Estimation of Dynamic Discrete Choice Model with Unobserved State Variables Using Reinforcement Learning

Yingyao Hu Fangzhu Yang

Johns Hopkins University

Machine Learning Department
Mohamed bin Zayed University of Artificial Intelligence
Jan 27, 2025

Outline

Outline

Introduction

- Economic problems often involve solving dynamic discrete choice models (DDCs), where agents make optimal decisions over time
- DDCs pose computational challenges due to high-dimensional state spaces with continuous variables
  - Discretizing continuous variables leads to inaccurate approximations and coding burdens
- Looping over grid points for state variables makes computation infeasible for complex models
- Many dynamic models include unobservable heterogeneity, further complicating computation

What This Paper Does

- Propose an estimation framework, embedding policy gradient methods from reinforcement learning into indirect inference methods to estimate DDCs with
  - a large state space
  - various types of unobserved state variables
- Our method contains two layers of loops
  - In the outer loop, use the Simulated Method of Moments to update deep parameters
  - In the inner loop, solve for the optimal policy using policy gradient methods
- Build on identification results in Hu and Shum (2012), propose a simple algorithm to estimate DDCs with continuous and time-varying unobserved state variables
  - The simulation-based algorithm makes it suitable to be combined with indirect inference
  - Reach identification by targeting the moments implied by the identification results in Hu and Shum (2012)

Contributions and Findings

- Proposed a convenient inference framework for dynamic models with continuous and time-varying unobserved variables
  - Solving the issue that it is nontrivial to estimate this type of models
  - Discretization of variables not needed for our method
- This method significantly reduces the computational burden for dynamic models with large state space and many continuous variables
  - Maintain similar level of precision
  - Allow for estimation of more complicated models
- Form a unified framework for estimating dynamic discrete choice models of various types
  - Easy to implement
  - Similar structure for models with or without unobservables

Contributions and Findings (Cont'd)

Table: Comparison Between Existing Methods and Our Proposed Methods

Model	**Existing Methods**	**RL + Indirect inference**
S: continuous, observed	Full solution estimation using Indirect Inference (with discretization)	A dynamic fertility model with both discrete and continuous choices
S: continuous, unobserved		Rust bus model with continuous and time-varying unobservables
S: discrete, observed	NFXP Algorithm	Toy Rust bus model with discrete state variables
S: discrete, unobserved	Two-step EM algorithm	Rust bus model with unobserved heterogeneity

This Paper in the Related Literature

- Estimation of dynamic discrete choice models and identification of models with unobserved heterogeneity
  - Rust (1987), Hotz and Miller (1993), Hotz et al. (1994), Aguirregabiria and Mira (2007), Arcidiacono and Jones (2003), Arcidiacono and Miller (2011), Kasahara and Shimotsu (2009), Hu and Shum (2012), Hwang (2024), Gallant et al. (2018)
  - **Our paper:** focuses on the estimation of the dynamic discrete choice models, combining the policy gradient method from the reinforcement learning literature with indirect inference. Able to deal with complicated models with time-varying and continuous unobservables using existing identification results
- Reinforcement learning and policy gradient methods
  - Lange et al. (2012), Sutton et al. (1999), Kakade (2001), Silver et al. (2014), Peters and Schaal (2006), Yu et al. (2017), Li et al. (2022), Jin et al. (2023), Hong et al. (2023)
  - **Our paper:** focus on the estimation of the structural model, while these papers only care about figuring out the optimal policy

Outline

Model Environment

- A structural dynamic model with a process $\{\boldsymbol{X}_t, J_t\}$
  - $\boldsymbol{X}_t \in \mathbb{R}^d$: observed state variables that evolve over time
  - $J_t \in \{0, 1\}$: agent's choice variable in period t
- Researchers observe a panel dataset of an i.i.d sample of $\{(\boldsymbol{X}_1, J_1), (\boldsymbol{X}_2, J_2), ..., (\boldsymbol{X}_T, J_T)\}$ for many agents
- Transitional model $f_{\boldsymbol{X}_t|\boldsymbol{X}_{t-1}, J_{t-1}}$ is known or can be estimated using the data
- Flow utility/period reward of the agent: $U(\boldsymbol{X}_t, J_t; \boldsymbol{\theta})$
- The agent's objective: choose a sequence of $\{J_1, J_2, ..., J_T\}$ to maximize the lifetime utility

Reinforcement Learning + Indirect Inference Algorithm

- Intuition: our algorithm contains two loops
  - Outer loop: use Simulated Method of Moments to update structural parameters
  - Inner loop: conditional on the structural parameters, use Policy Gradient Method to update policy function parameters

- Key: parametrize the choice variable J_t as a function of the state variables: $\pi_{\gamma(\boldsymbol{\theta})}(J_t|\boldsymbol{X}_t)$:

$$\Pr(J_t = 1|\boldsymbol{X}_t; \gamma(\boldsymbol{\theta})) = \frac{\exp(\boldsymbol{X}_t\gamma(\boldsymbol{\theta}))}{1 + \exp(\boldsymbol{X}_t\gamma(\boldsymbol{\theta}))}$$

  - Choice of functional form is flexible; as long as the derivative w.r.t the parameters have closed-form
  - Highly nonlinear parametrization can be applied inside the logistic function

Reinforcement Learning + Indirect Inference Algorithm (Cont'd)

- $V(\gamma(\boldsymbol{\theta}))$: reward function that depends on the policy parameter $\gamma(\boldsymbol{\theta})$
- Move γ toward the direction suggested by the gradient $\nabla_\gamma V(\gamma(\boldsymbol{\theta}))$:

$$\gamma_{q+1} = \gamma_q + s_q \nabla_\gamma V(\gamma_q(\boldsymbol{\theta}))$$

- Nontrivial to calculate $\nabla_\gamma V(\gamma_q(\boldsymbol{\theta}))$ as it involves the action and the stationary distribution of states
- Use Policy Gradient Theorem (Sutton and Barto (2018)):

$$\nabla_\gamma V(\gamma) = E_X\Big[E_J\big[Q^{\pi_\gamma}(\boldsymbol{X}, J)\nabla_\gamma \log \pi_\gamma(J|\boldsymbol{X})\big]\Big]$$

  - $Q^{\pi_\gamma}(\boldsymbol{X}, J)$: state-action value function of the policy π_γ
  - Gradient of the value function transferred to the gradient of the policy function

