From AI to Wisdom: Optimal Monetary Policy via Interpretable Reinforcement Learning

Yifan Wang
Renmin University of China
and
Hanfang Yang
Renmin University of China

September 12, 2024

Abstract

We use interpretable Reinforcement Learning (RL) to explore and understand optimal monetary policy rules. By conducting statistical inference on policy functions within offline RL in a discrete action space, we construct an interpretable RL approach. Additionally, we provide methods to extend this approach to continuous action spaces and online learning. We illustrate the effectiveness of our method through the analysis of U.S. data, by identifying and analyzing the Federal Reserve's optimal monetary policy. We find that when actual policies align with optimal policies, lower inflation, reduced unemployment, and higher economic growth can be achieved. Our approach provides interpretative tools for the optimal policy, thereby assisting policymaker and even humanity in gaining knowledge and wisdom from artificial intelligence (AI) and RL method.

Keywords: Economic policy; Markov decision processes; Confidence interval

1 Introduction

Monetary policy rule is a formula that provides the policy objectives, tools, and methods of policy formulation under existing economic conditions. After discretionary policymaking led to stagflation in Western economies, Milton Friedman proposed adhering to a fixed, singular quantitative rule to constrain policymakers (Friedman, 1969). One of the most widely accepted frameworks among scholars is the Taylor rule (Taylor, 1993), which dictates that the interest rate is determined by the central bank's target rate, inflation gap, and output gap. However, in reality, central banks consider a vast array of information when setting interest rates (Brash et al., 2001; Bernanke and Boivin, 2003). Therefore,Svensson (2003) argues that the Taylor rule, which focuses solely on inflation and output gaps, is not optimal. He suggests that the Taylor rule is more applicable in the United States because inflation and output gaps are the most critical economic variables for a large, trade-independent, and strong-currency country. In contrast, other economies, such as small, trade-dependent, or weak-currency countries, must consider additional variables like exchange rates, imports and exports, employment, financial stability, and balance of payments when formulating monetary policies (Xu, 2017). Existing research on interest rates attempts to explain the factors central banks focus on in monetary policy, such as capital (Adler et al., 2016), exchange rate fluctuations (Alstadheim et al., 2021; Lubik and Schorfheide, 2007), and asset prices (Bernanke and Gertler, 2001).

Most of these studies rely on regression analysis to examine the influencing factors of current monetary policies, which are not necessarily optimal. Therefore, most existing research on optimal monetary policy focuses on constructing theoretical macroeconomic models and analyzing representative agents (Khan et al., 2003; Angeletos and La'o, 2020), with less use of actual data. However, theoretical models are simplifications of reality,

making it difficult to capture all real-world situations, and for complex models, finding the optimal policy-making rules is also challenging.

As a result, in recent years, researchers have started attempting to use machine learning (ML) and artificial intelligence (AI) to deeply analyze existing data or utilize economic models in an effort to identify optimal economic policies. Within these methods, Reinforcement Learning (RL) transcends the limitations of data by learning through interaction with the environment, thus acquiring the ability to surpass existing knowledge. Specifically, in RL, an agent interacts with the environment, receiving rewards based on the current state and action. Through continual trial and error, the agent aims to maximize the cumulative rewards (value function) over time, ultimately discovering the optimal policy rule in the given environment (Sutton and Barto, 2018; Nian et al., 2020). The AI Economist framework proposed by Zheng et al. (2022) is an excellent example of applying RL to economic decision-making. They designed deep multi-agent RL to formulate tax policies, discovering that the resulting policies effectively improve social welfare and balance between equality and productivity. Similarly, Chen et al. (2021) found that deepRL could effectively solve the steady-state values of the classic dynamic stochastic general equilibrium (DSGE) model in economics, thus extending the DSGE model to more complex structures.

However, the optimal policies formulated by AI are difficult to apply in practice, as the decision-making process is hard for users to understand. Given the wide-reaching impact of economic policies, policymakers are unlikely to risk implementing them without fully comprehending the rationale behind the AI's decisions. Therefore, humans must first understand and learn from the optimal policies generated by AI. But as AI progresses towards wisdom, its ability to assist humans in this domain remains limited. For instance, AI products like AlphaGo (Silver et al., 2017) and ChatGPT (Radford et al., 2018, 2019),

after aggregating vast amounts of data, can provide information and create new knowledge to some extent (such as new chess strategies), but they fall short of helping humans attain wisdom. In recent years, explainable AI and interpretable RL have shown potential in contributing to the formation of wisdom (Adadi and Berrada, 2018). For example, building on AI Economist framework, Trott et al. (2021) provides an interpretable policy design for the stringency of US state policies and Federal subsidies during COVID-19.

Our paper deepens the above research from the perspective of statistical inference. Specifically, we propose a framework for finding optimal monetary policies based on RL and suggest conducting statistical inference on related parameters to analyze the factors of optimal policies from a statistical perspective, thereby enhancing the understanding of optimal monetary policy, and generating knowledge even wisdom through RL rather than merely creating products like AlphaGo.

Our contributions can be summarized in three points. First, utilizing interpretable RL, we examined the optimal monetary policy-making rules using the United States as a case study. Based on key variables from the U.S. financial market and macroeconomy, we analyzed the optimal monetary policy rules and the corresponding coefficients for three policy goals: stabilizing inflation, increasing employment, and promoting economic growth (Svensson, 2014; Loungani and Sheets, 1997). This allowed us to identify the factors and their weights that the Federal Reserve should consider when formulating optimal rules. Unlike traditional studies based on economic models (Khan et al., 2003; Angeletos and La'o, 2020), we used real-world data from suboptimal policies, employing RL to explore optimal monetary policies by learning from the Federal Reserve's past successes and failures rather than relying on case studies.

Second, we extended the existing theoretical research on interpretable RL. To under-

stand the optimal policies generated by RL, existing studies can be categorized into two approaches. One approach simplifies the policy function, adopting forms like logistic regression to obtain interpretable coefficients (Trott et al., 2021). The other approach uses deep learning to derive the optimal strategy and further refines it using models like decision trees (Milani et al., 2024). However, these studies lack statistical inference for the parameters, meaning they do not account for the impact of data randomness. In fact, RL, which relies on agent-environment interaction, does not satisfy many assumptions required for traditional confidence interval computation (White and White, 2010), making statistical inference in RL challenging. Current statistical inference in RL primarily focuses on value functions and simplifies computation by linearizing the value or policy functions (Luckett et al., 2020; Shi et al., 2022; Ramprasad et al., 2022). Based on the above researches, we perform statistical inference on the parameters of the policy function in RL. Specifically, we build on V-learning framework by linearizing the state value function and using a sigmoid-form policy function to provide confidence intervals for the estimated optimal policy parameters obtained from samples. Conducting statistical inference on policy function parameters in RL can significantly enhance our understanding of the operational logic behind policies, particularly optimal policies.

Third, we proposed a scalable interpretable RL framework that can be applied to broader policy-making and business decision-making. Existing research on using RL for economic decisions often relies on discrete action spaces (Zheng et al., 2022; Trott et al., 2021), assuming that policy decisions, such as interest rates, are discrete. However, in reality, interest rates, government spending, and corporate investments often change continuously. Therefore, our research provides theoretical insights for both discrete and continuous action spaces, offering a foundation for future studies. Additionally, we introduced

a method for extending the framework to online learning, allowing for real-time updates of optimal decisions based on current conditions.

We validated the feasibility and effectiveness of using interpretable RL to formulate optimal monetary policies through numerical simulations and an empirical study of U.S. monetary policy. The numerical simulations confirmed the theoretical results regarding the asymptotic distribution of policy function parameters in RL. In the empirical study, we simplified monetary policy decisions into interest rate hikes and cuts, and found that from 1954 to 2023, only about 57% of the Federal Reserve's decisions aligned with the optimal policies derived through RL. Furthermore, when the Federal Reserve's actual policy choices matched the RL-derived optimal policies, the U.S. economy experienced lower inflation, higher employment, and faster economic growth, regardless of the Fed's specific objective (whether stabilizing inflation, increasing employment, or promoting economic growth). Additionally, the coefficient results of the optimal monetary policy indicated that, compared to the Fed's current policy-making rules, the optimal policy placed less emphasis on economic growth and prioritized employment conditions.

In the following paper, Section 2 estimates the confidence intervals for the parameters of the policy function in a discrete action space under offline learning. Section 3 extends our research to continuous action spaces and online learning. Section 4 describes the simulation experiment setup and analyzes the results. Section 5 uses macroeconomic data in US to study the optimal monetary policies. Section 6 concludes the paper and discusses further implications.

2 Confidence Inverval Estimation in Discrete Action Spaces under Offline Learning

2.1 Settings

Assume that at time $t \in \mathbb{Z}_+$ $(t = 1, \ldots, T)$, the observable state vector in the environment is $\boldsymbol{S}_t \in \boldsymbol{\mathcal{S}} \subset \mathbb{R}^p$, following a distribution F_S. The action taken by the agent is $A_t \in \boldsymbol{\mathcal{A}}$, where the action space $\boldsymbol{\mathcal{A}}$ is a discrete space with q elements. The observed reward is $R_t \in \mathbb{R}$. The agent's policy for selecting action $A_t = a_t$ based on the current state $\boldsymbol{S}_t = \boldsymbol{s}_t$ is given by $\boldsymbol{\pi}(a_t \mid \boldsymbol{s}_t)$.

For the observed data sequence $(\boldsymbol{S}_1, A_1, R_1, \boldsymbol{S}_2, A_2, R_2, \ldots, \boldsymbol{S}_T, A_T, R_T, \boldsymbol{S}_{T+1})$, we assume it follows a Markov process and satisfies time homogeneity, that is:

$$P\left(\boldsymbol{S}_{t+1} = \boldsymbol{s}' \mid \boldsymbol{S}_t = \boldsymbol{s}, A_t = a, \{\boldsymbol{S}_i, A_i, R_i\}_{1 \le i < t}\right) = p\left(\boldsymbol{s}' \mid \boldsymbol{s}, a\right),$$

and the observed rewards are assumed to satisfy the Conditional Mean Independence Assumption (CIMA):

$$\mathbb{E}\left(R_t \mid \boldsymbol{S}_t = \boldsymbol{s}, A_t = a, \{\boldsymbol{S}_i, A_i, R_i\}_{1 \le i < t}\right) = \mathbb{E}\left(R_t \mid \boldsymbol{s}, a\right).$$

Given the discount factor $\gamma \in [0, 1)$, following Sutton and Barto (1998, 2018), the state value function is defined as

$$V(\boldsymbol{\pi}, \boldsymbol{s}_t) = \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i R_{t+i} | \boldsymbol{S}_t = \boldsymbol{s}_t\right]. \tag{1}$$

Combining the CIMA condition, we can deduce that there exists at least one optimal policy $\boldsymbol{\pi}^{opt}$ such that for all $\boldsymbol{\pi}$ and $\boldsymbol{s} \in \boldsymbol{\mathcal{S}}$, we have $V(\boldsymbol{\pi}^{opt}, \boldsymbol{s}) \ge V(\boldsymbol{\pi}, \boldsymbol{s})$ (Puterman, 1994, Theorem 6.2.12).

2.2 Parameterization of the Policy Function

First, we need to parameterize the policy function, i.e., $\boldsymbol{\pi}(a \mid \boldsymbol{s}) = \boldsymbol{\pi}(a \mid \boldsymbol{s}, \boldsymbol{\beta})$, which can also be written in simplified form as $\boldsymbol{\pi}$ or $\boldsymbol{\pi}_{\boldsymbol{\beta}}$. The policy parameterization is fundamental to policy gradient algorithms (Sutton et al., 2000). It can be modeled using generalized linear models or more complex neural network models, the latter being part of deep RL (DRL) (Mnih et al., 2015). However, the objective of this study is to perform statistical inference on the parameters of the policy function for interpretability. Therefore, we employ a linear parameterization of the policy function, rather than more complex but less interpretable models.

