Reinforcement learning for household finance: designing policy via responsiveness

Arka P. Bandyopadhyay,[1] Lilia Maliar[2]

[1]New York University; [2]CUNY Graduate Center and CEPR

January 19, 2024

Motivation

- Asymmetric information between debt servicers and borrowers stands in the way of efficient contract modification.
  $\Rightarrow$ A need for policy intervention.
- During Home Affordable Modification Program (HAMP), after the 2008 financial crisis,
  – policymakers have attempted to give incentives for servicers to gather information dynamically from borrowers;
  – however, takeup rates for such policies were low (Agarwal, et. al., 2017, JPE).
- During the 2020 COVID pandemic,
  – policymakers implemented a blanket forbearance;
  – since it was not targeted, it was inefficient and encouraged strategic forbearance (Bandyopadhyay, 2023).

This paper: a novel targeted quantitative solution to the problem of efficient contract modification under asymmetric information.

Our Framework

- We use a model-free methodology (no assumptions are made).
- We derive an optimal reinforcement-learning (RL) policy by maximizing the mortgage servicer's lifetime reward purely based on past servicer's actions given a certain delinquency state of the borrower.
- We treat *the borrower* as an adversary in the RL paradigm.
- *The servicer*
    - uses soft information about the borrower's current circumstances
    - chooses an optimal strategy for the most efficient contract modification for better outcome
    - preempts moral hazard emanating from the borrower's adversarial behavior
      $\Rightarrow$ *The borrower's* cooperation increases.

Our Findings

- We show that by using soft information, the servicer can provide targeted relief for the most efficient contract modification.
- Our novel responsiveness score helps the servicer to target borrowers with higher propensity to communicate and negotiate.
  $\Rightarrow$ Ad hoc conventional "sticks and carrots" approach can be avoided.
- Cooperation from responsive borrowers enables a final resolution.
- With a very low discount rate, a higher learning rate leads to a faster convergence and implements the optimal RL policy.
- Given a high learning rate, the discount rate does not affect the rate of convergence or the optimal RL policy.

Related Literature on Optimal RL policy

- Barberis and Jin (2022) is the only paper that considers a RL policy in finance.
- They compare a RL policy of *investor behavior* with and without model assumptions.
- In their model, the investor allocates wealth between two assets, a risk-free asset and the stock market.
- They find that the model-based system puts heavy weight on recent returns, while the model-free system puts substantially more weight on distant past returns.

Relation to Literature on Renegotiation

1. *Optimality of contracts and their outcomes:*
   - Aghion et. al. (1994), Hart and Moore (1998).
2. *Asymmetric information and moral hazard and their role in renegotiation:*
   - Roberts, Sufi (2008), Garleanu (2009)
3. *Debt renegotiation as a bargaining game between debtholders and management (shareholders)*:
   - Bergman (1991)
4. *Frictions from covenant violations leading to renegotiation*
   - Anderlini and Felli (2001)

Reinforcement Learning

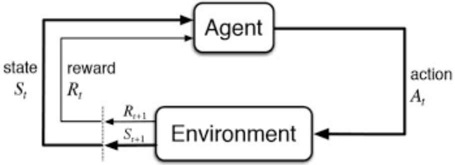
- RL is an area of machine learning concerned with how software agents ought to take actions in an environment in order to maximize some notion of cumulative reward.
- RL is one of three basic machine learning paradigms, alongside supervised learning and unsupervised learning.
- It differs from supervised learning in that it needs not label input/output pairs and correct sub-optimal actions explicitly. Instead the focus is on finding a balance between exploration (of uncharted territory) and exploitation (of current knowledge).
- The environment is typically stated in the form of a Markov decision process (MDP), because many RL algorithms for this context utilize dynamic programming techniques.
- Main difference between classical dynamic programming and RL algorithms: RL does not assume knowledge of an exact mathematical model of the MDP and targets large MDPs where exact methods become infeasible.

