Applied Mathematics & Optimization (2025) 92:56
https://doi.org/10.1007/s00245-025-10318-7

Inference of Utilities and Time Preference in Sequential Decision-Making

Haoyang Cao[1] · **Zhengqi Wu**[2] · **Renyuan Xu**[3]

Received: 18 May 2024 / Accepted: 26 August 2025 / Published online: 24 October 2025

Abstract

This paper introduces a novel stochastic control framework to enhance the capabilities of automated investment managers, or robo-advisors, by accurately inferring clients' investment preferences from past activities. Our approach leverages a continuous-time model that incorporates utility functions and a generic discounting scheme with a time-varying rate, which can be tailored to each client's risk tolerance, valuation of daily consumption, and significant life goals; this general discounting scheme is also referred to as the time preference. Through state augmentation, the new stochastic control framework allows for the joint inference of utilities and time preferences. We establish the corresponding dynamic programming principle and the verification theorem. Additionally, we provide sufficient conditions for the identifiability of client investment preferences. To complement our theoretical developments, we propose a learning algorithm based on maximum likelihood estimation within a discrete-time Markov Decision Process framework, augmented with entropy regularization. We prove that the Hessian matrix of the log-likelihood function is negative semi-definite at the client's investment parameters, facilitating fast convergence of our proposed algorithm. Practical effectiveness and efficiency are showcased through two numerical examples, including Merton's problem and an investment problem with unhedgeable risks. Our proposed framework not only advances financial technology by improving personalized investment advice but also contributes broadly to other fields such as healthcare, economics, and artificial intelligence, where understanding individual preferences is crucial.

Keywords Inference · Utility · Time preference · Sequential decision-making · Inverse optimal control · Identifiability

Extended author information available on the last page of the article

1 Introduction

Automated investment managers, commonly known as robo-advisors, have emerged as a modern alternative to traditional financial advisors in recent years [17, 28, 68]. The effectiveness and viability of robo-advisors depend significantly on their ability to provide customized financial guidance tailored to the unique collection of needs for each client. To provide impactful personalized advice, two critical steps must be undertaken: first, accurately estimate the client's investment preferences, and second, formulate investment recommendations that align with these preferences. This paper focuses on the first step, involving a detailed analysis of the client's investment preferences.

More often than not, it is difficult for the automated investment manager to have full access to clients' investment preferences. Therefore, it is worth exploring whether it is possible to infer relevant information by observing the clients' past investment activities. Nevertheless, inferring a client's investment preferences is typically challenging, as it involves several complex aspects that vary from individual to individual. For example, clients may have short-term or long-term investment objectives [43]. Additionally, they might exhibit varying utility functions [57, 82], reflecting distinct risk tolerances related to profit-and-loss (PnL) outcomes and valuations of daily consumption. Furthermore, individuals often demonstrate diverse time preferences in terms of the trade-off between immediate and deferred outcomes [7]. Finally, clients may have specific life goals [16], such as saving for their children's education or building a retirement nest egg, rather than focusing solely on generating the highest possible portfolio return or beating the market.

The inference of preferences in sequential decision-making is a critical component not only for financial investments but also in other fields, leveraging insights into individual behaviors to optimize decisions and predict outcomes. In economics, utility functions are inferred to model consumer behavior, guiding businesses in product development and pricing strategies [24, 74]. Healthcare professionals use inferred utility functions to evaluate patient preferences regarding different treatment options, which is essential for effective healthcare management and policy-making [18, 66]. Additionally, in artificial intelligence, particularly in areas like reinforcement learning (RL) and game theory, inferring utility functions helps in designing algorithms that can predict and mimic human decision-making processes, enhancing the interaction between humans and machines [13, 19].

Our framework, results, and contributions We propose a novel stochastic control framework in continuous time that incorporates all the aforementioned investment preferences. This framework includes two utility functions that allow a client to specify the risk tolerance related to the PnL outcomes and the valuations of daily consumption. Additionally, it allows for a generic discounting scheme with a time-varying rate, enabling the client to balance immediate and deferred outcomes. This time-varying discounting scheme, which we also refer to as the time preference of the client, further incorporates specific life goals by assigning greater importance to times of significant expenditures, such as college tuition fees for children. Lastly, we address control problems on both finite-time and infinite-time horizons to accommodate clients' preferred investment duration. To ultimately achieve an accurate joint

inference of the client's utility and time preference, we augment the state space to enunciate the impact of the time preference on the optimal strategy. We study the well-definedness of the augmented control framework by establishing the regularity of the value function, the dynamic programming principle (DPP), and the verification theorem (see Propositions 1, 2, 3 for finite-time horizon and Propositions 4, 5, 6 for infinite-time horizon). In addition, we identify sufficient conditions for identifying both the utility functions and the discounting scheme by *solely* observing the optimal policies provided by the client (see Theorem 1 for finite-time horizon and Theorem 2 for infinite-time horizon).

To complement the above theoretical framework, we propose an inference procedure based on maximum likelihood estimation. Motivated by practical implementation and computational efficiency, we focus on a discrete-time Markov Decision Process (MDP) with Shannon entropy regularization over a finite-time horizon, which is a setting often considered in the RL literature. The discrete-time framework is particularly well-suited for applications in statistical inference and machine learning, where data is often observed at discrete intervals and decision-making naturally occurs sequentially. The entropy term encourages full exploration of the state-action space and simultaneously introduces smoothness into the analysis [36]. We employ a parametric framework where the client uses an exponential discounting scheme, parameterized by $\bar{\rho}$, and a utility function parameterized by $\bar{\theta}$. Both sets of parameters are unknown to the automated investment manager. Mathematically, we show that the true preference parameter $(\bar{\rho}, \bar{\theta})$ is a stationary point of the log-likelihood function and the Hessian matrix of the log-likelihood function is negative semi-definite at that point; see Proposition 7 and Theorem 3. This landscape property facilitates the design of a gradient-based algorithm to update the inferred preference parameter. We demonstrate the promising performance of our algorithm through two examples–Merton's problem and an investment problem under unhedgeable risks.

Considering the wide-ranging applications and the versatility of our proposed framework, we use the term "inference agent" instead of "automated investment manager" to describe the individual who interacts with the clients and infers their preferences.

Related literature and comparisons to our results Our developments are associated with several lines of literature as follows.

Utility inference Back in 1964, Kalman [46] asked the question of whether it is possible to recover the quadratic cost by observing an optimal linear policy; a similar question was also considered by Boyd et al. [14]. In fact, economists have long been interested in such questions within the context of determining utility functions from observations such as Samuelson [70] and Richter [67]. For instance, Keeney and Raiffa [48] studied the proper rank of actions based on some deterministic evaluations under a static setting. Sargent [71] later extended this question into a dynamic setting where the actions were specified as labor demand and evaluations as wages. Dybvig and Rogers [26] paid special attention to the *recoverability* or *identifiability* of utility and showed that Von Neumann–Morgenstern preferences over terminal consumption can be inferred from wealth process of a discrete-time, binomial model or continuous-time Gaussian model.

Black proposed the inverse version of the classical Merton's problem in the note [10]; see also [11]. Among various studies of Black's inversion problem, some took a "backward-looking" perspective and modeled the problem as an inverse optimal control problem. For instance, Cox et al. [20] adopted this approach to study a particular continuous-time Black's inverse problem via the client's investment–consumption profile. They also observed the degeneracy of this inverse problem: there are infinitely many utility functions compatible with the investment profile under a given dynamic environment. This backward-looking approach was also adopted in [37]. El Karoui and Mrad [30], on the other hand, took a "forward-looking" perspective to study the connection between the observable process (i.e., the characteristic process), and the corresponding utility process (i.e., the dynamic utility). This concept of forward utility was proposed by Musiela and Zariphopoulou [58] and was subsequently adopted in, for example, [59] to study the Black's inverse problem. Different than the backward-looking perspective where the connection between the observable and utility is governed by some Markovian decision-making rule, the authors of [30] interpreted such a connection as the martingale property which can be fully characterized via the Itô–Ventzel formula; this set of analytical tools was introduced in [31] and [61]. In [32], the authors also extended the result of [30] to allow an exogenous default time τ. The forward utility approach was also adopted by Källblad et al. [45], Källblad [44], Angoshtari et al. [5], He et al. [38], among many others. Apart from the above two approaches, Monin [56] adopted the distribution builder method, which was initially proposed in [73], to infer the utility dynamically.

In recent years, utility inference has been integrated with machine learning to embrace the potential of the big data era (and the progress is summarized in the next paragraph). In addition, inference problems in sequential decision-making for modern applications are more complex than inferring solely the utility function. Other preferences such as time preferences and specific investment goals should also be included, leading to the main formulation of our paper.

Theory of inverse optimal control Inverse optimal control aims at inferring the underlying reward function that motivates the observed behavior of a rational agent in a sequential decision-making framework; within the context of MDP, inverse optimal control is also known as inverse reinforcement learning (IRL). In this area, Ng et al. [60] considered a particular setting that the true reward function is some linear combination of several action-free basis functions and that the true reward function maximally distinguishes the observed policy from the rest. They reformulated this question into a constrained linear programming problem eventually leading to a well-defined solution. In [1], the reward was assumed to be a linear combination of several features that best distinguish the demonstrated policy from other policies. The key assumption in both works is that the true reward function should maximize the margin between observations and the other policies. It also played a central role in the model of the well-known generative adversarial imitation learning (GAIL) algorithm [40]. Other than the "maximum margin" setting, another commonly adopted setting in IRL is to assume that an observed randomized policy should maximize the causal entropy of an underlying regularized MDP. For instance, Ziebart et al. [85] studied the maximum entropy IRL based on known features. They assumed that the reward is a linear function of such features. Ziebart [84] extended this approach to a selected

set of non-linear rewards; see also [13] and [51] for similar settings. Wulfmeier et al. [79] followed this approach but with rewards represented by neural networks. Finn et al. [33] combined the idea of adversarial training and IRL. They trained a discriminator to recover the reward function. Reddy et al. [64] proposed a soft Q imitation learning algorithm to imitate the expert's policy by learning the Q function. Garg et al. [35] proposed an algorithm to learn the soft Q function which implicitly represents both the reward function and the policy. Zeng et al. [83] adopted the maximum likelihood estimator and showed that their algorithm converges to a stationary point under a finite-time guarantee.

Back to our preference inference problem, since it is to infer the utility functions and the time preferences of the client simultaneously, these existing IRL algorithms are *not* directly applicable. Such a *multi-facet* inference problem motivates our main algorithm. Furthermore, we are able to provide a loss landscape analysis that facilitates fast convergence of our proposed algorithm; see Proposition 7 and Theorem 3.

Identifiability issues in IRL In 1998, Russell [69] pointed out the ill-posedness of inverse optimal control or IRL problems under a generic setting. Both the "maximum margin" and the "maximum entropy" settings mentioned above are reasonable assumptions to ameliorate this ill-posedness. Nonetheless, without prior access to the underlying true reward function, it is difficult to verify either one of them. To guarantee identifiability in IRL, alternative and more verifiable conditions are required. Under an entropy-regularized MDP setting, Cao et al. [15] pointed out two possible remedies for the identifiability issue. One way is to provide additional observations of the same agent (i.e., keeping the underlying reward function the same) under different environments; see also a repeated IRL setting proposed in [3] and [4]. It was shown in [15] that under proper technical conditions on the transition kernels, observations from two distinct environments would suffice. Another approach is to provide additional structural assumptions on the MDP environment or the family of candidate reward functions based on prior domain knowledge; see also the identification of an action-free reward in [34]. Both Cao et al. [15] and Kim et al. [50] provided sufficient structural conditions for the MDP environment that guarantee identifiability.

However, as pointed out by Schlaginhaufen and Kamgarpour [72], the identifiability may no longer hold without the entropy regularization. In addition, the majority of these previous studies rely on the *full disclosure of the MDP environment*, including the transition kernel, time horizon, and the rate of an *exponential discounting scheme*. Though Dong and Wang [25] provided a mathematical formulation and an algorithm for the partial information setting, it remains to be explored whether identifiability of both the unknown MDP information and the true reward function is viable. In this paper, we establish such identifiability for our preference inference problem, which is also one of the major theoretical contributions; see Theorems 1 and 2.

Stochastic control beyond exponential discount Stochastic control problems with a discounting scheme other than the exponential type are indeed nontrivial. In economics, one of the earliest studies to recognize this nontriviality is [75], where the author studied a dynamic utility maximization with a generic discount function. Later, Pollak [63] proposed a game-theoretic consistent planning approach to address the additional challenge of a non-exponential discount function for the discrete-time problem, where the game is among decision makers at different time steps and the

optimal decision path is considered to be the Nash equilibrium. There has been a line of works following this consistent planning approach under both discrete- and continuous-time settings; see, for instance, [8, 9, 29, 41, 42, 80], and more recently, [23, 39]. Apart from this game-theoretic approach, Karnam et al. [47] introduced the idea of "dynamic utility" to a family of utility optimization problems over a *finite-time horizon*. By modeling the utility as the solution to a backward stochastic differential equation (BSDE), the DPP could be revived. For an *infinite-time horizon* setting which is suitable to model a long-run investment planning problem though, this BSDE approach can no longer be applied. Hence we propose a state augmentation approach to revive DPP; see Propositions 1 and 4 in Sect. 2.

