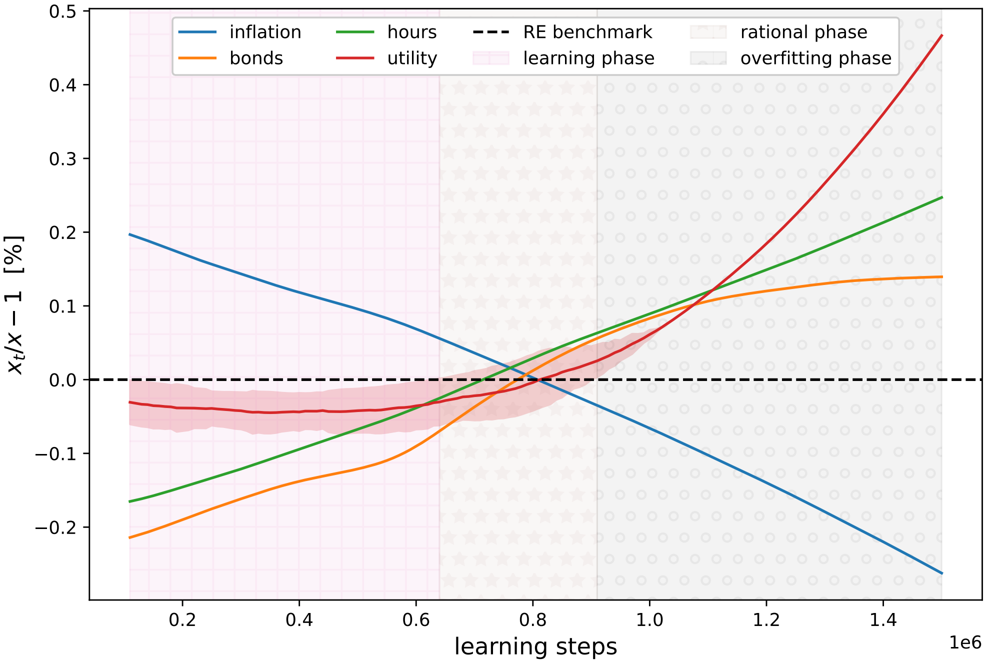
There can be volatility in learning outcomes across the test episodes due to the randomization of their initial state values. This is captured by the shaded areas that represent three standard deviations, approximating 99% confidence intervals. This variation is negligible for most quantities, but may not be for money and utility, with the latter guiding learning. The reason is that money and utility, being the product of a state transition, absorb all sources of uncertainty.

If the rational expectations solution is learned, we expect all lines to intersect with the horizontal line at about the same time. This occurs after about 750,000 steps. We define this rational phase (star hatches) as the interval of learning steps where the confidence intervals of utility realizations overlap with the horizontal line. This is marked by checkered hatches. However, the agent does not stay at the rational expectations solution when the duration of learning is extended, instead it diverges from it and does not return by the time 1,500,000 learning steps are reached.

To explore whether this divergence is permanent or temporary, we extend learning to 10,000,000 steps in the lower part of Figure 5. We observe that the household agent moves in long cycles around the steady state.[12] We label this the overfitting phase, which is marked by circled hatches.[13]

[12] It is not clear whether the leveling off at about 7,000,000 learning steps is due to a numerical breakdown of the learning algorithm or potentially an especially long cycle.

[13] Overfitting in the DRL context is not associated with fitting to noise but with a deterioration of generalization performance (Zhang et al. (2018)). We do not test this in the current work, but rather refer



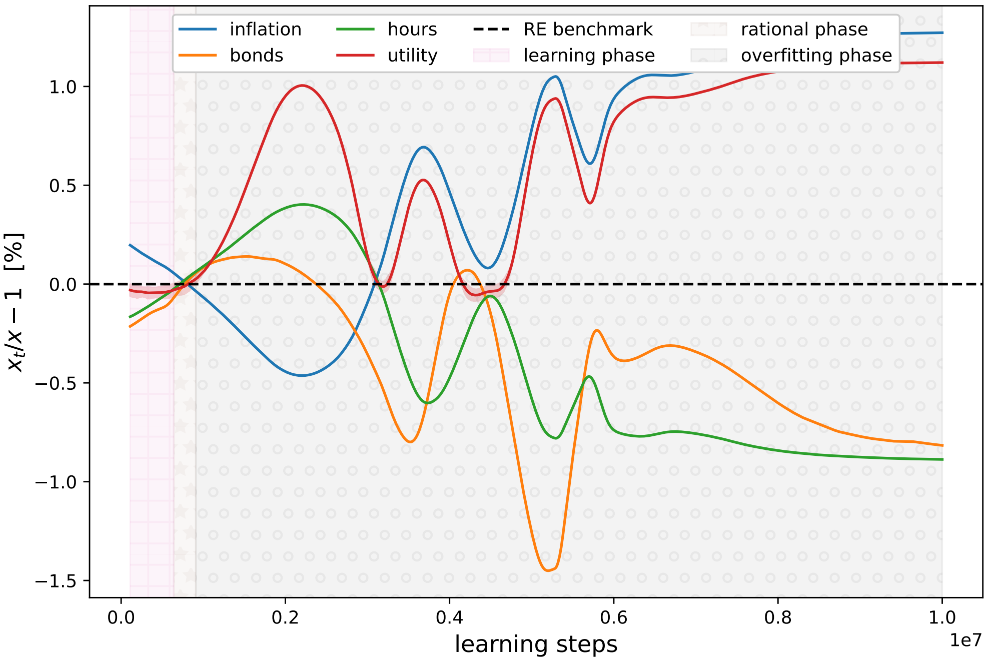


Figure 5 Learning under Unconstrained DRL in the AMP-PFP Regime. Upper part: 1,500,000 learning steps. Lower part: 10,000,000 learning steps.

The reason for this fragility is that the agent continuously tries to improve on the rational expectations solution while being unaware of its existence. This reflects the fact that the agent is not guided by the dynamics of the linearized system but rather by its actual preferences as encoded in its utility function.

The initial distances from steady state equal around 0.2% in the upper part of Figure 5, which compares to action bounds in Table A.1 of 0.5%-1.0%. We will allow for much larger distances and action bounds in the global learning settings of Section 3.5.

3.3.2 Constrained DRL

We study the same learning problem as in the previous section, but we now apply early stopping as described Section 2.5. Once an individual action such as hours worked enters its absorbing area, the household's actions remain fixed at the steady state value, and learning stops entirely when all action values have attained their steady state values. The convergence of household actions is shown in Figure 6 for the AMP-PFP regime.[14] In this figure the vertical axis shows the absolute distance to steady state relative to the maximal distance observed until convergence. This allows for a uniform representation and the comparison of learning for different state variables. We again indicate the learning and rational phases, the latter being absorbing under early stopping.

Compared to Figure 5, the learning curves for inflation and bonds in Figure 6 show kinks when the household's choice of hours triggers its early stopping criterion. This is due to the interactive nature of artificial neural networks with respect to input processing. We generally cannot identify the network weights corresponding to particular actions, and therefore cannot stop the updating of only such weights when a stopping criterion is reached for one of several actions. Instead we override the corresponding action with its steady state value. This constitutes a discontinuous change in behavior, to which learning subsequently adapts. The degree to which this behavior can be problematic will depend on the application.[15] We are here concerned with learnability, for which the shape of learning curves is less relevant. We see that all actions converge to their respective steady state values after about 1,200,000 learning steps.

to overfitting as optimization beyond the known optimal solution.

[14] As before, this represents the moving average of 100,000 learning steps averaged over the 50 last states of test episodes, and the confidence bands mark three standard deviations.

[15] Ways to address this include smoothing the early stopping criterion or its implementation. We will later see that noise in the learning process has precisely this effect.

