Model-free and Model-based Learning as Joint Drivers of Investor Behavior

Nicholas Barberis
Lawrence Jin

Yale University and Caltech

June 2022

Overview

- in the past decade, psychologists and neuroscientists have increasingly embraced a new framework for thinking about human decision-making in experimental settings
  - work of Daw; Niv; Gershman; Dayan; O'Doherty...
- the framework combines two algorithms, or systems
  - model-free learning
  - model-based learning
- computer scientists have contributed significantly to the development of these algorithms
  - use them to solve complex dynamic problems
  - e.g. Backgammon and Go
- psychologists are also very interested in these algorithms
  - because of neural evidence that they reflect the brain's actual computations when evaluating different possible courses of action

Overview, ctd.

In this paper:

- we import this framework into a simple financial setting
- examine its properties and implications
- use it to account for a range of empirical facts about investor behavior

Overview, ctd.

Models

- model-free system
- a portfolio-choice setting
- model-based system
- hybrid system

Properties

- despite its simplicity, the model-free system has rich implications and delivers novel intuitions

Overview, ctd.

Applications

- extrapolative demand
- experience effects
- the disconnect between investor beliefs and investor allocations in both the frequency domain and the cross-section
- dispersion and inertia in investor allocations
- non-participation in the stock market
- persistent investment mistakes

Broader theme:

- try to make sense of investor behavior using a framework rooted in algorithms the brain appears to use when evaluating different courses of action

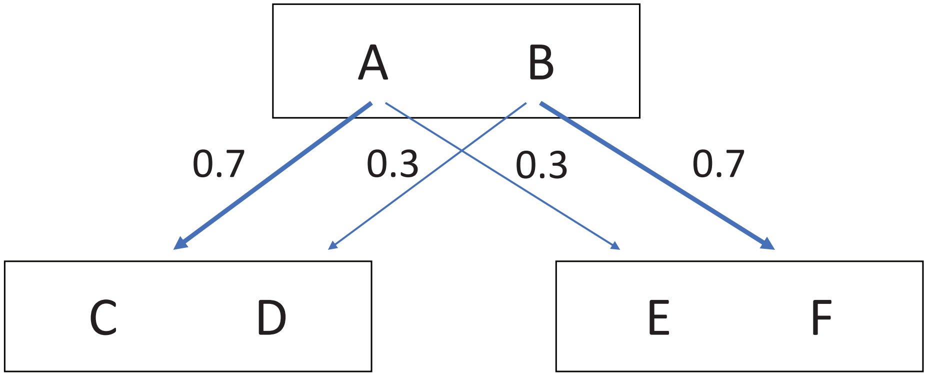
Overview, ctd.

- full name of “model-free learning” is “model-free reinforcement learning”
  - reinforcement learning has received much less attention in finance and economics than in psychology and neuroscience
  - closest antecedent in economics is in behavioral game theory
- model-based learning is closer to traditional frameworks in economics
  - novelty in this paper is model-free learning
  - and on how it compares to model-based learning

Psychological background

- psychologists have increasingly adopted a new framework for studying human decision-making in experimental settings
  - Daw, Niv, and Dayan (2005); Daw (2014)
- combines two algorithms, or systems
  - model-free learning
  - model-based learning
- the framework has found support in both behavioral and neural data
  - e.g., in the "two-step task" (Daw et al., 2011)

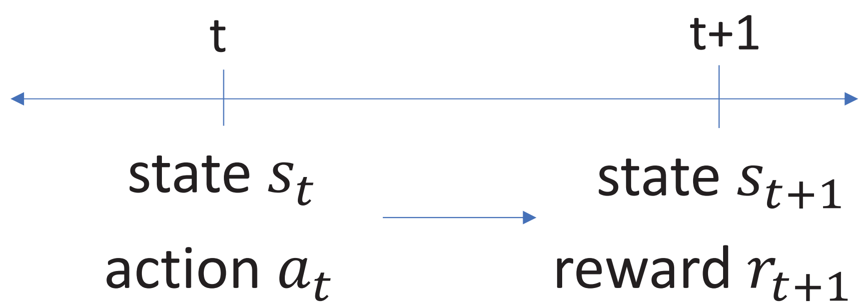
Psychological background, ctd.



- participant behavior in this experiment points to both model-free and model-based influences
  - as does neural activity

Models

- model-free and model-based algorithms are both intended to solve dynamic decision problems of the following form:



  - probability distribution $p(s_{t+1}, r_{t+1} | s_t, a_t)$ and Markov structure

- goal is to

$$\max_{\{a_t\}} E_0 \left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \right]$$

Models, ctd.

- economists almost always tackle problems of this kind using dynamic programming (DP)
  - and often use the DP solution to interpret observed behavior
- however, it is not clear how people would come to act according to the DP solution
- goal here: to explain observed behavior with a framework rooted in algorithms the brain appears to use when estimating the value of different courses of action

Models, ctd.

- there is growing evidence from psychology research that the way people tackle these problems is with a combination of model-free and model-based algorithms
- we discuss the model-free algorithm first
  - two prominent model-free algorithms that psychologists have focused on are Q-learning and SARSA
  - we work with Q-learning here

Model-free learning

- goal of both model-free and model-based approaches is to estimate $Q^*(s_t, a_t)$
  - the value of taking an action a_t at time t in state s_t, and then continuing optimally thereafter
- suppose that we take action a_t at time t in state s_t and then observe a reward r_{t+1} at time $t+1$ and land in state s_{t+1}
- the model-free algorithm updates its estimate of $Q^*(s_t, a_t)$ as follows

$$\begin{aligned} Q_{t+1}(s_t, a_t) &= Q_t(s_t, a_t) \\ &+ \alpha^{MF}[r_{t+1} + \gamma \max_{a'} Q_t(s_{t+1}, a') - Q_t(s_t, a_t)] \end{aligned}$$

  - the quantity in square brackets is the "reward prediction error" (RPE)
  - α^{MF} is the learning rate
- there is substantial evidence that the brain computes such RPEs
  - Montague, Dayan, Sejnowski (1996), Schultz, Dayan, Montague (1997), McClure, Berns, Montague (2003), O'Doherty et al. (2003)

Model-free learning, ctd.