Robo-advising Robo-advising has emerged over the last two decades as an alternative to traditional human financial advising, addressing limitations such as the human advisors' limited knowledge and high service fees [17, 27, 28]. Here, we mainly review some papers that explore the machine learning and inference aspects of this subject. One of the first RL algorithm for a robo-advisor was proposed by Alsabah et al. [2], where the authors designed an exploration-exploitation algorithm to learn a constant risk appetite parameter and then applied a follow-the-leader type of algorithm to invest. Wang and Yu [78] introduced a framework consisting of two agents: the first, an inverse portfolio optimization agent, infers a risk preference parameter and the corresponding expected return; the second aggregates the learned information to formulate a new multi-period portfolio optimization problem solved by deep learning. To transcend the rather single-facet inference settings above, the theoretical framework and the numerical procedure in our paper are designed to capture the multiple investment needs of a client.

2 Continuous-Time Framework

In this section, we study a continuous-time framework of the joint consumption-allocation problem of an investing client. The client's wealth consists of a risk-free asset and a risky asset. What distinguishes this framework from the classical ones is that the client holds a *general preference of time*, that is, the discounting scheme is *not necessarily exponential*. Consequently, we adopt a state augmentation approach to capture the explicit influence of a general discounting scheme on the optimal investment strategy. First, for the *optimal control* problem, we analyze the dynamical decision-making problem for such a client, assuming the client's utility functions of consumption and wealth as well as the time preference are fully disclosed. The optimal decision relies on reviving a suitable DPP under this framework. Then, for the *inverse optimal control* problem, we establish an identifiability result for both the utility functions and the time preference of the client, assuming instead only the optimal joint consumption-allocation plan is disclosed. Such an identifiability result provides inspirations for the algorithm proposed in Sect. 3.

2.1 Finite-Time Horizon

We first focus on a finite-time horizon setting. This setting is suitable to study investment strategies with a clearly defined end time.

Market Dynamics and Client's Wealth Let $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ be a filtered probability space satisfying the usual conditions, supporting a one-dimensional $\mathbb{F}$ -Brownian motion W. Assume there is a bond and a stock in the investment universe. The price of the bond follows

$$dS_t^0 = rdt, \tag{1}$$

and the price of the stock follows

$$dS_t = S_t(\mu dt + \sigma dW_t). \tag{2}$$

Assume the client chooses an allocation process $\boldsymbol{\alpha} = \{\alpha_t\}_{t\in[0,T]}$ and a consumption process $\boldsymbol{c} = \{c_t\}_{t\in[0,T]}$ with $c_t \geq 0$. Namely, the client allocates α_t proportion of wealth to the stock and $1-\alpha_t$ proportion of wealth to the bond at time t. In addition, the client is also consuming the wealth with a dollar-amount consumption rate c_t to achieve certain satisfaction in life.

Fixing a sufficiently large constant $M \in \mathbb{R}^+ := (0, +\infty)$ and introducing a compact space $\mathcal{K} = [-M, M] \times [0, M]$, define

$$\begin{aligned}\mathcal{A} := \Bigg\{ (\boldsymbol{\alpha}, \boldsymbol{c}) = \{(\alpha_t, c_t)\}_{t\geq 0} \Bigg| \, & (\alpha_t, c_t) \in \mathcal{K},\ (\alpha_t, c_t) \in \bar{\mathcal{F}}_t \\ & := \sigma\left(\sigma(\beta_s, 0 \leq s \leq t) \times \mathcal{F}_t^W\right) \Bigg\}\end{aligned} \tag{3}$$

as the admissible set of all possible joint consumption-allocation processes. Hence the wealth process follows [21, 54, 77, 81]:

$$dX_t^{\boldsymbol{\alpha},\boldsymbol{c}} = \left\{X_t^{\boldsymbol{\alpha},\boldsymbol{c}} \left[\alpha_t \mu + (1-\alpha_t) r\right] - c_t\right\} dt + \sigma \alpha_t X_t^{\boldsymbol{\alpha},\boldsymbol{c}} dW_t. \tag{4}$$

Note that under $(\boldsymbol{\alpha}, \boldsymbol{c}) \in \mathcal{A}$, we have $\mathbb{E}\left[|X_t^{\boldsymbol{\alpha},\boldsymbol{c}}|^2\right] < \infty$ for any $t \geq 0$.

Client's Preference In the finite-time horizon, the preference of the client can be characterized by a pair of utility functions and a discount scheme. More specifically, consider utility functions U_1, U_2 that belong to the following class

$$\begin{aligned}\mathcal{U} := \Big\{ U : \mathbb{R} \to [-\infty, +\infty) \Big| \, & U \text{ is finite, strictly concave and increasing on } \mathbb{R}^+, \\ & U \in \mathcal{C}^2(\mathbb{R}^+), \\ & U(0) = \lim_{x\to 0^+} U(x),\ U(x) = -\infty \text{ for } x < 0 \Big\}.\end{aligned} \tag{5}$$

Note that U_1 quantifies the client's evaluation regarding the consumption whereas U_2 quantifies the evaluation regarding the terminal wealth at the end of the investment plan, which will be specified below.

General Discounting Scheme We are particularly interested in a client that is subject to a general discounting scheme $\beta = \{\beta_t = \beta(t)\}_{t\geq 0}$, where

- $\beta_t \in [0,1]$ for all $t \in [0,T]$; and
- there exists $\dot{\beta} : [0,\infty) \to \mathbb{R}$ such that $\dot{\beta}$ is integrable on $[0,t]$ with $\beta_t = \int_0^t \dot{\beta}_s ds + \beta_0$ for any $t > 0$.

Such a discounting scheme $\{\beta_t\}_{t\geq 0}$ reflects a generic *time preference* of the client. A time-varying discounting rate could account for different levels of appreciation for the immediate outcome and the delayed fulfillment. It could also provide the flexibility of assigning greater importance to times of significant expenditures, such as college tuition for children and down-payment of a house.

Over a finite horizon $[0,T]$, we consider the problem

$$\sup_{\boldsymbol{\alpha},\boldsymbol{c}} \mathbb{E}\left[\int_0^T \beta_t U_1(c_t)dt + \beta_T U_2\big(X_T^{\boldsymbol{\alpha},\boldsymbol{c}}\big)\right].$$

To enunciate the impact of discounting scheme β which is not necessarily exponential, define the total reward as, for any $(t,x,z) \in [0,T]\times\mathbb{R}\times[0,1]$,

$$J(t,x,z,\boldsymbol{\alpha},\boldsymbol{c}) := \mathbb{E}\left[\int_t^T \beta_s U_1(c_s)ds + \beta_T U_2\big(X_T^{\boldsymbol{\alpha},\boldsymbol{c}}\big)\,\middle|\, X_t^{\boldsymbol{\alpha},\boldsymbol{c}} = x, \beta_t = z\right] \tag{6}$$

subject to the wealth process (4) and

$$d\beta_t = \dot{\beta}_t dt. \tag{7}$$

For any $(x,z) \in \mathbb{R}\times[0,1]$, define the value function as follows,

$$V(t,x,z) = \sup_{(\boldsymbol{\alpha},\boldsymbol{c})\in\mathcal{A}} J(t,x,z,\boldsymbol{\alpha},\boldsymbol{c}),\ t\in[0,T);\quad V(T,x,z) = zU_2(x) \tag{8}$$

subject to (4) and (7).

In this section, we start by revisiting the DPP to the above utility optimization problem (8) over $(X^{\boldsymbol{\alpha},\boldsymbol{c}},\beta)$, where the state space is enlarged with the general discounting scheme. Different than the BSDE characterization of dynamic utility in [47], the problem formulation (8) (inspired by [6]) enables us to depict a clearer connection between the discounting scheme β and the optimal joint consumption-allocation process $(\boldsymbol{\alpha}^*,\boldsymbol{c}^*)$. Subsequently, it lays the foundation for the study of the corresponding joint inverse problem.

2.1.1 Preliminary Analysis

First, we establish the well-definedness of the control problem (4)–(7) and introduce some analytical properties associated with it.

The first result concerns with the client's wealth under optimal consumption-allocation policies under reasonable utility functions.

Lemma 1 If utility functions $U_1, U_2 \in \mathcal{U}$ satisfy that $U_1(0) \in \mathbb{R}$ and $U_2(0) = -\infty$, then for any $(t, x, z) \in [0, T] \times \mathbb{R}^+ \times [0, 1]$, given that policy $\boldsymbol{\alpha}^*, \boldsymbol{c}^*$ satisfies that $J(t, x, z, \boldsymbol{\alpha}^*, \boldsymbol{c}^*) = V(t, x, z)$, it holds almost surely that

$$X_s^{\boldsymbol{\alpha}^*, \boldsymbol{c}^*} \in \mathbb{R}^+ \quad \text{for all} \quad s \in [t, T], \tag{9}$$

where $X^{\boldsymbol{\alpha}^*, \boldsymbol{c}^*}$ solves (4) on $[t, T]$ given $(\boldsymbol{\alpha}, \boldsymbol{c}) = (\boldsymbol{\alpha}^*, \boldsymbol{c}^*)$ and $X_t^{\boldsymbol{\alpha}, \boldsymbol{c}} = x$.

Proof Since $U_1 \in \mathcal{U}$ and therefore $U_1(0) = \lim_{x \to 0^+} U_1(x)$ and U_1 is increasing on $\mathbb{R}^+$, we have that $U_1(0) = \min_{x \in [0, +\infty)} U_1(x)$. Let $\tilde{U}_1(x) = U_1(x) - U_1(0)$ for all $x \in \mathbb{R}$, then $\tilde{U}_1 \in \mathcal{U}$ with $\tilde{U}_1(0) = 0$. For any $t, x, z \in [0, T] \times \mathbb{R}^+ \times [0, 1]$ and any $(\boldsymbol{\alpha}, \boldsymbol{c}) \in \mathcal{A}$, denote

$$\tilde{J}(t, x, z, \boldsymbol{\alpha}, \boldsymbol{c}) = \mathbb{E}\left[\int_t^T \beta_s \tilde{U}_1(c_s) ds + \beta_T U_2\big(X_T^{\boldsymbol{\alpha}, \boldsymbol{c}}\big) \middle| X_t^{\boldsymbol{\alpha}, \boldsymbol{c}} = x, \beta_t = z \right].$$

Then the total reward $J(t, x, z, \boldsymbol{\alpha}, \boldsymbol{c})$ defined in (6) can be rewritten as

$$J(t, x, z, \boldsymbol{\alpha}, \boldsymbol{c}) = U_1(0) \int_t^T \beta_s ds + \tilde{J}(t, x, z, \boldsymbol{\alpha}, \boldsymbol{c}),$$

and consequently, the value function defined in (8) satisfies

$$V(t, x, z) = \sup_{(\boldsymbol{\alpha}, \boldsymbol{c}) \in \mathcal{A}} J(t, x, z, \boldsymbol{\alpha}, \boldsymbol{c}) = U_1(0) \int_t^T \beta_s ds + \sup_{(\boldsymbol{\alpha}, \boldsymbol{c}) \in \mathcal{A}} \tilde{J}(t, x, z, \boldsymbol{\alpha}, \boldsymbol{c}).$$

Thus, without loss of generality, we can assume that $U_1(0) = 0$.

For any $(\boldsymbol{\alpha}, \boldsymbol{c}) \in \mathcal{A}$, we have $X_s^{\boldsymbol{\alpha}, \boldsymbol{c}} \leq X_s^{\boldsymbol{\alpha}, \mathbf{0}}$ for all $s \in [t, T]$ almost surely. Notice that $X_{t'}^{\boldsymbol{\alpha}, \mathbf{0}} = X_t^{\boldsymbol{\alpha}, \mathbf{0}} \exp\left\{ \int_t^{t'} \alpha_l(\mu - r) + r - \frac{\sigma^2 \alpha_l^2}{2} dl + \int_t^{t'} \sigma \alpha_l dW_l \right\}$ for $t' \geq t$. Then

$$\begin{aligned} J(t, x, z, \boldsymbol{\alpha}, \mathbf{0}) &= \beta_T \mathbb{E} U_2 \left(x \exp\left\{ \int_t^T \alpha_l(\mu - r) + r - \frac{\sigma^2 \alpha_l^2}{2} dl + \int_t^T \sigma \alpha_l dW_l \right\} \right) \\ &> -\infty. \end{aligned}$$

On the other hand, if $\mathbb{P}\{\exists t' \in (t,T] \text{ s.t. } X_{t'}^{\boldsymbol{\alpha},\boldsymbol{c}} \leq 0\} > 0$, then $\mathbb{P}\{X_T^{\boldsymbol{\alpha},\boldsymbol{c}} \leq 0\} > 0$, and hence, $J(t,x,z,\boldsymbol{\alpha},\boldsymbol{c}) = -\infty$. Thus, if $J(t,x,z,\boldsymbol{\alpha}^*,\boldsymbol{c}^*) = V(t,x,z) \geq J(t,x,z,\boldsymbol{\alpha},\mathbf{0}) > -\infty$, then $X_s^{\boldsymbol{\alpha}^*,\boldsymbol{c}^*} \in \mathbb{R}^+$ for all $s \in [t,T]$ almost surely. □

Remark 1 The class of utility functions given by (5) include many commonly-seen HARA utility functions such as exponential utility $\frac{1-e^{-\alpha x}}{\alpha}$ for some $\alpha > 0$, power utility $\frac{x^\theta}{\theta}$ for some $\theta \in (0,1)$, and log utility $\log x$, when $x > 0$. With additional assumptions that $U_1(0) \in \mathbb{R}$ and $U_2(0) = -\infty$, the consumption utility $U_1(c)$ can be an exponential utility or a power–log utility whose coefficient of relative risk aversion $R_1(c) = -\frac{cU_1''(c)}{U_1'(c)}$ satisfies $R_1 := \sup_{c>0} R_1(c) \in (0,1)$; the terminal wealth utility $U_2(x)$ can be a power-log utility whose coefficient of relative risk aversion $R_2(x) = -\frac{xU_2''(x)}{U_1'(x)}$ satisfies $R_2 := \sup_{x>0} R_2(x) \in [1,+\infty)$.