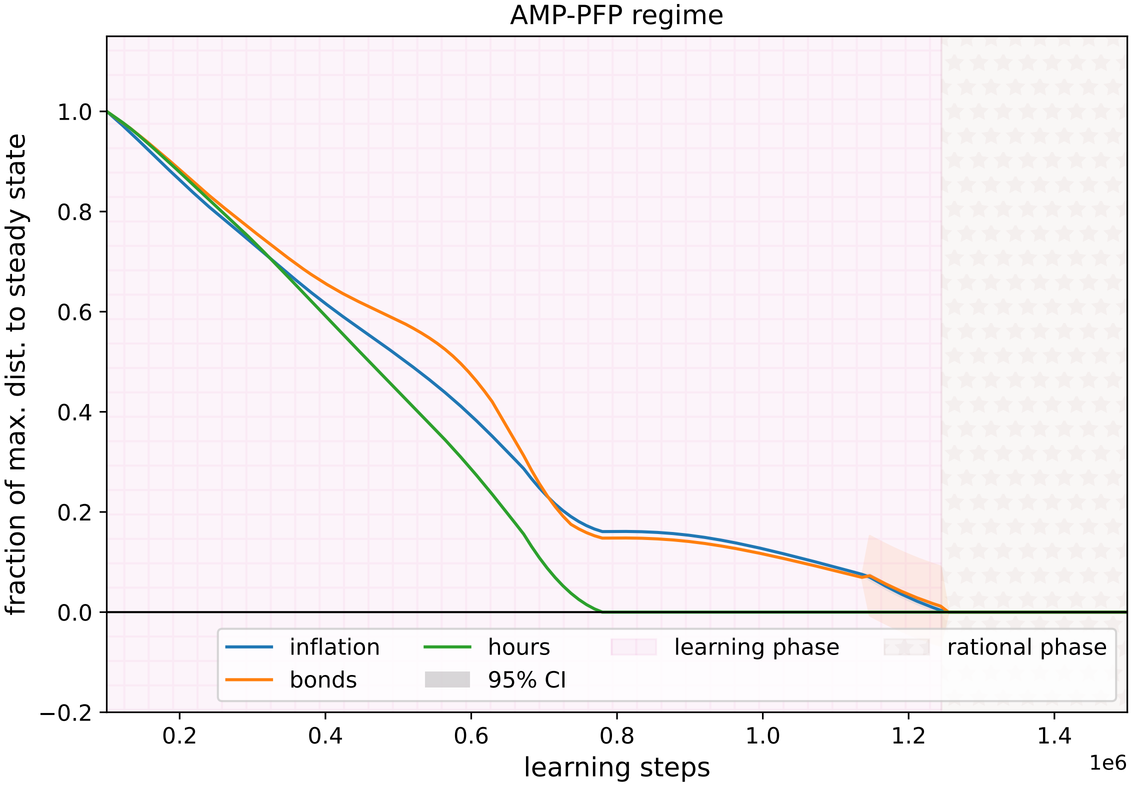


Figure 6 Learning with Early Stopping in the AMP-PFP Regime

A notable feature of DRL in this setting is its slow convergence of hundreds of thousands of steps.[16] These results can be interpreted in a social learning context (Bandura 1971, DeGroot 1974, Mobius & Rosenblat 2014). In the simplest social learning setting the actual learning time in quarters is the number of steps taken by the representative agent divided by the underlying population size. The idea is that everybody observes everybody else's state transitions or an informative summary of them and learns accordingly. When speaking of a country, this implies much faster learning. Specifying mechanisms through which such social learning takes place is beyond the scope of the current study. However, it is worth noting that experimentation (exploration) and learning from others have been documented in the social learning context. However, as a caveat, social learning does not necessarily lead to convergence or stability (Kirman 1993, Bikhchandani et al. 2024).

The action learning curves with early stopping for all four regimes are shown in Figure 7. We observe that learning dynamics is qualitatively similar for all four policy regimes, and that all four are learnable by DRL. This means that the dynamic stability properties

[16] The precise number of steps is a function of the algorithm used and of its parameterization. However, different choices do not change the number of required steps by orders of magnitude.

of the linearized model around its deterministic steady state are not necessarily a selection criterion if that steady state can be physically attained by the agent. Economically, this means that all policy regimes can become entrenched if the household spends enough time close to the corresponding steady state. More generally, this shows that DRL is a "global" solution technique in that it is able to retrieve multiple steady state solutions without the need to specify a localized approximation - a point we will discuss in more detail below.

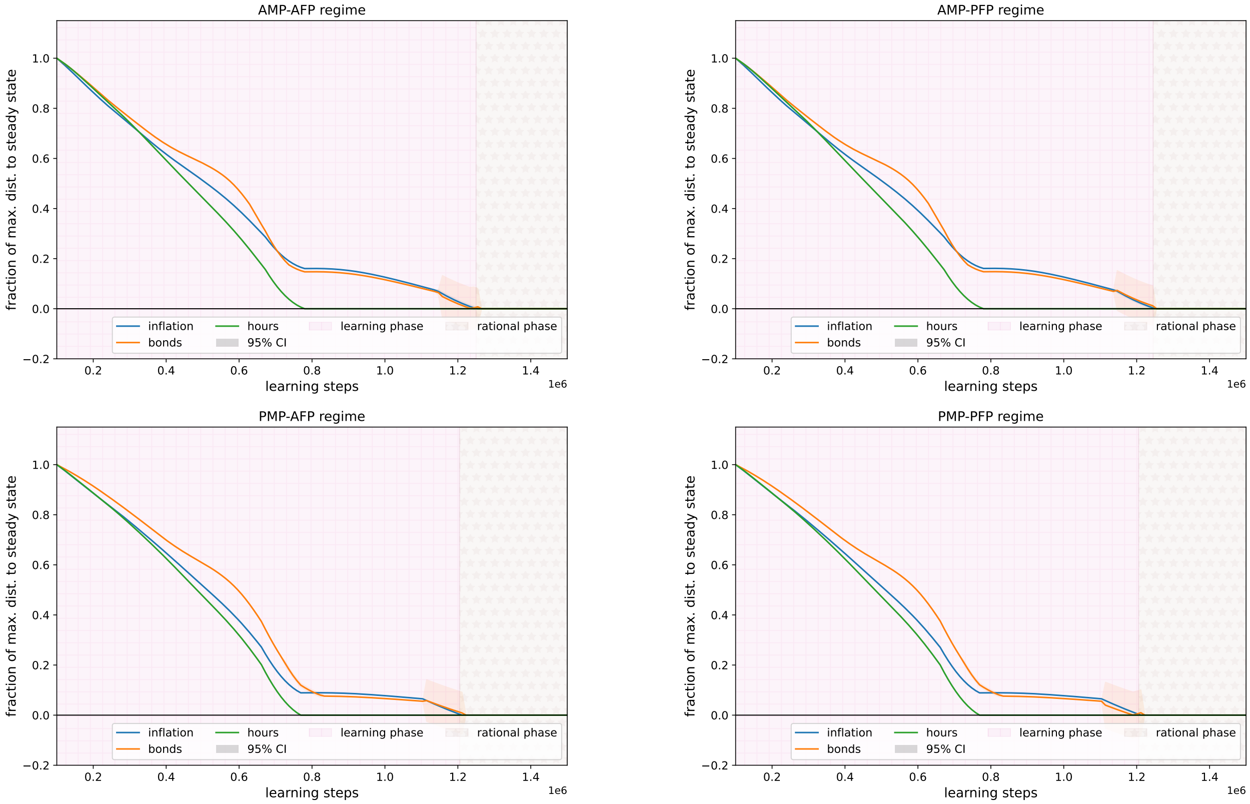


Figure 7 Learning with Early Stopping in All Four Regimes

3.4 Measuring Bounded Rationality

In this subsection we quantify bounded rationality by measuring how closely the agent's actions align with model-implied optimality conditions.

