

# Deep Reinforcement Learning in a Monetary Model

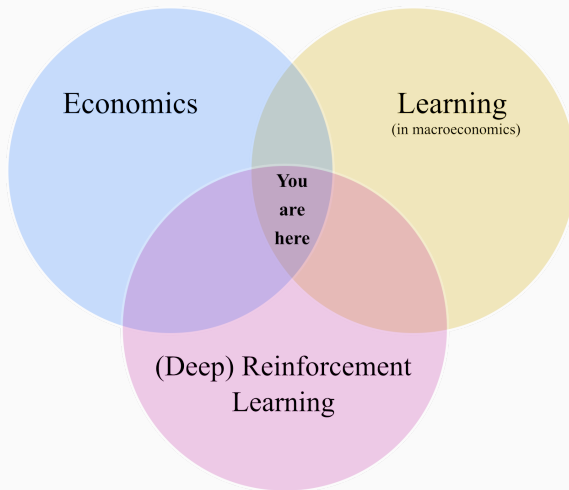
---

**Mingli Chen**

University of Warwick

Joint work with Andreas Joseph (Bank of England) and Michael Kumhof (Bank of England),  
Xinlei Pan (UC Berkeley), and Xuan Zhou (Reserve Bank of Australia)

Seminar at Fudan University, April 2025



# Table of contents

---

1. Overview & motivation
  2. Deep Reinforcement Learning (a.k.a. Artificial Intelligence)
  3. The Model Environment
  4. Results
- Appendix & References 41

## Overview & motivation

---

- Rational Expectation (RE) convenience choice to solve a model, but not necessarily how people and businesses actually behave
- Learning approach to **bounded rationality (BR)**: specify agent knowledge and behaviour away from RE (often ad hoc)
- BR broadens available state space. See also Moll (2024).
- Example: Monetary policy reaction functions possibly very different under learning, such as forward guidance or the stability of Taylor rules

See Eusepi and Preston (2018) and Hommes (2021) for recent reviews.

## Example: Adaptive learning

Agents are “econometricians” trying to estimate expected quantities via

$$x_{t+1}^e = x_t^e + \phi_t(x_t - x_t^e), \quad (1)$$

with a gain series  $\phi_t$ .

Under least-squares learning it is usually taken to be  $1/t$ . Together with the (optimal) behavioural rules, i.e. linearised FOCs, this leads to a set of ordinary differential equations determining the expectations (E-)stability of the model.

That is, if a steady state is **stable under learning**, which then serves as a selection criterion.

See, Sargent (1993) and Evans and Honkapohja (2001).

Models populated with *Adaptively Learning Agents* put the agents on an equal footing with the econometrician who is observing data from the model.

- However, this type of *parametric* recursive method assumes that agents correctly specify the laws of motion and other relevant functional relationships of the model

We work with models populated by *Deep Reinforcement Learning Agents (a.k.a. Artificially Intelligent Agents)* who

- have no a priori knowledge about the structure of the economy
- use their utility realisations in response to their actions in order to learn nonlinear decision rules via deep artificial neural networks

We adopt a policy-based deep reinforcement learning approach that can deal with high dimensional continuous action spaces.

Our approach enables agents to learn flexibly, as our learning algorithms are *nonparametric* and recursive, reducing the risk of misspecification

Allowing for misspecification and learning via expelling rational expectation agents and replacing them with “artificially intelligent” ones is also **reminiscent Sargent (1993)**



# Applications of Deep (Reinforcement) Learning in macroeconomics

- Global solution technique with no need of linearisation or other approximations
- Principled way to (bounded) rationality, i.e. agent behaviour and knowledge (this paper)
- General approach to heterogeneity, e.g. household income or age distribution (Hill et al., 2021).

⇒ Loads of potential applications in (macro)economics!