Remark 2 What Lemma 1 provides is a **sufficient condition** on utility functions that avoids exhausting the total wealth before the investing horizon T (almost surely). For instance, if U_1 takes the power utility and U_2 takes the log utility $\log x$, then the assumptions in Lemma 1 will be satisfied. Other conditions to almost surely avoid $X_T^{\boldsymbol{\alpha},\boldsymbol{c}} \leq 0$ include the case where $U_1 \equiv 0$ and $U_2(0) = \lim_{x\to 0^+} U_2(x) \leq 0$, and the corresponding proof follows similar arguments in the above proof.

Lemma 1, together with the discussion in Remark 2, ensures that the set of "rational utility pairs", $\mathbb{U}^R \subset \mathcal{U} \times \mathcal{U}$, is nonempty, where for any $(U_1,U_2) \in \mathbb{U}^R$, it is almost surely sub-optimal to have $X_T^{\boldsymbol{\alpha},\boldsymbol{c}} \leq 0$ subject to wealth process (4) and discounting scheme (7). Hence, for convenience, we denote the following notations:

$$\mathcal{D} := [0,T] \times \mathbb{R}^+ \times [0,1] \text{ and } \mathcal{D}^\circ := (0,T) \times \mathbb{R}^+ \times (0,1).$$

The next result concerns with total reward under optimal consumption–allocation policies.

Lemma 2 Under wealth process (4) and discounting scheme (7), assume that the utility functions $(U_1,U_2) \in \mathbb{U}^R$ satisfy that $U_1(0) \in \mathbb{R}$. Then the value function $V : \mathcal{D} \to \mathbb{R}$ defined in (8) is strictly concave and strictly increasing in $x \in \mathbb{R}^+$ given any $(t,z) \in [0,T] \times [0,1]$.

Proof Fix any $(t,x,z) \in \mathcal{D}$.

1. *Strictly concave and finite properties*. Since $(\boldsymbol{\alpha},\boldsymbol{c}) = \{(0,0)\}_{t\in[0,T]} \in \mathcal{A}$, by definition, $V(t,x,z) \geq \beta^T U_2(xe^{r(T-t)}) > -\infty$. By concavity and monotonicity of $U_1,U_2 \in \mathcal{U}$ and $U_1(0) \in \mathbb{R}$, for any $(\boldsymbol{\alpha},\boldsymbol{c}) \in \mathcal{A}$, there exists constant $C > 0$ that depends only on U_1 such that

$$\mathbb{E}\left[\int_t^T \beta_s U_1(c_s)ds\right] \le C\mathbb{E}\left[\int_t^T \beta_S(1+c_s)ds\right] \le C(1+M)\int_t^T \beta_s ds < +\infty,$$

and

$$\begin{aligned}\mathbb{E}\big[\beta_T U_2\big(X_T^{\boldsymbol{\alpha},\boldsymbol{c}}\big)\big] &\le \beta_T \mathbb{E}\big[U_2\big(X_T^{\boldsymbol{\alpha},0}\big)\big] \le \beta_T U_2\big(E\big[X_T^{\boldsymbol{\alpha},0}\big]\big)\\ &\le \beta_T U_2\big(xe^{(T-t)[M(\mu-r)+r]}\big) < +\infty.\end{aligned}$$

Take $y \in \mathbb{R}^+\backslash\{x\}$ and $\lambda \in (0,1)$. Define $u = \lambda x + (1-\lambda)y$ and $u \in \mathbb{R}^+$. Take any $(\boldsymbol{\alpha}^x, \boldsymbol{c}^x), (\boldsymbol{\alpha}^y, \boldsymbol{c}^y) \in \mathcal{A}$ and let $X^{\boldsymbol{\alpha}^x,\boldsymbol{c}^x}$ (resp. $X^{\boldsymbol{\alpha}^y,\boldsymbol{c}^y}$) be the solution to the SDE (4) over $[t,T]$ given $(\boldsymbol{\alpha},\boldsymbol{c}) = (\boldsymbol{\alpha}^x,\boldsymbol{c}^x)$ (resp. $(\boldsymbol{\alpha},\boldsymbol{c}) = (\boldsymbol{\alpha}^y,\boldsymbol{c}^y)$) and $X_t^{\boldsymbol{\alpha},\boldsymbol{c}} = x$ (resp. $X_t^{\boldsymbol{\alpha},\boldsymbol{c}} = y$). Define $\boldsymbol{\alpha}^u$ such that for any $s \in [t,T]$, $\alpha_s^u = \Gamma_s \alpha_s^x + (1-\Gamma_s)\alpha_s^y$ with $\Gamma_s = \frac{\lambda X_s^{\boldsymbol{\alpha}^x,\boldsymbol{c}^x}}{\lambda X_s^{\boldsymbol{\alpha}^x,\boldsymbol{c}^x} + (1-\lambda)X_s^{\boldsymbol{\alpha}^y,\boldsymbol{c}^y}} \in (0,1)$ almost surely, and $\boldsymbol{c}^u = \lambda \boldsymbol{c}^x + (1-\lambda)\boldsymbol{c}^y$. Then it immediately follows that $(\boldsymbol{\alpha}^u,\boldsymbol{c}^u) \in \mathcal{A}$. Let $X^{\boldsymbol{\alpha}^u,\boldsymbol{c}^u}$ be the solution to the SDE (4) over $[t,T]$ given $(\boldsymbol{\alpha},\boldsymbol{c}) = (\boldsymbol{\alpha}^u,\boldsymbol{c}^u)$ and $X_t^{\boldsymbol{\alpha},\boldsymbol{c}} = u$. Then we have

$$X_s^{\boldsymbol{\alpha}^u,\boldsymbol{c}^u} = \lambda X_s^{\boldsymbol{\alpha}^x,\boldsymbol{c}^x} + (1-\lambda)X_s^{\boldsymbol{\alpha}^y,\boldsymbol{c}^y}, \quad s \in [t,T].$$

Without loss of generality, we can assume that both $X_T^{\boldsymbol{\alpha}^x,\boldsymbol{c}^x}$ and $X_T^{\boldsymbol{\alpha}^y,\boldsymbol{c}^y}$ are strictly positive. Since $U_1, U_2 \in \mathcal{U}$, then the strict concavity implies that

$$J(t,u,z,\boldsymbol{\alpha}^u,\boldsymbol{c}^u) > \lambda J(t,x,z,\boldsymbol{\alpha}^x,\boldsymbol{c}^x) + (1-\lambda)J(t,y,z,\boldsymbol{\alpha}^y,\boldsymbol{c}^y).$$

Taking the supremum over both $(\boldsymbol{\alpha}^x,\boldsymbol{c}^x)$ and $(\boldsymbol{\alpha}^y,\boldsymbol{c}^y)$,

$$V(t,u,z) > \lambda V(t,x,z) + (1-\lambda)V(t,y,z).$$

2. *Strictly increasing property*. Fix any $\Delta x > 0$ take any $(\boldsymbol{\alpha},\boldsymbol{c}) \in \mathcal{A}$ such that $X_s^{\boldsymbol{\alpha},\boldsymbol{c}} > 0$ for $s \in [t,T]$ almost surely. Let $\widehat{X^{\boldsymbol{\alpha},\boldsymbol{c}}}$ be the solution to (4) given $X_t^{\boldsymbol{\alpha},\boldsymbol{c}} = x + \Delta x$. Then,

$$\begin{aligned}\Delta_T := \widehat{X_T^{\boldsymbol{\alpha},\boldsymbol{c}}} - X_T^{\boldsymbol{\alpha},\boldsymbol{c}} = \Delta x \exp\Bigg\{ &\int_t^T \alpha_l(\mu-r) + r\\ &- \frac{\sigma^2\alpha_l^2}{2}dl + \int_t^T \sigma\alpha_l dW_l\Bigg\} > 0 \text{ a.s.},\end{aligned}$$

and the monotonicity of U_2 implies that

$$J(t,x+\Delta x,z,\boldsymbol{\alpha},\boldsymbol{c}) - J(t,x,z,\boldsymbol{\alpha},\boldsymbol{c}) = \beta_T \mathbb{E}\left[U_2\left(\widehat{X_T^{\boldsymbol{\alpha},\boldsymbol{c}}}\right) - U_2(X_T^{\boldsymbol{\alpha},\boldsymbol{c}})\right] > 0.$$

Hence, $V(t,x+\Delta x,z) > V(t,x,z)$. □

Having established some preliminary properties of the value function, we first show a necessary condition for the value function (8).

Proposition 1 (Dynamic programming principle (DPP)) Take the same assumptions as in Lemma 2. For any $(t,x,z)\in[0,T)\times\mathbb{R}^+\times[0,1]$ and $\tau\in\mathbb{T}_t$ where $\mathbb{T}_t$ denotes all $\{\bar{\mathcal{F}}_t\}_{t\geq 0}$-adapted stopping times τ such that $\tau\in[t,T]$ a.s., the value function V defined in (8) satisfies

$$\begin{aligned} V(t,x,z)=\sup_{(\boldsymbol{\alpha},\boldsymbol{c})\in\mathcal{A}}\mathbb{E}\Bigg[&\int_t^\tau \beta_s U_1(c_s)ds \\ &+V\left(\tau,X_\tau^{\boldsymbol{\alpha},\boldsymbol{c}},\beta_\tau\right)\Big|X_t^{\boldsymbol{\alpha},\boldsymbol{c}}=x,\beta_t=z\Bigg], \end{aligned} \tag{DPP}$$

with $V(T,x,z)=zU_2(x)$.
For any $\alpha\in\mathbb{R}$ and $c\in\mathbb{R}^+$, define the following operator

$$\mathcal{L}^{\alpha,c}\phi(t,x,z)=\{[\alpha(\mu-r)x]-c\}\,\partial_x\phi(t,x,z)+\frac{\sigma^2\alpha^2}{2}x^2\partial_x^2\phi(t,x,z),$$

for any test function $\phi\in\mathcal{C}_b^\infty(\mathcal{D}^\circ)\bigcap\mathcal{C}_b^0(\mathcal{D})$. Following the DPP under a generic discounting scheme (DPP), we have the following Hamilton-Jacobi-Bellman (HJB) equation

$$\begin{cases} \partial_t V(t,x,z)+\dot{\beta}_t\partial_z V(t,x,z)+rx\partial_x V(t,x,z) \\ \quad+\sup\limits_{(\alpha,c)\in\mathcal{K}}\{zU_1(c)+\mathcal{L}^{\alpha,c}V(t,x,z)\}=0,\quad t\in(0,T); \\ V(T,x,z)=zU_2(x). \end{cases} \tag{HJB}$$

The next result provides sufficient conditions for the value function in (8) regarding classical solutions to (HJB).

Proposition 2 Take the same assumptions as in Proposition 1. Let $w:\mathcal{D}\to\mathbb{R}$ be a function such that

$$w\in\mathcal{C}^{1,2,1}\left(\mathcal{D}^\circ\right)\bigcap\mathcal{C}^0\left(\mathcal{D}\right),$$

and there exists a constant $C>0$ with

$$|w(t,x,z)|\leq C(1+|x|^2),\quad\forall(t,x,z)\in\mathcal{D}.$$

1. Assume that for any $(\alpha, c) \in \mathcal{K}$,

$$\begin{cases} \partial_t w(t,x,z) + \dot{\beta}_t \partial_z w(t,x,z) + rx\partial_x w(t,x,z) + zU_1(c) + \mathcal{L}^{\alpha,c} w(t,x,z) \le 0, \\ \quad \forall (t,x,z) \in (0,T) \times \mathbb{R}^+ \times (0,1); \\ w(T,x,z) \ge zU_2(x), \quad \forall (x,z) \in \mathbb{R} \times [0,1]. \end{cases}$$

Then $w \ge V$ on $\mathcal{D}$.

2. Assume further that there exists $\hat{\alpha} : \mathcal{D} \to [-M, M]$ and $\hat{c} : \mathcal{D} \to [0, M]$ such that

$$\begin{cases} \partial_t w(t,x,z) + \dot{\beta}_t \partial_z w(t,x,z) + rx\partial_x w(t,x,z) + zU_1(\hat{c}(t,x,z)) \\ \quad + \mathcal{L}^{\hat{\alpha}(t,x,z),\hat{c}(t,x,z)} w(t,x,z) = 0, \forall (t,x,z) \in (0,T) \times \mathbb{R}^+ \times [0,1]; \\ w(T,x,z) = zU_2(x), \quad \forall (x,z) \in \mathbb{R}^+ \times [0,1], \end{cases}$$

also, with $\beta_t = \beta_0 + \int_0^t \dot{\beta}_s ds \in [0,1]$ for all $t \in [0,T]$, the following SDE,

$$dX_t = \{X_t [\hat{\alpha}(t, X_t, \beta_t)(\mu - r) + r] - \hat{c}(t, X_t, \beta_t)\} dt + \sigma \hat{\alpha}(t, X_t, \beta_t) X_t dW_t,$$

admits a unique solution $X^{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{c}}}$ given $X_0 = x$ for any $x \in \mathbb{R}$, and

$$\left(\hat{\boldsymbol{\alpha}} = \{\hat{\alpha}_t\}_{t \in [0,T]} = \left\{ \hat{\alpha}\left(t, X_t^{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{c}}}, \beta_t\right) \right\}_{t \in [0,T]}, \right.$$
$$\left. \hat{\boldsymbol{c}} = \{\hat{c}_t\}_{t \in [0,T]} = \left\{ \hat{c}\left(t, X_t^{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{c}}}, \beta_t\right) \right\}_{t \in [0,T]} \right) \in \mathcal{A}.$$

Then $w = V$ on $\mathcal{D}$, with $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{c}})$ being an optimal joint allocation-consumption process.