Defining a RL problem

- Markov property: Current state completely characterized\s the state of the world
- Define a tuple of objects (S, A, R, P, γ)
  – S: set of possible states *(capital, productivity)*
  – A: set of possible actions *(consumption choices)*
  – R: distribution of reward given (state, action) pair *(utility level)*
  – P: transition probability, i.e., distribution over next state given (state, action) pair *(next period capital and shock)*
  – γ: discount factor

Markov Decision Process

- Markov decision process is a mathematical formulation of RL problem



- At time $t = 0$, environment samples initial state $s_0 \sim p(s_0)$
- Then, for $t = 0, T$
  - Agent selects action a_t
  - Environment samples reward $r_t \sim R(\cdot \mid s_t, a_t)$
  - Environment samples next state $s_{t+1} \sim P(\cdot \mid s_t, a_t)$
  - Agent receives reward r_t and next state s_{t+1}
- Google DeepMind learns to play Atari.
  https://www.youtube.com/watch?v=V1eYniJ0Rnk

Policy Function

- A policy π is a function from S to A that specifies what action to take in each state
- *Objective:* find policy π^* that maximizes cumulative discounted reward $\sum_{t>0} \gamma^t r_t$.
- Formally,

$$\pi^* = \arg\max_{\pi} E\left[\sum_{t>0} \gamma^t r_t \mid \pi\right],$$

  where $s_0 \sim p(s_0)$, $a_t \sim \pi(\cdot \mid s_t)$, $s_{t+1} \sim p(\cdot \mid s_t)$.
- Following a policy produces sample trajectories (or paths) s_0, a_0, r_0, s_1, a_1, r_1,...

Value Function and Q-Learning

- *How good is a state?*
- The value function at state s is the expected cumulative reward from following the policy from state s:

$$V^{\pi}(s) = E\left[\sum_{t>0} \gamma^t r_t \mid s_0 = s, \pi\right]$$

- *How good is a state-action pair?*
- The Q-value function at state s and action a is the expected cumulative reward from taking action a in state s and then following the policy

$$Q^{\pi}(s, a) = E\left[\sum_{t>0} \gamma^t r_t \mid s_0 = s, a_0 = a, \pi\right]$$

Bellman Equation

- The optimal Q-value function Q^* is the maximum expected cumulative reward achievable from a given (state, action) pair:

$$Q^*(s, a) = \max_{\pi} E\left[\sum_{t>0} \gamma^t r_t \mid s_0 = s, a_0 = a, \pi\right]$$

- Q^* satisfies the Bellman equation

$$Q^*(s, a) = E_{s' \sim \mathcal{E}}\left[r + \gamma \max_{a'}\left[Q^*\left(s', a'\right) \mid s, a\right]\right]$$

- *Intuition:* if the optimal state-action values for the next time step $Q^*(s', a')$ are known, then the optimal strategy is to take the action that maximizes the expected value of $r + \gamma \max_{a'} Q^*(s', a')$.
- The optimal policy π^* corresponds to taking the best action in any state as specified by Q^*.

Solving for Optimal Policy

- Value iteration algorithm: Use Bellman equation as an iterative update

$$Q_{i+1}(s,a) = E_{s' \sim \mathcal{E}}\left[r + \gamma \max_{a'}\left[Q_i(s',a') \mid s,a\right]\right]$$

  Q_i will converge to Q^* as $i \to$ infinity.

- *What is the problem with this?*
  Not scalable. Must compute $Q(s,a)$ for every state-action pair. If state is current game state pixels, computationally infeasible to compute for entire state space!

Solving for Optimal Policy: Q-Learning

- Use a function approximator to estimate the action-value function
$$Q\left(s, a; \theta\right) \approx Q^{*}\left(s, a\right)$$
θ: function parameters, weights.
- If the function approximator is a deep neural network $\Rightarrow$ deep Q-learning!
- *Remember:* want to find a Q-function that satisfies the Bellman equation.

RL in Economics

- *Surveys:* Arthur (1991), Singh (1991), Charpentier et al. (2020), Mosavi et al. (2020).