# Deep Reinforcement Learning (a.k.a. Artificial Intelligence)

---

# DRL at centre of recent advances in Artificial Intelligence

## ARTICLE

doi:10.1038/nature16961

### Mastering the game of Go with deep neural networks and tree search

David Silver<sup>1</sup>\*, Aja Huang<sup>1\*</sup>, Chris J. Maddison<sup>1</sup>, Arthur Guez<sup>1</sup>, L. Julian Schrittwieser<sup>1</sup>, Ioannis Antonoglou<sup>1</sup>, Veda Panneershelvam<sup>1</sup>, John Nham<sup>1</sup>, Nal Kalchbrenner<sup>1</sup>, Ilya Sutskever<sup>2</sup>, Timothy Lillicrap<sup>1</sup>, Thore Graepel<sup>1</sup> & Demis Hassabis<sup>1</sup>

### Playing Atari with Deep Reinforcement Learning

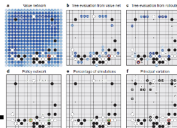
AUTOMATE EXPLORE CUSTOMIZE

**STRONG**  
Carry and power up to 14kg of impact on equipment.

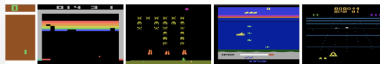
**EASY TO CONTROL**  
Control the robot from afar using an intuitive tablet application and built-in stereo cameras.



**SMART**  
Program repeatable autonomous missions to gather consistent data.



for Mnih Koray Kavukcuoglu David Silver Alex Graves Ioannis Antonoglou  
Daan Wierstra Martin Riedmiller

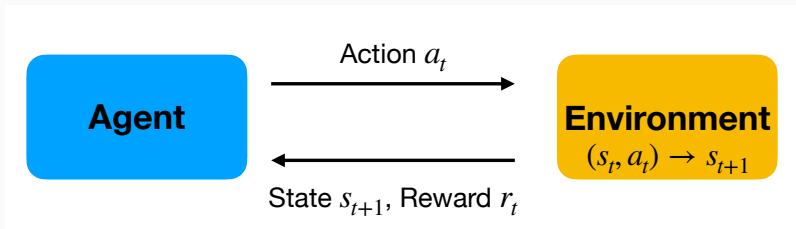


reen shots from five Atari 2600 Games: (Left-to-right) Pong, Breakout, Space Invaders  
am Rider

Sources: Nature, arXiv, Boston Dynamics



# The Reinforcement Learning (RL) setting



1. Agent observes state of the world  $s_t$
2. Agent takes actions  $a_t(s_t)$
3. Agent receives reward  $r_t$  from environment
4. Actions and state lead to state transition of the environment  $s_{t+1}$

This setting is very general. See Sutton and Barto (2018) for a comprehensive introduction.

## Formal RL definition

The agent aims to maximise expected cumulative lifetime reward, or **expected return**,

$$\max_{\mathcal{P}} \mathbb{E}_t[G_t] \quad \text{with} \quad G_t \equiv \sum_{k=0}^{\infty} \beta^k r_{t+1+k}(s), \quad (2)$$

following a **behavioural policy**  $\mathcal{P} : s_t \rightarrow a_t$ , with  $s_t \in \mathcal{S} \subset \mathbb{R}^{n_s}$  (state space) and  $a_t \in \mathcal{A} \subset \mathbb{R}^{n_a}$  (action space).

The **environment** the agents interaction with returns a reward and a new state, i.e.  $\mathcal{E} : (s_t, a_t) \rightarrow (s_{t+1}, r_t)$ , with  $r_t \in \mathbb{R}$ .

The **state transitions** is modelled as a Markov decision process (MDP)  $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \text{Pr}(s_{t+1}|s_t, a_t) \in [0, 1]$ .

**Problem:** Writing down  $\mathcal{T}$  is simple, knowing  $\mathbb{E}_t[G_t]$  and  $\text{Pr}(s_{t+1}|s_t, a_t)$  is hard (dynamic programming, value function iteration, etc.).

# State and action values

The expected return is maximised by finding the policy  $\mathcal{P}^*$ , which maximises the **values function**

$$V_{\mathcal{P}}(s) = \max_{a \in \mathcal{A}} \mathbb{E}_{\mathcal{P}} [G_t | s = s_t, a = a_t] \quad (3)$$

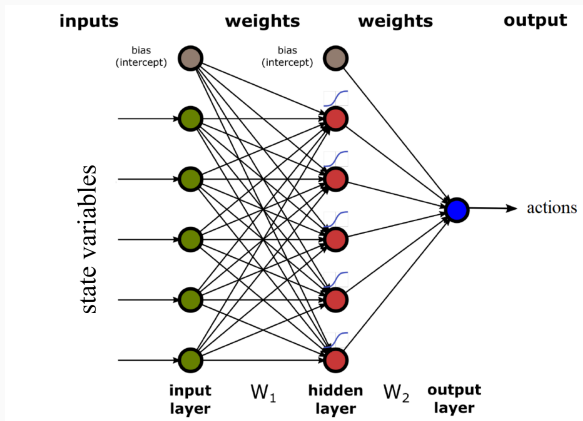
$$= \max_{a \in \mathcal{A}} Q(s, a), \quad (4)$$

with  $Q(s, a)$  the **state action value function**. We are done if we know  $\mathcal{P}^*$  and  $V^*/Q^*$ .