Without assuming the existence of a classical solution to (HJB), we could instead consider its viscosity solution. Recall the following definition of viscosity solution to (HJB) from classical literature such as [22].

Definition 1 1. A upper semi-continuous function $\underline{v} : \mathcal{D} \to \mathbb{R}$, with $\underline{v}(T,x,z) \le zU_2(x)$ for all $x > 0$ and $z \in [0,1]$, is a viscosity subsolution to (HJB) if for any $(t_0, x_0, z_0) \in \mathcal{D}^\circ$ and any $\phi \in \mathcal{C}^{1,2,1}(\mathcal{D}^\circ)$ such that

$$\min_{(t,x,z) \in B(t_0,x_0,z_0)} (\phi - \underline{v})(t,x,z) = (\phi - \underline{v})(t_0,x_0,z_0) = 0$$

for some open neighborhood $B(t_0, x_0, z_0) \subset \mathcal{D}$,

$$\begin{aligned} &- \partial_t \phi(t_0,x_0,z_0) - \dot{\beta}_t \partial_z \phi(t_0,x_0,z_0) - rx_0 \partial_x \phi(t_0,x_0,z_0) \\ &\quad - \sup_{(\alpha,c) \in \mathcal{K}} \left\{ z_0 U_1(c) + \mathcal{L}^{\alpha,c} \phi(t_0,x_0,z_0) \right\} \le 0. \end{aligned} \tag{10}$$

2. An lower semi-continuous function $\overline{v}: \mathcal{D} \to \mathbb{R}$, with $\overline{v}(T,x,z) \geq zU_2(x)$ for all $x > 0$ and $z \in [0,1]$, is a viscosity supersolution to (HJB) if for any $(t_0, x_0, z_0) \in \mathcal{D}^\circ$ and any $\psi \in \mathcal{C}^{1,2,1}(\mathcal{D}^\circ)$ such that

$$\max_{(t,x,z)\in B(t_0,x_0,z_0)} (\psi - \underline{v})(t,x,z) = (\phi - \underline{v})(t_0,x_0,z_0) = 0$$

for some open neighborhood $B(t_0,x_0,z_0) \subset \mathcal{D}$,

$$\begin{aligned} &-\partial_t \psi(t_0,x_0,z_0) - \dot{\beta}_t \partial_z \psi(t_0,x_0,z_0) - r x_0 \partial_x \psi(t_0,x_0,z_0) \\ &\quad - \sup_{(\alpha,c)\in\mathcal{K}} \Big\{ z_0 U_1(c) + \mathcal{L}^{\alpha,c} \psi(t_0,x_0,z_0) \Big\} \geq 0. \end{aligned} \tag{11}$$

3. A continuous function $v: \mathcal{D} \to \mathbb{R}$, with $v(T,x,z) = zU_2(x)$ for all $x > 0$ and $z \in [0,1]$, is a viscosity solution to (HJB) if it is both a viscosity subsolution and a viscosity supersolution to (HJB).

Proposition 3 Take the same assumptions as in Proposition 2, and in addition, assume that $U_2(x)$ satisfies a polynomial growth condition for all $x \in \mathbb{R}^+$, i.e., $\exists p \in [0, +\infty)$ and $C \in \mathbb{R}^+$ s.t. $|U_2(x)| \leq C(1 + |x|^p)$. The value function V in (8) is the unique viscosity solution to (HJB) over the any $\bar{\mathcal{D}} = [0,T] \times \mathcal{D}_1 \times \mathcal{D}_2 \subset \mathcal{D}$ with $\mathcal{D}_i$ compact, $i = 1, 2$.

Assuming that $V \in \mathcal{C}^{1,2,1}(\mathcal{D}^\circ) \bigcap \mathcal{C}(\mathcal{D})$, define the Hamiltonian as

$$H(t,x,z,\alpha,c,p,q) := zU_1(c) + \{x[\alpha\mu + (1-\alpha)r] - c\}p + \frac{\sigma^2\alpha^2}{2}x^2 q.$$

Then, we have that for any $(t,x,z) \in [0,T) \times \mathbb{R}^+ \in [0,1]$,

$$\begin{aligned} (\alpha_t^*, c_t^*) &= (\alpha^*(t,x,z), c^*(t,x,z)) \\ &= \operatorname{argmax}_{(\alpha,c)\in\mathcal{K}} H\Big(t,x,z,\alpha,c,\partial_x V(t,x,z), \partial_x^2 V(t,x,z)\Big), \end{aligned}$$

where

$$\alpha_t^* = -\frac{\partial_x V(t,x,z)}{\partial_x^2 V(t,x,z)} \cdot \frac{(\mu - r)}{\sigma^2 x} \wedge M, \tag{12}$$

$$c_t^* = \operatorname{argmax}_{c\in[0,M]} \Big\{ -\partial_x V(t,x,z)c + zU_1(c) \Big\}; \tag{13}$$

note that by Lemma 2, $\alpha_t^* > 0$.

2.1.2 The Inverse Problem: Identifiability of the Utility Functions

In this section, we focus on the "inverse" problem with respect to the optimal asset allocation-consumption scenario and study the "identifiability" of the utility functions as well as the discounting scheme out of the optimal investment policies. More specifically, we assume the client reveals the decision policies (in the sense of the allocation-consumption processes) to the inference agent.

Following practical protocols, we assume that the inference agent does not know the discounting scheme $\dot{\beta}$ nor the utility functions U_1 and U_2. Nevertheless, the inference agent tries to infer these characteristic functions based on available information, namely the joint allocation-consumption process (i.e., control policy) provided by the client.

To start, let

$$(\bar{\alpha}, \bar{c}) : \mathcal{D} \to \mathcal{K} \tag{14}$$

be some allocation and consumption policies of a client such that

$$\left(\bar{\boldsymbol{\alpha}} = \{\bar{\alpha}_t\}_{t\in[0,T]} = \left\{ \bar{\alpha}\big(t, X_t^{\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{c}}}, \beta_t\big) \right\}_{t\in[0,T]}, \right.$$
$$\left. \bar{\boldsymbol{c}} = \{\bar{c}_t\}_{t\in[0,T]} = \left\{ \bar{c}\big(t, X_t^{\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{c}}}, \beta_t\big) \right\}_{t\in[0,T]} \right) \in \mathcal{A},$$

where $\left(\{\beta_t\}_{t\in[0,T]}, \left\{ X_t^{\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{c}}} \right\}_{t\in[0,T]} \right)$ solves

$$\begin{aligned} d\beta_t &= \dot{\beta}_t dt, \\ dX_t^{\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{c}}} &= \left\{ X_t^{\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{c}}} \left[\bar{\alpha}\Big(t, X_t^{\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{c}}}, \beta_t\Big)(\mu - r) + r \right] - \bar{c}\Big(t, X_t^{\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{c}}}, \beta_t\Big) \right\} dt \\ &\quad + \sigma\bar{\alpha}\Big(t, X_t^{\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{c}}}, \beta_t\Big) X_t^{\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{c}}} dW_t, \end{aligned}$$

for any given $(\beta_0, X_0^{\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{c}}}) \in [0,1] \times \mathbb{R}^+$, and

$$(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{c}}) \in \operatorname{argmax}_{(\boldsymbol{\alpha},\boldsymbol{c})\in\mathcal{A}} J(t, x, z, \boldsymbol{\alpha}, \boldsymbol{c}), \quad \forall (t, x, z) \in \mathcal{D}$$

subject to (4) and (7) . Write

$$\bar{V}(t, x, z) = J(t, x, z, \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{c}}), \quad \forall (t, x, z) \in \mathcal{D}. \tag{15}$$

Note that the inference agent has full access to (14).

Remark 3 (Access to allocation and consumption policies) Note that we assume the inference agent has access to the *functional forms* of the allocation and consump-

tion policies $\bar{\alpha}$ and $\bar{c}$ as (14), which may not be realistic in practice. However, in reality, robo-advisor platforms routinely collect detailed information about clients' past investment behavior and consumption patterns through various tools. A common approach involves administering detailed interactive online questionnaires, which include hypothetical scenario-based questions (e.g., "How would you respond if market scenario A occurs?") as well as inquiries about prior investment history [76]. In addition, with user consent, robo-advisors may access clients' brokerage accounts (e.g., Robinhood) to observe historical trading behavior [65]. Furthermore, APIs such as Plaid or Yodlee can be used-again with user authorization-to retrieve transaction-level financial data [55]. This allows the inference of key financial metrics such as monthly saving and spending rates, cash flow stability, and the incidence of unexpected expenses.

Theorem 1 (Identifiability) Assume that

1. $\bar{\alpha}, \bar{c} \in \mathcal{C}^{1,1,1}(\mathcal{D}^\circ) \bigcap \mathcal{C}^0(\mathcal{D})$;
2. $\bar{\alpha}(t,x,z) \in (0,M)$ for all $(t,x,z) \in \mathcal{D}$;
3. for any $(t,z) \in [0,T] \times [0,1]$,

$$\bar{c}(t,x,z) < x, \quad \forall x > 0;$$

4. both $\bar{\alpha}(T,\cdot,\cdot)$ and $\bar{c}(T,\cdot,\cdot)$ are "z-free", denoted by

$$\bar{\alpha}(T,x,z) \equiv \bar{\alpha}_T(x), \quad \bar{c}(T,x,z) \equiv \bar{c}_T(x), \quad \forall (x,z) \in \mathbb{R}^+ \times [0,1];$$

5. for $x \in \mathbb{R}^+$, $\bar{\alpha}_T(x) > 0$, $\bar{c}_T$ is invertible, and the following difference for any $(t,z) \in [0,T) \times (0,1]$,

$$\Delta(t,x,z) := \int_x^1 \frac{dy}{y\bar{\alpha}(t,y,z)} - \int_{\bar{c}_T^{-1}(\bar{c}(t,x,z))}^1 \frac{dy}{y\bar{\alpha}_T(y)}$$

depends only on (t,z), namely $\Delta(t,x,z) \equiv \Delta(t,z)$.

Then both the discounting scheme characterized by $\dot{\beta}$ and the utility functions $U_i \in \mathcal{U}$ for $i = 1,2$, with $U_1(0) = 0$ and $U_2(0) = -\infty$, are identifiable up to an affine transform.

Proof First, by Assumption 1, for all $(t,x,z) \in \mathcal{D}$, (HJB) is equivalent to

$$\begin{cases} \partial_t V(t,x,z) + \dot{\beta}_t \partial_z V(t,x,z) + rx\partial_x V(t,x,z) \\ \quad - \dfrac{(r-\mu)^2}{2\sigma^2} \dfrac{[\partial_x V(t,x,z)]^2}{\partial_x^2 V(t,x,z)} = zU_1^* \left(\dfrac{\partial_x V(t,x,z)}{z} \right), \quad t \in (0,T); \\ V(T,x,z) = zU_2(x), \end{cases} \tag{16}$$

where U_1^* is the Legendre transform of the concave utility function $U_1 : [0,\infty) \to \mathbb{R}$,

$$U_1^*(\kappa) := \inf_{c\in[0,M]} \{\kappa c - U_1(c)\}, \quad \forall \kappa \in \mathbb{R}.$$

Now, we construct the value function $\bar{V}$ in (15) from (16). By (12) and Assumption 1,

$$\bar{\alpha}_T(x) = -\frac{U_2'(x)}{U_2''(x)}\frac{\mu - r}{\sigma^2 x} \implies U_2(x) = k_1 \int_1^x \exp\left\{\int_y^1 \frac{\mu - r}{\sigma^2 u \bar{\alpha}_T(u)} du\right\} dy + k_2,$$

for some $k_1, k_2 > 0$; in particular,

$$U_2'(x) = k_1 \exp\left\{\int_x^1 \frac{\mu - r}{y\bar{\alpha}_T(y)} dy\right\}.$$

By (13), for any $x > 0$,

$$U_1'(\bar{c}_T(x)) = U_2'(x) \implies U_1'(x) = U_2'(\bar{c}_T^{-1}(x)) = k_1 \exp\left\{\int_{\bar{c}_T^{-1}(x)}^1 \frac{\mu - r}{\sigma^2 y \bar{\alpha}_T(y)} dy\right\}$$

and

$$U_1(x) = k_1 \int_1^x \exp\left\{\int_{\bar{c}_T^{-1}(y)}^1 \frac{\mu - r}{\sigma^2 u \bar{\alpha}_T(u)} du\right\} dy + k_3$$

for some $k_3 > 0$.