3.4.1 FOC-Learning

We gauge the rationality of the DRL agent at different stages of learning by quantifying the proximity of household actions and realized state variables to optimal or rational behavior. Specifically, we assess whether household actions are in line with the first-order conditions (13)–(15). To evaluate deviations in a standardized way, we divide a first-order condition by its left-hand side to obtain $FOC(x)$, subtract one, and take absolute values, to define the agent's absolute *FOC-distance* during learning:

$$d_x^{FOC} \equiv \left|FOC(x) - 1\right|. \tag{36}$$

A value of zero implies that the agent satisfies the corresponding first order condition. The explicit expression for the Euler equation (13), which we call the *Euler distance*, is

$$d_\pi^{FOC} = \left|\beta\,\mathbb{E}_t\big[\left(\frac{c_{t+1}}{c_t}\right)^{-\sigma}\frac{R_t}{\pi_{t+1}}\big] - 1\right|, \tag{37}$$

and similarly for equations (14) and (15). Expected values are set equal to next-period realized values. We label the process of minimizing measures of the form (37) as *FOC-learning*.

The normalized FOC-learning curves for the Euler equation, money demand and labor supply under the AMP-PFP regime are shown in Figure 8. The pattern is similar to the convergence of household actions in Figure 6. This is in line with the GPI framework of Section 2.4, in that there is joint convergence to the RE equilibrium of behavior (FOC-distances) and state value learning. This also means that equations such as (36) quantify *bounded rationality* in this class of models. We also see that any meaningful uncertainty comes from money demand, while it is negligible for the Euler equation or labor supply.

3.4.2 An Example: Household Inflation Expectations

We apply FOC-learning to household inflation expectations, which are not directly observable. Specifically, we investigate the relationship between the current interest rate R_t and next-period inflation π_{t+1} as determined by the household's consumption choice. Next-period inflation is $\mathbb{E}_t[\pi_{t+1}]$, where the expectations operator refers to agent expectations that are consistent with its own state of learning $\mathcal{P}_\phi$ and $\mathcal{Q}_\theta$ at time t.

We isolate the learned relation between inflation and interest rates by fixing the other actions (h_t, b_t) at their steady state values. That is, we set $h_t = h_{ss}\ \forall t$, which fixes $c_t = c_{ss}\ \forall t$ through goods market clearing, and determines π_t via c_t^{act} according to (32). The Euler equation (13) then simplifies to the Fisher equation

$$\mathbb{E}_t[\pi_{t+1}] = \beta R_t\,. \tag{38}$$

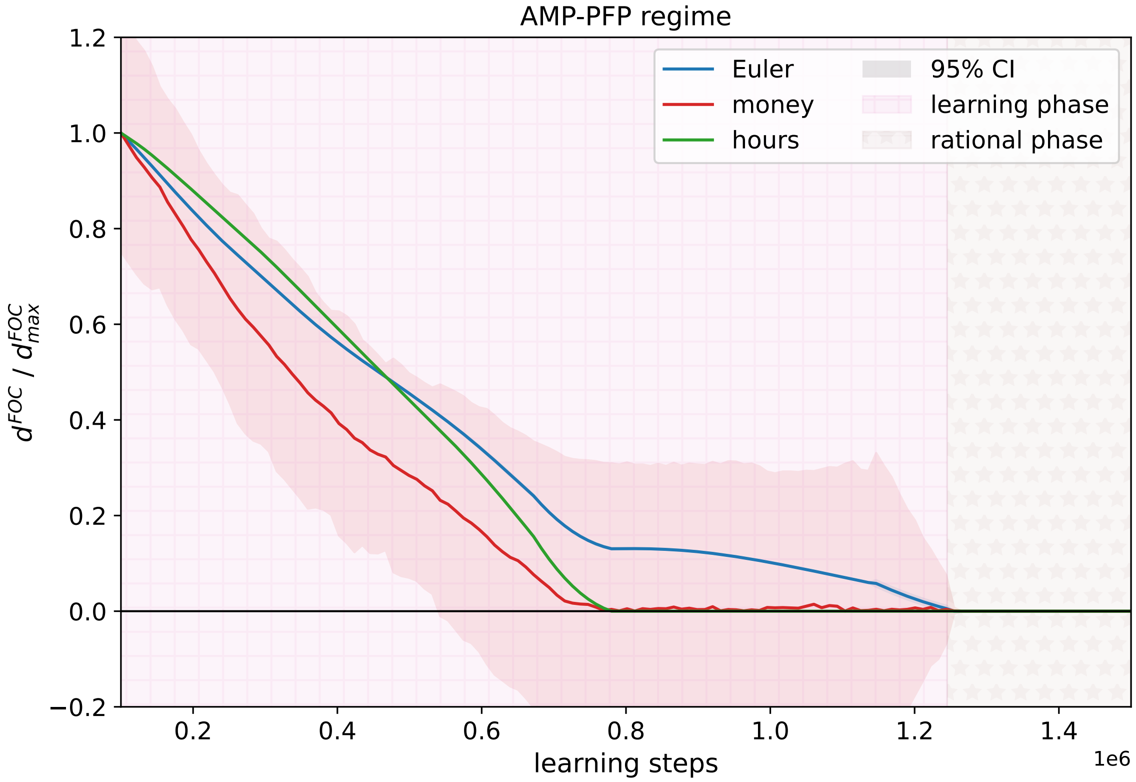


Figure 8 FOC-Learning in the AMP-PFP Regime

We reexamine household behavior at each test stage[17] by rerunning test cycles recording all state transitions until the early stopping criterion for inflation expectations is triggered. The results are shown in Figure 9. The horizontal axis shows net inflation expectations implied by the current interest rate R_t and the Fisher equation (38), and the vertical axis shows net inflation expectations measured through household consumption choices according to equation (31). The dotted diagonal line describes rational behavior. Household actions during each test transition are given by the scatter points at different times of learning as indicated by the color coding. The dashed line traces out household learning. The vertical distance between this line and the diagonal measures the deviation of expectations from rational expectations and is an explicit measure of bounded rationality.

We observe that implied inflation expectations at the start of learning (darker color) are about 0.2 percentage points below the optimal ones, during the learning phase expectations converge (lighter colors), and during the rational phase agent actions coincide with the Fisher equation (bright color). As in previous simulations, the starting position

[17] Algorithm 1 saves household policy rules at each test loop at different stages of learning, such that we can now reload the agent whose learning is only partially complete.

is a function of the experimental settings, which we do not investigate further. The initial divergence is again small because we have so far limited state and action spaces to areas of 0.5%-1% around the steady state (see Table A.1). We will analyze global learning next.