There are different ways to address this problem, which is an area of active AI research.

# Deep Learning + Reinforcement Learning = DRL

In DRL, functions  $\mathcal{P}$  and  $V/Q$  are parameterised using **deep artificial neural networks** (Goodfellow et al., 2016), i.e. neural nets with several hidden layers,  $\mathcal{P}_\phi$  and  $Q_\theta$ :



$\mathcal{P}$  and  $Q$  fulfil the **Bellman equation**

$$Q(s_t, a_t) = r(s_t) + \beta \mathbb{E}_{\mathcal{P}} [Q(s_{t+1}, a_{t+1})]. \quad (5)$$

using sampled state transitions as observations, i.e. interactions of the agent and the environment, and standard optimisation techniques like stochastic gradient descent, the policy and action-value function networks can be trained by iteratively minimising the Bellman residuum,

$$L(\phi, \theta) = \mathbb{E}_{s_t, a_t, r_t} \left[ \frac{1}{2} (Q_{\theta}(s_t, a_t) - \hat{Q}_{\theta}(s_t, a_t))^2 \right], \quad (6)$$

$$\text{with } \hat{Q}_{\theta}(s_t, a_t) = r_t(a_t, s_t) + \beta \mathbb{E}_{\mathcal{P}} [Q_{\theta}(s_{t+1}, \mathcal{P}_{\phi}(s_{t+1}))]. \quad (7)$$

We use Haarnoja et al. (2018). The code we used for optimisation is available at <https://github.com/pranz24/pytorch-soft-actor-critic>.



## General DRL setting for (macro)economics

- Write down model (environment and state)
- Specify **learning** agents, e.g. households, firms, etc., and their **actions**
- Specify state transitions as MDP
- Learning using DRL algorithm (e.g. Haarnoja et al. (2018)):
  1. sample state transition(s) and store in memory
  2. train  $\mathcal{P}_\phi$  and  $Q_\theta$  from memory
  3. test  $\mathcal{P}_\phi$  and  $Q_\theta$  with new state transitions and metric of choice

# Household learning protocol

---

**Algorithm 1** Training and testing protocol of household agent

---

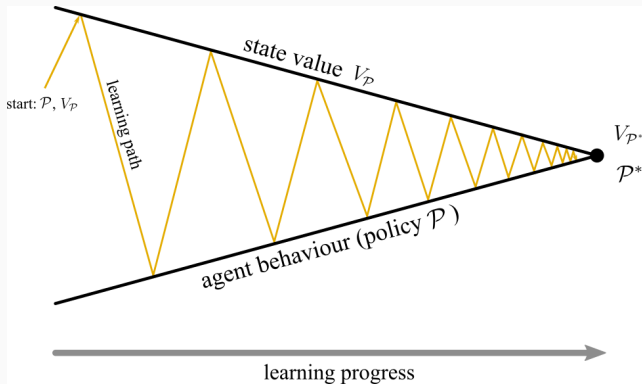
Initialise: Environment  $\mathcal{E}$  (parameterised model), agent (parameterised by  $\mathcal{P}_\phi$ ,  $Q_\theta$ )

```
for steps = 1 to  $N_{train}$  do
  initialise training episode with random state  $s_t$ 
  while training episode is not done do
    if steps  $\leq N_{burn}$  then
      Take allowed random action  $a_t$ 
    else
      Draw exploration action  $a_t = \mathcal{P}_\phi^{exp}(s_t)$ 
    end if
    Environment returns  $(r_t, s_{t+1}) = \mathcal{E}(s_t, a_t)$ 
    Add transition  $(s_t, a_t, r_t, s_{t+1})$  to memory
    Update  $\mathcal{P}_\phi$ ,  $Q_\theta$  using batch gradient descent from memory
    if  $mod(\text{steps}, N_{interval}) = 0$  then
      for test episode = 1 to  $N_{test}$  do
        Record state transitions (*)
      end for
      Save current agent  $(\mathcal{P}_\phi^{steps}, Q_\theta^{steps})$ 
    end if
    State update  $s_t \leftarrow s_{t+1}$ 
    Test episode termination criteria  $(N_{epi}^{max}, d_u^{min})$ 
  end while
end for
Save final agent  $(\mathcal{P}_\phi^{final}, Q_\theta^{final})$ 
```

---

# Generalised policy iteration (GPI)

GPI connects economics and learning, and conventional learning approaches with RL



$V^*$ : steady state values,  $\mathcal{P}^*$ : FOC.