- *Financial-market simulator* (Wiese et al., 2019)
- *Portfolio choice with atomistic investors* (Li and Hoi, 2014)
- *Portfolio choice with non-atomistic investors* (Spooner et al., 2018)
- *Bounded rationality*
  – RL leads to bounded rationality (Leimar & McNamara, 2019);
  – RL is suitable for studying boundedly rational agents (Abel, 2019);
  – Local thinking (Gabaix, 2014)
- *Single firm dynamics* (Erev and Roth, 1998)

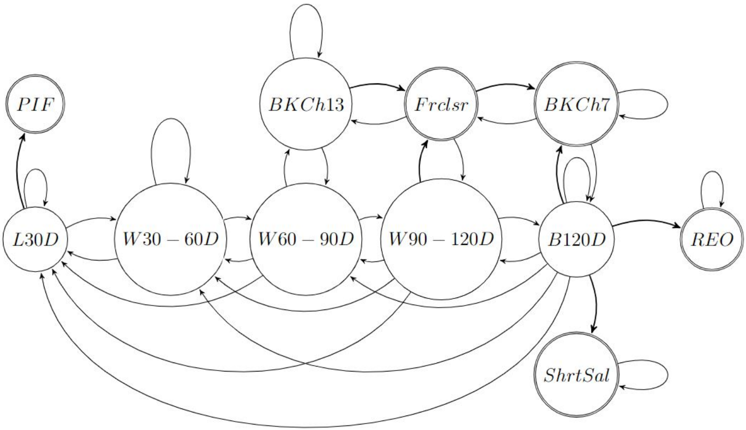
RL in Economics (cont.)

- *Stochastic games*
  - Zero-sum games with two players (Littman, 1994)
  - One-parameter RL (Erev and Roth, 1998)
- *Auctions and real-time bidding*
  - RL for describing the bid decision process (Schwind, 2007)
  - RL for designing a bidding strategy (Cai et al., 2017, Zhao et al., 2018)
  - RL for designing optimal auctions (Feng et al., 2018).
- *Oligopoly and dynamic games*
  - Experience-based equilibrium (Fershtman and Pakes, 2012)
  - Repeated Cournot games (Waltman and Kayman, 2008)
- *Computational economics* (Chen et al., 2021)

Data

- Proprietary administrative data for 23,693 loans from 09/2017 to 3/2020.
  – detailed information on residential mortgage performance collected from daily mortgage servicing logs.
  – also includes text communications between the borrowers and servicers.
- This data set is from a servicer which has 15% of the national market share of Ginnie Mae Early Buyout loans in terms of deal flow. $\Rightarrow$ Sizable proportion of all loans.
- In addition, we use proprietary data from Epsilon at the household level (100 million US households in total).
  – allows us to capture the spending patterns, demography, relocation, and several other aspects of these borrowers.

Possible Transitions During Life of Mortgage



Deliquency States

State	Loans
L30D	Current or less than 30 days delinquent.
W30-60D	Within 30 to 60 days of delinquency.
W60-90D	Within 60 and 90 days of delinquency.
W90-120D	In default after 90 days of delinquency with ongoing payments after missing 3 months of payments
B120D	Already beyond 120 days of delinquency.
BK	The borrower has filed for bankruptcy.
Frclsr	Have entered the foreclosure (FC) proceedings.
PIF	Already paid in full.
REO	Repossessed by the original lender/servicer representing the lender.
ShrtSal	Auctioned in public market for short sale.