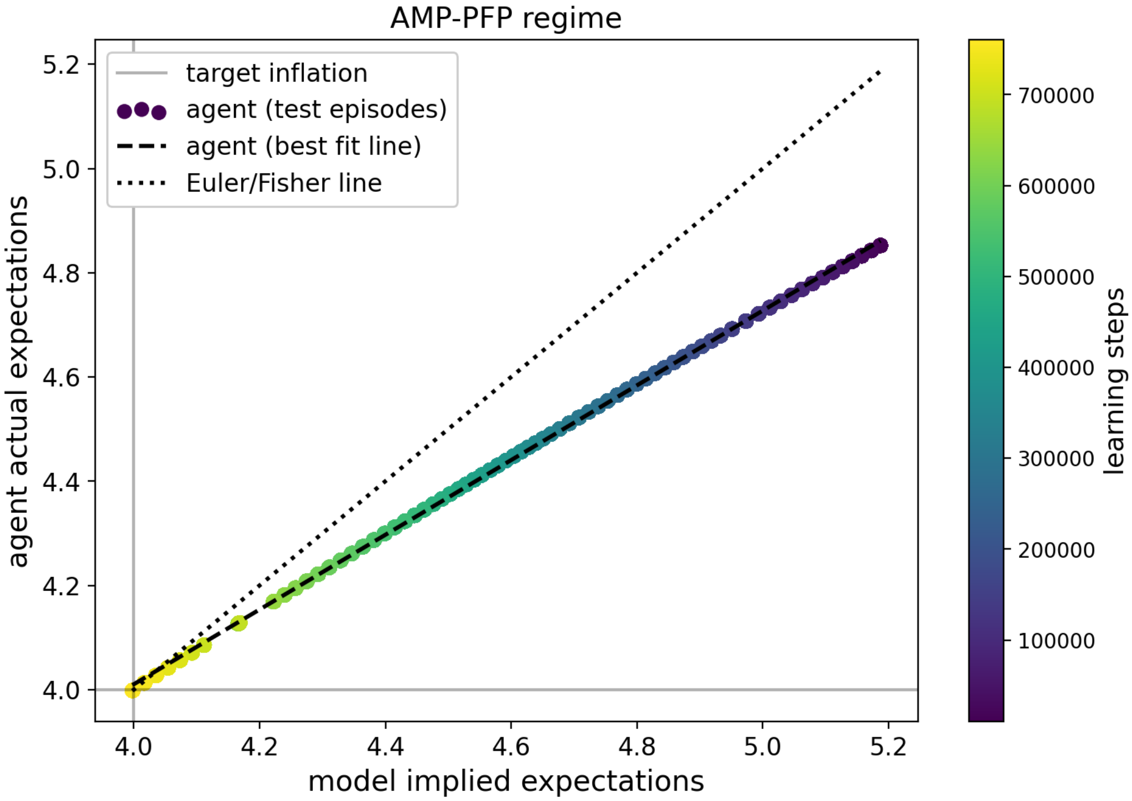


Figure 9 Evaluation of Fisher Equation (38) for the AMP-PFP Regime

3.5 Global Learning

DRL is a global solution method in that it maximizes the full objective function without reference to local approximations while incorporating exploration to mitigate the risk of getting stuck at a local extremum. However, we have so far not exploited this fully because we have only considered limited state and action spaces in the neighborhood of one of the RE steady states. We now test the household's ability to learn in global settings, by considering state and action spaces which include both of the model's steady states roughly symmetrically with some distance to the boundaries, for both active and passive fiscal policies.[18] The question is whether the household learns one of the steady states, and

[18] The monetary policy regime is determined by the location in the state space around the inflation steady states, while fiscal policy is determined by the parameter γ.

if so, which one and at what speed, and whether this depends on the fiscal policy regime. The results are summarized in Figure 10, where early stopping is implemented using the same criteria as before for each of the steady states separately.[19]

We find that the household agent learns the high-inflation steady state, with a strong monetary policy response to inflation, in both fiscal policy regimes, despite the fact that this has a lower utility than the low-inflation solution. In other words, the economy converges to the AMP-AFP or AMP-PFP regimes. This suggests that monetary policy can have stabilizing effects across large regions of the state space. However, it now takes significantly longer until the agent converges to the steady state, between 5,000,000 and 10,000,000 learning steps. This means that, despite eventual learning, the agent can spend a long time away from either steady state.

Interestingly, learning is considerably faster in the passive fiscal policy regime. The feedback provided by taxation seems to be a useful signal to the household, even though this feedback only arrives through money and utility realizations. Economically, the interpretation is that monetary dominance may support economic stabilization in this setting.

4 Robustness Analysis

The outcomes of reinforcement learning are often sensitive to parameters of the environment or of the optimization algorithm. Here we discuss four such quantities, learning rate, memory size, uncertainty, and initialization.

4.1 The Learning Rate

The learning rate ζ_{learn} controls the step size for updating the network weights $(\mathcal{P}_\phi, Q_\theta)$. While higher values enable faster convergence, excessively large values may prevent the agent from finding optimal solutions or even cause complete learning failure.

We find that our learning problem is sensitive to the learning rate. The default value provided in Raffin et al. (2021) for the implementation of the SAC algorithm (Haarnoja et al. 2018) is $3e-4$. We lowered this in steps of one order of magnitude until we observed convergence at a value of $\zeta^0_{learn} = 3e-8$. Furthermore, we implement learning rate decay that is dependent on the number of the current learning interval. That is, with $N_{learn} = 1,500,000$ and $N_{interval} = 10,000$, we have $n^{max}_{learn} = N_{learn}/N_{interval} = 150$, $n_{learn} = steps/N_{interval}$ (see Algorithm 1), and $\zeta_{learn} = \zeta^0_{learn}/n_{learn}$.

[19] The state and action bounds for the global learning experiments are given in Table A.2. Note that each fiscal policy regime has two steady states corresponding to π^* and π_L.

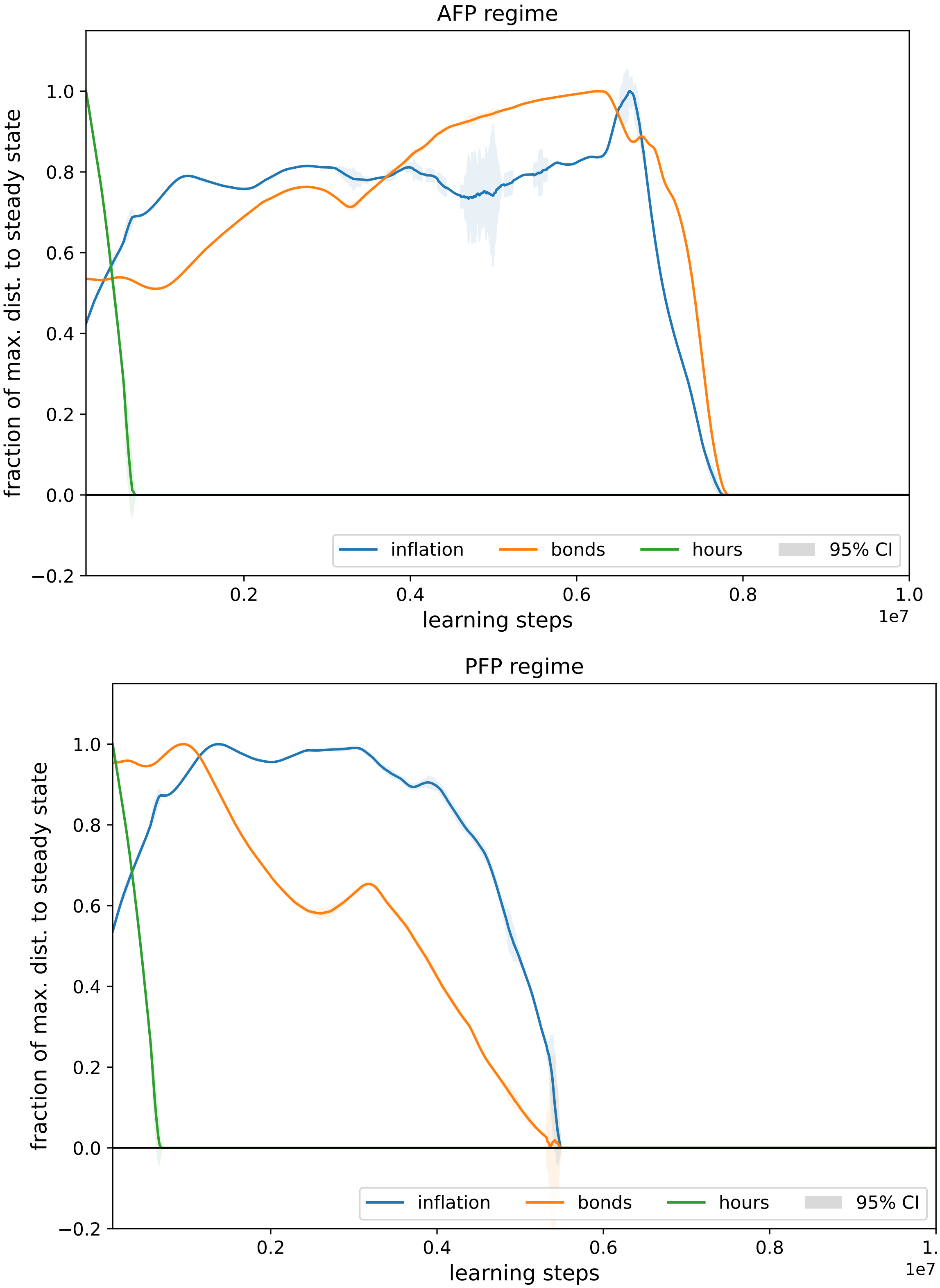


Figure 10 Global Learning with Early Stopping
(the high-inflation steady state is used as the reference point for distance measures)

There are two reasons for the observed sensitivity of learning outcomes to the learning rate. First, household utility (3) is quite flat, especially with respect to changes in money holdings. Second, the optimization problem is constrained, while the household only observes an unbounded utility function. Both factors, which are common in the economics context, make the learning problem challenging. This means that, while compared to most problems in the computer science literature our problem (and many others in economics and finance) is relatively simple in terms of the complexity of the state space, it nevertheless poses challenges for learning agents.