Algorithm for DDCs with Fully observed State Space: Flow Chart

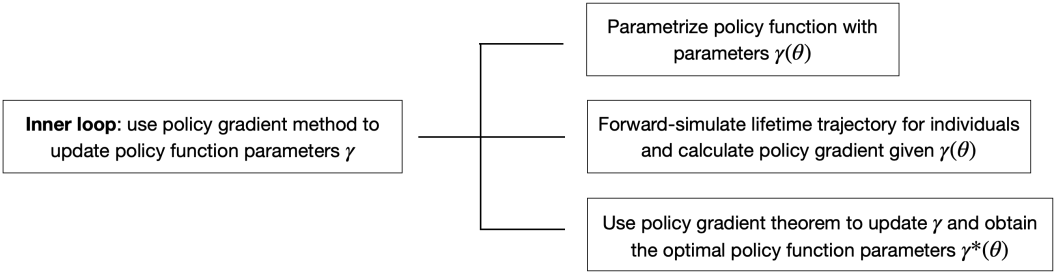


Figure: Flow chart for the Inner Loop of the RL + Indirect inference Algorithm for Models With Fully Observed State Space

Algorithm for DDCs with Fully observed State Space (Part I)

Algorithm A.1 Forward Simulation for Obtaining Individual i's Lifetime Trajectory

1: **Input:** policy parameters $\boldsymbol{\gamma}(\boldsymbol{\theta})$, utility parameters $\boldsymbol{\theta}$, initial state $\boldsymbol{X}_{i1}$ for individual i
2: Sample $\widehat{J}_{i1}$ using Equation 1, given $\boldsymbol{\gamma}(\boldsymbol{\theta})$ and $\boldsymbol{X}_{i1}$
3: **for** $t = 2, \ldots, T$ **do**
4: Sample $\widehat{\boldsymbol{X}}_{it}$ using $f_{\boldsymbol{X}_t|\boldsymbol{X}_{t-1},J_{t-1}}$
5: Sample $\widehat{J}_{it}$ Equation 1, conditional on $\boldsymbol{\gamma}(\boldsymbol{\theta})$ and $\widehat{\boldsymbol{X}}_{it}$
6: **end for**
7: Obtain the final dataset $\widehat{\boldsymbol{D}}_i = (\widehat{\boldsymbol{X}}_i, \widehat{\boldsymbol{J}}_i)$
8: **Output:** lifetime trajectory $\widehat{\boldsymbol{D}}_i$ for individual i

$$\Pr(J_t = 1|\boldsymbol{X}_t; \boldsymbol{\gamma}(\boldsymbol{\theta})) = \frac{\exp(\boldsymbol{X}_t\boldsymbol{\gamma}(\boldsymbol{\theta}))}{1 + \exp(\boldsymbol{X}_t\boldsymbol{\gamma}(\boldsymbol{\theta}))} \tag{1}$$

Algorithm for DDCs with Fully observed State Space (Part II)

Algorithm A.2 Stochastic Gradient Descent for optimizing $\boldsymbol{\gamma}$

Input: initial value $\boldsymbol{\gamma}_0$ for the policy parameters $\boldsymbol{\gamma}$, deep parameters $\boldsymbol{\theta}$, step size s_q, batch size I, initial state $\boldsymbol{X}_{i1}$ from the batch data D_I.
Initialize: $\boldsymbol{\gamma}_1 \leftarrow \boldsymbol{\gamma}_0$
for $q = 1, \ldots, Q$ **do**
  for $i = 1, 2, \ldots, I$ **do**
    Obtain $\widehat{\boldsymbol{D}}_i = (\widehat{\boldsymbol{X}}_i, \widehat{\boldsymbol{J}}_i)$ following the procedure in Algorithm A.1
    Calculate the lifetime value $V_i(\boldsymbol{\gamma}_q)$ and gradient $\nabla_{\boldsymbol{\gamma}} \log(\prod_{t>=1} \pi(J_{it}|X_{it}, t; \boldsymbol{\gamma}))$
  end for
  Average the lifetime value and gradient across individuals: $V(\boldsymbol{\gamma}_q) := \frac{1}{I} \sum_{i=1}^{I} V_i(\boldsymbol{\gamma}_q)$; $\nabla_{\boldsymbol{\gamma}} \log(\prod_{t>=1} \pi(J_t|X_t; \boldsymbol{\gamma})) := \frac{1}{I} \sum_{i=1}^{I} \nabla_{\boldsymbol{\gamma}} \log(\prod_{t>=1} \pi(J_{it}|X_{it}; \boldsymbol{\gamma}))$
  $\nabla_{\boldsymbol{\gamma}} V(\boldsymbol{\gamma}_q) \leftarrow V(\boldsymbol{\gamma}_q) \nabla_{\boldsymbol{\gamma}} \log(\prod_{t>=1} \pi(J_t|X_t; \boldsymbol{\gamma}))$
  $\boldsymbol{\gamma}_{q+1} \leftarrow \boldsymbol{\gamma}_q + s_q \nabla_{\boldsymbol{\gamma}} V(\boldsymbol{\gamma}_q)$
end for
Output: $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_{q^*}$

Algorithm for DDCs with Fully observed State Space (Part III)

Algorithm A.3 Indirect Inference for estimating $\boldsymbol{\theta}$

1: **Input:** initial value $\boldsymbol{\theta}_0$, initial state $\{\boldsymbol{X}_{11}, \boldsymbol{X}_{21}, ..., \boldsymbol{X}_{N1}\}$ from data D; data moments $\boldsymbol{\kappa}$.
2: **Initialize:** $\boldsymbol{\theta}_1 \leftarrow \boldsymbol{\theta}_0$
3: **for** $k = 1, \ldots, K$ **do**
4: Calculate $\boldsymbol{\gamma}^*(\boldsymbol{\theta}_k)$ following Algorithm A.2
5: **for** $i = 1, 2, \ldots, N$ **do**
Obtain $\widehat{\boldsymbol{D}}_i = (\widehat{\boldsymbol{X}}_i, \widehat{\boldsymbol{J}}_i)$ using $\boldsymbol{\gamma}^*(\boldsymbol{\theta}_k)$ following Algorithm A.1
6: **end for**
7: Calculate simulated moments $\widehat{\boldsymbol{\kappa}}$ using $(\widehat{\boldsymbol{X}}, \widehat{\boldsymbol{J}})$ according to targeted moments
8: $\boldsymbol{\theta}_{k+1} \leftarrow \min \text{Dis}(\boldsymbol{\kappa}, \widehat{\boldsymbol{\kappa}})$
9: **end for**
10: **Output:** $\boldsymbol{\theta}^* = \boldsymbol{\theta}_{k^*}$