Specifically, for q possible actions, the probability of the agent selecting each action can be expressed in the form of multinomial logistic regression:

$$\pi_{\boldsymbol{\beta},j} = \pi(a^{(j)} \mid \boldsymbol{s}, \boldsymbol{\beta}) = \begin{cases} \dfrac{1}{1 + \sum_{i=2}^{q} \exp(\boldsymbol{s}^\top \boldsymbol{\beta}_i)}, & j = 1 \\ \dfrac{\exp(\boldsymbol{s}^\top \boldsymbol{\beta}_j)}{1 + \sum_{i=2}^{q} \exp(\boldsymbol{s}^\top \boldsymbol{\beta}_i)}, & j = 2, 3, \ldots, q \end{cases}, \tag{2}$$

where $\boldsymbol{\beta}_j = (\beta_{1j}, \beta_{2j}, \ldots, \beta_{pj})^\top \in \boldsymbol{\mathcal{B}} \subset \mathbb{R}^p$ represents the parameters corresponding to the j-th action, and the baseline action is the first action, i.e., $\boldsymbol{\beta}_1 = (0, 0, \ldots, 0)^\top$, resulting in the parameter matrix as follows:

$$\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_q] = \begin{bmatrix} 0 & \beta_{12} & \cdots & \beta_{1q} \\ 0 & \beta_{22} & \cdots & \beta_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \beta_{p2} & \cdots & \beta_{pq} \end{bmatrix}_{p \times q}.$$

To estimate the parameters $\boldsymbol{\beta}$, we need to parameterize the state value function $V(\boldsymbol{\pi}, \boldsymbol{s}_t)$ and obtain its confidence interval, and subsequently derive the confidence interval for $\boldsymbol{\beta}$ through its partial derivatives. If the policy gradient theorem (Sutton and Barto, 2018) were used directly to obtain the partial derivatives, the action value function would still need to

be parameterized and estimated. For example, Shi et al. (2022) considers parameterizing the action value function due to the possibility that the policy function $\boldsymbol{\pi}(a \mid \boldsymbol{s})$ may not be a continuous function of the state $\boldsymbol{s}$. However, since this study assumes that the policy function satisfies continuity (2), we will directly parameterize the state value function and subsequently solve for the confidence interval of $\boldsymbol{\beta}$ [1].

2.3 Offline Confidence Interval Estimation

Reinforcement learning can be categorized into offline and online learning. Offline learning searches for the optimal policy based on existing data, while online learning interacts with the environment and trains simultaneously, acquiring new data in real-time through interaction. This section first constructs the confidence interval estimation in the offline setting, and the next section will extend this to the online setting.

Assume there are n independent and identically distributed sequences in the dataset, with each sequence having a total time of T_i $(i = 1, 2, \ldots, n)$:

$$\{(\boldsymbol{S}_{i,1}, A_{i,1}, R_{i,1}, \boldsymbol{S}_{i,2}, A_{i,2}, R_{i,2}, \ldots, \boldsymbol{S}_{i,T_i}, A_{i,T_i}, R_{i,T_i}, \boldsymbol{S}_{i,T_i+1})\}_{i=1,2,\ldots,n}.$$

These sequences are generated by the behavior policy $b_t(a_t \mid \boldsymbol{s}_t)$, which is typically known or can be estimated from the data and may differ from the optimal policy $\boldsymbol{\pi}^{opt}$. Let the total sample size be $N = \sum_{i=1}^{n} T_i$. The objective of this paper is to derive the optimal policy $\boldsymbol{\pi}^{opt}$ and its parameters, as well as the corresponding confidence intervals, based on the existing data sequences.

First, to construct the objective function, the following assumptions are needed:

[1] The policy gradient theorem (Sutton et al., 2000) provides the first partial derivative, but this derivative also includes the action value function $Q(\boldsymbol{\pi}_{\boldsymbol{\beta}}, \boldsymbol{s}, a)$, which itself contains the parameter $\boldsymbol{\beta}$. Therefore, to solve for the second partial derivative, it is still necessary to derive the derivative of $Q(\boldsymbol{\pi}_{\boldsymbol{\beta}}, \boldsymbol{s}, a)$ with respect to $\boldsymbol{\beta}$, which also requires linearization. Hence, this study directly linearizes the state value function.

Assumption 1. ***State and Action Distribution:*** *The distribution of the state $\boldsymbol{S}_t$ has finite mean and variance, and the action space for A_t is discrete and finite.*

Assumption 2. ***Bounded Rewards:*** *The reward function R_t is bounded.*

Assumption 3. ***No Omitted Variables:*** *Given t, the choice of A_t depends only on $\boldsymbol{S}_t$.*

Assumption 4. ***Time Homogeneity:*** *For any t, $P(\boldsymbol{S}_{t+1}, R_t \mid A_t = a, \boldsymbol{S}_t = s) = P(\boldsymbol{S}_t, R_{t-1} \mid A_{t-1} = a, \boldsymbol{S}_{t-1} = s)$.*

Assumption 5. ***Action Reachability:*** *For all $a_t \in \mathcal{A}$, $\boldsymbol{s}_t \in \boldsymbol{\mathcal{S}}$, and t, $P(A_t = a_t \mid \boldsymbol{S}_t = \boldsymbol{s}_t) > 0$.*

Assumptions 1 and 2 provide the necessary properties that the distribution of states and actions, as well as the rewards, must satisfy to ensure the solvability of the model. The assumption of a finite set of actions is derived from the form of the parameterized policy function (2). Assumption 3 indicates that there are no unobserved confounding variables affecting the agent's decisions, which depend only on the current state. In practice, this assumption requires decision-makers to include all their decision bases in the model, which is challenging in traditional economic model-based research (e.g., DSGE). However, in data-driven offline RL, this assumption becomes feasible due to the availability of extensive data.

Assumption 4 requires that states and actions exhibit time homogeneity, which, along with Assumption 3, satisfies the requirements of a Markov Decision Process (MDP). Assumption 5 states that at any time t, for any state $\boldsymbol{s}_t$, the probability of taking any action a_t is non-zero, implying that the policy is stochastic. This assumption ensures that the observed policy $b_t(a_t \mid \boldsymbol{s}_t)$ is always greater than zero, thereby maintaining a certain level of exploration, which can be achieved through policies such as ϵ-greedy.

Under Assumptions 3 to 5, and referencing Theorem 2.1 by Luckett et al. (2020), the state value function (1) can be further expressed as:

$$\mathbb{E}\left[\frac{\pi\left(A_t \mid \boldsymbol{S}_t\right)}{b_t\left(A_t \mid \boldsymbol{S}_t\right)}\left\{R_t+\gamma V\left(\boldsymbol{\pi}, \boldsymbol{S}_{t+1}\right)-V\left(\boldsymbol{\pi}, \boldsymbol{S}_t\right)\right\} \mid \boldsymbol{S}_t\right]=\mathbf{0}.$$

Consequently, for any function ψ defined on $\boldsymbol{\mathcal{S}}$, we have:

$$\mathbb{E}\left[\frac{\pi\left(A_t \mid \boldsymbol{S}_t\right)}{b_t\left(A_t \mid \boldsymbol{S}_t\right)}\left\{R_t+\gamma V\left(\boldsymbol{\pi}, \boldsymbol{S}_{t+1}\right)-V\left(\boldsymbol{\pi}, \boldsymbol{S}_t\right)\right\} \psi\left(\boldsymbol{S}_t\right)\right]=\mathbf{0} \tag{3}$$

Then, we can linearize the state value function and approximate it using basis functions (Huang, 1998; Hastie et al., 2009):

$$V(\boldsymbol{\pi}, \boldsymbol{s}_t) \approx V(\boldsymbol{\pi}, \boldsymbol{s}_t; \boldsymbol{\theta}_{\boldsymbol{\pi}})=\boldsymbol{\Phi}_m^{\top}\left(\boldsymbol{s}_t\right) \boldsymbol{\theta}_{\boldsymbol{\pi}}, \tag{4}$$

where $\boldsymbol{\Phi}_m\left(\boldsymbol{s}_t\right)=\left\{\phi_1\left(\boldsymbol{s}_t\right), \ldots, \phi_m\left(\boldsymbol{s}_t\right)\right\}^{\top}$ represents a given m-dimensional vector of basis functions, and $\boldsymbol{\theta}_{\boldsymbol{\pi}} \subset \mathbb{R}^m$ are the parameters. To simplify the solution process, we define $\psi\left(\boldsymbol{S}_t\right)=\boldsymbol{\Phi}_m^{\top}\left(\boldsymbol{S}_t\right)$, then the equation (3) can be further expressed as:

$$\mathbb{E}\left(\frac{\pi\left(A_t \mid \boldsymbol{S}_t\right)}{b_t\left(A_t \mid \boldsymbol{S}_t\right)}\left[R_t \boldsymbol{\Phi}_m\left(\boldsymbol{S}_t\right)+\boldsymbol{\Phi}_m\left(\boldsymbol{S}_t\right)\left\{\gamma \boldsymbol{\Phi}_m^{\top}\left(\boldsymbol{S}_{t+1}\right)-\boldsymbol{\Phi}_m^{\top}\left(\boldsymbol{S}_t\right)\right\} \boldsymbol{\theta}_{\boldsymbol{\pi}}\right]\right)=\mathbf{0}, \tag{5}$$

which can be simplified as:

$$\boldsymbol{v}_2(\boldsymbol{\pi})-\boldsymbol{v}_1(\boldsymbol{\pi}) \boldsymbol{\theta}_{\boldsymbol{\pi}}=\mathbf{0}.$$

That is, $\boldsymbol{\theta}_{\boldsymbol{\pi}}=\boldsymbol{v}_1^{-1}(\boldsymbol{\pi}) \boldsymbol{v}_2(\boldsymbol{\pi})$, where:

$$\begin{aligned}
\boldsymbol{v}_1(\boldsymbol{\pi}) &= \mathbb{E}\left[\frac{\pi\left(A_t \mid \boldsymbol{S}_t\right)}{b_t\left(A_t \mid \boldsymbol{S}_t\right)} \boldsymbol{\Phi}_m\left(\boldsymbol{S}_t\right)\left\{\boldsymbol{\Phi}_m^{\top}\left(\boldsymbol{S}_t\right)-\gamma \boldsymbol{\Phi}_m^{\top}\left(\boldsymbol{S}_{t+1}\right)\right\}\right] \in \mathbb{R}^{m \times m}, \\
\boldsymbol{v}_2(\boldsymbol{\pi}) &= \mathbb{E}\left\{\frac{\pi\left(A_t \mid \boldsymbol{S}_t\right)}{b_t\left(A_t \mid \boldsymbol{S}_t\right)} R_t \boldsymbol{\Phi}_m\left(\boldsymbol{S}_t\right)\right\} \in \mathbb{R}^{m \times 1}.
\end{aligned}$$

Given n data sequences, we can obtain the empirical version of the left-hand side of equation (5):

$$\begin{aligned}
\boldsymbol{D}_n\left(\boldsymbol{\pi}, \boldsymbol{\theta}_{\boldsymbol{\pi}}\right) =& \frac{1}{n} \sum_{i=1}^n \sum_{t=1}^{T_i} \frac{\pi\left(A_{i,t} \mid \boldsymbol{S}_{i,t}\right)}{b_t\left(A_{i,t} \mid \boldsymbol{S}_{i,t}\right)}\Big[R_{i,t} \boldsymbol{\Phi}_m\left(\boldsymbol{S}_{i,t}\right) \\
&+\boldsymbol{\Phi}_m\left(\boldsymbol{S}_{i,t}\right)\left\{\gamma \boldsymbol{\Phi}_m^{\top}\left(\boldsymbol{S}_{i,t+1}\right)-\boldsymbol{\Phi}_m^{\top}\left(\boldsymbol{S}_{i,t}\right)\right\} \boldsymbol{\theta}_{\boldsymbol{\pi}}\Big] \cdot \\
=:& \hat{\boldsymbol{v}}_2(\boldsymbol{\pi})-\hat{\boldsymbol{v}}_1(\boldsymbol{\pi}) \boldsymbol{\theta}_{\boldsymbol{\pi}}
\end{aligned} \tag{6}$$

where

$$\hat{\boldsymbol{v}}_1(\boldsymbol{\pi}) = \mathbb{E}_n\left[\frac{\pi(A_t \mid \boldsymbol{S}_t)}{b_t(A_t \mid \boldsymbol{S}_t)}\boldsymbol{\Phi}_m(\boldsymbol{S}_t)\left\{\boldsymbol{\Phi}_m^\top(\boldsymbol{S}_t) - \gamma\boldsymbol{\Phi}_m^\top(\boldsymbol{S}_{t+1})\right\}\right],$$

$$\hat{\boldsymbol{v}}_2(\boldsymbol{\pi}) = \mathbb{E}_n\left\{\frac{\pi(A_t \mid \boldsymbol{S}_t)}{b_t(A_t \mid \boldsymbol{S}_t)}R_t\boldsymbol{\Phi}_m(\boldsymbol{S}_t)\right\},$$

and $\mathbb{E}_n(\cdot) = \sum_{i=1}^{n}\sum_{t=1}^{T_i}(\cdot)/n$ represents the empirical expectation based on the sample.