- the algorithm chooses the action a_t at time t probabilistically:

$$p(a_t = a) = \frac{\exp[\beta Q_t(s_t, a)]}{\Sigma_{a'} \exp[\beta Q_t(s_t, a')]}$$

  - allows for "exploration"
  - as $\beta \to \infty$, choose action with the highest Q value

Model-free learning, ctd.

Why is Q-learning sensible?

- recall that $Q^*(s_t, a_t)$ satisfies

$$Q^*(s_t, a_t) = E_t[r_{t+1} + \gamma \max_{a'} Q^*(s_{t+1}, a')]$$

- we can rewrite the updating equation as

$$\begin{aligned} Q_{t+1}(s_t, a_t) \; &= \; (1 - \alpha^{MF}) Q_t(s_t, a_t) \\ &\quad + \alpha^{MF} [r_{t+1} + \gamma \max_{a'} Q_t(s_{t+1}, a')] \end{aligned}$$

Model-free learning, ctd.

- psychologists often make an adjustment to the basic Q-learning update equation
  - allow for different learning rates for positive and negative RPEs

$$
\begin{aligned}
Q_{t+1}(s_t, a_t) &= Q_t(s_t, a_t) + \alpha_+^{MF}(\mathrm{RPE}), \qquad \mathrm{RPE} \geq 0 \\
Q_{t+1}(s_t, a_t) &= Q_t(s_t, a_t) + \alpha_-^{MF}(\mathrm{RPE}), \qquad \mathrm{RPE} < 0
\end{aligned}
$$

A portfolio-choice setting

- infinite horizon, and two assets
  - risk-free asset with constant gross return R_f
  - risky asset with lognormal return $R_{m,t}$

$$R_{m,t} = e^{\mu+\sigma\varepsilon_t}, \quad \varepsilon_t \sim N(0,1), \text{ i.i.d.}$$

- an investor maximizes the expected log utility of wealth at some future horizon
- if an investor is still in financial markets entering time t
  - with probability $1-\gamma$, he receives a liquidity shock, leaves the markets at time t, and derives log utility of wealth at that time
- his objective then reduces to

$$\max_{\{a_t\}} E_0\left[\sum_{t=1}^{\infty} \gamma^{t-1} \log R_{p,t}\right]$$

  - where $R_{p,t}$ is the portfolio return from $t-1$ to t
  - and a_t is the fraction of wealth in the risky asset

A portfolio-choice setting, ctd.

- we can solve this mathematically
  - solution is to allocate a constant fraction of wealth a^* to the risky asset

$$a^* = \arg\max_a E_t \log((1-a)R_f + aR_{m,t+1})$$

- however, it is not clear how ordinary investors would find their way to this solution
- we want to investigate the implications, in this setting, of a decision-making algorithm that reflects the brain's actual computations
  - e.g. a model-free algorithm like Q-learning

Model-free learning, ctd.

- we could apply earlier Q-learning equation directly to this problem

$$\begin{aligned}Q_{t+1}(s_t, a_t) &= Q_t(s_t, a_t)\\ &\quad +\alpha^{MF}[\log R_{p,t+1} + \gamma \max_{a'} Q_t(s_{t+1}, a') - Q_t(s_t, a_t)]\end{aligned}$$

- instead, assume that investors take there to be only one state, and drop dependence of $Q(s, a)$ on s
  - a simplification on the part of the investor
- investor's goal is then to estimate $Q^*(a)$
- after trying action a at time t, update estimate of $Q^*(a)$ at time $t + 1$:

$$\begin{aligned}Q_{t+1}^{MF}(a) &= Q_t^{MF}(a)\\ &\quad +\alpha_{+/-}^{MF}[\log R_{p,t+1} + \gamma \max_{a'} Q_t^{MF}(a') - Q_t^{MF}(a)]\end{aligned}$$

Model-free learning, ctd.

- the correct $Q^*(a)$ is

$$\begin{aligned} Q^*(a) \;=\; & E_t \log((1-a)R_f + aR_{m,t+1}) \\ & + \frac{\gamma}{1-\gamma} E_t \log((1-a^*)R_f + a^* R_{m,t+1}) \end{aligned}$$

Model-free learning, ctd.

Generalization

- in the basic version of model-free learning, the algorithm updates only the value of the most recently-chosen action
- research in both psychology and computer science has studied "model-free generalization"
  - the algorithm generalizes from its experience of action a to also update the values of other actions
- we have implemented such generalization using the notion of similarity
  - the algorithm uses the RPE from taking allocation a to also update, to a lesser extent, the Q values of similar allocations

$$Q_{t+1}^{MF}(\widehat{a}) = Q_t^{MF}(\widehat{a}) + \alpha_{t,\pm}^{MF} \kappa(\widehat{a})[\log R_{p,t+1} + \gamma \max_{a'} Q_t^{MF}(a') - Q_t^{MF}(a)]$$

$$\kappa(\widehat{a}) = \exp(-\frac{(\widehat{a} - a)^2}{2b^2})$$

Model-free learning, ctd.

- the model-free algorithm uses no information about the structure of asset returns
  - it does not know what a "risk-free asset" is or what the "stock market" is
- nonetheless, it may still be an important driver of decisions in financial markets
  - the model-free system is a fundamental part of human decision-making
  - many investors may be unfamiliar with the structure of asset returns

Model-based learning

- psychologists use a framework that combines model-free and model-based learning
- dynamic programming is one possible model-based framework
  - we use an alternative motivated by neural evidence on the brain's computations
  - Glascher et al. (2010), Lee, Shimojo, O'Doherty (2014), Dunne et al. (2016)
- after observing the market return $R_{m,t} = R$ at time t, the algorithm updates the probability distribution using

$$p_t(R_m = R) = p_{t-1}(R_m = R) + \alpha^{MB}[1 - p_{t-1}(R_m = R)]$$

  - the quantity in square brackets is again a prediction error
  - and α^{MB} is a learning rate
  - there is evidence that such prediction errors are encoded in the brain (Glascher et al., 2010)

Model-based learning, ctd.