Reinforcement Learning Estimation: A Toy Model Monte Carlo Study

- Use the proposed approach to estimating a simplified Rust bus engine replacement model (Rust 1987)
- The transition of the mileage when $J_{it} = 0$ is deterministic:

$$X_{it} = X_{it-1} + 1 \quad \text{if } X_{it-1} < M$$
$$X_{it} = X_{it-1} \quad \text{if } X_{it-1} = M,$$

  - When $J_{it} = 1, X_{it} = 0$
- The flow utility in each period is:

$$\begin{aligned} U(X_{it}, J_{it}) &= u(X_{it}, J_{it}; \theta_1, \theta_2) + \epsilon_{jt} \\ &= -\theta_1 X_{it} - \theta_2 \mathbb{I}(J_{it} = 1) + \epsilon_{jt}, \end{aligned}$$

- The agent's problem:

$$V(X_{it}, \epsilon_{jt}) = \max_{J_{it} \in \{0,1\}} \Big\{ u(X_{it}, J_{it}; \theta_1, \theta_2) + \epsilon_{jt} + \beta E\big[V(X_{it+1}, \epsilon_{jt+1}) | X_{it}, J_{it}\big] \Big\}.$$

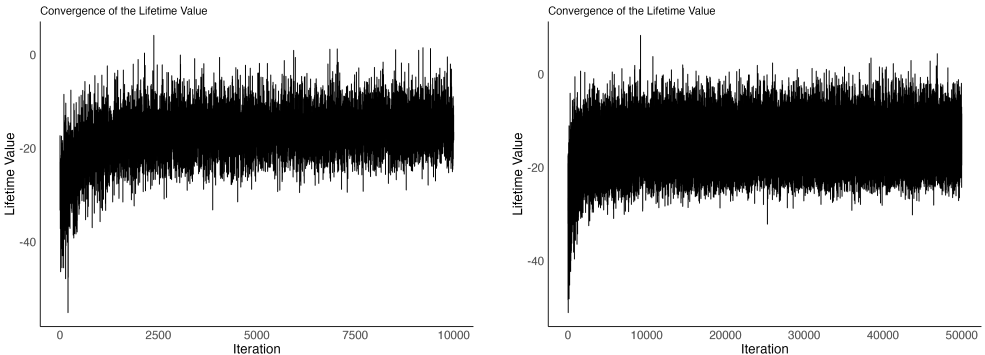
Simulation and Estimation Details

- Use Algorithm 1–Algorithm 3. Parametrize the choice probability as a function of the state variables:

$$\Pr(J_t = 1|X_t, t; \gamma) = \frac{\exp(\gamma_0 + \gamma_1 t + \gamma_2 X_t + \gamma_3 t X_t)}{1 + \exp(\gamma_0 + \gamma_1 t + \gamma_2 X_t + \gamma_3 t X_t)}.$$

- Moments to match: the empirical probability of bus engine replacement conditional on the time and mileage pair: (t, X)
- Simulate data for 500 buses who live for 20 periods and make decisions in each period
- Structural parameters (θ_1, θ_2) to estimate

Convergence of Lifetime Reward Over Iterations



(a) Iteration number=10000

(b) Iteration number=50000

Figure: Convergence of Lifetime Reward Under Different Iterations

Estimation Results for Simulation Study

Table: Comparison of Optimal Policy Function Between Different Methods

Method	**Mean(Rep.)**	**Mean(Rep.)** $X < 6$	**Mean(Rep.)** $X \geq 6$	**Lifetime Value**
Reinforcement learning	0.1642	0.0959	0.4730	-15.88
Backward induction	0.1696	0.1278	0.5148	-15.34

Table: Estimation Results for the Simulation Study

Parameter	**True Value**	**Est. Value (RL+II)**	**Est. Value (NFXP)**
θ_1	0.3	0.2922	0.2619
		(0.0545)	(0.0084)
θ_2	5.2	5.1302	5.2623
		(0.2345)	(0.1198)

Estimating a Dynamic Fertility Model with High-Dimensional State Space

- Focus on a dynamic collective fertility model with fully observed state space
- Couple in a household j interacts using a cooperative framework with limited commitment
- Use RL + indirect inference to show the computational advantage

Model setup

- Preferences:

$$u(c_t^g, l_t^g, n_t, \widetilde{n}_t^g) = \alpha_1 \ln c_t^g + \alpha_2 \ln l_t^g - \alpha_3((n_t - \widetilde{n}_t^g))^2 - p_t \times \mathbb{I}(n_t > 1)$$

- Household flow utility:

$$U(c_t^f, l_t^f, c_t^m, l_t^m, b_t, n_t, \widetilde{n}_t^f, \widetilde{n}_t^m, \theta_t) = \theta_t u(c_t^f, l_t^f, n_t, \widetilde{n}_t^f) + (1-\theta_t) u(c_t^m, l_t^m, n_t, \widetilde{n}_t^m) + \epsilon_{bt}$$

Model Setup (Cont'd)

- Pareto Weight Updating: $\theta_t = \{0.4, 0.3\}$, before and after the policy change
- The Couple's Problem

$$\mathbf{\Omega_t} = \{A_t^f, E^f, E^m, t_M, t_P, S, w_t^f, w_t^m, n_{t-1}, \widetilde{n}_t^f, \widetilde{n}_t^m, \theta_t\},$$
$$q_t = \{c_t^f, c_t^m, h_t^f, h_t^m, b_t\}.$$

$$V_t(\mathbf{\Omega_t}) = \max_{q_t} \theta_t u(c_t^f, l_t^f, n_t, \widetilde{n}_t^f) + (1-\theta_t) u(c_t^m, l_t^m, n_t, \widetilde{n}_t^m) + \epsilon_{bt} + \beta E_t[V_{t+1}(\mathbf{\Omega_{t+1}})]$$

$$c_t^f + c_t^m = \left(w_t^f h_t^f + w_t^m h_t^m - C_t(n_t)\right) \cdot e(n)$$
$$l_t^g + h_t^g = \bar{h}^g - x_t^g(n_t), \quad g \in \{f, m\}.$$

Simulation and Estimation Details

- Follow Algorithm A.1–Algorithm A.3, parametrize the fertility choice variable b_t as a function of the state space $\boldsymbol{\Omega_t}$:

$$\Pr(b_t = 1|\boldsymbol{\Omega_t}; \boldsymbol{\gamma}) = \frac{\exp(\gamma_0 + \boldsymbol{\Omega_t}\boldsymbol{\gamma})}{1 + \exp(\gamma_0 + \boldsymbol{\Omega_t}\boldsymbol{\gamma})},$$

  - $\boldsymbol{\gamma} = \{\gamma_1, ..., \gamma_{12}\}$: the policy function parameters for the 12 state variables.