To avoid the issue of overfitting and to ensure that the empirical version of equation (6) is as close as possible to its theoretical value of zero, we can set:

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}} = \arg\min_{\boldsymbol{\theta}_{\boldsymbol{\pi}}\in\boldsymbol{\Theta}} \boldsymbol{D}_n^\top(\boldsymbol{\pi}, \boldsymbol{\theta}_{\boldsymbol{\pi}})\boldsymbol{\Sigma}\boldsymbol{D}_n(\boldsymbol{\pi}, \boldsymbol{\theta}_{\boldsymbol{\pi}}) + \lambda_n\|\boldsymbol{\theta}_{\boldsymbol{\pi}}\|_2^2, \tag{7}$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times m}$ is a given positive definite matrix, λ_n is the regularization parameter and $\|\cdot\|_2$ is L_2 nrom. Let $\boldsymbol{E}_1(\boldsymbol{\pi}) = \boldsymbol{\Sigma}^{1/2}\hat{\boldsymbol{v}}_1(\boldsymbol{\pi})$ and $\boldsymbol{E}_2(\boldsymbol{\pi}) = \boldsymbol{\Sigma}^{1/2}\hat{\boldsymbol{v}}_2(\boldsymbol{\pi})$, then, following Luckett et al. (2020), we can solve for:

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}} = \left\{\boldsymbol{E}_1(\boldsymbol{\pi}) + \lambda_n\boldsymbol{E}_1^{-\top}(\boldsymbol{\pi})\right\}^{-1}\boldsymbol{E}_2(\boldsymbol{\pi}) := \boldsymbol{w}_1^{-1}(\boldsymbol{\pi})\,\boldsymbol{w}_2(\boldsymbol{\pi}). \tag{8}$$

For the the true parameter $\boldsymbol{\theta}_{\boldsymbol{\pi}}$, the true state value function is:

$$V_{\boldsymbol{\mathcal{S}}}(\boldsymbol{\pi}) = \int V(\boldsymbol{\pi}, \boldsymbol{s}_t; \boldsymbol{\theta}_{\boldsymbol{\pi}})dF_S = \int \boldsymbol{\Phi}_m^\top(\boldsymbol{S}_t)\,\boldsymbol{\theta}_{\boldsymbol{\pi}}dF_S.$$

Based on the sample and the optimization problem (7), the empirical state value function is:

$$\hat{V}_{\hat{\boldsymbol{\mathcal{S}}}}(\boldsymbol{\pi}) = \int \boldsymbol{\Phi}_m^\top(\boldsymbol{S}_t)\,\hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}}d\hat{F}_S = \mathbb{E}_n\left\{\boldsymbol{\Phi}_m^\top(\boldsymbol{S}_{i,t})\,\hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}}\right\} \tag{9}$$

Then, the true policy function parameter is $\boldsymbol{\beta}_0 = \arg\max_{\boldsymbol{\beta}\in\boldsymbol{\mathcal{B}}} V_{\boldsymbol{\mathcal{S}}}(\boldsymbol{\pi}_{\boldsymbol{\beta}})$, and its estimated value based on the sample is $\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}\in\boldsymbol{\mathcal{B}}} \hat{V}_{\hat{\boldsymbol{\mathcal{S}}}}(\boldsymbol{\pi}_{\boldsymbol{\beta}})$. To avoid overfitting the coefficients, a penalty term can be added, resulting in the final optimization problem:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}\in\boldsymbol{\mathcal{B}}} \hat{V}_{\hat{\boldsymbol{\mathcal{S}}}}(\boldsymbol{\pi}_{\boldsymbol{\beta}}) - \lambda_{n,\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_2^2. \tag{10}$$

Here, $\lambda_{n,\boldsymbol{\beta}}$ may differ from λ_n in equation (7), but both should be $o_p(n^{-0.5})$.

Let $\mathbf{\Pi} = \{\boldsymbol{\pi}_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \boldsymbol{\mathcal{B}}\}$ be the class of policies. Since $\{\boldsymbol{S}_t\}$ is a non-independent sequence, to construct the confidence interval for $\boldsymbol{\beta}_0$[2], we also need the following assumptions.

Assumption 6. *There exists $2 < \rho < \infty$ such that:*

1. $\mathbb{E}|R_t|^{3\rho} < \infty$, $\mathbb{E}\left\|\boldsymbol{\Phi}_m(\boldsymbol{S}_t)\right\|_2^{3\rho} < \infty$, *and* $\mathbb{E}\left\|\boldsymbol{S}_t\right\|_2^{3\rho} < \infty$.

2. *The sequence $(\boldsymbol{S}_t, A_t)$ satisfies $\sum_{k=1}^{\infty} k^{2/(\rho-2)}\zeta(k) < \infty$, where*

$$\zeta(k) = \mathbb{E}\left[\sup_{m\geq 1}\left\{\left|P\left(B \mid \mathcal{M}_1^m\right) - P(B)\right| : B \in \mathcal{M}_{m+k}^{\infty}\right\}\right],$$

   and $\mathcal{M}_b^c$ is the σ-algebra generated by $\{(\boldsymbol{S}_b, A_b), \ldots, (\boldsymbol{S}_c, A_c)\}$.

3. $J_{[]}\left\{\infty, \mathbf{\Pi}, L_{3\rho}(P)\right\} = \int_0^{\infty} \sqrt{\log N_{[]}\left\{\epsilon, \mathbf{\Pi}, L_{3\rho}(P)\right\}}\mathrm{d}\epsilon < \infty$. *where $N_{[]}$ is the bracketing number (Kosorok, 2008).*

Assumption 7. *For any $\boldsymbol{\pi} \in \mathbf{\Pi}$, there exists $a > 0$ such that for any $\boldsymbol{c} \in \mathbb{R}^m$, we have:*

$$\boldsymbol{c}^{\top}\mathbb{E}\left[\frac{\pi\left(A_t \mid \boldsymbol{S}_t\right)}{b_t\left(A_t \mid \boldsymbol{S}_t\right)}\left\{\boldsymbol{\Phi}_m\left(\boldsymbol{S}_t\right)\boldsymbol{\Phi}_m^{\top}\left(\boldsymbol{S}_t\right) - \gamma^2\boldsymbol{\Phi}_m\left(\boldsymbol{S}_{t+1}\right)\boldsymbol{\Phi}_m^{\top}\left(\boldsymbol{S}_{t+1}\right)\right\}\right]\boldsymbol{c} \geq a\left\|\boldsymbol{c}\right\|^2$$

.

Assumption 8. *The mapping from $\boldsymbol{\beta}$ to the policy value function $V_{\boldsymbol{\mathcal{S}}}(\boldsymbol{\pi}) = \int V(\boldsymbol{\pi}, \boldsymbol{s}_t; \boldsymbol{\theta}_{\boldsymbol{\pi}})dF_S$ is unique.*

Assumption 9. *The policy function $\boldsymbol{\pi}_{\boldsymbol{\beta}}$ has first and second derivatives with respect to $\boldsymbol{\beta}$.*

Assumption 10. *The parameters $\boldsymbol{\beta}$ in the policy function $\boldsymbol{\pi}_{\boldsymbol{\beta}}$ does not vary with time or other factors.*

[2]We cannot simply estimate the Hessian matrix and then invert it because this method is not MLE, and the Cramér-Rao lower bound is the lower bound of the variance for unbiased estimators. This method is not unbiased and does not reach the lower bound. Using the Hessian would severely underestimate.

Assumption 6 is derived from Kosorok (2008) Theorem 11.24, which is used to construct the Central Limit Theorem for non-independent samples, thereby ensuring the normality of the estimator. Assumptions 7 and 8 guarantee that the subsequent estimates for $\boldsymbol{\theta}_{\boldsymbol{\pi}_{\boldsymbol{\beta}}}$ and the optimal $\boldsymbol{\beta}$ are both existent and unique. Assumption 9 ensures that the policy function is differentiable with respect to $\boldsymbol{\beta}$, which is necessary for calculating confidence intervals. Finally, Assumption 10 ensures that $\boldsymbol{\beta}$ remains constant. Relaxing this assumption could lead to a more general varying coefficient model, but for the sake of simplicity, we assume that $\boldsymbol{\beta}$ does not change with any factors. Equation (2) represents a policy function form that satisfies the aforementioned assumptions, other forms may also be chosen.

Luckett et al. (2020) Theorems 4.2 and 4.3 provide the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}_{\hat{\boldsymbol{\beta}}}}$ given $\hat{\boldsymbol{\beta}}$, and show that $\hat{\boldsymbol{\beta}}$ converges in probability to the true $\boldsymbol{\beta}_0$. However, they do not investigate the asymptotic distribution of $\hat{\boldsymbol{\beta}}$. To perform statistical inference on $\boldsymbol{\beta}$, it is necessary to derive the first and second partial derivatives of $\hat{V}_{\hat{\boldsymbol{S}}}\left(\boldsymbol{\pi}_{\boldsymbol{\beta}}\right)$ with respect to $\boldsymbol{\beta}$ under the given assumptions. First, to simplify the theorem, we define the following notation:

1. $\boldsymbol{K}\left(A_t, \boldsymbol{S}_t\right) = \boldsymbol{e}_j / b_t\left(A_t = a^{(j)} \mid \boldsymbol{S}_t\right)$,
2. $\boldsymbol{H}_1\left(\boldsymbol{S}_t\right) = \boldsymbol{\Sigma}^{\frac{1}{2}}\left[\boldsymbol{\Phi}_m\left(\boldsymbol{S}_t\right)\left\{\boldsymbol{\Phi}_m^{\top}\left(\boldsymbol{S}_t\right) - \gamma \boldsymbol{\Phi}_m^{\top}\left(\boldsymbol{S}_{t+1}\right)\right\}\right]$,
3. $\boldsymbol{H}_2\left(\boldsymbol{S}_t\right) = \boldsymbol{\Sigma}^{\frac{1}{2}}\left\{R_t \boldsymbol{\Phi}_m\left(\boldsymbol{S}_t\right)\right\}$,
4. $\boldsymbol{H}_3\left(\boldsymbol{\pi}, A_t, \boldsymbol{S}_t\right) = -\left\{\boldsymbol{w}_1\left(\boldsymbol{\pi}\right)^{-\top} \otimes \boldsymbol{w}_1\left(\boldsymbol{\pi}\right)^{-1}\right\}$,
5. $\boldsymbol{H}_4\left(\boldsymbol{\pi}, A_t, \boldsymbol{S}_t\right) = -\left[\left\{\boldsymbol{w}_2\left(\boldsymbol{\pi}\right)^{\top} \boldsymbol{w}_1\left(\boldsymbol{\pi}\right)^{-\top}\right\} \otimes \boldsymbol{w}_1\left(\boldsymbol{\pi}\right)^{-1}\right]$,
6. $\nabla_{\boldsymbol{\pi}} \boldsymbol{E}_1\left(\boldsymbol{\pi}\right) = \mathbb{E}_n\left\{\boldsymbol{K}\left(A_t, \boldsymbol{S}_t\right) \operatorname{vec}\left\{\boldsymbol{H}_1\left(\boldsymbol{S}_t\right)\right\}^{\top}\right\}$,
7. $\nabla_{\boldsymbol{\pi}} \boldsymbol{E}_2\left(\boldsymbol{\pi}\right) = \mathbb{E}_n\left\{\boldsymbol{K}\left(A_t, \boldsymbol{S}_t\right) \boldsymbol{H}_2^{\top}\left(\boldsymbol{S}_t\right)\right\}$,