- in a continuous-distribution setting, can simplify the above to

$$p_t(R_m = R) = \alpha^{MB}$$

- after observing three returns R_1, R_2, and R_3 in sequence, update perceived distribution as follows

$$\begin{array}{l}(R_1, 1) \\ (R_1, 1-\alpha^{MB}; R_2, \alpha^{MB}) \\ (R_1, (1-\alpha^{MB})^2; R_2, \alpha^{MB}(1-\alpha^{MB}); R_3, \alpha^{MB})\end{array}$$

- we allow for different learning rates for positive and negative returns

$$\begin{aligned} p_t(R_m = R) &= \alpha_+^{MB} \text{ for } R \geq 1 \\ p_t(R_m = R) &= \alpha_-^{MB} \text{ for } R < 1 \end{aligned}$$

Model-based learning, ctd.

- given this return distribution, the investor estimates $Q^*(a)$ using the correct formula, but where the expectation is taken using his perceived distribution

$$\begin{aligned} Q_t^{MB}(a) &= E_t^p \log((1-a)R_f + aR_{m,t+1}) \\ &\quad + \frac{\gamma}{1-\gamma} E_t^p \log((1-a^*)R_f + a^* R_{m,t+1}) \end{aligned}$$

$$a^* = \arg\max_a E_t^p \log((1-a)R_f + aR_{m,t+1})$$

Hybrid system

- following the psychology literature, we use a framework that combines the two algorithms
  - Glascher et al. (2010), Daw et al. (2011)

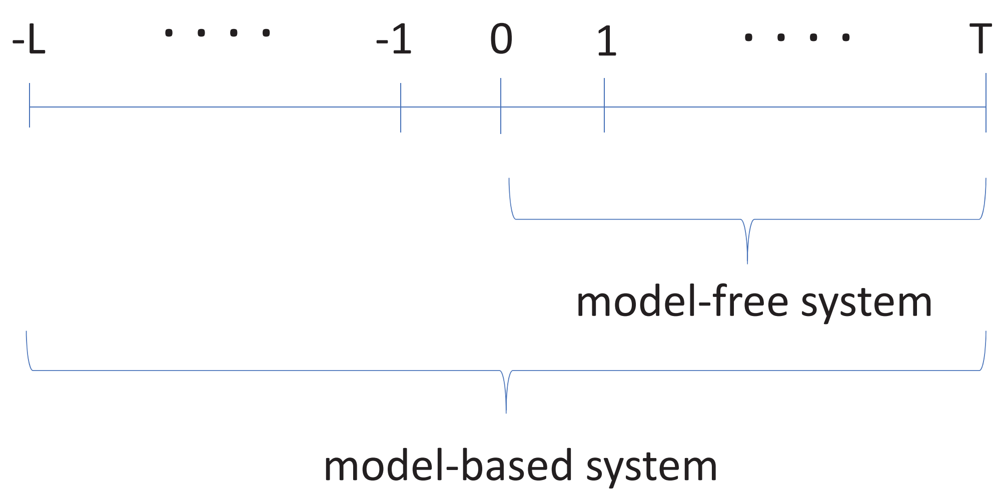
$$Q_t^{HYB}(a) = (1-w)Q_t^{MF}(a) + wQ_t^{MB}(a)$$

$$p(a_t = a) = \frac{\exp[\beta Q_t^{HYB}(a)]}{\Sigma_{a'} \exp[\beta Q_t^{HYB}(a')]}$$

- one difference between the two algorithms is that they likely apply to different intervals
  - if an investor starts participating in financial markets at time 0, the model-free system starts operating at that point
  - but before entering, the investor can observe prior data going back to time $t = -L$, which the model-based system can learn from
- this is consistent with experimental evidence (Dunne et al., 2016)

Properties

- we use the following structure
  - each investor enters financial markets at time 0
  - we track their behavior until time T
  - before entering, each investor observes data going back to $t = -L$
  - we take each period to be one year, and set $L = 30$ and $T = 30$
  - at each date from 0 to T, each investor chooses from the 11 allocations $\{0\%, 10\%, \ldots, 90\%, 100\%\}$



Properties, ctd.

- focus on learning rates that are constant over time
  - initially, learning rates are also the same across investors, but later allow for dispersion
- parameters:

parameter	value
$\alpha_+^{MF}, \alpha_-^{MF}, \alpha_+^{MB}, \alpha_-^{MB}$	0.5
β	30
γ	0.97
w	0.5
μ	0.01
σ	0.2

Properties, ctd.

The mechanics of each system

- consider an investor who observes a sequence of returns over time
- to understand how the two systems work, we first consider the cases where behavior is determined *only* by the model-free system
  - or *only* by the model-based system

Properties, ctd.

The mechanics of each system, ctd.

Model-free Q values

date	0	1	2	3	4	5
net return		-17.4%	18.3%	-1.3%	12.8%	-16.6%
0%	0	0	0	0	0	0
10%	0	0	0	0	0	0
20%	0	0	0.006	0.006	0.01	0.01
30%	0	0	**0.027**	0.027	**0.045**	0.041
40%	0	0	0.006	0.006	0.01	**-0.007**
50%	0	0	0	0	0	-0.004
60%	0	-0.015	-0.015	-0.015	-0.015	-0.015
70%	0	**-0.065**	-0.065	-0.065	-0.065	-0.065
80%	0	-0.015	-0.015	-0.014	-0.014	-0.014
90%	0	0	0	0.001	0.001	0.001
100%	0	0	0	**0.006**	0.006	0.006

Properties, ctd.

The mechanics of each system, ctd.