- Simulate for 1583 couples with 11 periods
- High-dimensional state space: over 25 millions number of states
- Step size $= 1 \times e^{-4}$, batch size $I = 1$, iteration number $= 5000$
- Targeting a total of 58 statistics in the outer loop SMM

Estimation Results for Simulation Study: Fully Observed State Variables

Parameters	**Symbol**	**DGP**	**Est. (DP)**	**Est. (RL+II)** Iteration = 5000
Utility Function parameters				
Utility from $\ln(c)$	α_1	8	7.4154	8.1119
			(0.1570)	(0.3761)
Utility from $\ln(I)$	α_2	5	5.1471	5.0845
			(0.1861)	(0.4484)
Dis-utility from not ideal num. of child	α_3	5	4.4581	4.8470
			(0.3568)	(0.3468)
Policy parameters				
Penalty on excess birth in strict provinces	p_1	2	1.9454	1.9831
			(0.2213)	(0.2314)
Penalty on excess birth in loose provinces	p_2	1	0.9372	1.0990
			(0.2227)	(0.1798)
Time and Criterion Function Values				
Time (minutes)			15.32	2.79
			(2.9453)	(0.4460)
Criterion function value			0.0072	0.0116
			(0.0023)	(0.0088)

Outline

Model Environment

- A structural dynamic model with a process $\{X_t, S_t^*, J_t\}$
  - X_t: observed state variables
  - S_t^*: time-varying continuous unobserved state variables
  - $J_t \in \{0, 1\}$: agent's choice variable in period t
- Researchers observe a panel data set $\{(X_1, J_1), (X_2, J_2), ..., (X_T, J_T)\}$ for many agents
- For agent i, $\{(X_1, J_1, S_1^*), (X_2, J_2, S_2^*), ..., (X_T, J_T, S_T^*)\}$ i.i.d drawn from

$$f_{(X_1,J_1,S_1^*),(X_2,J_2,S_2^*),...,(X_T,J_T,S_T^*)},$$

which is a bounded continuous distribution

Identification Assumptions

Assumption 1 (First-order Markov and limited feedback)

(i) First-order Markov : $f_{X_t,J_t,S_t^*|X_{t-1},J_{t-1},S_{t-1}^*,\Omega_{<t-1}} = f_{X_t,J_t,S_t^*|X_{t-1},J_{t-1},S_{t-1}^*}$, where $\Omega_{<t-1} = \{X_{t-2}, ..., X_1, J_{t-2}, ..., J_1, ..., S_{t-2}^*, ..., S_1^*\}$.
(ii) *Limited feedback*: $f_{X_t,J_t|X_{t-1},J_{t-1},S_t^*,S_{t-1}^*} = f_{X_t,J_t|X_{t-1},J_{t-1},S_t^*}$.

Assumption 2 (Invertibility)

Let $V_t \equiv g_t(W_t)$, where $W_t = \{X_t, J_t\}$. The function $g_t : \mathbb{R}^2 \to \mathbb{R}$. Denote the supports of V_t and W_t as $\mathcal{V}_t$ and $\mathcal{W}_t$, respectively. Let $L_{V_{t-2},\overline{w}_{t-1},\overline{w}_t,V_{t+1}}$ denote the linear operator that maps from the $\mathcal{L}^P$ space of functions of V_{t+1} to the $\mathcal{L}^P$ splace of functions of V_{t-2}. There exists variable(s) V_t such that
(i) for any $w_t \in \mathcal{W}_t$, there exists a $w_{t-1} \in \mathcal{W}_{t-1}$ and a neighborhood $\mathcal{N}^2$ around (w_t, w_{t-1}) such that, for any $(\overline{w}_t, \overline{w}_{t-1}) \in \mathcal{N}^2$, $L_{V_{t-2},\overline{w}_{t-1},\overline{w}_t,V_{t+1}}$ is one-to-one.
(ii) for any $w_t \in \mathcal{W}_t$, $L_{V_{t+1}|w_t,S_t^*}$ is one-to-one.
(iii) for any $w_{t-1} \in \mathcal{W}_{t-1}$, L_{V_{t-2},w_{t-1},V_t} is one-to-one.

Identification Assumptions (Cont'd)

Assumption 3 (Uniqueness of spectral decomposition)

For any $w_t \in \mathcal{W}_t$ and any $\overline{s}_t^* \neq \widetilde{s}_t^*$,
there exists a $w_{t-1} \in \mathcal{W}_{t-1}$ and corresponding neighborhood $\mathcal{N}^r$ with
$\overline{w}_t \neq w_t, \overline{w}_{t-1} \neq w_{t-1}$: (i) $0 < k(w_t, \overline{w}_t, w_{t-1}, \overline{w}_{t-1}, s_t^*) < C < \infty$ for any $s_t^* \in \mathcal{S}_t^*$ and some constant C;
(ii) $k(w_t, \overline{w}_t, w_{t-1}, \overline{w}_{t-1}, \overline{s}_t^*) \neq k(w_t, \overline{w}_t, w_{t-1}, \overline{w}_{t-1}, \widetilde{s}_t^*)$, where

$$k(w_t, \overline{w}_t, w_{t-1}, \overline{w}_{t-1}, s_t^*) = \frac{f_{W_t|W_{t-1},S_t^*}(w_t|w_{t-1}, s_t^*) f_{W_t|W_{t-1},S_t^*}(\overline{w}_t|\overline{w}_{t-1}, s_t^*)}{f_{W_t|W_{t-1},S_t^*}(\overline{w}_t|w_{t-1}, s_t^*) f_{W_t|W_{t-1},S_t^*}(w_t|\overline{w}_{t-1}, s_t^*)}.$$

Assumption 4 (Monotonicity and normalization)

For any $w_t \in \mathcal{W}_t$, there exists a known functional G such that $G[f_{V_{t+1}|W_t,S_t^*}(\cdot|w_t, s_t^*)]$ is monotonic in s_t^*. We normalize $s_t^* = G[f_{V_{t+1}|W_t,S_t^*}(\cdot|w_t, s_t^*)]$.