8.

$$\left(\nabla_{\boldsymbol{\pi}} \boldsymbol{H}_4^{\top}\right)^{\top} = \boldsymbol{C}_{m,m^2}\left[-\left\{\boldsymbol{I}_m \otimes \operatorname{vec}\left(\boldsymbol{w}_1^{-1}\right)\right\}\left\{\boldsymbol{w}_1^{-1}\left(\nabla_{\boldsymbol{\pi}} \boldsymbol{E}_2\right)^{\top} + \boldsymbol{H}_4\left(\nabla_{\boldsymbol{\pi}} \boldsymbol{E}_1\right)^{\top}\right\}\right.$$
$$\left. - \left\{\operatorname{vec}\left(\boldsymbol{w}_2^{\top} \boldsymbol{w}_1^{-\top}\right) \otimes \boldsymbol{I}_m\right\} \boldsymbol{H}_3\left(\nabla_{\boldsymbol{\pi}} \boldsymbol{E}_1\right)^{\top}\right].$$

For simplicity of notation, the last equation above does not include the contents in the parentheses of the variables. Here, $\boldsymbol{e}_j$ is a $q \times 1$ vector with 1 at the j-th position and 0 elsewhere, and $\boldsymbol{C}_{mm}$ is the commutation matrix [3]. Based on the estimation of $\boldsymbol{\theta}$ from equation (8), Lemmas 1 and 2 provide the first and second partial derivatives of $\hat{V}_{\hat{\boldsymbol{S}}}\left(\boldsymbol{\pi}_{\boldsymbol{\beta}}\right)$ with respect to $\boldsymbol{\beta}$ under the linearization (4).

Lemma 1. *Given Assumptions 1-10, and the policy function $\boldsymbol{\pi}_{\boldsymbol{\beta}}$, we have:*

$$\hat{V}'_{\hat{\boldsymbol{S}}}\left(\boldsymbol{\pi}_{\boldsymbol{\beta}}\right) = \nabla_{\boldsymbol{\beta}} \hat{V}_{\hat{\boldsymbol{S}}} = \nabla_{\boldsymbol{\beta}} \hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}_{\boldsymbol{\beta}}}\left\{\mathbb{E}_n \boldsymbol{\Phi}_m\left(\boldsymbol{S}_t\right)\right\},$$

where for the $\hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}_{\boldsymbol{\beta}}}$ obtained from equation (8), we have:

$$\begin{aligned}\nabla_{\boldsymbol{\beta}} \hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}_{\boldsymbol{\beta}}} &= \nabla_{\boldsymbol{\beta}} \boldsymbol{\pi} \nabla_{\boldsymbol{\pi}} \hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}_{\boldsymbol{\beta}}} \\ &= \nabla_{\boldsymbol{\beta}} \boldsymbol{\pi}\left[\left\{\nabla_{\boldsymbol{\pi}} \boldsymbol{E}_1\left(\boldsymbol{\pi}\right)\right\} \boldsymbol{H}_4^{\top}\left(\boldsymbol{\pi}, A_t, \boldsymbol{S}_t\right) + \left\{\nabla_{\boldsymbol{\pi}} \boldsymbol{E}_2\left(\boldsymbol{\pi}\right)\right\} \boldsymbol{w}_1\left(\boldsymbol{\pi}\right)^{-\top}\right] + o_p(1).\end{aligned}$$

Lemma 2. *Given Assumptions 1-10, and the policy function $\boldsymbol{\pi}_{\boldsymbol{\beta}}$, we have:*

$$\hat{V}''_{\hat{\boldsymbol{S}}}\left(\boldsymbol{\pi}_{\boldsymbol{\beta}}\right) = \nabla^2_{\boldsymbol{\beta}} \hat{V}_{\hat{\boldsymbol{S}}} = \nabla^2_{\boldsymbol{\beta}} \hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}_{\boldsymbol{\beta}}}\left[\left\{\mathbb{E}_n \boldsymbol{\Phi}_{\mathrm{m}}\left(\boldsymbol{S}_t\right)\right\}^{\top} \otimes \boldsymbol{I}_{q-1}\right]^{\top},$$

where:

$$\begin{aligned}\left(\nabla^2_{\boldsymbol{\beta}} \hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}_{\boldsymbol{\beta}}}\right)^{\top} = &\left\{\left(\nabla_{\boldsymbol{\pi}} \hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}_{\boldsymbol{\beta}}}\right)^{\top} \otimes \boldsymbol{I}_{p(q-1)}\right\}\left(\nabla^2_{\boldsymbol{\beta}} \boldsymbol{\pi}\right)^{\top} + \mathbb{E}_n\left\{\left\{\boldsymbol{I}_m \otimes \left(\nabla_{\boldsymbol{\beta}} \boldsymbol{\pi}\right)\right\}\right. \\ &\times\left(\left[\boldsymbol{I}_m \otimes\left\{\boldsymbol{K} \operatorname{vec}\left(\boldsymbol{H}_1\right)^{\top}\right\}\right]\left(\nabla_{\boldsymbol{\pi}} \boldsymbol{H}_4^{\top}\right)^{\top}\left(\nabla_{\boldsymbol{\beta}} \boldsymbol{\pi}\right)^{\top}\right. \\ &\left.\left.+\left\{\boldsymbol{I}_m \otimes\left(\boldsymbol{K} \boldsymbol{H}_2^{\top}\right)\right\} \boldsymbol{C}_{mm} \boldsymbol{H}_3\left(\nabla_{\boldsymbol{\pi}} \boldsymbol{E}_1\right)^{\top}\left(\nabla_{\boldsymbol{\beta}} \boldsymbol{\pi}\right)^{\top}\right)\right\} + o_p(1).\end{aligned}$$

For simplicity, we omit the parameters in variable parentheses of the above formula.

[3] For any matrix $\boldsymbol{M}_{p\times q}$, the commutation matrix transforms $\operatorname{vec}\left(\boldsymbol{M}\right)$ to $\operatorname{vec}\left(\boldsymbol{M}^{\top}\right)$: $\boldsymbol{C}_{pq} \operatorname{vec}\left(\boldsymbol{M}\right) = \operatorname{vec}\left(\boldsymbol{M}^{\top}\right)$

Theorem 1, based on Lemmas 1 and 2 and Luckett et al. (2020), further provides the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ and the corresponding covariance matrix, completing the theoretical basis for the confidence interval of $\hat{\boldsymbol{\beta}}$.

Theorem 1. *Under Assumptions 1-10, given the policy function $\boldsymbol{\pi}_{\boldsymbol{\beta}}$, let the covariance matrix of $\nabla_{\boldsymbol{\beta}_0}\hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}_{\boldsymbol{\beta}_0}}$ be $\boldsymbol{\Omega}_{\boldsymbol{\theta}'} \otimes \boldsymbol{\Sigma}_{\boldsymbol{\theta}'}$. For the parameterized policy function (2), we have:*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}),$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = V''_{\boldsymbol{\mathcal{S}}}\left(\boldsymbol{\pi}_{\boldsymbol{\beta}_0}\right)^{-1}\left[\boldsymbol{\Omega}_{\boldsymbol{\theta}'} \otimes \left\{\boldsymbol{\Phi}_m\left(\boldsymbol{S}\right)^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}'}\boldsymbol{\Phi}_m\left(\boldsymbol{S}\right)\right\}\right] V''_{\boldsymbol{\mathcal{S}}}\left(\boldsymbol{\pi}_{\boldsymbol{\beta}_0}\right)^{-\top}.$$

The consistent estimator for the $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is:

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = \hat{V}''_{\hat{\boldsymbol{\mathcal{S}}}}\left(\boldsymbol{\pi}_{\hat{\boldsymbol{\beta}}}\right)^{-1}\left(\hat{\boldsymbol{\Omega}}_{\boldsymbol{\theta}'} \otimes \left[\left\{\mathbb{E}_n\boldsymbol{\Phi}_m\left(\boldsymbol{S}_{i,t}\right)^\top\right\}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}'}\left\{\mathbb{E}_n\boldsymbol{\Phi}_m\left(\boldsymbol{S}_{i,t}\right)\right\}\right]\right)\hat{V}''_{\hat{\boldsymbol{\mathcal{S}}}}\left(\boldsymbol{\pi}_{\boldsymbol{\beta}_0}\right)^{-\top},$$

where

$$\hat{\boldsymbol{\Omega}}_{\boldsymbol{\theta}'} \otimes \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}'} = \frac{1}{n}\sum_{i=1}^{n}\left\{\operatorname{vec}\left(\nabla_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}_{\hat{\boldsymbol{\beta}}}}\right)\right\}^2 - \left\{\mathbb{E}_n \operatorname{vec}\left(\nabla_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}_{\hat{\boldsymbol{\beta}}}}\right)\right\}^2.$$

In Theorem 1, $\hat{\boldsymbol{\Omega}}_{\boldsymbol{\theta}'} \otimes \left[\left\{\mathbb{E}_n\boldsymbol{\Phi}_m\left(\boldsymbol{S}_{i,t}\right)^\top\right\}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}'}\left\{\mathbb{E}_n\boldsymbol{\Phi}_m\left(\boldsymbol{S}_{i,t}\right)\right\}\right]$ can be obtained by blocking $\hat{\boldsymbol{\Omega}}_{\boldsymbol{\theta}'} \otimes \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}'}$ according to the dimensions of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}'}$, and then left-multiplying by $\mathbb{E}_n\boldsymbol{\Phi}_m\left(\boldsymbol{S}_{i,t}\right)^\top$ and right-multiplying by $\mathbb{E}_n\boldsymbol{\Phi}_m\left(\boldsymbol{S}_{i,t}\right)$. The values of $\nabla_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}_{\hat{\boldsymbol{\beta}}}}$ and $\hat{V}''_{\hat{\boldsymbol{\mathcal{S}}}}\left(\boldsymbol{\pi}_{\hat{\boldsymbol{\beta}}}\right)$are obtained from Lemmas 1 and 2.

The form of $\boldsymbol{\pi}_{\boldsymbol{\beta}}$ can be chosen in various ways, but it is essentially a multinomial regression in the discrete case. In Lemmas 1 and 2, the first and second derivatives of $\boldsymbol{\pi}_{\boldsymbol{\beta}}$ with respect to $\boldsymbol{\beta}$ ($\nabla_{\boldsymbol{\beta}}\boldsymbol{\pi}$ and $\nabla^2_{\boldsymbol{\beta}}\boldsymbol{\pi}$) also depend on the form of $\boldsymbol{\pi}_{\boldsymbol{\beta}}$. Theorem 2 provides the form of these derivatives when multinomial logistic regression is used.