Model-based Q values

date	0	1	2	3	4	5
net return		-17.4%	18.3%	-1.3%	12.8%	-16.6%
0%	0.72	0	**1.352**	0.464	2.179	0
10%	0.723	-0.007	1.357	0.466	**2.187**	-0.005
20%	0.726	-0.015	1.362	0.468	2.194	-0.01
30%	0.729	-0.022	1.367	0.47	2.201	-0.015
40%	0.731	-0.03	1.372	0.472	2.208	**-0.02**
50%	0.733	**-0.039**	1.376	0.473	2.215	-0.026
60%	0.736	-0.047	1.38	0.475	2.222	-0.031
70%	0.737	-0.056	1.384	0.476	2.228	-0.037
80%	0.739	-0.065	1.387	0.477	2.234	-0.044
90%	0.741	-0.075	1.39	**0.478**	2.241	-0.05
100%	0.742	-0.085	1.393	0.479	2.247	-0.057

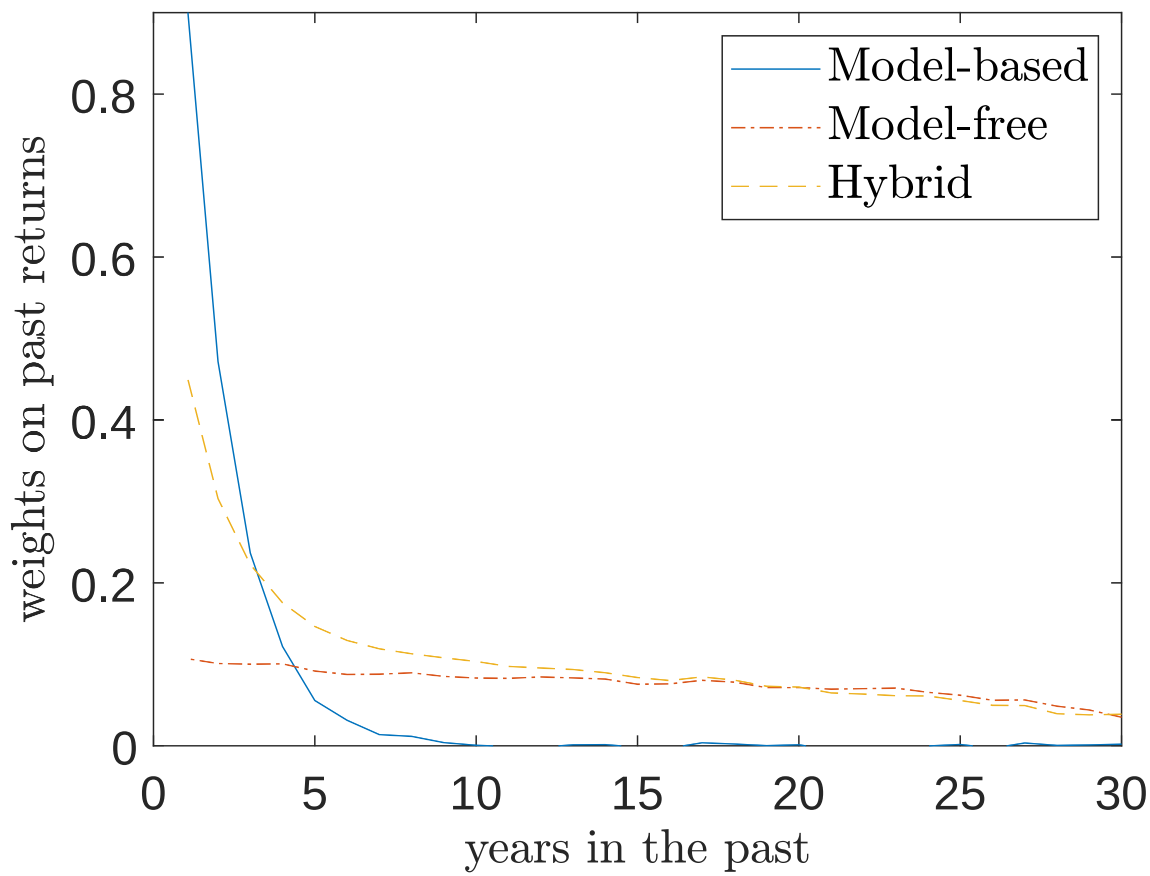
Properties, ctd.

Dependence on past returns

- consider many investors, each of whom is exposed to a different sequence of stock market returns
  - examine how investors' date T allocation a_T depends on the past market returns investors have been exposed to
- for both systems:
  - the allocation puts weights on past stock market returns that are positive and that decline, the further back we go into the past
- importantly, the decline is much faster in the case of the model-based system

Properties, ctd.

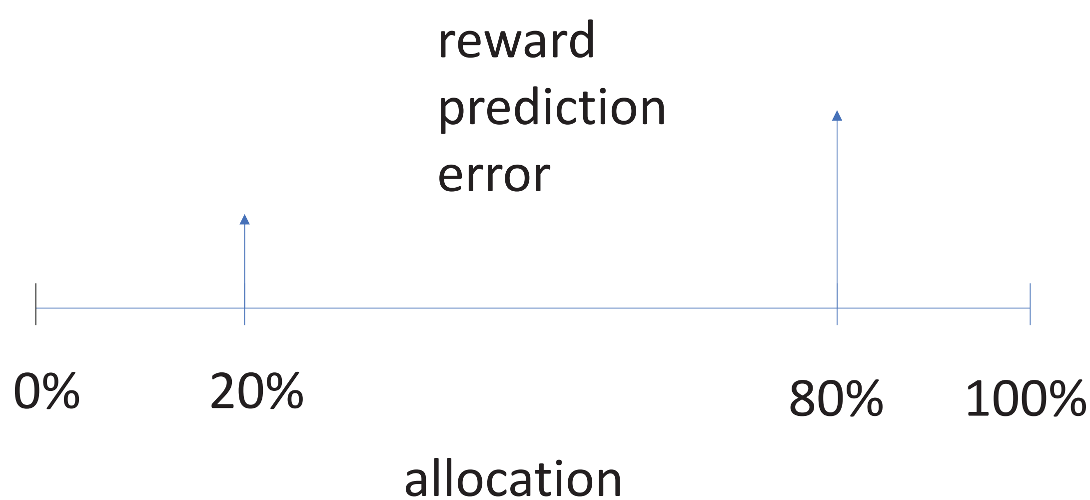
Dependence on past returns, ctd.