Identification

Applying Theorem 1 in Hu and Shum (2012) , we have that

Lemma

Under Assumptions 1-4, and under the additional assumption that the Markov law of motion of the state variables is time-invariant, we have that:
(1) $f_{X_t,S_t^*|J_{t-1},X_{t-1},S_{t-1}^*}$ *is identified from* $f_{J_{t+1},X_{t+1},J_t,X_t,J_{t-1},X_{t-1},J_{t-2},X_{t-2}}$.
(2) Initial condition $f_{X_{t-2},S_{t-2}^*,J_{t-2}}$ *is identified from* $f_{J_{t+1},X_{t+1},J_t,X_t,J_{t-1},X_{t-1},J_{t-2},X_{t-2}}$.

Monte Carlo Study: A Rust Model with Extensions

- Rust (1987) bus engine replacement model with an unobserved state variable
  - $X_t \in \mathbb{R}$: accumulated mileage for bus i
  - $J_t \in \{0, 1\}$: bus company's engine replacement decisions in each period
  - $S_t^* \in \mathbb{R}$: condition of the bus i at time t
- S_t^* unobserved to econometricians, affecting both the transition of X_t and the flow utility in each period:
  - A higher S_t^* leads to faster accumulation in mileage
  - A higher S_t^* leads to a lower cost for maintaining the bus engine
  - S_t^* generates direct utility due to a better condition of the bus

Transitional Models and Initial Conditions

- The transitional process for S_t^* follows an AR(1) process:

$$S_t^* = \lambda_1 S_{t-1}^* + \nu_t; \quad \nu_t \sim N(0, \sigma_2^2)$$

- The transitional process for the mileage X_t if $J_t = 0$ is:

$$X_{t+1} = X_t\left[1 + \lambda_3 \exp(\eta_{t+1} + \lambda_2 S_{t+1}^*)\right], \quad f_{\eta_{t+1}}(\eta) = \exp(\eta - e^{\eta})$$

- The transitional process for mileage X_t if $J_t = 1$ is:

$$p(X_{t+1}|X_t, J_t, \beta_2) = \beta_2 \exp(-\beta_2 X_{t+1})$$

- The initial values of state variables (S_1^*, X_1) are independent:

$$S_1^* \sim N(\alpha_1, \sigma_1^2); \quad X_1 \sim \exp(\beta_1)$$

Flow Utility and the Bus Company's Problem

- The flow utility in period t is specified as follows:

$$\begin{aligned} U(X_t, J_t, S_t^*, \epsilon_{jt}) &= u(X_t, J_t, S_t^*; \theta_1, \theta_2, \theta_3) + \epsilon_{jt} \\ &= -\theta_1 X_t \exp(-S_t^*) - \theta_2 \mathbb{I}(J_t = 1) + \theta_3 S_t^* + \epsilon_{jt} \end{aligned}$$

- The value function of the bus company is:

$$\begin{aligned} V(X_t, S_t^*, \epsilon_{jt}) = \max_{J_t \in \{0,1\}} \big\{ & u(X_t, J_t, S_t^*; \theta_1, \theta_2, \theta_3) + \epsilon_{jt} \\ & + \beta E[V(X_{t+1}, S_{t+1}^*, \epsilon_{jt+1}) | X_{it}, S_{it}, J_t] \big\} \end{aligned}$$

Validating Assumptions

Focus on the case when $J_t = 0$, since when $J_t = 1$ the transitional model of X_t can be directly estimated

Assumption 1 (first-order Markov and limited feedback)

- X_t and S_t^* have a Markov structure for the law of motions
- Limited feedback is satisfied: $S_{t-1}^* \perp\!\!\!\perp X_t \mid S_t^*$

Assumption 2 (Invertibility)

- X_4 is a convolution of S_4^*: $\log\big[X_4 - X_3\big] - \log(\lambda_3 X_3) = \lambda_2 S_4^* + \eta_4$
- S_3^* is a convolution of S_2^* for fixed w_2: $S_3^* = \lambda_1 S_2^* + \nu_3$
- Initial values of the state variables (S_1^*, X_1) are independently distributed

Validating Assumptions (Cont'd)

Assumption 3 (Uniqueness of spectral decomposition)

$$k(w_3, \overline{w}_3, w_2, \overline{w}_2, s_3^*) = \exp\Big(- e^{-\lambda_2 s_3^*} \cdot \frac{-(\overline{x}_3 - x_3)(\overline{x}_2 - x_2)}{\lambda_3 x_2 \overline{x}_2} \Big), \quad \text{when } j_3 = 0.$$

Then we have $0 < k(w_3, \overline{w}_3, w_2, \overline{w}_2, s_3^*) < C$ for some finite C, and $k(w_3, \overline{w}_3, w_2, \overline{w}_2, \overline{s}_3^*) \neq k(w_3, \overline{w}_3, w_2, \overline{w}_2, \widetilde{s}_3^*)$

Assumption 4 (Monotonicity and normalization) We can set G to be

$$G(x_3, j_3, s_3^*) = E\Big[\log \frac{X_4 - x_3}{\lambda_3 x_3} | x_3, j_3, s_3^*\Big],$$

and we normalize $s_3^* = E\Big[\log \frac{X_4 - x_3}{\lambda_3 x_3} | x_3, j_3, s_3^*\Big]$.