## Examples of RL in economics and finance - a very selective literature

- Charpentier et al. (2020): Brief introduction to RL in a economics and finance background
- Zheng et al. (2020): Learning in large-scale geographic ABM
- Calvano et al. (2020): Investigate algorithmic collusion. AI can autonomously learn to set supracompetitive prices in a simulated oligopolistic market.
- Chaudhry and Oh (2020): Extract high-frequency expectations in financial markets to measure information effects
- Castro et al. (2021): Learn policy rules of banks participating in a high-value payments system

# The Model Environment

---

A single representative household maximises its expected lifetime utility, subject to an inter-temporal budget constraint:

$$\max_{c_t, m_t, n_t} \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t U(c_t, m_t, n_t) \quad \text{s.t.} \quad (8)$$

$$M_t + B_t + C_t = M_{t-1} + B_{t-1}R_{t-1} + W_t n_t - P_t \tau_t, \quad (9)$$

with  $P_t$  the price level at time  $t$ ,  $x_t = \frac{X_t}{P_t}$ ,  $x \in \{M_t, B_t, C_t, W_t\}$  relate real and nominal money, government bonds, consumption and wages, and  $\tau_t$  is a real lump-sum tax to the government each period.

We take the utility (Evans and Honkapohja, 2005)

$$U(c_t, m_t, n_t) = \frac{c_t^{1-\sigma}}{1-\sigma} + \chi \frac{m_t^{1-\sigma}}{1-\sigma} - \frac{n_t^{1+\varphi}}{1+\varphi}. \quad (10)$$

A single representative firm produces according to

$$y_t = \varepsilon_t^y n_t, \quad (11)$$

with technology (shock)  $\varepsilon_t^y$ , maximising profits

$$\max_{w_t} y_t - w_t n_t, \quad (12)$$

by setting setting the optimal wage

$$w_t = \varepsilon_t^y. \quad (13)$$

Markets clear every period, i.e.

$$y_t = c_t \quad (\text{goods}), \quad (14)$$

and

$$c_t^\sigma n_t^\varphi = \varepsilon_t^y \quad (\text{labour}). \quad (15)$$



The government issues interest-bearing bonds and non-interest-bearing currency (money), and collects taxes under the real inter-temporal *government budget constraint* (GBC)

$$m_t + b_t + \tau_t = \frac{m_{t-1}}{\pi_t} + R_{t-1} \frac{b_{t-1}}{\pi_t}, \quad (16)$$

subject to the transversality condition

$$\lim_{j \rightarrow \infty} \prod_{k=0}^j \left( \frac{\pi_{t+k}}{R_{t+k-1}} \right) b_{t+j} = 0. \quad (17)$$

*Fiscal policy* takes the linear tax rule as in Leeper (1991)

$$\tau_t = \gamma_0 + \gamma b_{t-1} + \varepsilon_t^\tau, \quad (18)$$

where  $\varepsilon_t^\tau$  is an exogenous random shock that is assumed to be i.i.d. with mean zero, and  $0 \leq \gamma \leq \beta^{-1}$ . We follow Leeper (1991) to define fiscal policy as being **active** if  $\gamma < \beta^{-1} - 1$  (AFP) and **passive** if  $\gamma > \beta^{-1} - 1$  (PFP).

We follow Benhabib et al. (2001) and Evans and Honkapohja (2005) with a global non-linear interest rate rule

$$R_t - 1 = \varepsilon_t^R f(\pi_t) \quad (\textit{Taylor rule}), \quad (19)$$

with  $f(\pi)$  assumed to be **non-negative and nondecreasing**, and  $\varepsilon_t^R$  is an exogenous, i.i.d. and positive random shock with a mean of one:

$$f(\pi_t) = (R^* - 1) \left( \frac{\pi_t}{\pi^*} \right)^{\frac{AR^*}{R^* - 1}}, \quad (20)$$

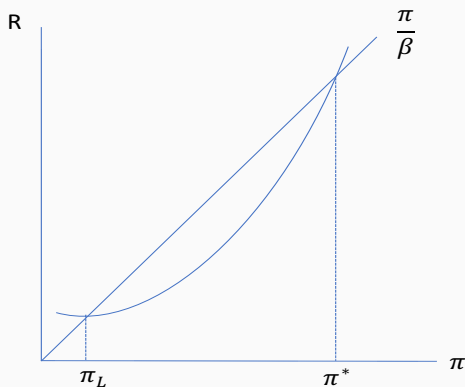
where  $A > 1$ , and  $\pi^* > 1$  is the inflation target.