Properties, ctd.

Dependence on past returns, ctd.

- it is clear why the model-based allocation depends positively on past returns
- the intuition for the model-free system is more novel
  - after a positive market return, the RPE is larger when the investor's starting allocation is high



Properties, ctd.

Dependence on past returns, ctd.

- it is clear why, for the model-based system, the weights on past returns decline as we go further into the past
- the model-free system exhibits the same pattern, but the decline is much slower
  - the model-free system learns slowly
  - at each time, it updates primarily the Q value of the action just taken

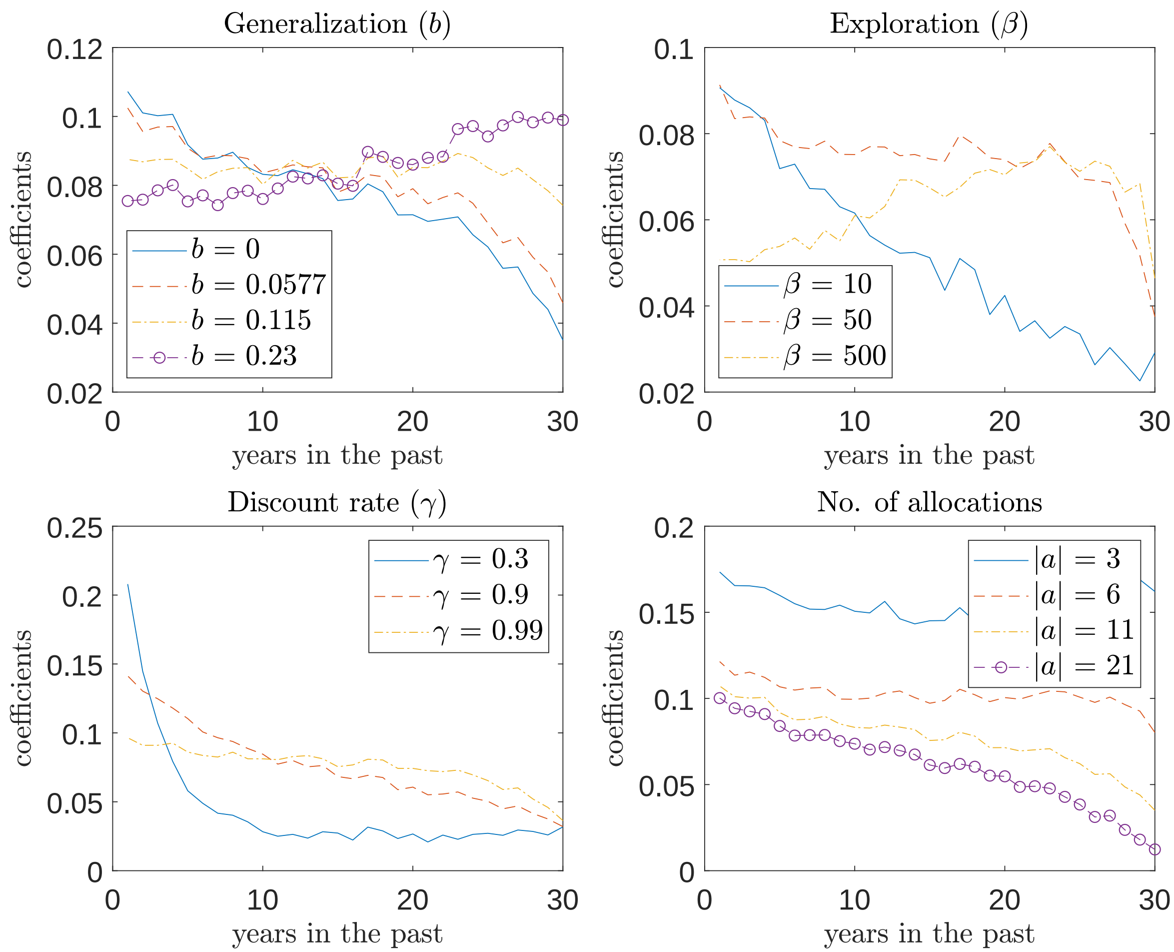
Properties, ctd.

Dependence on past returns, ctd.

- the model-free system can exhibit substantially richer behavior
- the relationship between allocations and past returns is affected by factors that play no role in the model-based system
  - exploration parameter β
  - discount rate γ
  - the number of allocation choices