Theorem 2. *When the policy function is in the form of multinomial logistic regression as given in equation (2), let* $\operatorname{diag}\left(\boldsymbol{\pi}_{-1}\right)$ *be the diagonal matrix with* $\boldsymbol{\pi}-1 = \left(\pi_2, \ldots, \pi_q\right)^\top$

on the diagonal, $\boldsymbol{I}_{(2:q),q}$ *be the identity matrix* $\boldsymbol{I}_q$ *with the first row removed, and* $\mathbf{1}_{q-1} = (1, 1, \ldots, 1)^\top$ *be a* $(q-1) \times 1$ *vector. Then we have:*

1. $\nabla_{\boldsymbol{\beta}} \boldsymbol{\pi} = \left\{ \operatorname{diag}(\boldsymbol{\pi}_{-1}) \boldsymbol{I}_{(2:q),q} - \boldsymbol{\pi}_{-1} \boldsymbol{\pi}^\top \right\} \otimes \boldsymbol{s}$,

2. $\nabla_{\boldsymbol{\beta}} \boldsymbol{\pi}_{-1} = \left\{ \operatorname{diag}(\boldsymbol{\pi}_{-1}) \boldsymbol{I}_{q-1} - \boldsymbol{\pi}_{-1} \boldsymbol{\pi}_{-1}^\top \right\} \otimes \boldsymbol{s}$,

3. $\nabla^2_{\boldsymbol{\beta}} \boldsymbol{\pi} = \left[\nabla_{\boldsymbol{\beta}} \boldsymbol{\pi}_{-1} \left\{ \mathbf{1}_{q-1} \operatorname{vec}^T(\boldsymbol{I}_{q-1}) \left(\boldsymbol{I}_{(2:q),q} \otimes \boldsymbol{I}_{q-1} \right) - \left(\boldsymbol{\pi}^\top \otimes \boldsymbol{I}_{q-1} \right) \right\} \right] \otimes \boldsymbol{s}^\top - \left\{ \nabla_{\boldsymbol{\beta}} \boldsymbol{\pi} \left(\boldsymbol{I}_q \otimes \boldsymbol{\pi}_{-1}^\top \right) \right\} \otimes \boldsymbol{s}^\top$.

To solve the optimization problem (10), we can use the REINFORCE algorithm (Williams, 1992) or the Actor-Critic algorithm (Konda and Tsitsiklis, 2000). Other options include the Newton method and its derivatives, such as BFGS and L-BFGS (Liu and Nocedal, 1989).

3 Extensions to Continuous Action Spaces and Online Learning

3.1 Confidence Interval Estimation Estimation in Continuous Action Spaces

We currently focus on discrete action spaces, but in practice, the action space for decision-making is often continuous. For example, in autonomous driving (Kiran et al., 2021), the action space is continuous, as is the choice space in economic decisions like interest rate setting, which can be seen as continuous—decisions are not merely about raising or lowering interest rates but determining the specific value of the change. With slight adjustments in the proof, our research can be extended to continuous action spaces, and the assumption about the action space in Assumption 1 can also be relaxed.

Unlike the discrete action space, in a continuous action space, the policy function is not the probability of choosing a particular action but rather a probability density function (Sutton and Barto, 2018). For instance, if actions follow a normal distribution, the policy function can be specified as:

$$\pi^c\left(a_t \mid \boldsymbol{s}_t, \boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\sigma\right) = \frac{1}{\sqrt{2\pi}\sigma\left(\boldsymbol{s}_t, \boldsymbol{\beta}_\sigma\right)} \exp\left[-\frac{\left\{a_t - \mu\left(\boldsymbol{s}_t, \boldsymbol{\beta}_\mu\right)\right\}^2}{2\sigma\left(\boldsymbol{s}_t, \boldsymbol{\beta}_\sigma\right)^2}\right], \tag{11}$$

where $\mu\left(\boldsymbol{s}_t, \boldsymbol{\beta}_\mu\right)$ and $\sigma\left(\boldsymbol{s}_t, \boldsymbol{\beta}_\sigma\right)$ are parameterized as the mean and standard deviation, respectively. The dimensions of the parameters $\boldsymbol{\beta}_\mu$ and $\boldsymbol{\beta}_\sigma$ are $p \times 1$ and align with the state vector, rather than being $\boldsymbol{\beta} \in \mathbb{R}^{p \times q}$ as in the discrete case, which significantly reduces computational complexity when action selection is more complex. For example, in interest rate setting, if the choice range is between a 5% increase and a 5% decrease with intervals of 25 basis points (0.25%), the number of parameters would be $2p$ in the continuous action space, compared to $41p$ in the discrete action space (including 0%).

In the continuous action space $\mathcal{A}^c$, if the behavior policy $b_t^c\left(a_t \mid \boldsymbol{S}_t = \boldsymbol{s}_t\right)$ also represents a probability density function, then Assumption 5 can be rewritten as

Assumption 11. *Action Reachability: For all $a_t \in \mathcal{A}^c$, $\boldsymbol{s}_t \in \boldsymbol{\mathcal{S}}$, and t, the probability density function $f\left(a_t \mid \boldsymbol{S}_t = \boldsymbol{s}_t\right)$ is greater than 0.*

Based on this assumption, Theorem 2.1 of Luckett et al. (2020) can be extended to Theorem 3.

Theorem 3. *Let $\pi^c(a_t \mid \boldsymbol{s}_t)$be any probability density function, $b_t^c\left(a_t \mid \boldsymbol{S}_t = \boldsymbol{s}_t\right)$ be the behavior policy at time t, and $\gamma \in [0, 1)$ be the discount factor. Under Assumptions 3, 4, and 11, we have:*

$$V\left(\pi^c, \boldsymbol{s}_t\right) = \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i R_{t+i} \left\{\prod_{j=0}^{i} \frac{\pi^c\left(a_{t+j} \mid \boldsymbol{s}_{t+j}\right)}{b_{t+j}^c\left(a_{t+j} \mid \boldsymbol{s}_{t+j}\right)}\right\} \mid \boldsymbol{S}_t = \boldsymbol{s}_t\right]$$

Let $\boldsymbol{K}(A_t, \boldsymbol{S}_t) = 1/b_t(A_t \mid \boldsymbol{S}_t)$, then the rest of the derivation leading to Theorem 1 remains the same as in the offline discrete action space case. As with the discrete case, there are many choices for the policy function corresponding to continuous action spaces. For instance, the normal distribution corresponds to equation (11), but other distributions like the Poisson distribution, exponential distribution, etc., can also be used depending on the characteristics of the action space. For interest rate decisions, since both increases and decreases are possible, we use the normal distribution for the study and adopts linear parameterization (Sutton and Barto, 2018), where $\mu(\boldsymbol{s}_t, \boldsymbol{\beta}_\mu) = \boldsymbol{s}_t^\top \boldsymbol{\beta}_\mu$ and $\sigma(\boldsymbol{s}_t, \boldsymbol{\beta}_\sigma) = \exp\left(\boldsymbol{s}_t^\top \boldsymbol{\beta}_\sigma\right)$. Theorem 4 provides the corresponding derivative forms.

Theorem 4. *When the policy function π^c corresponding to a continuous action space takes the form of a normal density function (11), with linear parameterizations $\mu(\boldsymbol{s}_t, \boldsymbol{\beta}_\mu) = \boldsymbol{s}_t^\top \boldsymbol{\beta}_\mu$ and $\sigma(\boldsymbol{s}_t, \boldsymbol{\beta}_\sigma) = \exp\left(\boldsymbol{s}_t^\top \boldsymbol{\beta}_\sigma\right)$, and letting $\boldsymbol{\beta}^c = (\boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\sigma)$, we have:*

1.

$$\nabla_{\boldsymbol{\beta}^c} \pi^c = \left\{ \begin{array}{c} \dfrac{\left(a_t - \boldsymbol{s}^\top \boldsymbol{\beta}_\mu\right)}{\exp^2\left(\boldsymbol{s}^\top \boldsymbol{\beta}_\sigma\right)} \\ \dfrac{\left(a_t - \boldsymbol{s}^\top \boldsymbol{\beta}_\mu\right)^2}{\exp^2\left(\boldsymbol{s}^\top \boldsymbol{\beta}_\sigma\right)} - 1 \end{array} \right\} \otimes \left(\boldsymbol{s}_t \pi^c\right)$$

2.

$$\nabla^2_{\boldsymbol{\beta}^c} \pi^c = \left[\begin{array}{cc} \left\{ \dfrac{\left(a_t - \boldsymbol{s}_t^\top \boldsymbol{\beta}_\mu\right)^2}{\exp^2\left(\boldsymbol{s}_t^\top \boldsymbol{\beta}_\sigma\right)} - 1 \right\} \dfrac{1}{\exp^2\left(\boldsymbol{s}_t^\top \boldsymbol{\beta}_\sigma\right)} & \dfrac{\left(a_t - \boldsymbol{s}_t^\top \boldsymbol{\beta}_\mu\right)}{\exp^2\left(\boldsymbol{s}_t^\top \boldsymbol{\beta}_\sigma\right)} \left\{ \dfrac{\left(a_t - \boldsymbol{s}_t^\top \boldsymbol{\beta}_\mu\right)^2}{\exp^2\left(\boldsymbol{s}_t^\top \boldsymbol{\beta}_\sigma\right)} - 3 \right\} \\ \dfrac{\left(a_t - \boldsymbol{s}_t^\top \boldsymbol{\beta}_\mu\right)}{\exp^2\left(\boldsymbol{s}_t^\top \boldsymbol{\beta}_\sigma\right)} \left\{ \dfrac{\left(a_t - \boldsymbol{s}_t^\top \boldsymbol{\beta}_\mu\right)^2}{\exp^2\left(\boldsymbol{s}_t^\top \boldsymbol{\beta}_\sigma\right)} - 3 \right\} & \left\{ \dfrac{\left(a_t - \boldsymbol{s}_t^\top \boldsymbol{\beta}_\mu\right)^4}{\exp^4\left(\boldsymbol{s}_t^\top \boldsymbol{\beta}_\sigma\right)} - 4\dfrac{\left(a_t - \boldsymbol{s}_t^\top \boldsymbol{\beta}_\mu\right)^2}{\exp^2\left(\boldsymbol{s}_t^\top \boldsymbol{\beta}_\sigma\right)} + 1 \right\} \end{array} \right] \otimes \left(\boldsymbol{s}_t \boldsymbol{s}_t^\top \pi^c\right)$$

3.2 Confidence Interval Estimation under Online Learning

Online RL utilizes interactions between the model and the environment, generating one or a group of data points each time, and updating the model with this data before continuing the process. Compared to offline learning, online learning is faster, but its drawback lies in its practical application, as real-world interactions with the environment can be costly.

Specifically, at time t, suppose we have already estimated the optimal policy $\hat{\boldsymbol{\pi}}^{(t)}$. Using this policy and the current state $\boldsymbol{S}_t$, we can select an action A_t, which then generates a reward R_t and a new state $\boldsymbol{S}_{t+1}$. Online learning updates the optimal policy using the data $(\boldsymbol{S}_t, A_t, R_t, \boldsymbol{S}_{t+1})$, thereby obtaining the policy $\boldsymbol{\pi}^{(t+1)}$. This section updates the model using all accumulated data by minimizing equation (12) to estimate $\boldsymbol{\theta}$, and subsequently $\boldsymbol{\beta}$:

$$\begin{aligned}\boldsymbol{D}_t\left(\boldsymbol{\pi}, \boldsymbol{\theta}_{\boldsymbol{\pi}}\right) =& \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{t} \frac{\pi\left(A_{i,j} \mid \boldsymbol{S}_{i,j}\right)}{\hat{\pi}^{(j-1)}\left(A_{i,j} \mid \boldsymbol{S}_{i,j}\right)} \Big[R_{i,j} \boldsymbol{\Phi}_m\left(\boldsymbol{S}_{i,j}\right) \\ &+ \boldsymbol{\Phi}_m\left(\boldsymbol{S}_{i,j}\right) \left\{ \gamma \boldsymbol{\Phi}_m^{\top}\left(\boldsymbol{S}_{i,j+1}\right) - \boldsymbol{\Phi}_m^{\top}\left(\boldsymbol{S}_{i,j}\right) \right\} \boldsymbol{\theta}_{\boldsymbol{\pi}} \Big],\end{aligned} \tag{12}$$

where $\hat{\pi}^{(0)}$ is the initial policy. The above process is effectively equivalent to replacing b_t in equation (6) with $\hat{\pi}^{(t)}$. In practice, the update interval can be increased to accumulate sufficient information for model updates, thereby improving computational efficiency. Additionally, to ensure some level of exploration, the initial policy is set to be a completely random policy.