## Steady states

The Taylor rule (19) implies **two steady states** at the intersection with the Euler/Fisher equation

$$\frac{\pi}{\beta} = 1 + (R^* - 1) \left( \frac{\pi}{\pi^*} \right)^{\frac{AR^*}{R^* - 1}}. \quad (21)$$

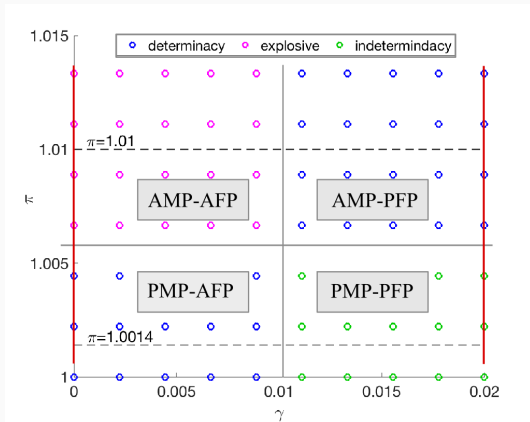
Monetary policy (MP) is said to be **active** at  $\pi^*$  ( $f'(\pi_t) > 1$ ; AMP) and **passive** at  $\pi_L$  ( $f'(\pi_t) < 1$ ; PMP).



This situation is very general and commonly investigated in learning in macroeconomics.

# Policy regimes

Using a standard parameterisation and local stability analysis we obtain four policy regimes



	AMP ( $\pi^*$ )		PMP ( $\pi_L$ )	
	PFP	AFP	PFP	AFP
$\pi_{ss}$	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
$\gamma_0$	-0.0566	0.0234	-0.0426	0.0375

# Joining the model and RL

## State representation

$$s_t = \left( m_{t-1}, b_{t-1}, \pi_{t-1}, c_{t-1}, n_{t-1}, \epsilon_t^\tau, \epsilon_t^R, \epsilon_t^y \right). \quad (22)$$

## Household agent actions

$$a_t = (c_t^{act}, b_t^{act}, n_t), \quad (23)$$

where  $x_t^{act} = X_t/P_{t-1}$ ,  $x \in \{c, b\}$ . Information flow and market clearing