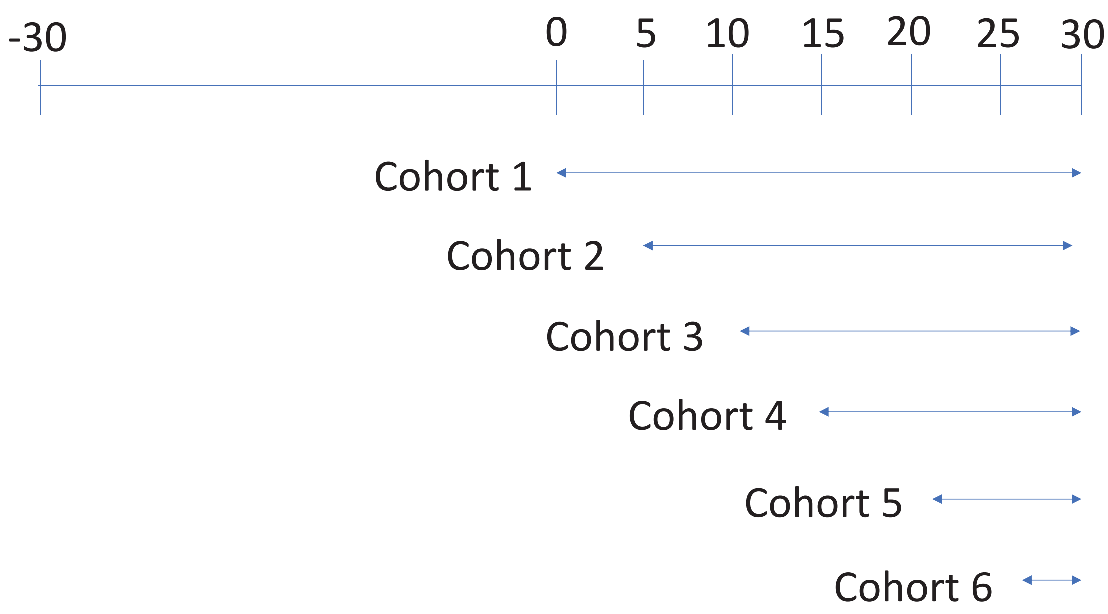
Properties, ctd.

Dependence on past returns, ctd.



Applications

- before considering applications, enrich the framework on two dimensions
  - allow for dispersion in learning rates across investors
  - allow for different cohorts of investors who enter financial markets at different times
  - six cohorts, which enter at $t = 0, 5, 10, 15, 20, 25$, respectively



Applications, ctd.

- show that, for a simple parameterization, obtain a qualitative and approximate quantitative fit to several empirical facts
- later, formally estimate key model parameters

parameter	value
L	30
T	30
$\alpha_+^{MF}, \alpha_-^{MF}, \alpha_+^{MB}, \alpha_-^{MB}$	$\sim [0.25, 0.75]$
β	30
γ	0.97
w	0.5
μ	0.01
σ	0.2

Applications, ctd.

Our framework is helpful for thinking about:

- extrapolative demand
- experience effects
- the disconnect between investor beliefs and investor allocations in both the frequency domain and the cross-section
- dispersion and inertia in investor allocations
- non-participation in the stock market
- persistent investment mistakes

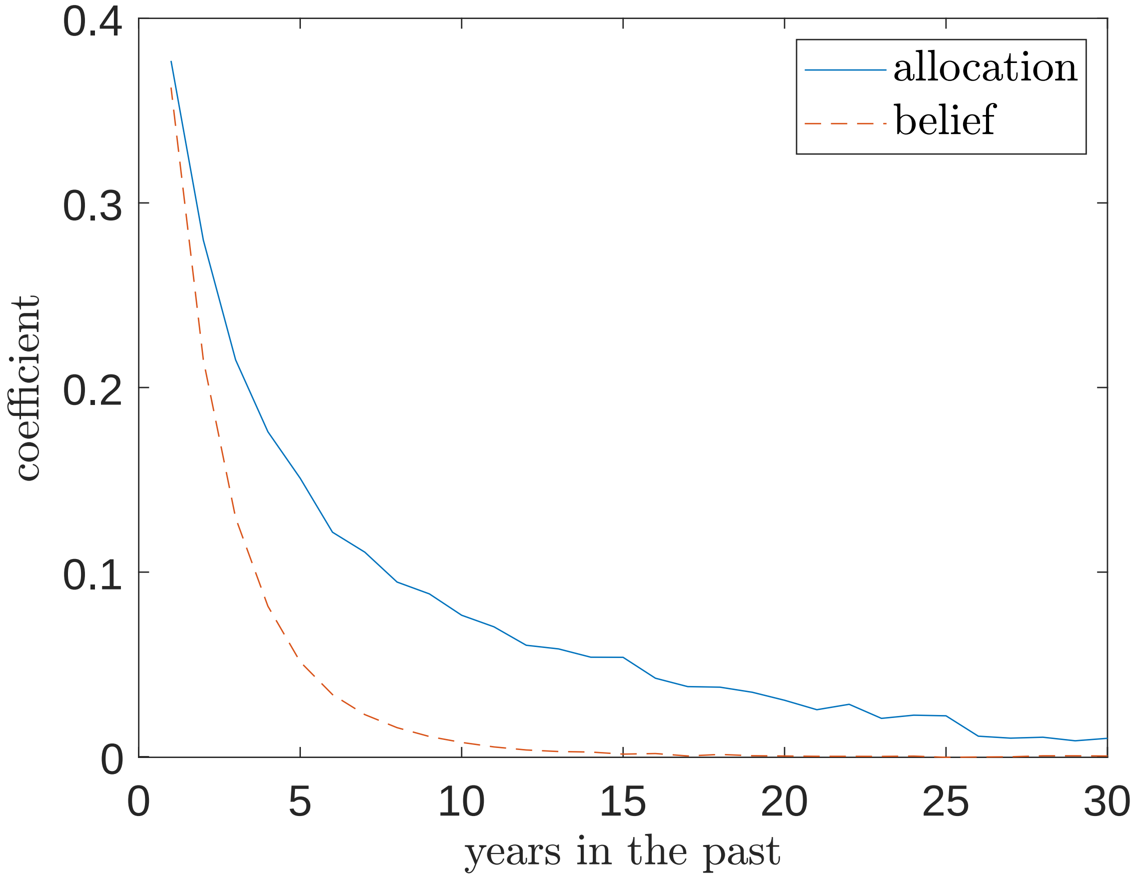
Applications, ctd.

Extrapolative demand

- many models assume that investors have extrapolative demand for risky assets
  - e.g. demand is based on a weighted average of past returns, with more weight on recent returns
- our framework provides a new foundation for such demand, through the mechanics of the model-free system
- it also says that extrapolative demand has two distinct sources operating at different frequencies
  - a model-based source that puts heavy weight on recent returns
  - a model-free source that puts substantial weight even on distant past returns

Applications, ctd.

Extrapolative demand



Applications, ctd.