Action Space

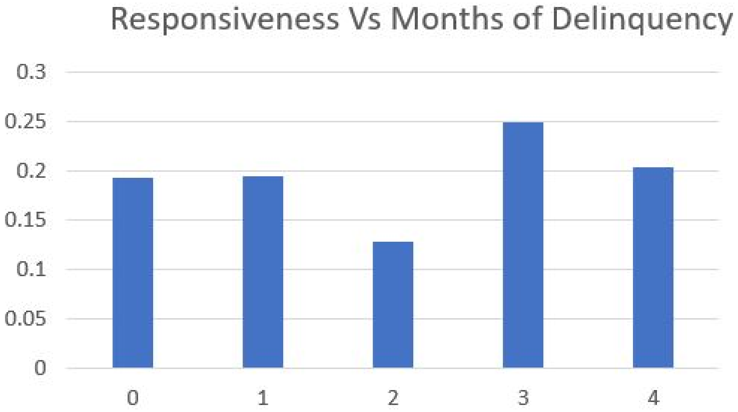
Examples of Actions	Definition
Pending claim (PC)	The servicer has filed for a HUD claim.
Modification in review (Mod)	The ongoing phase of active negotiation between borrowers and servicers.
No Action (NA)	The servicer has taken no action.
Pending foreclosure (PF) completion	A foreclosure process about to close in the near future.
Real Estate Owned (REO)	The process is which the lender or the has gained back possession of the property after offering deed-in-lieu (DIL).
Bankruptcy	The ongoing bankruptcy filed the borrower, servicer for a renegotiation or ch. 7 for a co liquidation of assets.
Not referred for short refinance	Not offering a loan modification to the borr based on the servicer's discretion.

Cross-Sectional Results and Motivation for RL

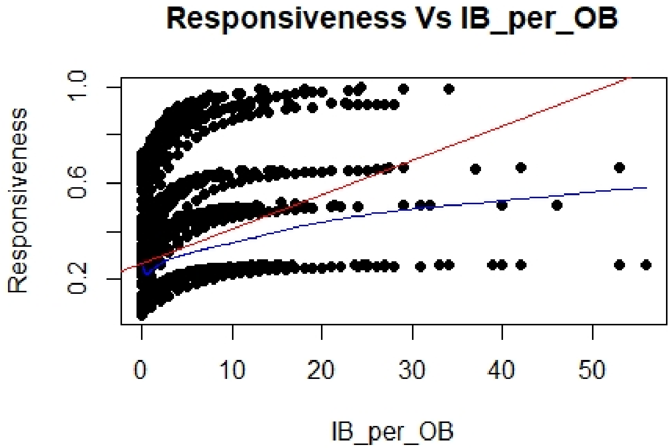
We created a time-invariant measure, *Responsiveness*, which is a cumulative distribution function of the following five random variables:

1. *Months of Delinquency*: higher scores for less deliquent:
   – Paid Ahead := 4, Current := 3, 1 month behind := 2, ...
2. *Loan Deliquency Status*: lower scores for more adverse status:
   – Current := 6, 30 days delinquent := 5, 60 days delinquent := 4, ...
3. *Known Inbound Calls*: sum of all known Inbound communications from inception.
4. *Inbound calls from borrowers as a return to the servicer's Outbound calls*: $\frac{\text{number of return Inbound calls by the borrower}}{\text{number of Outbound calls of the servicer}}$
5. *Information Content*: Reasons for the calls:
   – Forbearance, Foreclosure moratorium, Loan modification.