For any $(t, z) \in [0, T) \times (0, 1]$ and $x > 0$, by (12) and (13),

$$\begin{cases} \bar{\alpha}(t,x,z) = -\frac{\partial_x \bar{V}(t,x,z)}{\partial_x[\partial_x \bar{V}(t,x,z)]}\frac{\mu - r}{\sigma^2 x}, \\ \partial_x \bar{V}(t,x,z) = zU_1'(\bar{c}(t,x,z)); \end{cases} \implies \partial_x \bar{V}(t,x,z) = K_1(t,z) \exp\left\{\int_x^1 \frac{\mu - r}{\sigma^2 y \bar{\alpha}(t,y,z)} dy\right\},$$

where

$$K_1(t,z) = k_1 z \exp\{-\frac{\mu - r}{\sigma^2}\Delta(t,z)\}.$$

Rewrite (16) as

$$\begin{aligned} \partial_t \bar{V}(t,x,z) + \dot{\beta}_t \partial_z \bar{V}(t,x,z) = &-\left\{\left[r + \frac{\bar{\alpha}(t,x,z)(\mu - r)}{2}\right] x - \bar{c}(t,x,z)\right\} \\ &\partial_x \bar{V}(t,x,z) - zU_1(\bar{c}(t,x,z)). \end{aligned}$$

Differentiating with respect to x on both sides, we have

$$\dot{\beta}_t = -\frac{\partial_t[\partial_x \bar{V}(t,x,z)] + \partial_x\left\{\left\{\left[r + \frac{\bar{\alpha}(t,x,z)(\mu-r)}{2}\right]x - \bar{c}(t,x,z)\right\}\partial_x \bar{V}(t,x,z) + zU_1(\bar{c}(t,x,z))\right\}}{\partial_z[\partial_x \bar{V}(t,x,z)]}.$$

□

Remark 4 1. This result is consistent with the finding in [15] that the identifiability of the unknown utility function in an inverse optimal control problem is equivalent to the identifiability of the corresponding value function under the observed optimal policy.

2. In the literature that studies the identifiability in inverse optimal control/reinforcement learning problems, to resolve the degeneracy or the nonidentifiability issue, there are primarily two categories of attempts:

(a) collecting observed policies under the same utilities by varying stochastic environments [4, 15]; or

(b) adding technical assumptions on the family of candidate utilities and/or the stochastic environment [15, 34, 50]. Theorem 1 aligns with the second category: apart from the particular problem formulation that optimal consumption-allocation problem usually takes and the assumptions on the family utilities such as (5), assumptions specified in Theorem 1 serve as additional conditions for the stochastic environment which need to be verified through the observed policy $(\bar{\alpha}, \bar{c})$.

3. More specifically, Assumptions 1–5 allow us to make full use of PDE characterization of the value function (16) through the observation $(\bar{\alpha}, \bar{c})$, which subsequentially leads to the identifiability of the utilities.

We conclude the analysis on finite-time horizon by discussing a special case with an explicit solution.

Example 1 Set $\beta_0 = 1$, $\dot{\beta}_t = 0$ $(0 \le t \le T)$, $U_1(c) = 0$ and a CRRA (power) utility $U_2(x) = \frac{x^\theta}{\theta}$ with $0 < \theta < 1$. Also set the constant M such that $M > \frac{\mu - r}{\sigma^2}$. In this case, we face a classic control problem with an exponential discounting function. Hence state augmentation is not necessary. In addition, both the optimal control and the inverse problem have explicit representations. The goal here is to identify the parameter θ from the client.

Consequently, define the value function:

$$V(t,x) = \sup_{\boldsymbol{\alpha} \in \mathcal{A}} \mathbb{E}\Big[U_2(X_T) \,\Big|\, X_t = x\Big]. \tag{17}$$

The value function satisfies the following HJB equation:

$$-\partial_t V - \sup_{\boldsymbol{\alpha} \in \mathcal{A}} \Big[\mathcal{L}^{\boldsymbol{\alpha}} V(t,x)\Big] = 0, \tag{18}$$

with boundary condition $V(T,x) = U_2(x) = \frac{x^\theta}{\theta}$, where the generator is defined as $\mathcal{L}^\alpha V(t,x) = x\Big(\alpha\mu + (1-\alpha)r\Big)\partial_x V + \frac{1}{2}x^2\alpha^2\sigma^2\partial_x^2 V$. The optimal policy follows:

$$\bar{\alpha}(t,x) = -M \vee \left(-\frac{(\mu - r)\partial_x V}{x\sigma^2\partial_x^2 V}\right) \wedge M. \tag{19}$$

For now assume that $-M \leq \left(-\frac{(\mu-r)\partial_x V}{x\sigma^2\partial_x^2 V}\right) \leq M$ (to be checked later). Then plugging it back into the HJB equation, we have

$$-\frac{\partial V}{\partial t} = xr\partial_x V - \frac{1}{2}\frac{(\mu - r)(\partial_x V)^2}{\sigma^2\partial_x^2 V} \tag{20}$$

with boundary condition $V(T,x) = \frac{x^\theta}{\theta}$. We take the ansatz $V(t,x) = \phi(t)\frac{x^\theta}{\theta}$. Hence $\phi(t)$ satisfies:

$$\phi'(t) = \rho\phi(t), \quad \phi(T) = 1, \tag{21}$$

where $\rho = \theta \times \sup_{\alpha\in[-M,M]}[\alpha(\mu - r) + r - \frac{1}{2}a^2(1-\theta)\sigma^2]$. Hence $\bar{\alpha} = \frac{\mu - r}{\sigma^2(1-\theta)} \in [-M, M]$. In this case, it is obvious that condition $-M \leq \left(-\frac{(\mu-r)\partial_x V}{x\sigma^2\partial_x^2 V}\right) \leq M$ is satisfied. Therefore we can recover the preference parameter by using $1 - \frac{\mu - r}{\bar{\alpha}\sigma^2}$.

2.2 Infinite-Time Horizon

Now we shift our focus to an infinite-time horizon setting that accommodates a long-run investment planning scenario.

Recall that the investing client is holding a general discounting scheme $\beta = \{\beta_t\}_{t\geq 0}$ where

- $\beta_t \in [0,1]$ for all $t \in [0,\infty)$ such that $\lim_{t\to\infty}\beta_t = 0$; and
- there exists $\dot{\beta} : [0,\infty) \to \mathbb{R}$ such that $\dot{\beta}$ is integrable on $[0,t]$ with $\beta_t = \int_0^t \dot{\beta}_s ds + \beta_0$ for any $t > 0$.

The infinite-horizon counterpart of the optimal joint consumption-allocation problem is

$$\sup_{\boldsymbol{\alpha},\boldsymbol{c}} \mathbb{E}\left[\int_0^\infty \beta_t U_1(c_t)dt\right].$$

For any $(t,x,z) \in \mathcal{D}' := [0,\infty) \times \mathbb{R}^+ \times [0,1]$, with $\mathcal{D}'^\circ := \mathbb{R}^+ \times \mathbb{R}^+ \times (0,1)$, define the total reward function as

$$J_\infty(t,x,z,\boldsymbol{\alpha},\boldsymbol{c}) := \mathbb{E}\left[\int_t^\infty \beta_s U_1(c_s)ds \middle| X_t = x, \beta_t = z\right] \tag{22}$$

subject to (4) and (7), under a given allocation process $\boldsymbol{\alpha} = \{\alpha_t\}_{t\geq 0}$ and a given consumption process $\boldsymbol{c} = \{c_t\}_{t\geq 0}$ with $c_t \geq 0$. Note that this infinite-horizon problem is time-inhomogeneous, primarily due to the inhomogeneity of β, and has been less studied in the literature.

For any $(t, x, z) \in \mathcal{D}'$, define the value function as follows,

$$V_\infty(t,x,z) = \sup_{(\boldsymbol{\alpha},\boldsymbol{c})\in\mathcal{A}} J_\infty(t,x,z,\boldsymbol{\alpha},\boldsymbol{c}),\ t\in[0,t);\quad \lim_{t\to\infty} V_\infty(t,x,z) = 0, \tag{23}$$

subject to (4) and (7).

It is easy to show that the value function V_∞ in (23) will have similar results as specified in Sect. 2.1 therefore here we state these results without proofs. First, we have the necessary condition for V_∞.

Proposition 4 For any $(t, x, z) \in \mathcal{D}'$ and $\tau \in \bar{\mathbb{T}}_t$ where $\bar{\mathbb{T}}_t$ denotes all $\{\bar{\mathcal{F}}_t\}_{t\geq 0}$ -adapted stopping times τ such that $\tau \in [t, \infty)$ a.s. Then the value function V_∞ defined in (23) satisfies

$$\begin{aligned} V_\infty(t,x,z) = \sup_{(\boldsymbol{\alpha},\boldsymbol{c})\in\mathcal{A}} \mathbb{E}\Bigg[&\int_t^\tau \beta_s U_1(c_s)ds \\ &+ V_\infty(\tau, X_\tau^{\boldsymbol{\alpha},\boldsymbol{c}}, \beta_\tau)\,\Big|\, X_t^{\boldsymbol{\alpha},\boldsymbol{c}} = x, \beta_t = z\Bigg]. \end{aligned} \tag{DPP'}$$

The corresponding HJB equation is given by

$$\begin{cases} \partial_t V_\infty(t,x,z) + \dot{\beta}_t \partial_z V_\infty(t,x,z) + rx\partial_x V_\infty(t,x,z) \\ \qquad + \sup_{\alpha\in\mathbb{R}, c\geq 0} \left\{ zU_1(c) + \mathcal{L}^{\alpha,c} V_\infty(t,x,z)\right\} = 0,\ t\in(0,\infty) \\ \lim_{t\to\infty} V_\infty(t,x,z) = 0, \quad \forall (x,z)\in\mathbb{R}^+\times[0,1]. \end{cases} \tag{HJB'}$$

Likewise, combining Proposition 4 and Itô's formula, we have the following result.

Proposition 5 If the value function V_∞ in (23) is jointly continuous on $\mathcal{D}'$, then it is a viscosity solution to (HJB') over the domain $\mathcal{D}'$.

With a classical solution to (HJB'), we have the following verification theorem serving as sufficient conditions for V_∞.

Proposition 6 Suppose that $U_1 : [0, \infty) \to \mathbb{R}^+ \in \mathcal{U}$ is continuous at 0. Let $w : \mathcal{D}' \to \mathbb{R}$ be a function such that $w \in \mathcal{C}^{1,2,1}(\mathcal{D}'^\circ) \bigcap \mathcal{C}^0(\mathcal{D}')$, and there exists a constant $C > 0$ with

$$w(t,x,z) \le C(1+|x|^2), \quad \forall (t,x,z) \in \mathcal{D}'.$$

1. Assume that for any $(\alpha, c) \in \mathcal{K}$,

$$\begin{cases} \partial_t w(t,x,z) + \dot{\beta}_t \partial_z w(t,x,z) + rx\partial_x w(t,x,z) + zU_1(c) + \mathcal{L}^{\alpha,c} w(t,x,z) \le 0, \\ \quad \forall (t,x,z) \in (0,\infty) \times \mathbb{R}^+ \times (0,1); \\ \lim\limits_{t\to\infty} w(t,x,z) = \infty, \quad \forall (x,z) \in \mathbb{R}^+ \times [0,1]. \end{cases}$$

   Then $w \ge V_\infty$ on $\mathcal{D}'$.
2. Assume further that there exists $\hat{\alpha} : \mathcal{D}' \to [-M, M]$ and $\hat{c} : \mathcal{D}' \to [0, M]$ such that

$$\begin{cases} \partial_t w(t,x,z) + \dot{\beta}_t \partial_z w(t,x,z) + rx\partial_x w(t,x,z) + zU_1(\hat{c}(t,x,z)) \\ \quad + \mathcal{L}^{\hat{\alpha}(t,x,z),\hat{c}(t,x,z)} w(t,x,z) = 0, \forall (t,x,z) \in \mathcal{D}'^{\circ}; \\ \lim\limits_{t\to\infty} w(t,x,z) = \infty, \quad \forall (x,z) \in \mathbb{R}^+ \times [0,1], \end{cases}$$

   also, with $\beta_t = \beta_0 + \int_0^t \dot{\beta}_s ds \in [0,1]$ for all $t \ge 0$, the following SDE,

$$dX_t = \{X_t\,[\hat{\alpha}(t, X_t, \beta_t)(\mu - r) + r] - \hat{c}(t, X_t, \beta_t)\}\, dt + \sigma \hat{\alpha}(t, X_t, \beta_t) X_t dW_t,$$

   admits a unique solution $X^{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{c}}}$ given $X_0 = x$ for any $x \in \mathbb{R}$, and

$$\left(\hat{\boldsymbol{\alpha}} = \{\hat{\alpha}_t\}_{t\ge 0} = \left\{\hat{\alpha}(t, X_t^{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{c}}}, \beta_t)\right\}_{t\ge 0}, \hat{\boldsymbol{c}} = \{\hat{c}_t\}_{t\ge 0} = \left\{\hat{c}(t, X_t^{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{c}}}, \beta_t)\right\}_{t\ge 0}\right) \in \mathcal{A}.$$

   Then $w = V_\infty$ on $\mathcal{D}'$, with $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{c}})$ being an optimal joint allocation-consumption process.

Given that $V_\infty \in \mathcal{C}^{1,2,1}(\mathcal{D}')$ being a classical solution to (HJB'), then the optimal policy is given by

$$\bar{\alpha}^*(t,x,z) = -M \vee -\frac{\partial_x V_\infty(t,x,z)}{\partial_x^2 V_\infty(t,x,z)} \cdot \frac{(\mu - r)}{\sigma^2 x} \wedge M, \tag{24}$$

$$\bar{c}^*(t,x,z) = \operatorname{argmax}_{c\in[0,M]} \Big\{ -\partial_x V_\infty(t,x,z)c + zU_1(c) \Big\}. \tag{25}$$

Accordingly, for the inverse problem, we also have the following identifiability result.