4 Simulation Experiments

In each simulation scenario, this section repeats the entire simulation 100 times, using a discount factor of $\gamma = 0.95$. The given positive definite matrix $\boldsymbol{\Sigma}$ is the identity matrix for simplicity, and the initial $\boldsymbol{\beta}$ is a zero matrix, ensuring that each action is selected with equal probability. This study employs the L-BFGS algorithm (Byrd et al., 1995), and both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ include an intercept term. For the selection of basis functions, this section uses linear basis functions with an intercept term, $\boldsymbol{\Phi}(\boldsymbol{S}) = (1, \boldsymbol{S})^{\top}$, and exponential basis functions, $\boldsymbol{\Phi}(\boldsymbol{S}) = (1, e^{-\boldsymbol{S}^*})^{\top}$, referred to as "Linear basis" and "Exp basis" respectively. In the exponential basis function, $\boldsymbol{S}^*$ is obtained by applying the cumulative density function

of the standard normal distribution to $(\boldsymbol{S} - c)$, ensuring that $\boldsymbol{S}^*$ is bounded[4].

4.1 Action Selection Simulation

First, we introduces a simple action selection environment for simulation, where the state $\boldsymbol{S}_t$ at each time follows a joint normal distribution $N(\mathbf{0}, \boldsymbol{I}_2)$, and the action space is $\mathcal{A} = \{0, 1\}$. The probability of selecting each action is determined by equation (2). In this environment, the state is independent of the action, and the results of the model can be evaluated by setting a true $\boldsymbol{\beta}$. When the selected action is same with the action chosen based on the true $\boldsymbol{\beta}_0$, the reward R_t equals 1. Otherwise, it equals 0. This section sets the true $\boldsymbol{\beta}_0 = [(0, 0, 0)^\top, (0, 5, -5)^\top]$ and $c = 0$ for the basis function. The number of sequences is set to $n = 25$, 50, 75, 100, with each sequence simulating $T = 50$, 75, 100 steps. Since the first column of $\boldsymbol{\beta}$ is assumed to be zero, the unknown parameters to be estimated are the three elements in the second column of $\boldsymbol{\beta}$, with the first element being the intercept term. This section only presents the estimation results for the two elements β_1 and β_2.

Figure 1 shows the empirical coverage probabilities (ECP), the median width of confidence intervals, and the median log MSE for the two parameter estimates. Different colors and point types represent the proportion of empirical replay and the number of sequences n. It can be observed that as $\log(nT)$ increases, the results for both types of basis functions gradually approach and stabilize at 95%, while the width of the confidence intervals and $\log(MSE)$ decrease. The choice of basis functions has a minor effect on the results. Additionally, the impact of n and T can be compared when $\log(nT)$ is the same. It is evident that with the same total sample size, an increase in n results in smaller confidence inter-

[4]Since exponential basis functions can yield large absolute values, this section adopts the approach of Shi et al. (2022) to ensure that the basis function values do not become excessively large. The parameter c can be the expectation or sample mean of $\boldsymbol{S}$.

vals, leading to more accurate estimates. For example, at $\log(nT) = 8.52$, the confidence interval for $n = 100$ is narrower than that for $n = 50$, and $\log(MSE)$ is smaller. Similar situations occur at $\log(nT) = 7.54$, 8.92, etc. Despite the samples being independent in this simulation, the results suggest that the effect of increasing n is smaller than that of increasing T. In practical situations, this implies that going deeper in a single attempt is a better choice compared to multiple attempts.

4.2 Generative Simulation

The second simulation is a generative simulation, the specific setup of the simulation is:

$$\boldsymbol{S}_t = \begin{bmatrix} S_{t,1} \\ \\ S_{t,2} \end{bmatrix} = \begin{bmatrix} 1 - \frac{1}{2}A_t & 0 & 0 \\ 0 & 1 - \frac{1}{2}A_t & 0 \end{bmatrix} \boldsymbol{s}_t + \boldsymbol{\varepsilon}_t,$$

where $S_{0,k}$ follows a standard normal distribution, and $\varepsilon_{t,1}$ and $-\varepsilon_{t,1}$ are independently and identically distributed as an exponential distribution $Exp(1)$. A_t follows a Bernoulli distribution with an expectation of 0.5, meaning the policy generating the observed sequence is $b_t(A_t = 1 \mid \boldsymbol{S}t) = b_t(At = 0 \mid \boldsymbol{S}_t) = 0.5$. The reward function is

$$R_t = 1 - S_{t,1} - S_{t,2} - A_t.$$

In this setup, action A_t causes $S_{t,1}$ and $S_{t,2}$ to decrease. The reward function imposes penalties on states $S_{t,1}$, $S_{t,2}$, and action A_t. If no action A_t is taken, $S_{t,1}$ and $S_{t,2}$ will gradually increase over time. Therefore, the probability of A_t being 1 increases as $S_{t,1}$ and $S_{t,2}$ become larger. Thus, theoretically, the coefficients for the two states in A_t should both be greater than 0. In the generative simulation scenario, this section sets the number of sequences $n = 20$, 35, 40, 55, and each sequence simulates $T = 20$, 30, 40, 55 steps. For the exponential basis function, c is set to the sample mean (approximately 3.5). Additionally, unlike the action selection simulation, the generative simulation cannot theoretically obtain

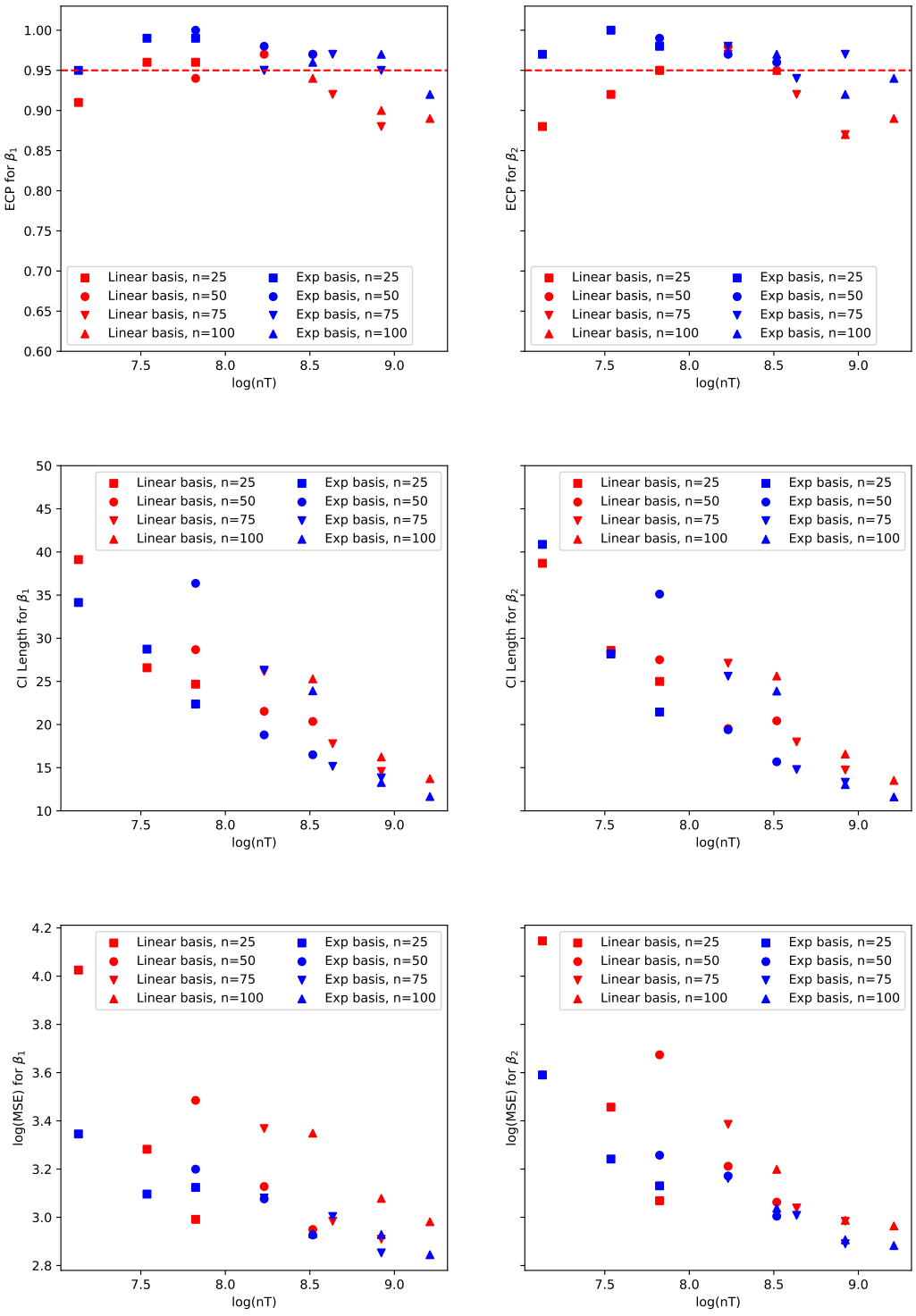


Figure 1: The estimation results of action selection simulation with $\boldsymbol{\beta}_0 = [(0, 0, 0^\top), (0, 5, -5)^\top]$. The left column and the right column contain the results for β_1 and β_2, respectively.

the true $\boldsymbol{\beta}$. Following Shi et al. (2022), we use the estimation results from $n = 10000$ to approximate the true values, thereby estimating the probability of the confidence interval containing the true parameters and the mean squared error (MSE).

Unlike the simple action selection simulation, the state in the generative simulation is not independent over time, making it closer to real-world situations. The characteristics of the various methods also differ from those in the action selection simulation. According to the setup in Section 4.2, theoretically, both β values should be greater than 0.

Figure 2 further presents the results of the offline estimation for the generative simulation, with the calculation methods for each metric the same as in the action selection simulation. It can be observed that as $\log(nT)$ increases, the ECP for the linear basis functions remains around 95%, while the ECP for the exponential basis functions approaches 95% as the sample size increases. The width of the confidence intervals and log MSE both show a decreasing trend.

4.3 Online Simulation

In real-life scenarios, it is often necessary to continuously optimize policies based on new data, which requires online RL. This section will perform statistical inference using V-learning-based online RL, building on the action selection simulation. Online RL updates policies after accumulating a certain amount of data. This section first sets a burn-in period of $T = 50$ to ensure sequence stability. Subsequently, the policy is updated every 5 data points, and the data for the next sequence is drawn based on the new optimal policy. Finally, this section presents simulation results for sequence numbers $n = 25$, 50, 75, 100 with updates at 5, 10, 15, and 20 times (i.e., $T = 25$, 50, 75, 100).