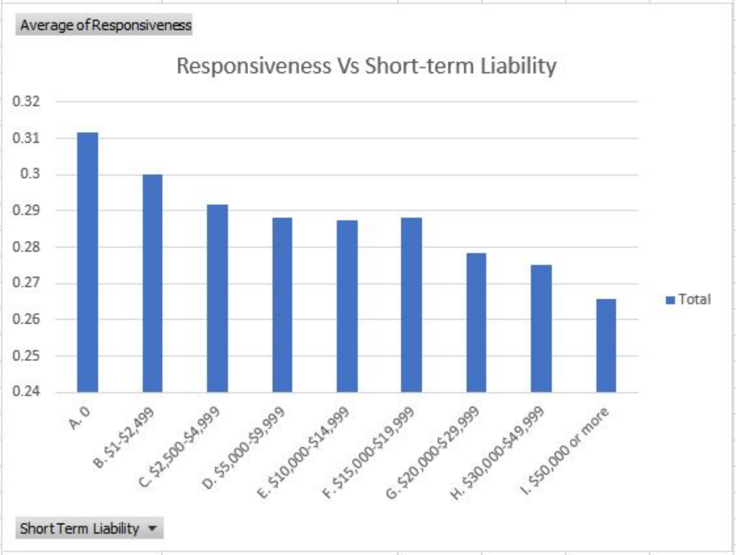
Responsiveness Vs Months of Delinquency



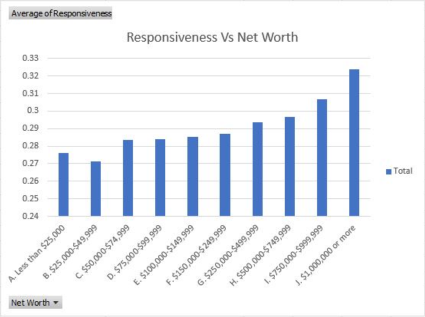
Responsiveness Vs Inbound per Outbound Calls

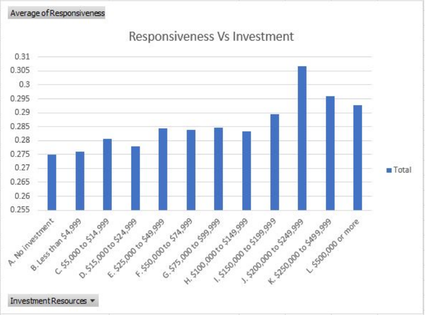


Responsiveness Vs Short Term Liability

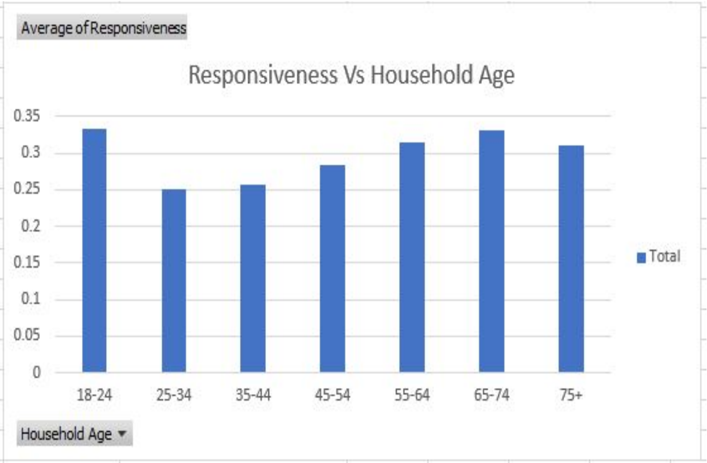


Responsiveness Vs Net Worth and Investment Resources





Responsiveness Vs Household Age



Variable Importance

	Responsiveness	Importance		
		Relative	Scaled	Percentage
1	**modification date**	1075.59	1.00	0.17
2	original fico	441.61	0.41	0.07
3	**current rate**	439.89	0.41	0.07
4	**current fico**	412.80	0.38	0.07
5	orig ltv	291.95	0.27	0.05
6	original rate	288.24	0.27	0.05
7	**foreclosure stage**	269.14	0.25	0.04
8	year home built	265.78	0.25	0.04
9	buy a house rank	257.29	0.24	0.04
10	**bankrupcy delay**	252.61	0.23	0.04
11	home loan rank	252.50	0.23	0.04
12	move residence rank	228.07	0.21	0.04
13	move residence date	209.98	0.20	0.03
14	...	...	...	...

Conventional Qualitative (Stick-Carrot) Policy

- *Steps:*
  1) Information related to title, foreclosure, bankruptcy, property is sequentially received.
  2) Combined legal grades are determined.
- Grades from A to E reflect the likelihood of loss, as well as the time/cost/complexity involved in addressing the concerns.
  - **Grade A**: non-issue from risk standpoint; no discount.
  - **Grade B**: no material risk of loss; covered by valid insurance.
  - **Grade C**: moderate risk of loss; a significant discount (10-25%).
  - **Grade D**: require litigation or significant expenditures to resolve; a substantial discount (50-90%).
  - **Grade E**: nearly certain to result in complete loss.
- "*Carrot":* an overly conservative grade (e.g., grade A) prices the servicer out of every trade.
- "*Stick":* an overly aggressive grade (e.g., grade E) results in undersized returns.