4.2 Memory Size

Learning is based on batches of observations that are sampled from a replay buffer that may be interpreted as the length of memory from which the agent can draw. Shorter memory can have the advantage of learning faster from new experience, while longer memory allows the agent to retain knowledge, which may be useful when there is a tendency to move away from good solutions as we saw in the unconstrained learning case of Section 3.3.1.

The default value for our experiments is $N_{mem} = 25,000$ state transitions. This is considerably shorter than the approximately 1,000,000 steps until convergence that we observe in Section 3.3.2. This raises the question of whether our convergence results in the absence of early stopping are due to an excessively small memory, which implies that the household agent easily 'forgets' about the rational expectations solution.

We test this by rerunning the experiment from section Section 3.3.1 while setting $N_{mem} = 1,000,000$. The results are qualitatively the same, with the household finding and later losing the optimal solution at about the same number of learning steps as in Figure 5. This suggests that the learning dynamics that we observe are robust to a wide range of memory sizes.

4.3 Uncertainty

We now repeat the exercises of Section 3.3.2 for all regimes, with the difference that uncertainty during learning is taken into account. Shocks are realized at each state transition during learning but not during testing. This is because we are interested in quantifying learning trends and under our maintained assumptions realized action values are unaffected by i.i.d. shocks.[20] The agent again follows the learning protocol in Algorithm 1. The results are summarized in Figure 11.[21]

We again see that learning is possible and qualitatively similar for all four policy regimes. Interestingly, learning is faster overall in the presence of shocks. The reason is that the kink in learning behavior after the hours worked early stopping is triggered largely disappears because the imposition of steady state hours is not distinguishable from noise from the point of view of the learning algorithm, resulting in a smoother adjustment.

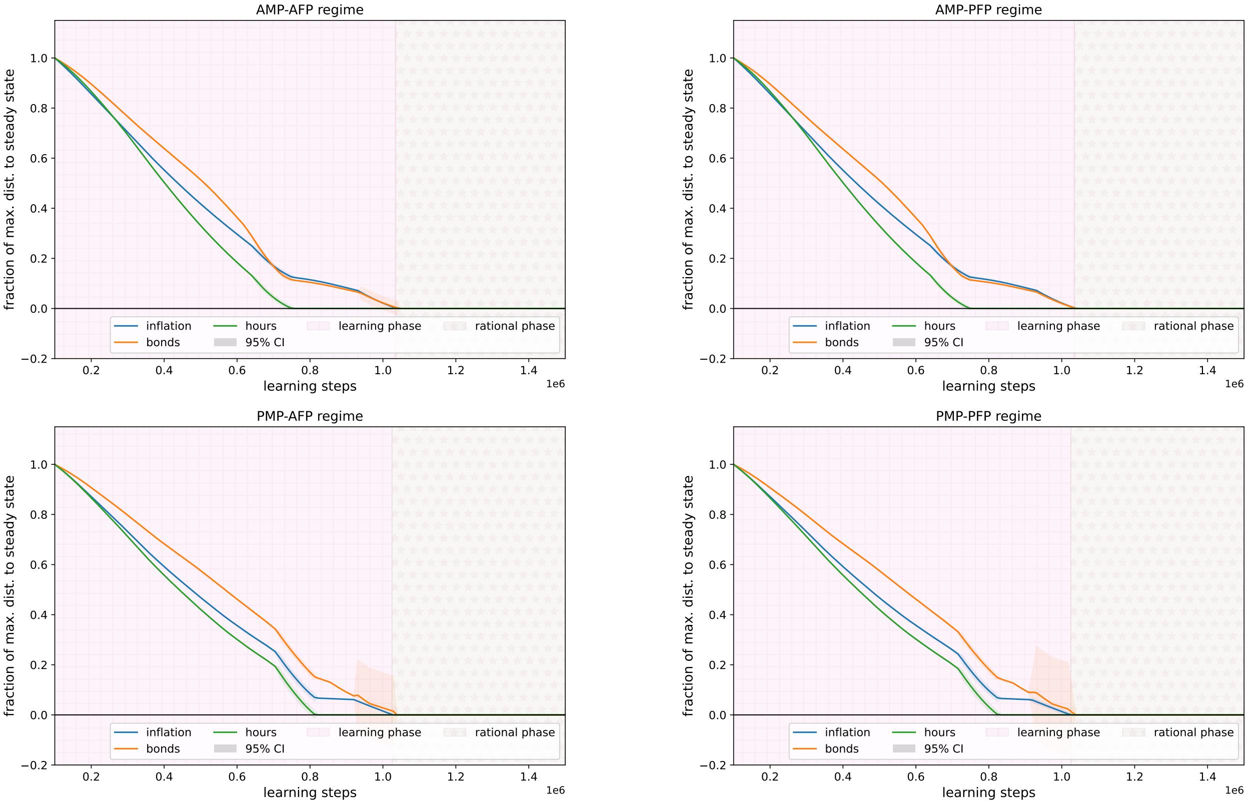


Figure 11 Learning with Early Stopping in All Four Regimes - with Shocks

[20] Shocks are assumed to be unobserved at the time actions are taken - see the step sequence for a single state transition in Section 2.3. Shocks can be made observable, in which case their realizations have to be included in the state vector.

[21] The shock standard deviations are shown in Table 1.

4.4 Initialization

All results so far have been based on a single random seed. However, results may vary depending on random initializations, both of the neural network weights and of the point in the state space where learning starts. To test the relevance and effects of random initializations, we now run our main experiment of learning in the AMP-PFP regime with early stopping for ten different random seeds.

The results of this exercise for the steady state learning of bond holdings and hours worked are shown in Figure 12. The different lines correspond to different seed values, where the value of 4 (purple line) is the default value used for all other results presented in this paper.

We see that the time to steady state convergence partly depends on the distance of the steady state values at the beginning of learning. However, some initializations, like seeds 2 and 8, imply considerably slower learning even though they do not start especially far away from the target. Some initializations may cross the steady state values at first. This happens when the algorithm is not aware of the distinctiveness of this point and crosses over it with high momentum, due to large learning updates, such that the early stopping criterion is not triggered.

These results suggest further brittleness in the use of DRL to solve macroeconomic models. However, they also can be used to simplify the learning setting using domain knowledge. In many situations, only a specific region of the state space may be of interest, and accounting for this in the choice of bounds then leads to more uniform learning dynamics.