Theorem 2 Assume that

1. $\bar{\alpha}, \bar{c} \in \mathcal{C}^{1,1,1}(\mathcal{D}'^{\circ}) \bigcap \mathcal{C}^0(\mathcal{D}')$;
2. $\bar{\alpha}(t,x,z) \in (-M, M)$ for all $(t,x,z) \in \mathcal{D}'$;
3. for any $(t,z) \in [0,\infty) \times [0,1]$,

$$\bar{c}(t,x,z) < x, \quad \forall x > 0;$$

4. $\exists (t_0, z_0) \in [0,\infty) \times (0,1]$ such that $\bar{c}_0(\cdot) := \bar{c}(t_0, \cdot, z_0)$ is invertible, and the following difference for any $(t,x,z) \in [0,T) \times \mathbb{R}^+ \times (0,1]$,

$$\Delta(t,x,z) := \int_x^1 \frac{dy}{y\bar{\alpha}(t,y,z)} - \int_{\bar{c}_0^{-1}(\bar{c}(t,x,z))}^1 \frac{dy}{y\bar{\alpha}_T(y)}$$

depends only on (t,z), namely, $\Delta(t,x,z) \equiv \Delta(t,z)$.

Then both the discounting scheme characterized by $\dot{\beta}$ and the utility function $U_1 \in \mathcal{U}$ with $U_1(0) = 0$ are identifiable up to an affine transform.

3 Learning and Inference with Entropy Regularization

The continuous-time framework in Sect. 2 establishes the mathematical foundation for modeling client behavior under a general time-varying discounting scheme and provides conditions for identifying both the utility functions and the discounting structure. While this offers valuable theoretical insight, direct application in real-world settings can be challenging due to the complexities of continuous-time inference and the discrete nature of observations in practice. To bridge this gap, we turn to a discrete-time MDP with Shannon entropy regularization over a finite-time horizon—a formulation frequently adopted in RL. The discrete-time framework not only enhances computational tractability but also aligns more naturally with practical implementation, where data is typically sampled at discrete intervals and actions are executed in sequential steps. This section therefore focuses on a more implementable inference procedure, where the agent aims to recover the client's discounting scheme and utility function from observed policy-driven behavior.

We adopt a parametric framework in which the client utilizes an exponential discounting scheme, parameterized by $\bar{\rho}$, alongside a utility function parameterized by $\bar{\theta} \in \mathbb{R}^d$, which is assumed to be twice differentiable with respect to $\bar{\theta}$ with bounded first and second-order derivatives. The client's preference parameter is summarized as $(\bar{\rho}, \bar{\theta})$, which is unknown to the inference agent. The inference agent employs a maximum likelihood estimation method to infer the parameters $(\bar{\rho}, \bar{\theta})$.

Mathematically, let us consider the entropy-regularized MDP with state space $\mathcal{S}$ and bounded action space $\mathcal{A}$. Let us consider a finite-time horizon problem, with T being the total number of time steps. For any time $t \in \{0, 1, 2, \ldots T\}$, conditioning on the current state and action being $(s_t, a_t) = (s, a) \in \mathcal{S} \times \mathcal{A}$, the next state s_{t+1} follows $s_{t+1} \sim P(\cdot|s,a)$, with $P : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{A})$ being the transition kernel that maps from the joint state-action space to the distribution over the state space. After taking action a at state s, we assume the client receives a deterministic reward

$R(s,a) \in [0,1]$. Throughout the remainder of this paper, we use the notation $\sum\!\!\!\!\!\!\int$ to denote summation over the action space when it is finite, and integration over the action space with respect to the Lebesgue measure when the action space is a bounded subspace of $\mathbb{R}^k$ for some $k \in \mathbb{N}_+$, emphasizing that our framework accommodates both finite action space and (possibly) continuous action space.

Under a generic preference parameter (ρ, θ), consider the entropy-regularized objective

$$\max_{\pi=\{\pi^t\}_{t=0}^T} \mathbb{E}^\pi \left[\sum_{t=0}^T \rho^t \left(U_\theta(R(s_t,a_t)) + \mathcal{H}(\pi^t(\cdot|s_t)) \right) \middle| s_0 = s \right],$$

where $\mathbb{E}^\pi$ denotes the expectation of trajectories under policy $\pi = \{\pi^t\}_{t=0}^T$ with the initial state s_0 following the distribution $\mu(\cdot)$, $\mathcal{H}\big(\pi^t(\cdot|s)\big) := -\sum\!\!\!\!\!\!\int \pi^t(a|s) \log\big(\pi^t(a|s)\big)$ is the Shannon's entropy, and U_θ is the utility function parameterized by θ. Define

$$\begin{aligned} Q^T_{\rho,\theta}(s,a) &:= U_\theta(R(s,a)), \\ Q^t_{\rho,\theta}(s,a) &:= U_\theta(R(s,a)) + \max_{\pi=\{\pi^k\}_{k=t+1}^T} \mathbb{E}^\pi \\ &\qquad \left[\sum_{k=t+1}^T \rho^{k-t} \left(U_\theta(R(s_k,a_k) + \mathcal{H}(\pi^k(\cdot|s_k))) \right) \middle| s_t = s, a_t = a \right], \\ &\qquad t = 0, \ldots, T-1. \end{aligned} \tag{26}$$

Correspondingly, define

$$\begin{aligned} V^t_{\rho,\theta}(s) := \max_{\pi=\{\pi^k\}_{k=t}^T} \mathbb{E}^\pi \left[\sum_{k=t}^T \rho^{k-t} \left(U_\theta(R(s_k,a_k) + \mathcal{H}(\pi^k(\cdot|s_k))) \right) \middle| s_t = s \right], \\ t = 0, \ldots, T. \end{aligned} \tag{27}$$

Let $\pi_{\rho,\theta} := \{\pi^t_{\rho,\theta}\}_{t=0}^T$ denote the optimal policy, with $\pi^t_{\rho,\theta}$ given by

$$\pi^t_{\rho,\theta}(a|s) = \frac{e^{Q^t_{\rho,\theta}(s,a)}}{\sum\!\!\!\!\!\!\int_{a' \in \mathcal{A}} e^{Q^t_{\rho,\theta}(s,a')}}, \tag{28}$$

and the soft Bellman equation holds: for any $t \in \{0,1,2,\ldots,T\}$, we have

$$\begin{aligned} &Q^t_{\rho,\theta}(s,a) = U_\theta(R(s,a)) + \rho \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[V^{t+1}_{\rho,\theta}(s') \right], \text{ with} \\ &V^t_{\rho,\theta}(s) = \log\Big(\sum\!\!\!\!\!\!\int_{a \in \mathcal{A}} e^{Q^t_{\rho,\theta}(s,a)} \Big). \end{aligned} \tag{29}$$

We note that (28) and (29) are well-defined when $\mathcal{A}$ is finite or bounded.

With an entropy-regularized objective, the optimal decision balances the trade-off between maximizing the cumulative expected reward (exploitation) and maximizing the entropy (exploration), resulting in a randomized policy. The soft Bellman equation establishes the relationship between the optimal value function and the optimal Q-function (under entropy regularization). Further details can be found in [36, 49].

3.1 Maximum Likelihood Estimation

With a trajectory $\xi = \{(s_t, a_t)\}_{t=0}^T$ following the trajectory distribution under the client's policy $\pi_{\bar{\rho},\bar{\theta}}$ with the initial state s_0 following the distribution $\mu(\cdot)$, we adopt a *maximum likelihood estimation method* to infer the client's preference parameter $(\bar{\rho}, \bar{\theta})$, which is unknown to the inference agent. Specifically, the discounted likelihood of a trajectory $\xi = \{(s_t, a_t)\}_{t=0}^T$ following the client's policy $\pi_{\bar{\rho},\bar{\theta}} = \{\pi^t_{\bar{\rho},\bar{\theta}}\}_{t=0}^T$ is defined as

$$\begin{aligned}
&\mathbb{E}_{\xi\sim\pi_{\bar{\rho},\bar{\theta}}}\left[\log\left(\prod_{t=0}^{T}(P(s_{t+1}|s_t,a_t)\pi^t_{\rho,\theta}(a_t|s_t))^{\gamma^t}\right)\right] \\
&\quad= \mathbb{E}_{\xi\sim\pi_{\bar{\rho},\bar{\theta}}}\left[\sum_{t=0}^{T}\gamma^t\log\pi^t_{\rho,\theta}(a_t|s_t)\right] \\
&\quad+ \mathbb{E}_{\xi\sim\pi_{\bar{\rho},\bar{\theta}}}\left[\sum_{t=0}^{T}\gamma^t\log P(s_{t+1}|s_t,a_t)\right],
\end{aligned} \tag{30}$$

where the notation $\mathbb{E}_{\xi\sim\pi_{\bar{\rho},\bar{\theta}}}$ represents the expectation over the trajectory ξ following policy $\pi_{\bar{\rho},\bar{\theta}}$ and γ is a discount factor specified by the inference agent, which is potentially *different* from $\bar{\rho}$.

Remark 5 Note that in our case $\gamma \neq \bar{\rho}$ because $\bar{\rho}$ is the client's discount factor and is unknown to the inference agent. This distinguishes us from the usual IRL literature, where $\bar{\rho}$ is always assumed to be known [12]. For example, Zeng et al. [83] studied the IRL problem using a maximum likelihood estimator by setting $\gamma = \bar{\rho}$ in (30) and showed that their algorithm converges to a stationary point with a finite-time guarantee. Note that this stationary point may not be the ground-truth solution.

The maximum likelihood inference problem can be written as:

$$\max_{(\rho,\theta)\in\Theta} \quad \mathcal{L}(\rho,\theta) := \mathbb{E}_{\xi\sim\pi_{\bar{\rho},\bar{\theta}}}\left[\sum_{t=0}^{T}\gamma^t\log\pi^t_{\rho,\theta}(a_t|s_t)\right], \tag{31}$$

where $\pi_{\rho,\theta} = \{\pi^t_{\rho,\theta}\}_{t=0}^T$ is the optimal policy under the preference parameter (ρ,θ) defined in (28). Here for simplicity we set $\Theta := (0,1)\times\mathbb{R}^d$. The maximum likeli-

hood problem is to find a preference parameter $(\bar{\rho}, \bar{\theta})$ that generates the client's trajectory with the highest likelihood.

The goal is to investigate the landscape of the log-likelihood function $\mathcal{L}(\rho, \theta)$ with respect to (ρ, θ) and understand the possibility of recovering $(\bar{\rho}, \bar{\theta})$. To proceed, we first show that $(\bar{\rho}, \bar{\theta})$ is a stationary point of the likelihood function (see Proposition 7) and then show that the Hessian matrix of the likelihood function is negative semi-definite at $(\bar{\rho}, \bar{\theta})$ (see Theorem 3). Interestingly, the results of the landscape analysis is *independent of* the choice of γ, making our proposed method robust and practical.

Proposition 7 It holds that

$$\nabla_\theta \mathcal{L}(\rho, \theta)\Big|_{(\rho,\theta)=(\bar{\rho},\bar{\theta})} = 0, \quad \nabla_\rho \mathcal{L}(\rho, \theta)\Big|_{(\rho,\theta)=(\bar{\rho},\bar{\theta})} = 0. \tag{32}$$

Proposition 7 suggests that the gradient of the likelihood function equals zero at the client's preference parameter value $(\bar{\rho}, \bar{\theta})$, and hence $(\bar{\rho}, \bar{\theta})$ is a stationary point of the likelihood function $\mathcal{L}(\rho, \theta)$.

Proof Our proof can be divided into three steps. We first prove the differentiability of $Q^t_{\rho,\theta}$ and $V^t_{\rho,\theta}$ in ρ and θ by induction, and provide some useful formulas regarding their first-order and second-order derivatives with respect to (ρ, θ). With such formulas, we next derive the derivatives of the log-likelihood function. Finally, we show that (32) holds.

Step 1 To begin with, for any $(\rho, \theta) \in \Theta$, for any $t \in \{0, 1, 2, \dots T-1\}$, and for any $(s_t, a_t) \in \mathcal{S} \times \mathcal{A}$, if $Q^{t+1}_{\rho,\theta}$ is differentiable with respect to θ, then $V^{t+1}_{\rho,\theta}$ is differentiable with respect to θ by (29). Then by the soft Bellman equation we have that $Q^t_{\rho,\theta}$ is differentiable with respect to θ. More specifically, we have

$$\begin{aligned}
\nabla_\theta Q^t_{\rho,\theta}(s,a) &= \nabla_\theta U_\theta(R(s,a)) + \rho \mathbb{E}_{\xi' \sim \pi_{\rho,\theta}}\Big[\nabla_\theta V^{t+1}_{\rho,\theta}(s'_{t+1})\Big| s'_t = s, a'_t = a\Big] \\
&= \nabla_\theta U_\theta(R(s,a)) + \rho \mathbb{E}_{\xi' \sim \pi_{\rho,\theta}}\Big[\nabla_\theta \log\Big(\sum_{a'} e^{Q^{t+1}_{\rho,\theta}(s'_{t+1},a')}\Big)\Big| s'_t = s, a'_t = a\Big] \qquad (33) \\
&= \nabla_\theta U_\theta(R(s,a)) + \rho \mathbb{E}_{\xi' \sim \pi_{\rho,\theta}}\Big[\sum_{a'} \pi^{t+1}_{\rho,\theta}(a'|s'_{t+1}) \nabla_\theta Q^{t+1}_{\rho,\theta}(s'_{t+1}, a')\Big| s'_t = s, a'_t = a\Big] \qquad (34) \\
&= \nabla_\theta U_\theta(R(s,a)) + \rho \mathbb{E}_{\xi' \sim \pi_{\rho,\theta}}\Big[\mathbb{E}_{\xi' \sim \pi_{\rho,\theta}}\Big[\nabla_\theta Q^{t+1}_{\rho,\theta}(s'_{t+1}, a'_{t+1})\Big| s'_{t+1}\Big]\Big| s'_t = s, a'_t = a\Big], \qquad (35)
\end{aligned}$$

where $\xi' = \{(s'_t, a'_t)\}_{t=0}^T$ follows the trajectory distribution under the optimal policy $\pi_{\rho,\theta} = \{\pi^t_{\rho,\theta}\}_{t=0}^T$, (33) holds by the soft Bellman equation and (34) holds due to the optimality of $\pi_{\rho,\theta}$. Moreover, note that $Q^T_{\rho,\theta}(s,a) = U_\theta\big(R(s,a)\big)$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, which is differentiable by the assumption on the utility function. Therefore, the differentiability of $Q^t_{\rho,\theta}$ and $V^t_{\rho,\theta}$ in θ is justified by induction for all $t \in \{0, 1, 2, \dots T\}$. Applying (35) recursively yields:

$$\nabla_\theta Q^t_{\rho,\theta}(s,a) = \mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\sum_{k=t}^{T}\rho^{k-t}\nabla_\theta U_\theta(R(s'_k,a'_k))\Big|s'_t=s,a'_t=a\right]. \quad (36)$$