Figure 3 shows the results for empirical coverage probability (ECP), confidence in-

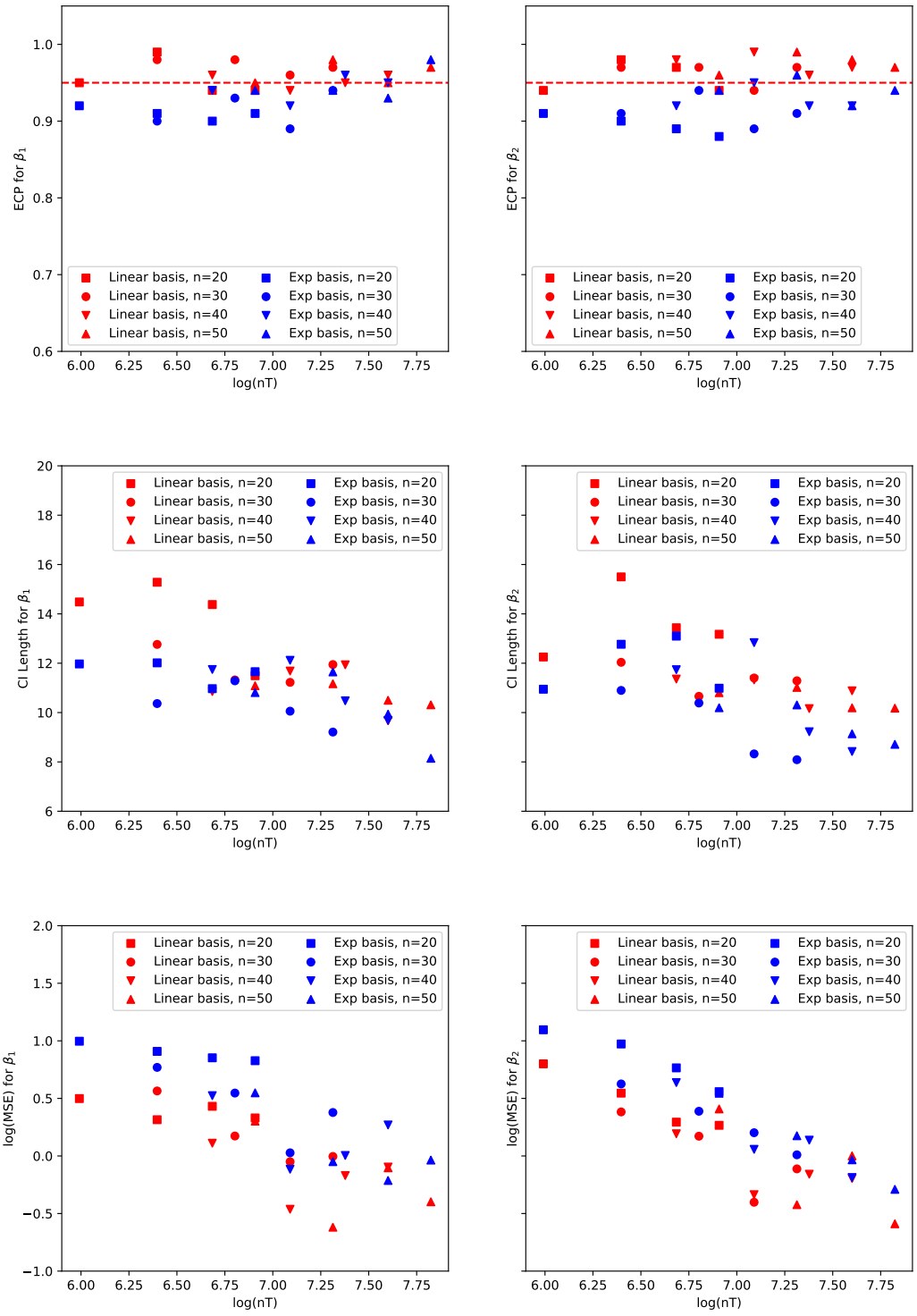


Figure 2: Results of Offline Estimation in Generative Simulation. The left column and the right column contain the results for β_1 and β_2, respectively.

terval width, and log MSE under different parameter settings. The true $\boldsymbol{\beta}$ is set to $[(0,0,0)^\top,(0,5,-5)^\top]$. Similar to the offline learning results, it can be observed that as the sample size $\log(nT)$ increases, the empirical coverage probability for both types of basis functions fluctuates around 95%, and both the confidence interval width and MSE decrease gradually, with no significant difference due to the choice of basis functions. Regarding confidence interval width and log MSE, it can be seen that for the same $\log(nT)$, a larger n corresponds to narrower confidence intervals and smaller log MSE. This conclusion is similar to that of offline learning, indicating that increasing n is more effective than increasing T.

5 Real Data

Central banks around the world often adjust interest rates to achieve the three primary goals of stabilizing inflation, increasing employment, and fostering economic growth (Svensson, 2014; Loungani and Sheets, 1997). Specifically, central banks typically combat inflation by raising interest rates, but this can lead to lower employment rates and slower economic growth. Thus, central banks need to consider multiple factors when adjusting interest rates. The classical approach to interest rate regulation follows the Taylor rule (Woodford, 2001; Taylor, 1993):

$$i_t = i_{ss} + \phi_\pi \left(\Pi_t - \Pi_{ss}\right) + \phi_y \left[\log\left(Y_t\right) - \log\left(Y_{ss}\right)\right],$$

where i, Π, and Y represent the interest rate, inflation, and real output, respectively. i_{ss}, Π_{ss}, and Y_{ss} are the target or steady-state values of these variables[5].

[5]In the traditional general equilibrium framework, the steady state represents the values of economic variables when the economy reaches a stable long-term condition. In this state, all variable values are constant (Baxter and King, 1993). Thus, the Taylor rule suggests that the government should adjust

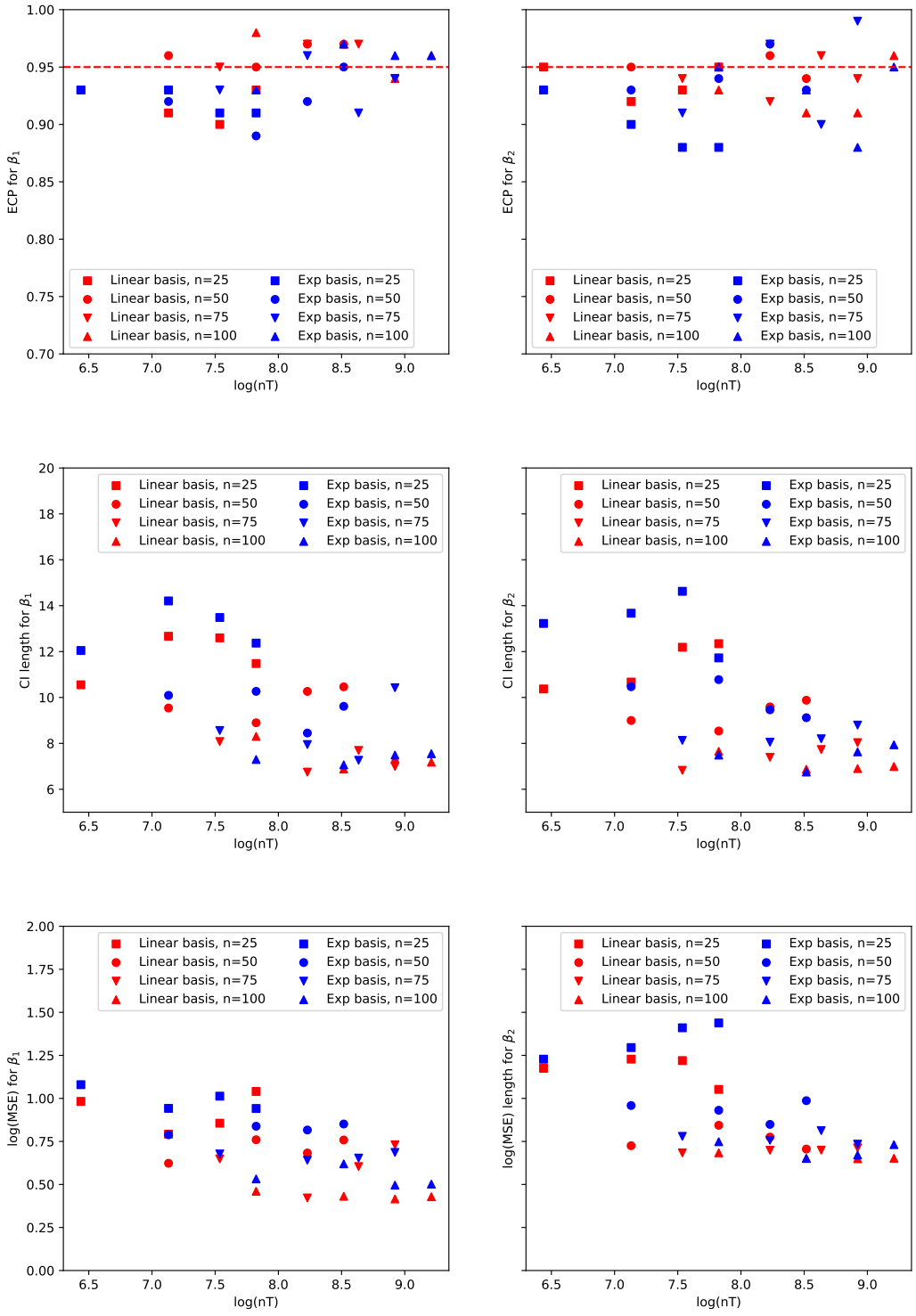


Figure 3: Online Estimation Results for Action Selection Simulation: True $\boldsymbol{\beta} = [(0,0,0)^\top, (0,5,-5)^\top]$. The left column and the right column contain the results for β_1 and β_2, respectively.

This section explores the potential for using RL to assist the Federal Reserve in making monetary policy decisions. It uses the federal funds rate (i) published by the Federal Reserve Board as the target for monetary policy adjustment, where the action variable represents the Fed's decision to raise ($A_t = 1$) or lower ($A_t = 0$) interest rates[6]. This section assumes that the Fed makes decisions each quarter based on financial market indicators and macroeconomic variables. Therefore, the state variable $\boldsymbol{S}_t$ includes the Dow Jones Industrial Average ($Stock$) to represent the financial market, and macroeconomic variables from the Bureau of Economic Analysis and the Bureau of Labor Statistics, including GDP (Y), imports (Imp), exports (Exp), personal savings (B), average employee compensation (W), the consumer price index (pi), and unemployment rate (Uem). All data are quarterly, spanning from Q3 1954 (U.S. Dollar Index: Q1 1971) to Q3 2023. Macroeconomic variables are adjusted for inflation by dividing by the consumer price index, and inflation is measured by the rate of change in the consumer price index. After the adjustment, we calculate the quarter-on-quarter growth rate of each economic variable treat them as state variables.

For the reward R_t in RL, different settings are used depending on the goal. When the goal is to stabilize inflation, $R_t = -|\Pi_t - \Pi_{ss}|$, where Π_t is the inflation rate at time t (in percent), and Π_{ss} is the target inflation rate, set at a quarterly target of $2\%/4 = 0.5\%$. When the goal is to increase employment, R_t is set as the negative unemployment rate at time t (in percent). When the goal is economic growth, R_t is the quarterly GDP growth

interest rates around their long-term steady-state values based on the deviations of inflation and output from their steady-state levels. However, because the steady-state values are unknown, central banks usually adjust interest rates based on policy targets (such as target inflation or GDP growth rates), substituting target values for steady-state values in equation (5).

[6]Since there were only five quarters during the sample period when the Fed did not change interest rates, leading to potential estimation errors for such rare actions, the behavior of keeping rates unchanged is merged with raising rates.

rate at time t (in percent). The basis function $\boldsymbol{\Phi}(\boldsymbol{S})$ is the state value $\boldsymbol{S}$ itself.

Table 1 shows the average economic indicators when actual policy and optimal policy match or differ under three different goals during the sample period. The data indicate that, regardless of the policy goal, the actual policy aligns with the optimal policy in about 57% of cases. In other instances, the actual policy tends to favor lowering interest rates, while the optimal policy leans towards raising them, suggesting that the Fed prefers a more accommodative policy than the optimal strategy.

A closer look at inflation, unemployment, and economic growth under different scenarios reveals that when actual policy aligns with the optimal policy, the economy exhibits the lowest inflation and unemployment rates and the highest economic growth. Conversely, when the optimal policy suggests raising interest rates, but the actual policy lowers them, inflation and unemployment are higher, with a notable impact on economic growth. This indicates that in such situations, the Fed prefers to lower interest rates to stimulate economic growth, while the optimal policy would raise rates to preserve future adjustment flexibility. Although the optimal policy might seem unreasonable in the short term, it is based on long-term effects, meaning current economic slowdown does not imply poor policy performance. When the optimal policy suggests lowering rates but the actual policy raises them, inflation is highest, unemployment is highest, and economic growth is lowest. This reflects that in a high inflation and high unemployment environment, the Fed is more likely to raise rates to curb inflation, while the optimal policy would lower rates to increase employment opportunities and stimulate economic growth.