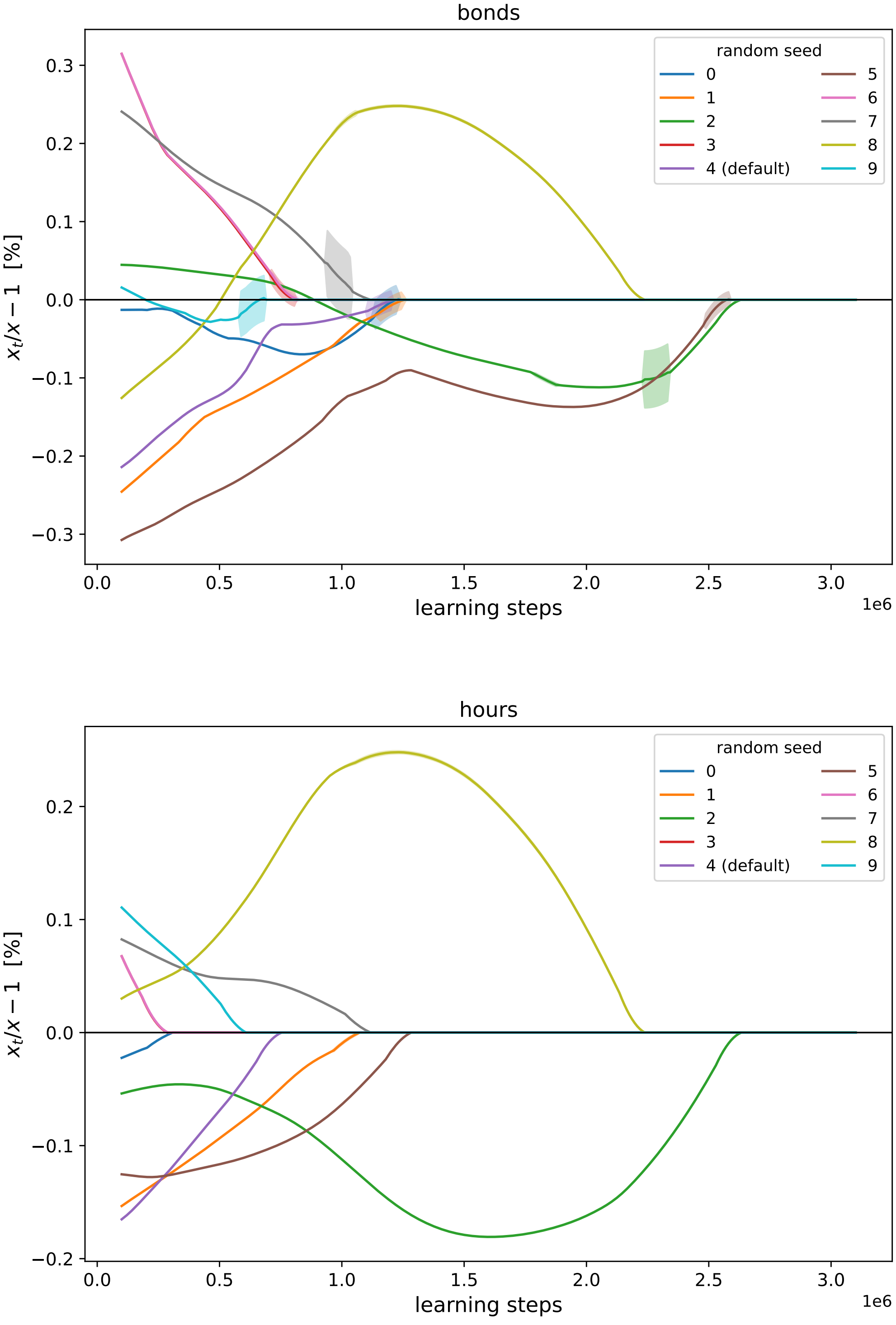


Figure 12 Testing the Impact of Different Random Seeds in the AMP-PFP Regime

5 Discussion

We explore the use of DRL to solve DSGE models. Unlike in traditional RE settings, where agents are assumed to have full knowledge of the model economy, DRL agents have no a priori knowledge except for their utility function. Instead, they learn by taking actions and observing the resulting utility rewards and state transitions. We apply DRL to a classical model from the adaptive learning literature, which studies the interaction of monetary and fiscal policy with a representative household agent. The model features two steady states, a high-inflation or inflation-target steady state and a low-inflation or liquidity-trap steady state.

Our study connects to a recent critique of the RE assumption (Moll 2024), which focused on heterogeneous agent models. We remain within the representative agent paradigm, and instead focus on studying whether DRL generates different outcomes from AL.

We find that DRL agents can learn steady states through utility maximization alone, without prior knowledge of the structure of the economy. Furthermore, unlike AL agents, DRL agents can learn locally unstable steady states. This is because they do not know of their existence as they are guided solely by utility realizations. Finally, in a global setting where the state space includes both the high-inflation and low-inflation steady states, the DRL agent converges to the high-inflation steady state irrespective of assumptions about fiscal policy, which suggests a relatively large basin of attraction for this solution.

While promising, DRL also poses challenges. Learning via deep artificial neural networks is sensitive to design choices such as learning rates, initializations, action and state space boundaries, and stopping criteria. Additionally, DRL agents can exhibit slow convergence to RE equilibria, and multi-agent settings can be difficult to implement. Moreover, DRL requires substantial computational resources and specialized programming skills, which may limit its practical application.

Agent expectations formation is one of the most important aspects of modern economic analysis. This study contributes to ongoing developments in the modeling of expectations formation, by demonstrating how DRL can offer new perspectives within representative agent models. DRL is a flexible and powerful technique that can be applied to a wide range of economic problems, provided they can be structured within its framework. This opens up a rich future research agenda.

References

Arifovic, J. (1995), 'Genetic algorithms and inflationary economies', *Journal of Monetary Economics* **36**(1), 219–243.

Bandura, A. (1971), *Social Learning Theory*, General Learning Press.

Benhabib, J., Schmitt-Grohe, S. & Uribe, M. (2001), 'The perils of taylor rules', *Journal of Economic Theory* **91**, 40–69.

Bikhchandani, S., Hirshleifer, D., Tamuz, O. & Welch, I. (2024), 'Information cascades and social learning', *Journal of Economic Literature* (3), 1040–93.

Blanchard, O. J. & Kahn, C. M. (1980), 'The solution of linear difference models under rational expectations', *Econometrica* **48**(5), 1305–1311.

Cybenko, G. (1989), 'Approximation by superpositions of a sigmoidal function', *Mathematics of Control, Signals, and Systems (MCSS)* **2**(4), 303–314.

DeGroot, M. (1974), 'Reaching a consensus', *Journal of the American Statistical Association* **69**(345), 118–121.

Eusepi, S. (2007), 'Learnability and monetary policy: A global perspective', *Journal of Monetary Economics* **54**, 1115–1131.

Eusepi, S. & Preston, B. (2018), 'The science of monetary policy: An imperfect knowledge perspective', *Journal of Economic Literature* **56**(1), 3–59.

Evans, G. W. & Honkapohja, S. (2001), *Learning and Expectations in Macroeconomics*, Princeton University Press.

Evans, G. W. & Honkapohja, S. (2005), 'Policy interaction, expectations and the liquidity trap', *Review of Economic Dynamics* **8**, 303–323.

Evans, G. W. & Honkapohja, S. (2007), 'Policy interaction, learning and the fiscal theory of prices.', *Macroeconomic Dynamics* **11**, 665–690.

Evans, G. W. & Honkapohja, S. (2008), 'Liquidity traps, learning and stagnation', *European Economic Review* **52**, 1438–1463.

Evans, G. W. & Honkapohja, S. (2009), 'Learning and macroeconomics', *Annu. Rev. Econ.* **1**(1), 421–449.

Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. (2016), *Deep learning*, Vol. 1, MIT press Cambridge.

Haarnoja, T., Zhou, A., Abbeel, P. & Levine, S. (2018), 'Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor', *arXiv-eprint* **1801.01290**.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, 2nd ed. edn, Springer.

Karwowski, J., Hayman, O., Bai, X., Kiendlhofer, K., Griffin, C. & Skalse, J. (2023), 'Goodhart's law in reinforcement learning'.