Experience effects

- Malmendier and Nagel (2011) find that stock market allocations can be explained in part by a weighted average of the stock market returns an investor has personally experienced

Two features:

- investors put more weight on returns they have experienced than on those they have not
  - e.g. if an investor enters the market at time t, he puts significantly more weight on $R_{m,t+1}$ as opposed to $R_{m,t}$
- weights on experienced returns decline the further back we go

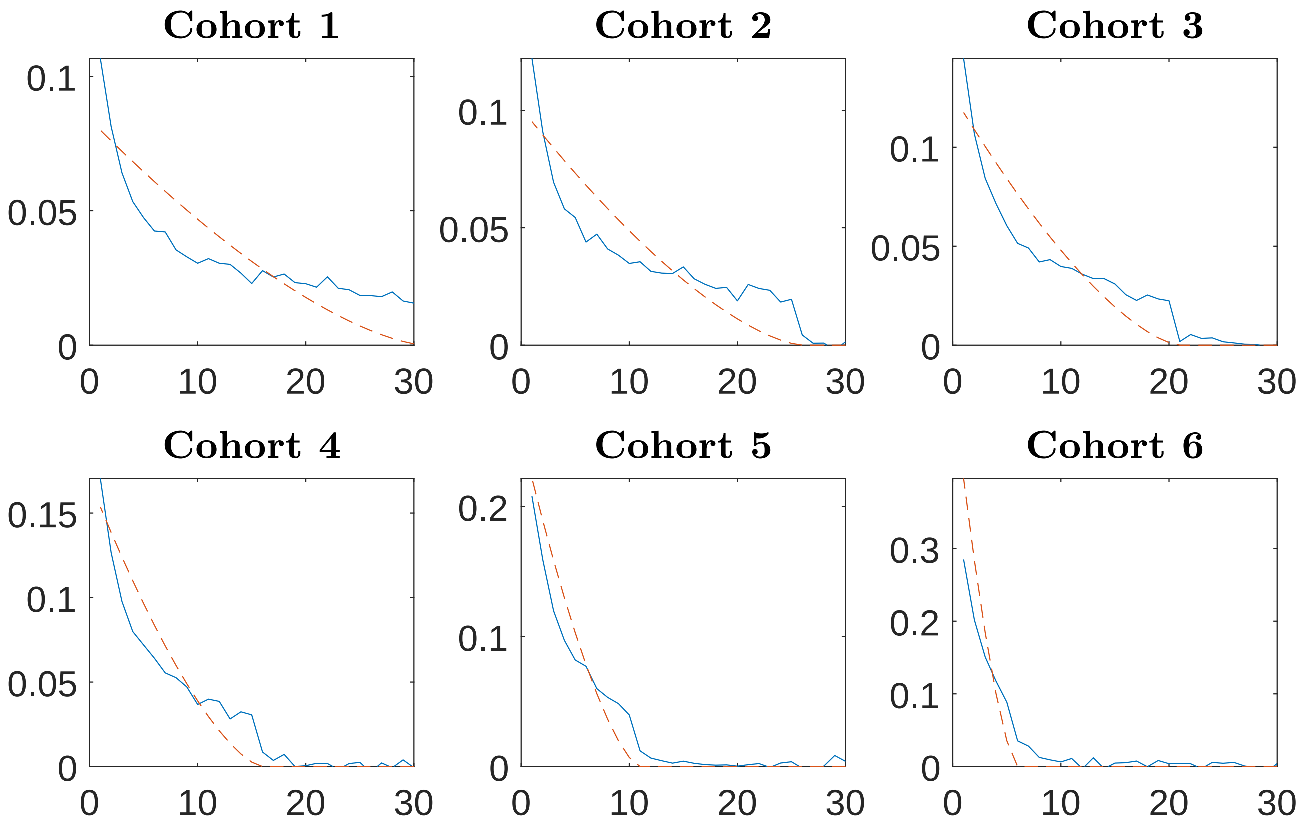
Applications, ctd.

Experience effects, ctd.

- our framework can capture both features
  - the model-free system puts weight only on experienced returns
  - both systems put less weight on more distant past returns
- to check this, we regress, for each cohort, the date T allocation a_T on past stock market returns
  - observe both features

Applications, ctd.

Experience effects, ctd.



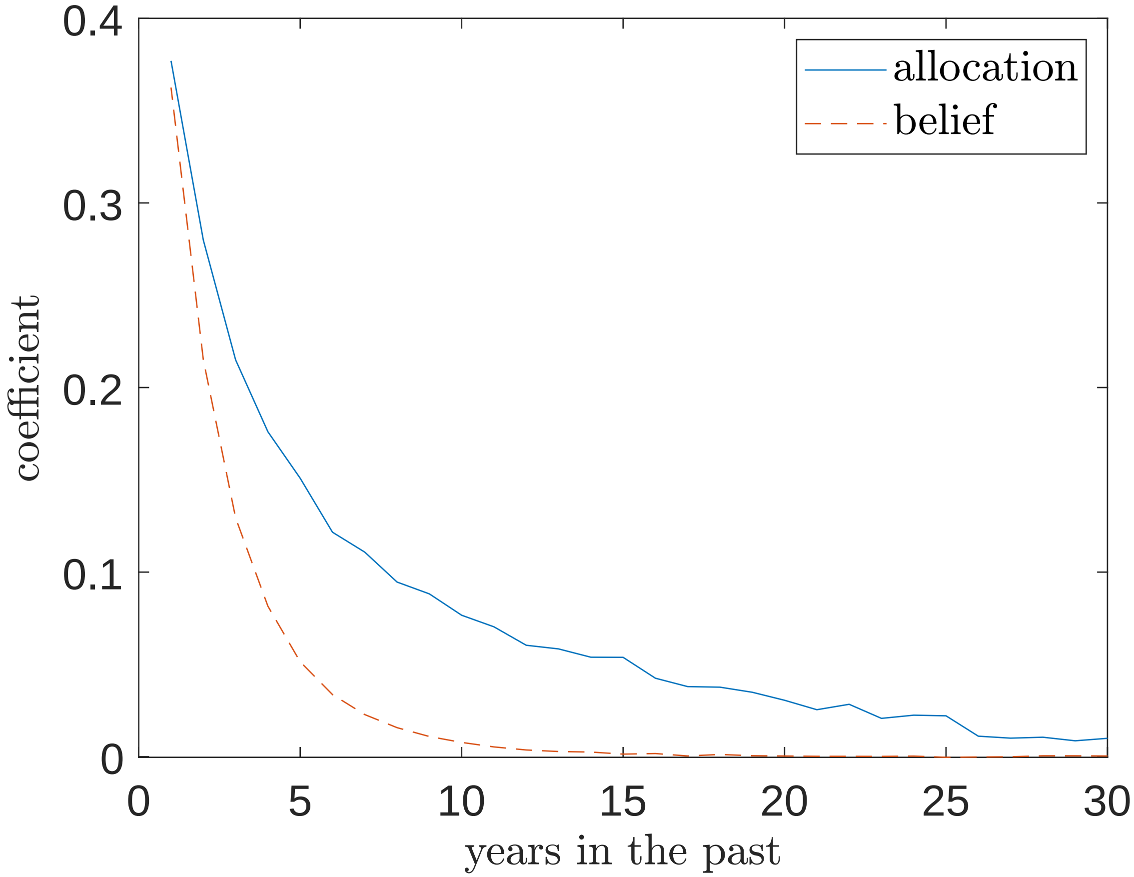
Applications, ctd.