Designing RL-optimal Policy

- We design an optimal policy which is stricter than carrots and more considerate than sticks.
- To derive optimal RL policy, we maximize profit of the servicer.
- RL can extract the best course of action towards the borrower assuming he is an adversary agent (Goodfellow et. al., 2014).
- We simulate *housing market environment* using our proprietary data about borrowers' spending habits, demography, income bracket, real-time unemployment status, etc.
- We compare our optimal policy with the current ad hoc qualitative methodology used by the servicer.
- For each loan and for each month, we have the actual action (strategy undertaken) by the servicer.
- A clear dollar difference in collections between our optimal RL policy and current heuristic servicer's action provides a direct support to our quantitative approach.

Servicer's Problem

- The servicer maximizes a Q-value

$$Q^*(s,a) = \max_{\{a_t\}} E_0 \left[\sum_{t=1}^{T} \gamma^t r_t \right]$$

- In our analysis, reward

$$r_t = \frac{\text{Servicer's collections in a given month } t}{\text{Original balance of loans}}$$

- We bucket the reward variable into groups of equal width.
  – We discretaize because Q learning cannot handle continuous variables.

Q-Leaning Algorithm

- Assume that at time t in state $s = s_t$, the algorithm takes an action $a_t = a$. This leads to reward r_{t+1} and state s_{t+1} at time $t + 1$.
- At time t, the algorithm's initial estimate of $Q^* (s, a)$ is $Q_t (s, a)$.
- At $t + 1$, we update $Q^* (s, a)$ as

$$Q_{t+1} (s, a) = Q_t (s, a) + \alpha_t \left[r_t + \gamma \max_{a'} Q_t \left(s_{t+1}, a' \right) - Q_t (s, a) \right]$$

  α_t : learning rate.
- An action $a_t = a$ in state $s = s_t$ at time t is chosen probabilistically: probability is an increasing function of its Q value

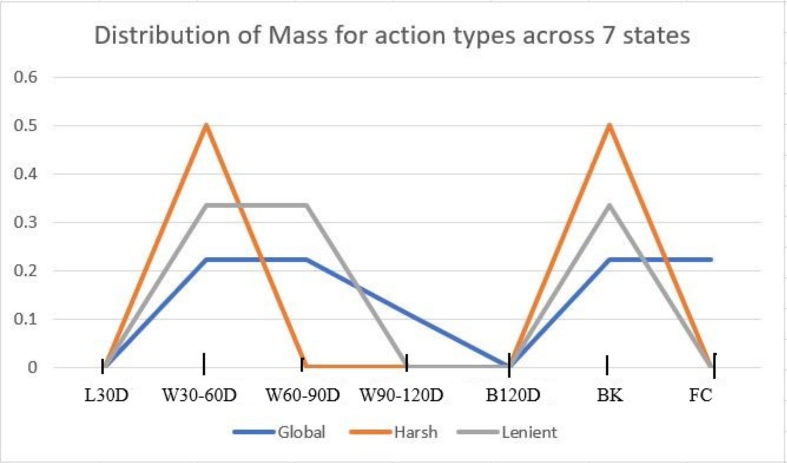
$$p (a_t = a, s_t = s) = \frac{\exp \left[\beta Q_t (s, a) \right]}{\sum_{a'} \exp \left[\beta Q_t (s, a') \right]}$$

  β : exploration parameter.