Kirman, A. (1993), 'Ants, rationality, and recruitment*', *The Quarterly Journal of Economics* **108**(1), 137–156.

Leeper, E. M. (1991), 'Equilibria under 'active'and 'passive'monetary and fiscal policies', *Journal of Monetary Economics* **27**(1), 129–147.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D. & Wierstra, D. (2015), 'Continuous control with deep reinforcement learning', *arXiv-eprint* **1509.02971**.

Liu, K., Wang, P., Wang, D., Du, W., Wu, D. O. & Fu, Y. (2021), Efficient reinforced feature selection via early stopping traverse strategy, *in* '2021 IEEE International Conference on Data Mining (ICDM)', IEEE, pp. 399–408.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. & Riedmiller, M. (2013), 'Playing atari with deep reinforcement learning'.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G. et al. (2015), 'Human-level control through deep reinforcement learning', *Nature* **518**(7540), 529–533.

Mobius, M. & Rosenblat, T. (2014), 'Social learning in economics', *Annual Review of Economics* **6**(1), 827–847.

Moll, B. (2024), ‘The trouble with rational expectations in heterogeneous agent models: A challenge for macroeconomics’, *London School of Economics, mimeo, available at https://benjaminmoll. com* .

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M. & Dormann, N. (2021), ‘Stable-baselines3: reliable reinforcement learning implementations’, *J. Mach. Learn. Res.* **22**(1).

Sargent, T. J. (1993), ‘Bounded rationality in macroeconomics: The arne ryde memorial lectures’, *OUP Catalogue* .

Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. (2017), ‘Proximal policy optimization algorithms’, *CoRR* **abs/1707.06347**.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. & Hassabis, D. (2016), ‘Mastering the game of Go with deep neural networks and tree search’, *Nature* **529**(7587), 484–489.

Sutton, R. & Barto, A. (2018), *Reinforcement Learning: An Introduction*, second edn, The MIT Press.

Woodford, M. (2013), ‘Macroeconomic analysis without the rational expectations hypothesis’, *Annu. Rev. Econ.* **5**(1), 303–346.

Xia, E., Khamaru, K., Wainwright, M. J. & Jordan, M. I. (2023), ‘Instance-dependent confidence and early stopping for reinforcement learning’, *Journal of Machine Learning Research* **24**(392), 1–43.

Zhang, C., Vinyals, O., Munos, R. & Bengio, S. (2018), ‘A study on overfitting in deep reinforcement learning’, *arXiv preprint arXiv:1804.06893* .

A Appendix

A.1 The Linearized Model

In the neighborhood of either steady state, our model can be described by a linear approximation for π_t and b_t of the form

$$\begin{bmatrix} \hat{\pi}_t \\ \hat{b}_t \end{bmatrix} = \mathbf{B} \begin{bmatrix} E_t \hat{\pi}_{t+1} \\ E_t \hat{b}_{t+1} \end{bmatrix} + \mathbf{C} \begin{bmatrix} \hat{\varepsilon}_t^R \\ \hat{\varepsilon}_t \tau \\ \hat{\varepsilon}_t^y \end{bmatrix}. \tag{A.1}$$

In this simple model, output (16) is exogenous, depending only on the technology shock. According to Blanchard & Kahn (1980), the solution to (A.1) is locally unique if and only if one eigenvalue is within the unit circle and the other eigenvalue is outside the unit circle. The two eigenvalues of the system (A.1) are given by $\frac{1}{\alpha\beta}$ and $\frac{1}{1/\beta-\gamma}$ (derivation given below). These eigenvalues are the inverses of the eigenvalues of the Blanchard-Kahn conditions. This formulation is common in the learning literature, with the expectations operator on the right-hand side.

When there is a non-stochastic steady state, it can be shown that a stochastic steady state exists in its neighborhood if the support of the exogenous shocks is sufficiently small. Furthermore, in this case the steady state is locally determinate, provided the corresponding linearization is determinate. Throughout the paper we assume that the shocks are small in the sense of having small support. Determinacy needs to be assessed separately for the two steady states at π^* and π_L. Following Evans & Honkapohja (2007), we can verify that in the linear system given by (A.1), if fiscal policy is passive, $|\gamma - \beta^{-1}| < 1$, the steady state π^* is locally determinate and the steady state π_L is locally indeterminate. If fiscal policy is active, $|\gamma - \beta^{-1}| > 1$, the steady state π^* is locally explosive and the steady state π_L is locally determinate.[22]

[22] With $\alpha = f'(\pi)$, it is easy to verify that at the higher steady state π^*, $|\alpha\beta| > 1$ and at the lower steady state π_L, $|\alpha\beta| < 1$. More details can be found in Evans & Honkapohja (2007), who prove that the linearization yields a locally unique asymptotically stationary rational expectations equilibrium if monetary policy is (locally) active and fiscal policy is passive, or if monetary policy is (locally) passive and fiscal policy is active.

In a neighborhood of a non-stochastic steady state π and c we derive the linear approximation

$$\text{Euler Equation:} \qquad \hat{R}_t = \beta^{-1} E_t \hat{\pi}_{t+1} + \frac{\sigma}{\beta} \frac{\pi}{c} (E_t \hat{c}_{t+1} - \hat{c}_t), \tag{A.2}$$

$$\text{Monetary Policy:} \qquad \hat{R}_t = \alpha \hat{\pi}_t + \delta \hat{\varepsilon}_t^R, \quad \text{where } \alpha = f'(\pi) \text{ and } \delta = f(\pi), \tag{A.3}$$

$$\text{Fiscal Policy \& GBC:} \quad \hat{b}_t + \hat{m}_t + \hat{\varepsilon}_t^\tau = \left(\frac{1}{\beta} - \gamma\right) \hat{b}_{t-1} - \frac{m + Rb}{\pi^2} \hat{\pi}_t + \frac{1}{\pi} \hat{m}_{t-1} + \frac{b}{\pi} \hat{R}_{t-1}, \tag{A.4}$$

$$\text{Money Demand:} \qquad \hat{m}_t = \frac{m}{c} \hat{c}_t - \frac{1}{\sigma} \frac{m}{R(R-1)} \hat{R}_t, \tag{A.5}$$

$$\text{Output:} \qquad \frac{\sigma + \varphi}{1 + \varphi} \frac{1}{c} \hat{c}_t = \hat{\varepsilon}_t^y. \tag{A.6}$$

Note that $\hat{x}_t$ denotes the deviation of variable x_t from steady state. To study determinacy, we rewrite (A.2)–(A.6) as a bivariate forward-looking system of the form

$$\begin{bmatrix} \hat{\pi}_t \\ \hat{b}_t \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} E_t \hat{\pi}_{t+1} \\ E_t \hat{b}_{t+1} \end{bmatrix} + \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \end{bmatrix} \begin{bmatrix} \hat{\varepsilon}_t^R \\ \hat{\varepsilon}_t \tau \\ \hat{\varepsilon}_t^y \end{bmatrix}. \tag{A.7}$$