Similarly, if $Q^{t+1}_{\rho,\theta}$ is differentiable with respect to ρ, then $V^{t+1}_{\rho,\theta}$ is differentiable with respect to ρ by its definition in (29). Then we also have that $Q^t_{\rho,\theta}$ is differentiable with respect to ρ. More specifically,

$$\begin{aligned}
\nabla_\rho Q^t_{\rho,\theta}(s,a) &= \nabla_\rho U_\theta(R(s,a)) + \nabla_\rho(\rho\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[V^{t+1}_{\rho,\theta}(s'_{t+1})\right]) && (37)\\
&= \mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[V^{t+1}_{\rho,\theta}(s'_{t+1})\right] + \rho\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\nabla_\rho\log(\sum_{a'} e^{Q^{t+1}_{\rho,\theta}(s'_{t+1},a')})\right] && (38)\\
&= \mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[V^{t+1}_{\rho,\theta}(s'_{t+1})\right] + \rho\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\sum_{a'}\pi^{t+1}_{\rho,\theta}(a'|s'_{t+1})\nabla_\rho Q^{t+1}_{\rho,\theta}(s'_{t+1},a')\right] && (39)\\
&= \mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[V^{t+1}_{\rho,\theta}(s'_{t+1})\right] + \rho\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\nabla_\rho Q^{t+1}_{\rho,\theta}(s'_{t+1},a'_{t+1})\Big|s'_{t+1}\right]\Big|s'_t=s,a'_t=a\right], && (40)
\end{aligned}$$

where (37) holds by the Bellman equation and (39) holds because $\pi_{\rho,\theta}$ is the optimal policy. Since $Q^T_{\rho,\theta}(s,a) = U_\theta\big(R(s,a)\big)$, the differentiability of $Q^t_{\rho,\theta}$ and $V^t_{\rho,\theta}$ in ρ is justified by induction for all $t\in\{0,1,2,\dots T\}$. Applying (40) recursively yields:

$$\nabla_\rho Q^t_{\rho,\theta}(s,a) = \mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\sum_{k=t+1}^{T}\rho^{k-t-1}V^k_{\rho,\theta}(s'_k)\Big|s'_t=a,a'_t=a\right]. \quad (41)$$

Furthermore, we have for any $s\in\mathcal{S}$,

$$\begin{aligned}
&\nabla_\theta V^t_{\rho,\theta}(s) = \nabla_\theta(\log\sum_a e^{Q^t_{\rho,\theta}(s,a)})\\
&= \mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t)\Big|s'_t=s\right] = \mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\sum_{k=t}^{T}\rho^{k-t}\nabla_\theta U_\theta(R(s'_k,a'_k))\Big|s'_t=s\Big],
\end{aligned} \quad (42)$$

where the last equation holds by (36). In addition,

$$\begin{aligned}
&\nabla_\rho V^t_{\rho,\theta}(s) = \nabla_\rho(\log\sum_a e^{Q^t_{\rho,\theta}(s,a)})\\
&= \mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\nabla_\rho Q^t_{\rho,\theta}(s'_t,a'_t)\Big|s'_t=s\right] = \mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\sum_{k=t+1}^{T}\rho^{k-t-1}V^k_{\rho,\theta}(s'_k)\Big|s'_t=s\Big],
\end{aligned} \quad (43)$$

where the last equation holds by (41).

In summary, it holds that for any $t\in\{0,1,2\dots T\}$, $(\rho,\theta)\in\Theta$, $s_t\in\mathcal{S}$, $a_t\in\mathcal{A}$,

$$\nabla_\theta Q^t_{\rho,\theta}(s,a) = E_{\xi'\sim\pi_{\rho,\theta}}\left[\sum_{k=t}^{T}\rho^{k-t}\nabla_\theta U_\theta(R(s'_k,a'_k))\Big|s'_t=s,a'_t=a\right], \quad (44)$$

$$\nabla_\rho Q^t_{\rho,\theta}(s,a) = E_{\xi'\sim\pi_{\rho,\theta}}\left[\sum_{k=t+1}^{T}\rho^{k-t-1}V^k_{\rho,\theta}(s'_k)\Big|s'_t=s,a'_t=a\right], \tag{45}$$

$$\nabla_\theta V^t_{\rho,\theta}(s) = \mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\sum_{k=t}^{T}\rho^{k-t}\nabla_\theta U_\theta(R(s'_k,a'_k))\Big|s'_t=s\right], \tag{46}$$

$$\nabla_\rho V^t_{\rho,\theta}(s) = \mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\sum_{k=t+1}^{T}\rho^{k-t-1}V^k_{\rho,\theta}(s'_k)\Big|s'_t=s\right]. \tag{47}$$

Step 2 Next, we derive the gradients of the log-likelihood function. For any $(\rho,\theta)\in\Theta$,

$$\mathcal{L}(\rho,\theta) = \mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{T}\gamma^t\log\pi^t_{\rho,\theta}(a_t|s_t)\Big] = \mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{T}\gamma^t\log\frac{e^{Q^t_{\rho,\theta}(s_t,a_t)}}{\int_a e^{Q^t_{\rho,\theta}(s_t,a)}}\Big] \tag{48}$$

$$= \mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{T}\gamma^t\Big(Q^t_{\rho,\theta}(s_t,a_t)-V^t_{\rho,\theta}(s_t)\Big)\Big]$$

$$= \mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{T}\gamma^t U_\theta(R(s_t,a_t))\Big] - \mathbb{E}_{s_0\sim\mu(\cdot)}\Big[V^0_{\rho,\theta}(s_0)\Big] + (\rho-\gamma)\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{T}\gamma^{t-1}V^t_{\rho,\theta}(s_t)\Big], \tag{49}$$

where μ is the distribution of the initial state s_0. (48) holds by the optimality of the policy, and (49) holds by the soft Bellman equation. Taking the gradient of (49) with respect to θ gives

$$\begin{aligned}\nabla_\theta\mathcal{L}(\rho,\theta) = {}& \mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=0}^{T}\gamma^t\nabla_\theta U_\theta(R(s_t,a_t))\right]\\ & - \mathbb{E}_{s_0\sim\mu(\cdot)}\left[\nabla_\theta V^0_{\rho,\theta}(s_0)\right]\\ & + (\rho-\gamma)\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=1}^{T}\gamma^{t-1}\nabla_\theta V^t_{\rho,\theta}(s_t)\right].\end{aligned} \tag{50}$$

Combining (50) with (46) gives:

$$\begin{aligned}\nabla_\theta\mathcal{L}(\rho,\theta) = {}& \mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=0}^{T}\gamma^t\nabla_\theta U_\theta(R(s_t,a_t))\right]\\ & - \mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\sum_{t=0}^{T}\rho^t\nabla_\theta U_\theta(R(s'_t,a'_t))\right] + (\rho-\gamma)\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\\ & \left[\sum_{t=1}^{T}\gamma^{t-1}\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\sum_{k=t}^{T}\rho^{k-t}\nabla_\theta U_\theta(R(s'_k,a'_k))\Big|s'_t=s_t\right]\right].\end{aligned} \tag{51}$$

Similarly, taking the gradient of (49) with respect to ρ gives

$$\begin{aligned}\nabla_\rho \mathcal{L}(\rho,\theta) = &-\mathbb{E}_{s_0\sim\mu(\cdot)}\left[\nabla_\rho V^0_{\rho,\theta}(s_0)\right] + \mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=1}^{T}\gamma^{t-1}V^t_{\rho,\theta}(s_t)\right]\\ &+(\rho-\gamma)\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=1}^{T}\gamma^{t-1}\nabla_\rho V^t_{\rho,\theta}(s_t)\right].\end{aligned} \tag{52}$$

Combining (52) with (47) yields:

$$\begin{aligned}\nabla_\rho \mathcal{L}(\rho,\theta) = &\,\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=1}^{T}\gamma^{t-1}V^t_{\rho,\theta}(s_t)\right] - \mathbb{E}_{\xi\sim\pi_{\rho,\theta}}\left[\sum_{t=1}^{T}\rho^{t-1}V^t_{\rho,\theta}(s_t)\right]\\ &+(\rho-\gamma)\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=1}^{T}\gamma^{t-1}\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\sum_{k=t+1}^{T}\rho^{k-t-1}V^k_{\rho,\theta}(s'_k)\,\middle|\,s'_t=s_t\right]\right].\end{aligned} \tag{53}$$

Step 3 Finally, when $(\rho,\theta)=(\bar\rho,\bar\theta)$, by (51) we have

$$\begin{aligned}\nabla_\theta \mathcal{L}(\rho,\theta)|_{(\rho,\theta)=(\bar\rho,\bar\theta)} =&\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=0}^{T}\gamma^t\nabla_\theta U_\theta(R(s_t,a_t))\right]\\ &-\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=0}^{T}\rho^t\nabla_\theta U_\theta(R(s_t,a_t))\right]\\ &+(\rho-\gamma)\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\\ &\left[\sum_{t=1}^{T}\gamma^{t-1}\sum_{k=t}^{T}\rho^{k-t}\nabla_\theta U_\theta(R(s_k,a_k))\right].\end{aligned} \tag{54}$$

Note that for the last line of the above equation,

$$\begin{aligned}&\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=1}^{T}\gamma^{t-1}\sum_{k=t}^{T}\rho^{k-t}\nabla_\theta U_\theta(R(s_k,a_k))\right]\\ &=\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=1}^{T}\left(\frac{\gamma}{\rho}\right)^{t-1}\sum_{k=t}^{T}\rho^{k-1}\nabla_\theta U_\theta(R(s_k,a_k))\right]\\ &=\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{k=1}^{T}\sum_{t=1}^{k}\left(\frac{\gamma}{\rho}\right)^{t-1}\rho^{k-1}\nabla_\theta U_\theta(R(s_k,a_k))\right]\end{aligned} \tag{55}$$

$$
\begin{aligned}
&= \mathbb{E}_{\xi\sim\pi_{\bar{\rho},\bar{\theta}}}\left[\sum_{k=1}^{T}\frac{(\gamma/\rho)^k-1}{\gamma/\rho-1}\rho^{k-1}\nabla_\theta U_\theta(R(s_k,a_k))\right] \\
&= \mathbb{E}_{\xi\sim\pi_{\bar{\rho},\bar{\theta}}}\left[\sum_{k=1}^{T}\frac{\gamma^k-\rho^k}{\gamma-\rho}\nabla_\theta U_\theta(R(s_k,a_k))\right],
\end{aligned}
\tag{56}
$$

where (55) holds by changing the order of summations. Plugging (56) back into (54), we have the desired result that $\nabla_\theta\mathcal{L}(\rho,\theta)|_{(\rho,\theta)=(\bar{\rho},\bar{\theta})}=0$.

Similarly we have $\nabla_\rho\mathcal{L}(\rho,\theta)|_{(\rho,\theta)=(\bar{\rho},\bar{\theta})}=0$. □

We next show results on the Hessian matrix.

Theorem 3 (Landscape analysis) It holds that

$$
\begin{aligned}
&\nabla^2_\theta\mathcal{L}(\rho,\theta)|_{(\rho,\theta)=(\bar{\rho},\bar{\theta})} = -\mathbb{E}_{\xi\sim\pi_{\bar{\rho},\bar{\theta}}} \\
&\left[\sum_{t=0}^{T}\gamma^t\mathbb{V}_{\xi'\sim\pi_{\bar{\rho},\bar{\theta}}}\left[\nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t)|_{(\rho,\theta)=(\bar{\rho},\bar{\theta})}\middle|s'_t=s_t\right]\right], \\
&\nabla^2_\rho\mathcal{L}(\rho,\theta)|_{(\rho,\theta)=(\bar{\rho},\bar{\theta})} = -\mathbb{E}_{\xi\sim\pi_{\bar{\rho},\bar{\theta}}} \\
&\left[\sum_{t=0}^{T}\gamma^t\mathbb{V}_{\xi'\sim\pi_{\bar{\rho},\bar{\theta}}}\left[\nabla_\rho Q^t_{\rho,\theta}(s'_t,a'_t)|_{(\rho,\theta)=(\bar{\rho},\bar{\theta})}\middle|s'_t=s_t\right]\right], \\
&\nabla_\theta\nabla_\rho\mathcal{L}(\rho,\theta)|_{(\rho,\theta)=(\bar{\rho},\bar{\theta})} = -\mathbb{E}_{\xi\sim\pi_{\bar{\rho},\bar{\theta}}} \\
&\left[\sum_{t=0}^{T}\gamma^t\mathrm{Cov}_{\xi'\sim\pi_{\bar{\rho},\bar{\theta}}}\left[\nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t)|_{(\rho,\theta)=(\bar{\rho},\bar{\theta})},\nabla_\rho Q^t_{\rho,\theta}(s'_t,a'_t)|_{(\rho,\theta)=(\bar{\rho},\bar{\theta})}\middle|s'_t=s_t\right]\right]
\end{aligned}
$$

in which we define

$$
\begin{aligned}
&\mathbb{V}_{\xi'\sim\pi_{\rho,\theta}}\left[\nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t)\middle|s'_t=s\right] := \mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t)\nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t)^\top\middle|s'_t=s\right] \\
&\quad-\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t)\middle|s'_t=s\right]\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t)\middle|s'_t=s\right]^\top,
\end{aligned}
\tag{57}
$$

and

$$
\begin{aligned}
&\mathrm{Cov}_{\xi'\sim\pi_{\rho,\theta}}\left[\nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t),\nabla_\rho Q^t_{\rho,\theta}(s'_t,a'_t)\middle|s'_t=s\right] \\
&:= \mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t)\nabla_\rho Q^t_{\rho,\theta}(s'_t,a'_t)\middle|s'_t=s\right] \\
&\quad-\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t)\middle|s'_t=s\right]\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\nabla_\rho Q^t_{\rho,\theta}(s'_t,a'_t)\middle|s'_t=s\right]\in\mathbb{R}^d.
\end{aligned}
\tag{58}
$$

In addition,

$$\mathcal{H}(\bar{\rho},\bar{\theta}) := \begin{pmatrix} \nabla_\theta^2 \mathcal{L}(\rho,\theta) & \nabla_\theta \nabla_\rho \mathcal{L}(\rho,\theta) \\ \nabla_\theta \nabla_\rho \mathcal{L}(\rho,\theta)^\top & \nabla_\rho^2 \mathcal{L}(\rho,\theta) \end{pmatrix}\Bigg|_{(\rho,\theta)=(\bar{\rho},\bar{\theta})}$$

is negative semi-definite.
In Theorem 3 an interesting finding is that the negative semi-definite property of the Hessian *does not rely on* the choice γ, making the likelihood estimation method particularly suitable for inference problems.