$$\pi_t = c_t^{act} / y_t, \quad (24)$$

$$c_t = c_t^{act} / \pi_t, \quad (25)$$

$$b_t = b_t^{act} / \pi_t. \quad (26)$$

**Model environment:** Production, market clearing, pricing, GBC, FP, MP.

**No** first-order conditions (FOC)

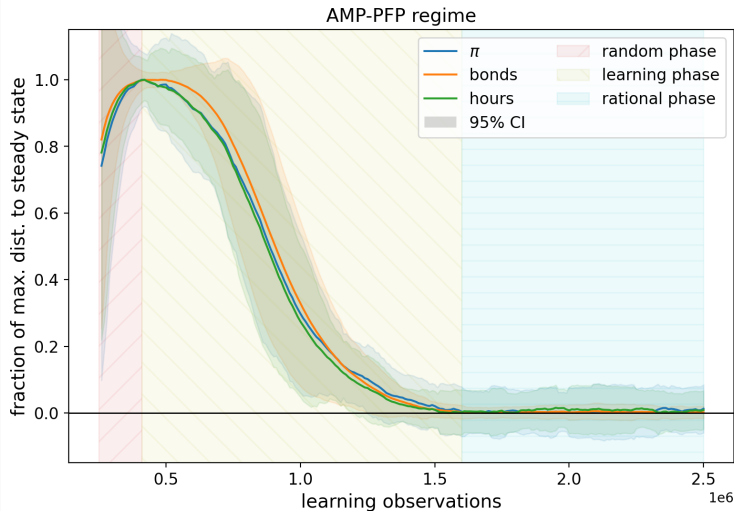
## State transition

1. Observe state  $s_t$
2. Take actions  $\mathcal{P}_\phi(s_t) = a_t = (b_t^{act}, c_t^{act}, n_t)$
3. Production (11) takes place and firm sets wages (13)
4. Markets clear: Inflation  $\pi_t$  is set by (24)
5. This determines real consumption  $c_t$  and bond holdings  $b_t$  (25)-(26)
6. Policy realisations:
  - The monetary authority sets the current gross interest rate  $R_t$  via the Taylor rule (19)
  - The government raises taxes  $\tau_t$  (18)
7. The money holdings  $m_t$  are realised from the GBC (16)
8. Agent obtains reward  $r_t = U(c_t, m_t, n_t)$
9. Next periods shocks are realised,  $(\epsilon_{t+1}^\tau, \epsilon_{t+1}^R, \epsilon_{t+1}^y)$
10. State update  $s_t \leftarrow s_{t+1} = (m_t, b_t, \pi_t, c_t, n_t, \epsilon_{t+1}^\tau, \epsilon_{t+1}^R, \epsilon_{t+1}^y)$

## Results

---

# Steady state learning in the AMP-PFP regime

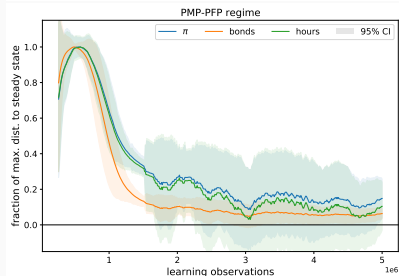
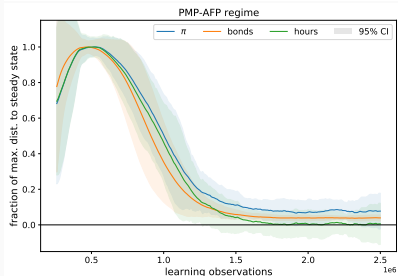
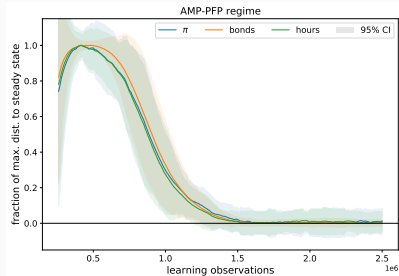
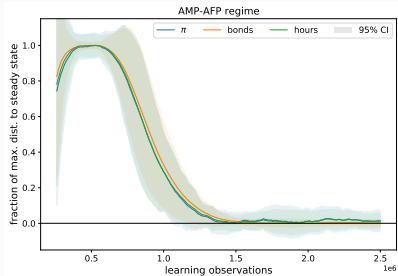


## Learning phases

- random (agent initialisation)
- learning
- rational



# Steady state learning in all regimes (charts)



## Steady state learning in all regimes (table)

	AMP ( $\pi^*$ )		PMP ( $\pi_L$ )	
	PFP	AFP	PFP	AFP
AL	yes	no	no	yes
RL	yes	yes	yes <sup>†</sup>	yes <sup>†</sup>
	$ \Delta_{ss} $ (%) for RL			
$\pi$	0.346	0.278	9.217	5.209
$b$	0.005	0.004	0.038	0.024
$n$	0.004	0.003	0.009	0.003
$m$	0.091	0.089	11.569	7.364
$u$	0.003	0.003	0.346	0.196

<sup>†</sup>imprecision in learning about inflation at  $\pi_L$ .

## Measuring bounded rationality

The household is said to behave rational if it follows FOC. During learning, we define the **FOC-distance** to measure deviations in a standardised way

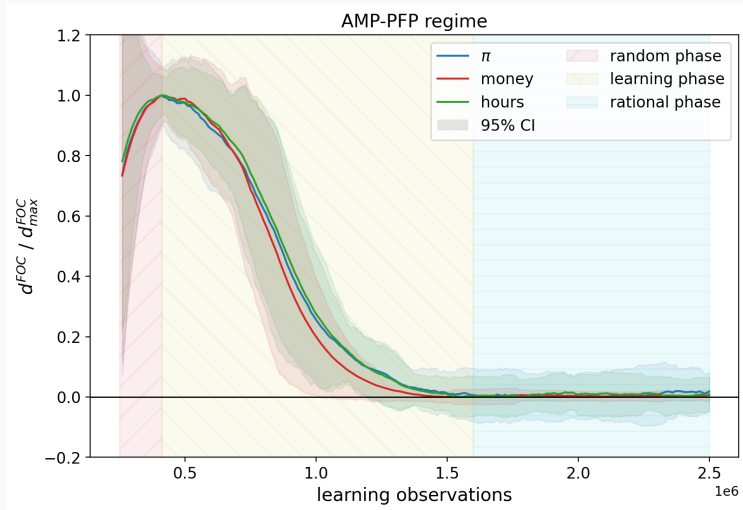
$$d_x^{FOC} \equiv |FOC(x) - 1|, \quad (27)$$