Estimating DDCs with Unobserved State Variables Using RL

- Parametrize J_t as a function of both the observed and unobserved state variables:

$$\Pr(J_t = 1|X_t, S_t^*, t; \gamma) = \frac{\exp(\gamma_0 + \gamma_1 t + \gamma_2 X_t + \gamma_3 S_t^*)}{1 + \exp(\gamma_0 + \gamma_1 t + \gamma_2 X_t + \gamma_3 S_t^*)}. \quad (2)$$

- Three sets of parameters to estimate:
  - γ: policy function parameters
  - θ: structural parameters
  - ξ: transitional model parameters

- Two-layered loop structure:
  - Outer loop searching for optimal parameters (θ, ξ) using SMM
  - Inner loop searching for optimal policy parameters γ using policy gradient

- Key for identification: target moments in the outer loop that satisfy the nonparametric identification results in Lemma 1

Estimating DDCs with Unobserved State Variables Using RL: Flow Chart

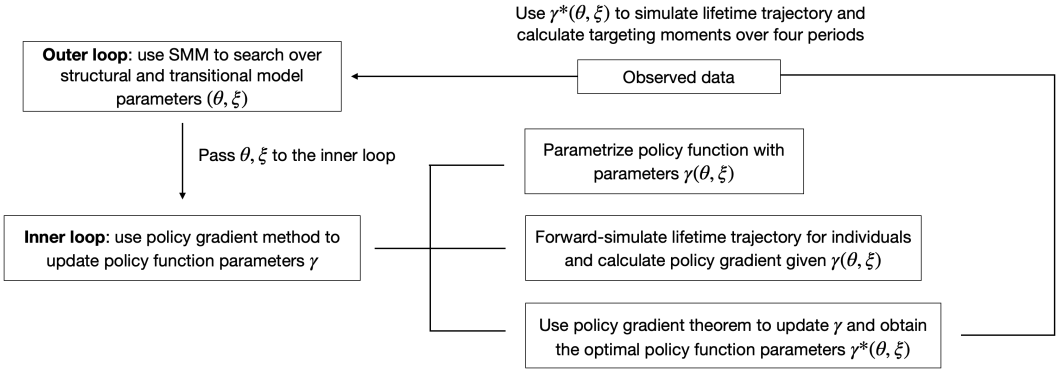


Figure: Flow chart for the Inner Loop of the RL + Indirect inference Algorithm for Models With Unobserved State Variables

Algorithm for DDCs with Unobserved State Variables (Part I)

Algorithm 1 Forward Simulation for Obtaining Individual's Lifetime Trajectory

1: **Input:** policy parameters $\boldsymbol{\gamma}(\boldsymbol{\theta}, \boldsymbol{\xi})$, utility parameters $\boldsymbol{\theta}$, transitional model parameter $\boldsymbol{\xi}$, initial state X_{i1} for individual i
2: Sample $\widehat{S}^*_{i1}$ using the initial distribution of S^* given $\boldsymbol{\xi}$
3: Sample $\widehat{J}_{i1}$ using Equation 2 given $(X_{i1}, \widehat{S}^*_{i1}; \boldsymbol{\gamma}(\boldsymbol{\theta}, \boldsymbol{\xi}))$
4: Sample $\widehat{S}^*_{i2}$ using the transitional model of S^* given $(\widehat{S}^*_{i1}; \boldsymbol{\xi})$
5: Sample $\widehat{X}_{i2}$ using the transitional model of X given $(\widehat{X}_{i1}, \widehat{J}_{i1}, \widehat{S}^*_{i2}; \boldsymbol{\xi})$
6: Sample $\widehat{J}_{i2}$ using Equation 2 given $(\widehat{X}_{i2}, \widehat{S}^*_{i2}; \boldsymbol{\gamma}(\boldsymbol{\theta}, \boldsymbol{\xi}))$
7: **for** $t = 3, \ldots, T$ **do**
8: Sample $\widehat{S}^*_{it}$, $\widehat{X}_{it}$, and $\widehat{J}_{it}$ conditional on $\boldsymbol{\theta}$, $\boldsymbol{\xi}$, and $\boldsymbol{\gamma}(\boldsymbol{\theta}, \boldsymbol{\xi})$
9: **end for**
10: Obtain the final dataset $\widehat{\boldsymbol{D}}_i = (\widehat{\boldsymbol{S}}_i, \widehat{\boldsymbol{X}}_i, \widehat{\boldsymbol{J}}_i)$
11: **Output:** lifetime trajectory $\widehat{\boldsymbol{D}}_i$ for individual i

Algorithm for DDCs with Unobserved State Variables (Part II)

Algorithm 2 Stochastic Gradient Descent for optimizing $\boldsymbol{\gamma}$

Input: initial value $\boldsymbol{\gamma}_0$ for the policy parameters $\boldsymbol{\gamma}$, deep parameters $\boldsymbol{\theta}$, transitional model parameter $\boldsymbol{\xi}$, step size s_q, batch size I, initial state $\{X_{11}, X_{21}, ..., X_{I1}\}$ from the batch data D_I.
Initialize: $\boldsymbol{\gamma}_1 \leftarrow \boldsymbol{\gamma}_0$
for $q = 1, \ldots, Q$ **do**
  for $i = 1, 2, \ldots, I$ **do**
    Obtain $\widehat{\boldsymbol{D}}_i = (\widehat{\boldsymbol{S}}_i, \widehat{\boldsymbol{X}}_i, \widehat{\boldsymbol{J}}_i)$ following the procedure in Algorithm 1
    Calculate the lifetime value $V_i(\boldsymbol{\gamma}_q)$ and gradient $\nabla_{\boldsymbol{\gamma}} \log(\prod_{t>=1} \pi(J_{it}|S^*_{it}, X_{it}, t; \boldsymbol{\gamma}))$
  end for
  Average the lifetime value and gradient: $V(\boldsymbol{\gamma}_q) := \frac{1}{I} \sum_{i=1}^{I} V_i(\boldsymbol{\gamma}_q)$;
  $\nabla_{\boldsymbol{\gamma}} \log(\prod_{t>=1} \pi(J_t|S^*_t, X_t; \boldsymbol{\gamma})) := \frac{1}{I} \sum_{i=1}^{I} \nabla_{\boldsymbol{\gamma}} \log(\prod_{t>=1} \pi(J_{it}|S^*_{it}, X_{it}; \boldsymbol{\gamma}))$
  Update $\nabla_{\boldsymbol{\gamma}} V(\boldsymbol{\gamma}_q) \leftarrow V(\boldsymbol{\gamma}_q) \nabla_{\boldsymbol{\gamma}} \log(\prod_{t>=1} \pi(J_t|S^*_t, X_t; \boldsymbol{\gamma}))$
  Update $\boldsymbol{\gamma}_{q+1} \leftarrow \boldsymbol{\gamma}_q + s_q \nabla_{\boldsymbol{\gamma}} V(\boldsymbol{\gamma}_q)$
end for
Output: $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_{q^*}$

Algorithm for DDCs with Unobserved State Variables (Part III)

Algorithm 3 Indirect Inference for estimating $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$