The frequency disconnect

- investor expectations of future stock market returns depend heavily on recent past market returns (Greenwood and Shleifer, 2014)
  - but investor allocations depend even on distant past returns (Malmendier and Nagel, 2011)
- our framework can help explain this
  - only the model-based system has an explicit role for beliefs
- when asked for their *beliefs*, investors consult the model-based system and give an answer that depends primarily on recent past returns
- *allocations* depend on both the model-based *and* model-free systems
  - and the model-free system puts substantial weight even on distant past returns

Applications, ctd.

The frequency disconnect, ctd.



Applications, ctd.

The cross-sectional disconnect

- Giglio et al. (2021) regress investors' allocations on their expectations of future stock market returns
  - find a positive relationship, but weaker than traditional models suggest
- our framework can account for this
  - beliefs are generated by the model-based system, which puts substantial weight on recent returns
  - allocations are also affected by the model-free system, which puts a lot of weight on distant past returns
- following a good stock market return
  - an investor's expected return, generated by the model-based system, goes up significantly
  - his allocation, which is also affected by the model-free system, goes up less

Applications, ctd.

The cross-sectional disconnect, ctd.

w	Sensitivity
0.2	0.7
0.5	1.25
1	1.91

Applications, ctd.

Dispersion in allocations

- there is substantial dispersion in investors' allocations to the stock market
- our framework points to two sources of this dispersion
  - differences in learning rates across investors
  - reinforcement of earlier probabilistic choices

Inertia in allocations

- there is also substantial inertia in investor allocations
- the model-free system can generate such inertia
  - it learns slowly: at each time, it updates primarily the Q value of the action taken

    $\Rightarrow$ from one period to the next, there is little variation in the Q values of the 11 possible allocations

Applications, ctd.

Non-participation

- the model-free system can help account for widespread non-participation in the stock market
- if there is a poor stock market return, this raises the likelihood that the investor will move to a 0% allocation
- once there, the model-free system updates only the Q value of the riskless asset
  - and so will fail to learn that the stock market has good properties
- through simulations, confirm that relative to the model-based system, the model-free system is much more likely to generate non-participation

Applications, ctd.

Persistent investment mistakes

- the framework can explain the persistence of investment mistakes
  - due to the slow learning of the model-free system
- consider a setting with ten risky assets
  - nine have the same low mean return
  - one has a substantially higher mean return
- we show, through simulations, that the model-free system is much slower in figuring out which of the ten assets has the higher mean
  - after 30 years, individuals using the model-free system are less likely to be invested in the higher-mean asset

Parameter estimation

- we estimate four key parameters of our framework
  - the mean model-free learning rate $\bar{\alpha}^{MF}$
  - the mean model-based learning rate $\bar{\alpha}^{MB}$
  - the exploration parameter β
  - the weight w on the model-based system
- we search for values of these parameters that best match:
  - the empirical relationship between investor beliefs and past market returns
  - experience effects, as summarized by Malmendier and Nagel (2011)
  - the sensitivity of allocations to beliefs, as measured by Giglio et al. (2021)
- we obtain $\bar{\alpha}^{MF} = 0.66$, $\bar{\alpha}^{MB} = 0.38$, $\beta = 20$, and $w = 0.46$

Extensions, ctd.

Other directions:

- allow for time-varying learning rates
- allow for time-varying weight w on the model-based system
- allow for state dependence
- allow investors to make inferences about beliefs from the model-free Q values

Broader themes

(1)

- the parameters that best fit the data put substantial weight on the model-free system
  - a system that uses little information about financial markets
- this is initially surprising, but may reflect:
  - how fundamental the model-free system is to human decision-making
  - and investors' unfamiliarity with the structure of asset returns

Broader themes, ctd.

(2)

- we usually start with beliefs and preferences as primitives, and derive a value function from them
  - in the model-free system, the value function is the primitive
  - the investor may infer beliefs from the value function

(3)

- the framework offers a way of thinking about investor behavior that is rooted in algorithms that the brain appears to use when estimating the value of different courses of action

Summary

- in the past decade, psychologists and neuroscientists have increasingly embraced a new framework for thinking about human decision-making in experimental settings
- the framework combines two algorithms, or systems
  - model-free learning
  - model-based learning

In this paper:

- we import this framework into a simple financial setting
- examine its properties and implications
- use it to account for a range of facts about investor behavior