According to Blanchard & Kahn (1980), the solution to (A.7) is locally unique if and only if one eigenvalue is within the unit circle and the other eigenvalue is outside the unit circle. To assess this we bring the above expressions into an explicit form

$$\begin{bmatrix} \frac{b\alpha}{\pi} - \frac{m\alpha}{\pi\sigma R(R-1)} & \frac{1}{\beta} - \gamma \\ \alpha & 0 \end{bmatrix} \begin{bmatrix} \hat{\pi}_t \\ \hat{b}_t \end{bmatrix} = \begin{bmatrix} \frac{m + \frac{1}{\beta}\pi b}{\pi^2} - \frac{m\alpha}{\sigma R(R-1)} & 1 \\ \frac{1}{\beta} & 0 \end{bmatrix} \begin{bmatrix} E_t \hat{\pi}_{t+1} \\ E_t \hat{b}_{t+1} \end{bmatrix} + \begin{bmatrix} \frac{m\delta}{\pi\sigma R(R-1)} - \frac{b\delta}{\pi} & 0 & -\frac{m}{\pi c \xi} \\ -\delta & 0 & -\frac{\sigma\pi}{\beta c \xi} \end{bmatrix} \begin{bmatrix} \hat{\varepsilon}_t^R \\ \hat{\varepsilon}_t \tau \\ \hat{\varepsilon}_t^y \end{bmatrix}, \tag{A.8}$$

where $\xi = \frac{\sigma + \varphi}{(1+\varphi)c}$. Therefore

$$\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} \frac{b\alpha}{\pi} - \frac{m\alpha}{\pi\sigma R(R-1)} & \frac{1}{\beta} - \gamma \\ \alpha & 0 \end{bmatrix}^{-1} \begin{bmatrix} \frac{m + \frac{1}{\beta}\pi b}{\pi^2} - \frac{m\alpha}{\sigma R(R-1)} & 1 \\ \frac{1}{\beta} & 0 \end{bmatrix}, \tag{A.9}$$

$$\begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \end{bmatrix} = \begin{bmatrix} \frac{b\alpha}{\pi} - \frac{m\alpha}{\pi\sigma R(R-1)} & \frac{1}{\beta} - \gamma \\ \alpha & 0 \end{bmatrix}^{-1} \begin{bmatrix} \frac{m\delta}{\pi\sigma R(R-1)} - \frac{b\delta}{\pi} & 0 & -\frac{m}{\pi c D} \\ -\delta & 0 & -\frac{\sigma\pi}{\beta c D} \end{bmatrix}. \tag{A.10}$$

The two eigenvalues are $\frac{1}{\alpha\beta}$ and $\frac{1}{1/\beta-\gamma}$. Then a unique solution takes the form:

$$\begin{bmatrix} \hat{\pi}_t \\ \hat{b}_t \end{bmatrix} = \begin{bmatrix} D_{11} & D_{12} & D_{13} \\ D_{21} & D_{22} & D_{23} \end{bmatrix} \begin{bmatrix} \hat{\varepsilon}_t^R \\ \hat{\varepsilon}_t\tau \\ \hat{\varepsilon}_t^y \end{bmatrix} \tag{A.11}$$

A.2 E-Stability

To examine the global dynamics of the system we return to the nonlinear model. The full rational expectations problem can be written as $F(\mathbb{E}y_{t+1}, \varepsilon_t) = y_t$ for state variables y_t and innovations ε_t. We replace rational expectations with point expectations in the model equations (13) and (7). As an example, we replace $E_t[c_{t+1}\pi_{t+1}^\sigma] = c_{t+1}^e {\pi_{t+1}^e}^\sigma$. This is a reasonable approximation for shocks with small bounded support. This leads to the nonlinear dynamic system F^e,

$$c_t = c_{t+1}^e \left(\frac{\pi_{t+1}^e}{\beta R_t}\right)^\sigma = (\varepsilon_t^y)^{\frac{1+\varphi}{\sigma+\varphi}}, \tag{A.12}$$

$$\chi^\sigma c_{t+1}^e \left(\frac{f(\pi_{t+1}^e)+1}{f(\pi_{t+1}^e)}\right)^{1/\sigma} + b_{t+1}^e + \gamma_0 + \gamma b_t + \varepsilon_{t+1}^\tau$$

$$= \chi^\sigma \frac{c_t}{\pi_{t+1}^e}\left(\frac{R_t - 1}{R_t}\right)^{-1/\sigma} + R_t \frac{b_t}{\pi_{t+1}^e}, \tag{A.13}$$

$$R_t - 1 = \varepsilon_t^R f(\pi_t). \tag{A.14}$$

The dynamics for π_t and b_t under learning is then given by equations (A.12)–(A.13) after substituting R_t using (A.14). Note that output c_t is exogenous. According to Evans & Honkapohja (2001), the local asymptotic stability of the ordinary differential equation

$$\frac{dx^e}{du} = EF_x^e(\pi^e, b^e, \varepsilon_t) - x^e\,, \tag{A.15}$$

again with $x \in \{\pi, b\}$, provides the relevant E-stability criterion for the stochastic model under steady state learning when the shocks are small. Here, u denotes notional time, and $EF_x^e(\cdot)$ is the mapping from the perceived law of motion to the corresponding actual law of motion. E-stability is determined by the Jacobian matrix of $EF_x^e(\cdot)$ at the steady state.

A.3 DRL Parameterization

parameter	AMP (π^*)	PMP (π_L)	description
action bounds			
c_{min}^{act}	1.005	1.000	minimal consumption choice
c_{max}^{act}	1.015	1.003	maximal consumption choice
b_{min}^{act}	4.000	3.965	minimal bond holdings
b_{max}^{act}	4.080	4.045	maximal bond holdings
n_{min}	0.990	0.990	minimal hours worked
n_{max}	1.010	1.010	maximal hours worked
initial state bounds			
m_{min}	1.670	2.010	minimal money holdings
m_{max}	1.750	2.110	maximal money holdings
b_{min}	3.960	3.960	minimal bond holdings
b_{max}	4.040	4.040	maximal bond holdings
c_{min}	0.995	0.997	minimal consumption
c_{max}	1.005	1.003	maximal consumption
π_{min}	1.005	1.000	minimal inflation
π_{max}	1.015	1.003	maximal inflation
n_{min}	0.990	0.990	minimal hours worked
n_{max}	1.010	1.010	maximal hours worked
learning algorithm			
ζ_{learn}^{0}	3.0e-8	3.0e-8	learning rate
d_{u}^{min}	1.0e-6	1.0e-6	utility difference (episode termination)
N_{learn}^{max}	1,500,000	1,500,000	learning steps (experiment)
$N_{interval}$	10,000	10,000	learning steps (between test episodes)
N_{test}	10	10	number of test episodes between learning intervals
N_{epi}^{max}	1,000	1,000	max. steps / episode (learning or testing)
N_{burn}	10,000	10,000	initial burn-in random actions
N_{mem}	25,000	25,000	max. memory of state transitions
N_{batch}	256	256	batch size for parameter updates
N_{layers}^{hidden}	(3,2)	(3,2)	number of hidden layers in $(\mathcal{P}_{phi}, Q_\theta)$
N_{nodes}^{hidden}	32	32	number of nodes in each hidden
δ_{ES}	1.0e-4	1.0e-4	early stopping threshold

Table A.1 Learning Parameters (Baseline)

parameter	AFP	PFP	description
		action bounds	
c^{act}_{min}	0.995	0.995	minimal consumption choice
c^{act}_{max}	1.015	1.015	maximal consumption choice
b^{act}_{min}	2.450	2.450	minimal bond holdings
b^{act}_{max}	4.150	4.150	maximal bond holdings
n_{min}	0.990	0.990	minimal hours worked
n_{max}	1.010	1.010	maximal hours worked
		initial state bounds	
m_{min}	1.660	1.660	minimal money holdings
m_{max}	2.110	2.110	maximal money holdings
b_{min}	2.450	2.450	minimal bond holdings
b_{max}	4.150	4.150	maximal bond holdings
c_{min}	0.990	0.990	minimal consumption
c_{max}	1.010	1.010	maximal consumption
π_{min}	0.995	0.995	minimal inflation
π_{max}	1.015	1.015	maximal inflation
n_{min}	0.990	0.990	minimal hours worked
n_{max}	1.010	1.010	maximal hours worked
		steady state values	
m	2.061	1.716	money holdings
π_L	1.001	1.001	inflation
b_L	2.610	4.000	bond holdings
π^*	1.010	1.010	inflation
b^*	4.000	2.582	bond holdings
c	1.000	1.000	consumption
n	1.000	1.000	hours worked

Table A.2 Learning Parameters (Global Learning). Other parameters are as in the Baseline.