1: **Input:** initial value $\boldsymbol{\theta}_0$, initial value $\boldsymbol{\xi}_0$; initial state $\{X_{11}, X_{21}, ..., X_{N1}\}$ from data D; data moments $\boldsymbol{\kappa}$.
2: **Initialize:** $\boldsymbol{\theta}_1 \leftarrow \boldsymbol{\theta}_0$; $\boldsymbol{\xi}_1 \leftarrow \boldsymbol{\xi}_0$
3: **for** $k = 1, \ldots, K$ **do**
4: Calculate $\boldsymbol{\gamma}^*(\boldsymbol{\theta}_k, \boldsymbol{\xi}_k)$ following Algorithm 2
5: **for** $i = 1, 2, \ldots, N$ **do**
Obtain $\widehat{\boldsymbol{D}}_i = (\widehat{\boldsymbol{S}}_i, \widehat{\boldsymbol{X}}_i, \widehat{\boldsymbol{J}}_i)$ using $\boldsymbol{\gamma}^*(\boldsymbol{\theta}_k, \boldsymbol{\xi}_k)$ following Algorithm 1
6: **end for**
7: Calculate simulated moments $\widehat{\boldsymbol{\kappa}}$ using $(\widehat{\boldsymbol{X}}, \widehat{\boldsymbol{J}})$ according to targeted moments
8: $(\boldsymbol{\theta}_{k+1}, \boldsymbol{\xi}_{k+1}) \leftarrow \min \text{Dis}(\boldsymbol{\kappa}, \widehat{\boldsymbol{\kappa}})$
9: **end for**
10: **Output:** $\boldsymbol{\theta}^* = \boldsymbol{\theta}_{k^*}$, $\boldsymbol{\xi}^* = \boldsymbol{\xi}_{k^*}$

Targeted Moments for SMM

- We match the moments based on $(J_{t+1}, X_{t+1}, J_t, X_t, J_{t-1}, X_{t-1}, J_{t-2}, X_{t-2})$
  - Because of the identification results in Hu and Shum (2012), we only need to match moments of observed state and choice variables to identify θ and ξ
  - We need to match the moments of four periods' observed data for identification
- We match on 28 coefficients for the regressions for $t \geq 3$

$$
\begin{aligned}
J_{t+1} &= \kappa_0 + \kappa_1 X_{t+1} + \kappa_2 J_t + \kappa_3 X_t + \kappa_4 J_{t-1} + \kappa_5 X_{t-1} + \kappa_6 J_{t-2} + \kappa_7 X_{t-2} + \epsilon_1 \\
J_t &= \kappa_8 + \kappa_9 X_{t+1} + \kappa_{10} J_{t+1} + \kappa_{11} X_t + \kappa_{12} J_{t-1} + \kappa_{13} X_{t-1} + \kappa_{14} J_{t-2} + \kappa_{15} X_{t-2} + \epsilon_2 \\
J_{t-1} &= \kappa_{16} + \kappa_{17} X_{t+1} + \kappa_{18} J_t + \kappa_{19} X_t + \kappa_{20} J_{t+1} + \kappa_{21} X_{t-1} + \kappa_{22} J_{t-2} + \kappa_{23} X_{t-2} + \epsilon_3 \\
J_{t-2} &= \kappa_{24} + \kappa_{25} X_{t+1} + \kappa_{26} J_t + \kappa_{27} X_t + \kappa_{28} J_{t-1} + \kappa_{29} X_{t-1} + \kappa_{30} J_{t+1} + \kappa_{31} X_{t-2} + \epsilon_3
\end{aligned}
$$

- We also match on the mean, standard deviation, and correlation matrix of $(J_{t+1}, X_{t+1}, J_t, X_t, J_{t-1}, X_{t-1}, J_{t-2}, X_{t-2})$ for $t \geq 3$
- A total of 72 moments to match in the SMM model

Simulation and Estimation Details

Simulation Setting

- Simulate data for 1000 buses who lives for 10 periods
- Using backward induction to solve for agent's optimal choice
  - Discretize S_t^* into 5 grid points
  - Discretize X_t into 20 grid points
  - 10 periods $\times$ 5 unobserved states $\times$ 20 mileages: 1000 states
- Forward simulation following the continuous model setting, given the obtained optimal policy function

Estimation setting

- Pre-estimate β_1 and β_2, normalize $\lambda_3 = 0.2$
- Step size $s_q = 1 \times e^{-4}$, batch size $I = 1$, policy iteration $= 5000$
- State variables normalized by mean and std. in the policy function
- Remaining 8 parameters to estimate

Simulation Results: Continuous Unobserved State Variables

Parameters	True (1)	Est. (1)	True (2)	Est. (2)	True (3)	Est. (3)
Utility parameters						
θ_1	0.5	0.454	1	1.041	1	1.059
		(0.034)		(0.080)		(0.065)
θ_2	5	4.696	5	5.023	10	9.425
		(0.285)		(0.769)		(0.714)
θ_3	2	1.740	2	1.840	2	1.705
		(0.302)		(0.350)		(0.269)
Transitional process parameters						
α_1	1	0.974	1	0.938	1	0.952
		(0.061)		(0.075)		(0.057)
σ_1	1	0.976	1	0.939	1	0.943
		(0.050)		(0.091)		(0.070)
λ_1	0.5	0.544	0.5	0.439	0.5	0.473
		(0.085)		(0.057)		(0.071)
λ_2	0.8	0.760	0.8	0.748	0.8	0.746
		(0.064)		(0.078)		(0.058)
σ_2	0.2	0.164	0.2	0.139	0.2	0.144
		(0.019)		(0.059)		(0.057)

DDCs with Discrete Unobserved Heterogeneity: A Special Case

- Focus on the popular setting: a finite mixture model
- Consider the Rust bus engine problem with a fixed and discrete unobserved variable
  - $S^* \in \{1, 2\}$: condition of the bus that is fixed over time
  - X_t: observed discrete accumulated mileage
  - $J_t \in \{0, 1\}$: engine replacement decision
- Compare with the results in Arcidiacono and Miller (2011)
  - Follow similar simulation setting as in their paper using the extended Rust model
- A special case under the general framework proposed for RL estimation for DDCs with unobserved state variables

Model Setup

- The flow utility is:

$$u(X_t, S^*) = \begin{cases} \theta_0 + \theta_1 \min\{X_t, 25\} + \theta_2 S^* & \text{if } J_t = 0 \\ 0 & \text{if } J_t = 1 \end{cases} .$$