RL-Optimal, Harsh and Lenient Policies

		Disposition strategies	
States	**RL-optimal**	**Harsh**	**Lenient**
L30D	Mod Review	Mod Review	Mod Review
W30-60D	No Action	Pend FC Complet	No Action
W60-90D	No Action	Mod Review	Not refer Short
W90-120D	Pend FC Complet	Mod Review	Mod Review
B120D	Mod Review	Mod Review	Mod Review
BK	Pending Claim	Pend FC Complet	No Action
FC	REO	Pend FC Complet	Mod Review

Comparison of Policies in Terms of Flexibility



Transition Matrix from a State-Action Pair

				States			
Actions	L30D	W30-60D	W60-90D	W90-120D	B120D	BK	FC
Pend FC	2.32%	2.40%	2.59%	**2.61%**	8.87%	2.55%	***77.57%***
Pend Claim	0.00%	0.00%	0.00%	0.00%	6.41%	**0.02%**	0.99%
Bankruptcy	0.97%	0.67%	0.11%	0.83%	2.57%	***87.65%***	5.72%
No Action	8.18%	9.70%	11.54%	11.62%	7.10%	7.29%	4.55%
Performing	***68.54%***	***50.39%***	***44.03%***	***42.81%***	**17.00%**	0.90%	0.81%
No Refi	3.48%	1.86%	0.90%	0.44%	0.17%	0.00%	0.33%
REO	0.00%	0.00%	0.00%	0.00%	0.02%	0.00%	**0.18%**
Mod In Rev	**0.17%**	0.30%	0.64%	1.28%	3.01%	0.23%	1.14%
Pend ShrtSale	0.00%	0.00%	0.00%	0.01%	0.56%	0.00%	0.62%
Pend D-In-L	0.00%	0.00%	0.00%	0.00%	0.20%	0.00%	0.17%
Pend Repurch	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%	0.11%
Cnsnt Judgm	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%
Mod Compltd	1.50%	0.89%	0.85%	0.55%	0.31%	0.06%	0.91%
Roll Delinq	0.00%	0.05%	0.11%	0.04%	0.02%	0.00%	0.00%
NOI: Not FC	14.83%	33.73%	39.24%	39.81%	***53.74%***	1.30%	6.88%

Learning Rate

Learning rate	α =0.99	α =0.95	α =0.9	α =0.8	α =0.7	α =0.6
L30D	Mod	Mod	Mod	Mod	Mod	Mod
W30-60D	NA	NA	Mod	NRSR	NA	NA
W60-90D	NA	NA	Mod	NA	NA	NA
W90-120D	FC	FC	Mod	Mod	FC	Mod
B120D	Mod	Mod	Mod	Mod	Mod	Mod
BK	FC	PC	PC	PC	PC	NA
FC	DIL	DIL	REO	DIL	DIL	DIL

Mod=modification in review; NA=no action; NRSR=not referred for short refinancing; FC=Pend FC completion; DIL=deed in lieu

Discounting

# iterations	$N = 10^4$	$N = 10^4$	$N = 10^5$	$N = 10^5$
Discount factor	$\gamma = 0.99$	$\gamma = 0.01$	$\gamma = 0.99$	$\gamma = 0.01$
L30D	Mod	Mod	Mod	Mod
W30-60D	NA	NA	NA	NA
W60-90D	NA	NA	NA	NA
W90-120D	FC	FC	FC	FC
B120D	Mod	Mod	Mod	Mod
BK	FC	PC	NA	NA
FC	DIL	DIL	DIL	DIL

Mod=modification in review; NA=no action; FC=Pend FC completion; DIL=deed in lieu

Conclusion

- Because of information asymmetry at the loan level, the servicers have anecdotally used a sticks or carrots approach.
- We measure the responsiveness of borrowers based on our unique administrative data set of text communications between the borrowers and servicers.
- We provide evidence that more responsive borrowers cooperate upon communication with them.
- This enables us to document the most efficient transition among delinquency states during the life of a loan.
- This requires a dynamic setting to evaluate an optimal servicer's strategy, mostly aligned with the lender.
- We show divergence in learning based on decreasing learning rates.
    - For high learning rates, the discount factor does not matter $\Rightarrow$ can mitigate differences in the existing viewpoints on beliefs.
    - Past experiences do not dominate the RL-optimal actions of an agent in a high learning environment.

Thank you!