The explicit expression for the **Euler equation**, or *Euler distance*, is

$$d_\pi^{FOC} = \left| \beta \mathbb{E}_t \left[ \left( \frac{c_{t+1}}{c_t} \right)^{-\sigma} \frac{R_t}{\pi_{t+1}} \right] - 1 \right|. \quad (28)$$

FOC distances evaluate agent actions  $\mathcal{P}(s)$ . Analogous measures for  $V/Q$  can be derived with respect to state values.

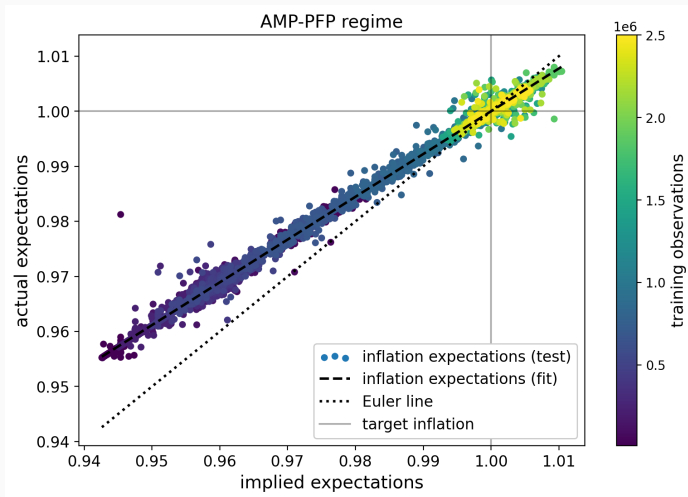
# FOC learning in the AMP-PFP regime



The same Learning phases  
as expected by GPI.

# Measuring inflation expectations during learning

Implied agent expectations can be extracted from realised values



- Improve and better understand learning **robustness**
- Aim for truly global learning
- Compare IRF with those from adaptive learning
- Conduct **experiments**: policy or regimes shifts

## Take-away messages

- DRL offers a **general approach** to solve structural macro models
- **Quantify bounded rationality** and learning in a principled way
- **Global** solution techniques which can also address heterogeneity
- **All policy regimes** are learnable under DRL
- **Promising toolbox** for (macro)economics
- Learning and convergence **technically challenging**

Thanks for listening

Q & A



## Model dynamic properties I

The deterministic steady states in the absence of random shocks is characterised by the following set of equations:

$$\text{Euler / Fisher Equation: } R = \frac{\pi}{\beta} \quad (29)$$

$$\text{Money Demand: } m = y \left( \frac{\pi - \beta}{\chi \pi} \right)^{-1/\sigma} \quad (30)$$

$$\text{Monetary Policy: } R = 1 + (R^* - 1) \left( \frac{\pi}{\pi^*} \right)^{\frac{AR^*}{R^* - 1}} \quad (31)$$

$$\text{Fiscal Policy \& GBC: } b = \left( \frac{1}{\beta} - 1 - \gamma \right)^{-1} \left[ \gamma_0 + \left( 1 - \frac{1}{\pi} \right) m \right] \quad (32)$$

$$\text{Output: } y^{\sigma + \varphi} = 1 \quad (33)$$

Equation (29) and (31) together determine the steady state of inflation:

$$\frac{\pi}{\beta} = 1 + (R^* - 1) \left( \frac{\pi}{\pi^*} \right)^{\frac{AR^*}{R^* - 1}} \quad (34)$$