Table 2 presents the coefficients and significance of the state variables associated with the optimal policy under different policy goals, compared with the actual policy coefficients obtained through multinomial logistic regression (essentially binary classification).

Table 1: Comparison of Actual and Optimal Policy Effects.

Target	Actual	Optimal	Count	Inflation	Unemployment	Growth
Stabilize Inflation		=	158	0.89%	5.70%	0.87%
		≠	117	0.90%	6.07%	0.38%
	Cut Rates	Raise Rates	114	0.90%	5.98%	0.44%
	Raise Rates	Cut Rates	3	1.18%	9.58%	-2.21%
Increase Employment		=	158	0.88%	5.70%	0.81%
		≠	117	0.91%	6.07%	0.47%
	Cut Rates	Raise Rates	116	0.89%	6.06%	0.46%
	Raise Rates	Cut Rates	1	2.65%	7.97%	1.94%
Economic Growth		=	157	0.86%	5.67%	0.86%
		≠	118	0.94%	6.11%	0.40%
	Cut Rates	Raise Rates	113	0.93%	5.98%	0.44%
	Raise Rates	Cut Rates	5	1.11%	9.07%	-0.57%

Lowering rates ($A_t = 0$) is the baseline category, and the table shows the coefficients corresponding to raising rates ($A_t = 1$). Logistic regression models the actual policy directly, so differences exist between these coefficients and those of the optimal policy. In logistic regression analysis, high interest rates (i) and high unemployment rates (Uem) generally prompt the central bank to lower rates, while rapid economic growth (Y), high inflation (pi), and high real wage levels (W) prompt rate increases.

In contrast, the optimal policy results indicate that regardless of the goal, policy decisions consistently consider current interest rates, inflation, and employment. These findings align with the logistic regression results: higher interest rates and unemployment lead to a preference for lowering rates, while high inflation prompts rate increases. Unlike the logistic regression results, the optimal policy does not emphasize current economic growth, consistent with the conclusions in Table 1. The optimal policy focuses more on controlling inflation and promoting employment, reflecting the central bank's primary function as a financial system manager, balancing inflation control and maximum employment rather than solely pursuing short-term economic growth. This policy choice highlights the central bank's commitment to long-term macroeconomic stability and the complexity of balancing different macroeconomic objectives.

Moreover, the results in Table 2 differ from traditional variable importance studies based on coefficient magnitudes (Trott et al., 2021). For instance, in the goal of stabilizing inflation, the coefficient for real economic growth rate (Y) is 7.6232, higher than those for real wage growth rate (W) and real export growth rate (Exp). However, considering the variability of coefficient estimates (i.e., standard errors), the real economic growth rate Y is not a significant variable.

Table 2: Coefficients of State Variables Corresponding to Optimal Policy.

	Optimal Policy			Multinomial
	Stabilize Inflation	Increase Employment	Economic Growth	Logistic Regression
i	-16.2938***	-21.6894	-17.2723*	-2.8904***
	(-3.595)	(-1.303)	(-1.839)	(-3.308)
Y	7.6232	-1.4089	3.2963	5.3367**
	(0.38)	(-0.37)	(0.160)	(-2.271)
Uen	-11.4902	-6.9758***	-9.9496***	-1.9002**
	(-0.575)	(-2.627)	(-6.655)	(-2.389)
pi	14.2718***	8.0377	15.6883***	5.3925***
	(4.046)	(1.371)	(5.278)	(3.151)
W	-7.0564***	-3.7698**	-9.8214	2.7439**
	(-4.039)	(-2.293)	(-1.008)	(2.487)
B	3.5986	13.8061	1.8811	1.8582
	(0.097)	(0.817)	(0.054)	(0.883)
Exp	-5.0785	6.9007	-2.1856	-0.3366
	(-0.249)	(1.228)	(-0.396)	(-0.258)
Imp	-3.4573	-1.3967	-1.8942	0.5284
	(-0.098)	(-1.023)	(-0.408)	(0.328)
Stock	2.3415***	-9.5215	3.4975**	1.114
	(5.56)	(-0.729)	(2.18)	(1.120)
Constant	8.3927	18.313	7.5292	-6.6129***
	(0.571)	(0.792)	(0.285)	(-3.185)

(1) *, **, *** represent the case when the p value is less than 10%, 5%, and 1%, respectively.

6 Conclusion

Determining optimal monetary policy is core for central banks and economists worldwide. Traditional research primarily focuses on understanding the factors influencing current economic decisions (Brash et al., 2001; Bernanke and Boivin, 2003). However, existing economic decisions may not always be optimal. The advent of reinforcement learning, extensively applied in recent years, has facilitated the search for optimal policies. Studies like AI Economist (Zheng et al., 2022) leverage RL to find optimal decisions. Governments often struggle to apply model-derived results in practice due to insufficient understanding of the underlying decision-making logic. We examine the asymptotic distribution of coefficients corresponding to optimal decisions, akin to traditional OLS models, providing insight into the significance of these coefficients and aiding in the comprehension of optimal decisions derived from RL.

Specifically, we utilize the V-learning algorithm to linearize the state value function V and solve for the optimal policy accordingly. Within this framework, we present the general asymptotic distribution for state variable coefficient estimates in differentiable policy functions, facilitating statistical inference on the coefficients of state variables in the optimal policy. Using multinomial logistic regression as a special case of the policy function, we derive the specific form of the asymptotic distribution and apply it to simulation experiments and real data, yielding favorable results. We effectively assess the importance of state variables in the optimal policy, allowing for a better understanding of optimal decisions derived from interpretable RL. This statistical inference approach to variable importance aids in extracting knowledge and even wisdom from AI-generated strategies. We conducted an empirical study on formulating optimal monetary policies using the U.S. as a case study, discovering the impact of these optimal policies on stabilizing inflation, in-

creasing employment, and promoting economic growth. Additionally, we gained a deeper understanding of the rules behind the formulation of these optimal policies.

For future extensions, we can conduct studies on different types of economies to compare and summarize the optimal monetary policy rules for developed versus developing countries, resource-based versus technology-based economies, and core versus peripheral countries. This will deepen the theoretical research on monetary policy and support practical policy formulation. Furthermore, our research can be extended to businesses and even individuals, assisting in various decision-making processes and helping to derive insights and wisdom from them.

7 BibTeX

References

Adadi, A. and M. Berrada (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access 6*, 52138–52160.

Adler, G., P. Castro, and C. Tovar (2016). Does central bank capital matter for monetary policy? *Open Economies Review 27*(1), 183–205.

Alstadheim, R., H. C. Bjørnland, and J. Maih (2021). Do central banks respond to exchange rate movements? a markov-switching structural investigation of commodity exporters and importers. *Energy Economics 96*, 105138.

Angeletos, G.-M. and J. La'o (2020). Optimal monetary policy with informational frictions. *Journal of Political Economy 128*(3), 1027–1064.

Baxter, M. and R. G. King (1993). Fiscal policy in general equilibrium. *The American Economic Review*, 315–334.

Bernanke, B. S. and J. Boivin (2003). Monetary policy in a data-rich environment. *Journal of Monetary Economics 50*(3), 525–546.

Bernanke, B. S. and M. Gertler (2001). Should central banks respond to movements in asset prices? *American economic review 91*(2), 253–257.

Brash, D. et al. (2001). Making monetary policy: A look behind the curtains. *Reserve Bank of New Zealand Bulletin 64*.

Byrd, R. H., P. Lu, J. Nocedal, and C. Zhu (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing 16*(5), 1190–1208.

Chen, M., A. Joseph, M. Kumhof, X. Pan, and X. Zhou (2021). Deep reinforcement learning in a monetary model. *arXiv preprint arXiv:2104.09368*.

Friedman, M. (1969). *The Optimum Quantity of Money*. London, UK: Macmillan.

Hastie, T., R. Tibshirani, J. H. Friedman, and J. H. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction* (2 ed.). New York: Springer.

Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional anova models. *The annals of statistics 26*(1), 242–272.

Khan, A., R. G. King, and A. L. Wolman (2003). Optimal monetary policy. *The Review of Economic Studies 70*(4), 825–860.

Kiran, B. R., I. Sobh, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogamani, and P. Pérez (2021). Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems 23*(6), 4909–4926.

Konda, V. and J. Tsitsiklis (2000). Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pp. 1008–1014.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. New York: Springer.

Liu, D. C. and J. Nocedal (1989). On the limited memory bfgs method for large scale optimization. *Mathematical programming 45*(1-3), 503–528.

Loungani, P. and N. Sheets (1997). Central bank independence, inflation, and growth in transition economies. *Journal of Money, Credit, and Banking*, 381–399.

Lubik, T. A. and F. Schorfheide (2007). Do central banks respond to exchange rate movements? a structural investigation. *Journal of Monetary Economics 54*(4), 1069–1087.

Luckett, D. J., E. B. Laber, A. R. Kahkoska, D. M. Maahs, E. Mayer-Davis, and M. R. Kosorok (2020). Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association 115*(530), 692–706.

Milani, S., N. Topin, M. Veloso, and F. Fang (2024). Explainable reinforcement learning: A survey and comparative review. *ACM Computing Surveys 56*(7), 1–36.

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

Nian, R., J. Liu, and B. Huang (2020). A review on reinforcement learning: Introduction and applications in industrial process control. *Computers & Chemical Engineering 139*, 106886.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York: John Wiley & Sons, Inc. A Wiley-Interscience Publication.

Radford, A., K. Narasimhan, T. Salimans, I. Sutskever, et al. (2018). Improving language understanding by generative pre-training. *OpenAI blog*.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog 1*(8), 9.

Ramprasad, P., Y. Li, Z. Yang, Z. Wang, W. W. Sun, and G. Cheng (2022). Online bootstrap inference for policy evaluation in reinforcement learning. *Journal of the American Statistical Association*, 1–14.

Shi, C., S. Zhang, W. Lu, and R. Song (2022). Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 84*(3), 765–793.

Silver, D., J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. (2017). Mastering the game of go without human knowledge. *Nature 550*(7676), 354–359.

Sutton, R. S. and A. G. Barto (1998). *Reinforcement Learning: An Introduction* (1st ed.). Cambridge, Massachusetts, USA: The MIT Press.

Sutton, R. S. and A. G. Barto (2018). *Reinforcement Learning: An Introduction* (2nd ed.). Cambridge, Massachusetts, USA: The MIT Press.

Sutton, R. S., D. McAllester, S. Singh, and Y. Mansour (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pp. 1057–1063.

Svensson, L. E. (2014). How to weigh unemployment relative to inflation in monetary policy? *Journal of Money, Credit and Banking 46*(S2), 183–188.

Svensson, L. E. O. (2003). What is wrong with taylor rules? using judgment in monetary policy through targeting rules. *Journal of Economic Literature 41*(2), 426–477.

Taylor, J. B. (1993). Discretion versus policy rules in practice. In *Carnegie-Rochester Conference Series on Public Policy*, Volume 39, pp. 195–214. Elsevier.

Trott, A., S. Srinivasa, D. van der Wal, S. Haneuse, and S. Zheng (2021). Building a foundation for data-driven, interpretable, and robust policy design using the ai economist. *arXiv preprint arXiv:2108.02904*.

White, M. and A. White (2010). Interval estimation for reinforcement-learning algorithms in continuous-state domains. In *Advances in Neural Information Processing Systems*, pp. 2433–2441.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning 8*, 229–256.

Woodford, M. (2001). The taylor rule and optimal monetary policy. *American Economic Review 91*(2), 232–237.

Xu, Z. (2017). The progress in monetary economics and practices in china. *Journal of Financial Research* (1), 1–21.

Zheng, S., A. Trott, S. Srinivasa, D. C. Parkes, and R. Socher (2022). The ai economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science advances 8*(18), eabk2607.