- When $J_t = 0$, the transitional model for X_t is:

$$f(X_{t+1}|X_t) = \begin{cases} \exp(-(X_{t+1} - X_t)) - \exp(-(X_{t+1} + 0.125 - X_t)) \text{ if } X_{t+1} \geq X_t \\ 0, \text{ otherwise} \end{cases} .$$

  - X_t accumulates in increments of 0.125
  - $X_{t+1} = 0$ when $J_t = 1$

- $S^* = 1$ with probability π_0, and $S^* = 2$ with probability $1 - \pi_0$

Estimation Using RL + Indirect Inference method

- Parametrize J_t as a function of both the observed and unobserved state variables:

$$\Pr(J_t = 1|X_t, S_t^*, t; \gamma) = \frac{\exp(\gamma_0 + \gamma_1 t + \gamma_2 X_t + \gamma_3 S^*)}{1 + \exp(\gamma_0 + \gamma_1 t + \gamma_2 X_t + \gamma_3 S^*)}. \quad (3)$$

- Estimate (γ, θ, ξ) following Algorithm 1-3
  - In the outer loop, use SMM to update $\{\theta_0, \theta_1, \theta_2, \pi_0\}$ by matching the 72 moments
  - In the inner loop
    - Conditional on $\{\theta_0, \theta_1, \theta_2, \pi_0\}$ in the outer loop
    - Conduct forward simulation using S's initial distribution, X's transitional process, and policy function equation (3)
    - Use policy gradient theorem to update $\{\gamma_0, \gamma_1, \gamma_2, \gamma_3\}$
- Identification is guaranteed by Lemma 1 and the choice of targeting moments

Simulation and Estimation Details

Simulation Setting

- Simulate data for 1000 buses who lives for 30 periods
  - Assume $X_{i1} = 0$ for all i
  - Keep the last 20 periods of data after simulation
- Using backward induction to solve for agent's optimal choice
  - 20 periods $\times$ 2 unobserved states $\times$ 201 mileages: 8,040 states

Estimation setting

- Step size $s_q = 1 \times e^{-4}$, batch size $I = 1$, try different policy iteration numbers
- State variables normalized by mean and std. in the policy function
- Report three structural utility parameters

Simulation Results: Discrete Unobserved State Variables

Parameters	DGP	Est. (CCP)	Est. (RL+II) 5000	Est. (RL+II) 20000
		Utility parameters		
θ_0 (intercept)	2	2.0344	1.9606	1.9888
		(0.1394)	(0.1803)	(0.1755)
θ_1 (mileage)	-0.15	-0.1481	-0.1579	-0.1565
		(0.0057)	(0.0242)	(0.0204)
θ_2 (unobs. state)	1	1.0412	1.0055	0.9910
		(0.1129)	(0.0868)	(0.1031)
Time (minutes)		0.6553	0.3983	1.15

Note: Mean and standard deviations for 50 simulations. The observed data consists of 1000 buses for 20 periods. The column CCP presents estimation results using Arcidiacono and Miller 2011's two-step EM algorithm. The rest columns shows the estimation results using RL+Indirect Inference methods with different iteration numbers for solving the optimal policy function. The initial values for the estimated parameters are (1.8, -0.17, 0.9, 0.45).

Empirical Application

A Dynamic Fertility Model with Time-Varying Unobserved Pareto Weights

- The transitional process of the Pareto weights θ_t becomes unobserved and follows

$$\begin{aligned}
\pi_1 &= \alpha_4(\omega_1^f - \omega_1^m) + \nu_1, \quad \nu_1 \sim N(0, \sigma_1^2), \\
\pi_t &= \alpha_5 \pi_{t-1} + \alpha_6\left[(w_t^f - \omega_t^f) - (w_t^m - \omega_t^m)\right] + \nu_2, \quad \nu_2 \sim N(0, \sigma_2^2), \\
\theta_t &= \exp(\pi_t)/(1 + \exp(\pi_t)),
\end{aligned}$$

  - The Pareto weight follows an AR(1) process inside the link function
- Other model elements the same as in the fully observed case
- Adopt Algorithm 1–Algorithm 3 for estimation
- Five utility parameters $(\alpha_1, \alpha_2, \alpha_3, p_1, p_2)$ and five transitional process parameters $(\alpha_4, \sigma_1, \alpha_5, \alpha_6, \sigma_2)$ to be estimated

Empirical Application Results

Table: Estimation Results for Empirical Study: Continuous Unobserved State Variables

Parameters	Symbol	Initial Value	Est. (RL) Iter. = 5000
Utility parameters			
Utility Function parameters			
Utility from $\ln(c)$	α_1	7.6	7.0041 (0.2348)
Utility from $\ln(l)$	α_2	5.4	4.9224 (0.2582)
Dis-utility from not ideal num. of child	α_3	2.7	1.8271 (0.1129)
Pareto weight parameters			
First period in initial wage diff.	α_4	0.1355	0.1473 (0.0239)
First period standard deviation	σ_1	0.5	0.4113 (0.0188)
Later periods AR(1) parameter	α_5	0.9	0.8370 (0.0188)
Later periods wage shock diff.	α_6	0.1770	0.1668 (0.0197)
Later periods standard deviation	σ_2	0.5	0.1224 (0.0188)

Parameters	Symbol	Initial Value	Est. (RL) Iter. = 5000
Utility parameters			
Policy parameters			
Penalty on excess birth in strict provinces	p_1	1	0.9547 (0.5596)
Penalty on excess birth in loose provinces	p_2	0.5	0.4405 (0.2370)
Time and Criterion Function Values			
Time (minutes)			5.05
Criterion function value		4.60	0.011

Outline

Conclusion

- Our framework merges reinforcement learning's policy gradient methods with economics' indirect inference, revolutionizing dynamic discrete choice model estimation
- It reduces computational burden linked to high-dimensional state spaces by directly parametrizing the policy function
- Adaptable to models with partially observed state variables, leveraging non-parametric identification results in Hu and Shum (2012)
- Empirically validated across diverse models
  - Yields accurate estimates for models with unobserved state variables
  - Outperforms traditional methods in computational efficiency
- Introduces a robust estimation framework for DDCs, blending reinforcement learning insights with econometric techniques to offer a comprehensive and efficient estimation approach