## Model dynamic properties II

In the neighbourhood of either steady state, our model can be described by a linear approximation for  $\pi_t$  and  $b_t$  of the form

$$\begin{bmatrix} \hat{\pi}_t \\ \hat{b}_t \end{bmatrix} = \mathbf{B} \begin{bmatrix} \hat{E}_t \pi_{t+1} \\ \hat{E}_t b_{t+1} \end{bmatrix} + \mathbf{C} \begin{bmatrix} \hat{\varepsilon}_t^R \\ \hat{\varepsilon}_t^\tau \\ \hat{\varepsilon}_t^y \end{bmatrix}. \quad (35)$$

**Proposition:**(Evans and Honkapohja, 2007)] In the linear system given by (35),

- (i) If fiscal policy is passive,  $|\gamma - \beta^{-1}| < 1$ , the steady state  $\pi^*$  is locally determinate and the steady state  $\pi_L$  is locally indeterminate.
- (ii) If fiscal policy is active,  $|\gamma - \beta^{-1}| > 1$ , the steady state  $\pi^*$  is locally explosive and the steady state  $\pi_L$  is locally determinate.

# Model parameters

parameter	value	description
$\beta$	0.9900	discount factor
$\sigma$	3.0000	inverse of intertemporal elasticity of consumption and money holdings
$\varphi$	1.0000	inverse of Frisch elasticity of labor supply
$\chi$	0.1000	relative preference weight of money holdings
$\gamma_P$	0.0200	passive fiscal policy (PFP) coefficient
$\gamma_A$	0.0000	active fiscal policy (AFP) coefficient
$A$	1.3000	Taylor rule coefficient
$\pi^*$	1.0100	target gross high-inflation rate (4% net per annum)
$\pi_L$	1.0014	implied gross low-inflation steady state (see Figure ??)
$\epsilon_t^\tau$	0.0005	monetary policy shock (std. dev.)
$\epsilon_t^R$	0.0005	fiscal policy shock (std. dev.)
$\epsilon_t^y$	0.0005	technology shock (std. dev.)

Baseline model parameterisation. The shock series  $\epsilon_t^\tau$ ,  $\epsilon_t^R$ ,  $\epsilon_t^y$  follow log-normal, normal and normal distributions, with means of one, zero and one, respectively.

## References i

- Benhabib, J., Schmitt-Grohe, S., and Uribe, M. (2001). The perils of taylor rules. *Journal of Economic Theory*, 91:40–69.
- Calvano, E., Calzolari, G., Denicolò, V., and Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–97.
- Castro, P. S., Desai, A., Du, H., Garratt, R., and Rivadeneyra, F. (2021). Estimating Policy Functions in Payments Systems Using Reinforcement Learning. Staff Working Papers 21-7, Bank of Canada.
- Charpentier, A., Elie, R., and Remlinger, C. (2020). Reinforcement learning in economics and finance. Technical report.
- Chaudhry, A. and Oh, S. (2020). High-frequency expectations from asset prices: A machine learning approach. Technical report.
- Eusepi, S. and Preston, B. (2018). The science of monetary policy: An imperfect knowledge perspective. *Journal of Economic Literature*, 56(1):3–59.

- Evans, G. W. and Honkapohja, S. (2001). *Learning and Expectations in Macroeconomics*. Princeton University Press.
- Evans, G. W. and Honkapohja, S. (2005). Policy interaction, expectations and the liquidity trap. *Review of Economic Dynamics*, 8:303–323.
- Evans, G. W. and Honkapohja, S. (2007). Policy interaction, learning and the fiscal theory of prices. *Macroeconomic Dynamics*, 11:665–690.
- Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.
- Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv-eprint*, 1801.01290.
- Hill, E., Bardoscia, M., and Turrell, A. (2021). Solving heterogeneous general equilibrium economic models with deep reinforcement learning. Technical report.

- Hommes, C. (2021). Behavioral and experimental macroeconomics and policy analysis: A complex systems approach. *Journal of Economic Literature*, 59(1):149–219.
- Leeper, E. M. (1991). Equilibria under ‘active’ and ‘passive’ monetary and fiscal policies. *Journal of Monetary Economics*, 27(1):129–147.
- 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>*.
- Sargent, T. J. (1993). Bounded rationality in macroeconomics: The arne ryde memorial lectures. *OUP Catalogue*.
- Sutton, R. and Barto, A. (2018). *Reinforcement Learning: An Introduction*. The MIT Press, second edition.
- Zheng, S., Trott, A., Srinivasa, S., Naik, N., Gruesbeck, M., Parkes, D. C., and Socher, R. (2020). The ai economist: Improving equality and productivity with ai-driven tax policies.