Proof Our proof consists of two parts. We first prove the second-order differentiability of $Q^t_{\rho,\theta}$ and $V^t_{\rho,\theta}$ by induction, and derive formulas for their second-order derivatives with respect to θ and ρ. Then we calculate the second-order derivatives of the log-likelihood function and study its Hessian matrix when $(\rho,\theta) = (\bar{\rho},\bar{\theta})$.

Step 1 To begin with, for any $(\rho,\theta) \in \Theta$, if $Q^t_{\rho,\theta}$ is twice differentiable with respect to θ, then $V^t_{\rho,\theta}$ is also twice differentiable with respect to θ by taking the derivative of (42). We have:

$$\begin{aligned}
\nabla_\theta^2 V^t_{\rho,\theta}(s) &= \nabla_\theta \Big(\sum_a \pi^t_{\rho,\theta}(a|s) \nabla_\theta Q^t_{\rho,\theta}(s,a)^\top \Big) \\
&= \sum_a \nabla_\theta \Big(e^{Q^t_{\rho,\theta}(s,a) - V^t_{\rho,\theta}(s)} \Big) \nabla_\theta Q^t_{\rho,\theta}(s,a)^\top + \sum_a \pi^t_{\rho,\theta}(a|s) \nabla_\theta^2 Q^t_{\rho,\theta}(s,a) \qquad (59) \\
&= \mathbb{E}_{\xi' \sim \pi_{\rho,\theta}} \Big[\nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t) \nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t)^\top \Big| s'_t = s \Big] \\
&\quad - \mathbb{E}_{\xi' \sim \pi_{\rho,\theta}} \Big[\nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t) \Big| s'_t = s \Big] \mathbb{E}_{\xi' \sim \pi_{\rho,\theta}} \Big[\nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t) \Big| s'_t = s \Big]^\top \\
&\quad + \mathbb{E}_{\xi' \sim \pi_{\rho,\theta}} \Big[\nabla_\theta^2 Q^t_{\rho,\theta}(s'_t,a'_t) \Big| s'_t = s \Big] \qquad (60) \\
&= \mathbb{V}_{\xi' \sim \pi_{\rho,\theta}} \Big[\nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t) \Big| s'_t = s \Big] + \mathbb{E}_{\xi' \sim \pi_{\rho,\theta}} \Big[\nabla_\theta^2 Q^t_{\rho,\theta}(s'_t,a'_t) \Big| s'_t = s \Big], \qquad (61)
\end{aligned}$$

where the covariance matrix $\mathbb{V}_{\xi' \sim \pi_{\rho,\theta}} \Big[\nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t) \Big| s'_t = s \Big]$ is defined in (57). In particular, (59) holds because:

$$\pi^t_{\rho,\theta}(a|s) = \frac{e^{Q^t_{\rho,\theta}(s,a)}}{\sum_{a'} e^{Q^t_{\rho,\theta}(s,a')}} = e^{Q^t_{\rho,\theta}(s,a) - V^t_{\rho,\theta}(s)},$$

and (60) holds by (42). In addition, we have:

$$\begin{aligned}
\nabla_\theta^2 Q^t_{\rho,\theta}(s,a) &= \nabla_\theta^2 U_\theta(R(s,a)) + \rho \mathbb{E}_{\xi' \sim \pi_{\rho,\theta}} \Big[\nabla_\theta^2 V^{t+1}_{\rho,\theta}(s'_{t+1}) \big| s'_t = s, a'_t = a \Big] \\
&= \nabla_\theta^2 U_\theta(R(s,a)) + \rho \mathbb{E}_{\xi' \sim \pi_{\rho,\theta}} \Big[\mathbb{V}_{\xi' \sim \pi_{\rho,\theta}} \Big[\nabla_\theta Q^{t+1}_{\rho,\theta}(s'_{t+1},a'_{t+1}) \big| s'_{t+1} \Big] \Big| s'_t = s, a'_t = a \Big] \qquad (62) \\
&\quad + \rho \mathbb{E}_{\xi' \sim \pi_{\rho,\theta}} \Big[\nabla_\theta^2 Q^{t+1}_{\rho,\theta}(s'_{t+1},a'_{t+1}) \big| s'_t = s, a'_t = a \Big],
\end{aligned}$$

where (62) holds by (61). Since $Q^T_{\rho,\theta}(s,a) = U_\theta\big(R(s,a)\big)$ is twice differentiable with respect to θ by the assumption of the utility function, the second-order differentiability of $Q^t_{\rho,\theta}$ and $V^t_{\rho,\theta}$ in θ is justified by induction for all $t \in \{0,1,2,\dots T\}$. Applying (62) recursively yields:

$$\begin{aligned}\nabla^2_\theta Q^t_{\rho,\theta}(s,a) &= \mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\sum_{k=t}^{T}\rho^{k-t}\nabla^2_\theta U_\theta(R(s'_k,a'_k))\Big|s'_t=s,a'_t=a\right]\\ &+\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\sum_{k=t}^{T}\rho^{k-t+1}\mathbb{V}_{\xi'\sim\pi_{\rho,\theta}}\left[\nabla_\theta Q^{k+1}_{\rho,\theta}(s'_{k+1},a'_{k+1})\Big|s'_{k+1}\right]\Big|s'_t=s,a'_t=a\right].\end{aligned} \tag{63}$$

Similarly, the second-order differentiability of $Q^t_{\rho,\theta}$ and $V^t_{\rho,\theta}$ with respect to ρ can be justified by induction for all $t \in \{0,1,2,\dots T\}$. For any $\rho \in (0,1)$, $\theta \in \Theta$, by taking the gradient of (43) with respect to ρ, we have

$$\begin{aligned}\nabla^2_\rho V^t_{\rho,\theta}(s) &= \nabla_\rho\Big(\sum_a \pi^t_{\rho,\theta}(a|s)\nabla_\rho Q^t_{\rho,\theta}(s,a)\Big)\\ &= \sum_a \nabla_\rho\Big(e^{Q^t_{\rho,\theta}(s,a)-V^t_{\rho,\theta}(s)}\Big)\nabla_\rho Q^t_{\rho,\theta}(s,a)+\sum_a \pi^t_{\rho,\theta}(a|s)\nabla^2_\rho Q^t_{\rho,\theta}(s,a)\\ &= \mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[(\nabla_\rho Q^t_{\rho,\theta}(s'_t,a'_t))^2\Big|s'_t=s\Big]-\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla_\rho Q^t_{\rho,\theta}(s'_t,a'_t)|s'_t=s\Big]^2 \qquad (64)\\ &\quad+\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla^2_\rho Q^t_{\rho,\theta}(s'_t,a'_t)|s'_t=s\Big]\\ &= \mathbb{V}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla_\rho Q^t_{\rho,\theta}(s'_t,a'_t)\Big|s'_t=s\Big]+\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla^2_\rho Q^t_{\rho,\theta}(s'_t,a'_t)\Big|s'_t=s\Big]. \qquad (65)\end{aligned}$$

Similarly, by taking the gradient of (37),

$$\begin{aligned}\nabla^2_\rho Q^t_{\rho,\theta}(s,a) &= \nabla_\rho\Big(\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[V^{t+1}_{\rho,\theta}(s'_{t+1})\Big|s'_t=s,a'_t=a\Big]\\ &\quad+\rho\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla_\rho V^{t+1}_{\rho,\theta}(s'_{t+1})\Big|s'_t=s,a'_t=a\Big]\Big)\\ &= 2\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla_\rho V^{t+1}_{\rho,\theta}(s'_{t+1})\Big|s'_t=s,a'_t=a\Big]\\ &\quad+\rho\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla^2_\rho V^{t+1}_{\rho,\theta}(s'_{t+1})\Big|s'_t=s,a'_t=a\Big]\\ &= 2\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla_\rho Q^{t+1}_{\rho,\theta}(s'_{t+1},a'_{t+1})\Big|s'_t=s,a'_t=a\Big]\\ &\quad+\rho\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\mathbb{V}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla_\rho Q^{t+1}_{\rho,\theta}(s'_{t+1},a'_{t+1})\Big|s'_{t+1}\Big]\Big|s'_t=s,a'_t=a\Big]\\ &\quad+\rho\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla^2_\rho Q^{t+1}_{\rho,\theta}(s'_{t+1},a'_{t+1})\Big|s'_t=s,a'_t=a\Big].\end{aligned} \tag{66}$$

Applying (66) recursively yields:

$$\begin{aligned}\nabla_\rho^2 Q_{\rho,\theta}^t(s,a) &= 2\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\sum_{k=t+1}^{T}\rho^{k-t-1}\nabla_\rho Q_{\rho,\theta}^k(s'_k,a'_k)\Big|s'_t=s,a'_t=a\right]\\ &+\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\sum_{k=t+1}^{T}\rho^{k-t}\mathbb{V}_{\xi'\sim\pi_{\rho,\theta}}\left[\nabla_\rho Q_{\rho,\theta}^k(s'_k,a'_k)\Big|s'_k\right]\Big|s'_t=s,a'_t=a\right].\end{aligned} \tag{67}$$

Furthermore, for any $(\rho,\theta)\in\Theta$, by taking the gradient of (43) with respect to θ , we have

$$\begin{aligned}&\nabla_\theta\nabla_\rho V_{\rho,\theta}^t(s)=\nabla_\theta\Big(\sum_a \pi_{\rho,\theta}^t(a|s)\nabla_\rho Q_{\rho,\theta}^t(s,a)\Big)\\ &=\sum_a \nabla_\theta\Big(e^{Q_{\rho,\theta}^t(s,a)-V_{\rho,\theta}^t(s)}\Big)\nabla_\rho Q_{\rho,\theta}^t(s,a)+\sum_a \pi_{\rho,\theta}^t(a|s)\nabla_\theta\nabla_\rho Q_{\rho,\theta}^t(s,a)\\ &=\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla_\theta Q_{\rho,\theta}^t(s'_t,a'_t)\nabla_\rho Q_{\rho,\theta}^t(s'_t,a'_t)\Big|s'_t=s\Big]+\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla_\theta\nabla_\rho Q_{\rho,\theta}^t(s'_t,a'_t)\Big|s'_t=s\Big]\\ &\quad-\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla_\theta Q_{\rho,\theta}^t(s'_t,a'_t)\Big|s'_t=s\Big]\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla_\rho Q_{\rho,\theta}^t(s'_t,a'_t)\Big|s'_t=s\Big]\\ &=\mathbf{Cov}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla_\theta Q_{\rho,\theta}^t(s'_t,a'_t),\nabla_\rho Q_{\rho,\theta}^t(s'_t,a'_t)\Big|s'_t=s\Big]+\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla_\theta\nabla_\rho Q_{\rho,\theta}^t(s'_t,a'_t)\Big|s'_t=s\Big],\end{aligned} \tag{68}$$

where the "covariance" between $\nabla_\theta Q$ and $\nabla_\rho Q$ is defined in (58). Note that $\mathrm{Cov}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla_\theta Q_{\rho,\theta}^t(s'_t,a'_t),\nabla_\rho Q_{\rho,\theta}^t(s'_t,a'_t)\Big|s'_t=s\Big]\in\mathbb{R}^d$, as we have $\theta\in\mathbb{R}^d$ and $\rho\in\mathbb{R}$.

Lastly, by taking the gradient of (37), we obtain

$$\begin{aligned}\nabla_\theta\nabla_\rho Q_{\rho,\theta}^t(s,a)&=\nabla_\theta\Big(\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[V_{\rho,\theta}^{t+1}(s'_{t+1})\Big|s'_t=s,a'_t=a\Big]\\ &\quad+\rho\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla_\rho V_{\rho,\theta}^{t+1}(s'_{t+1})\Big|s'_t=s,a'_t=a\Big]\Big)\\ &=\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla_\theta V_{\rho,\theta}^{t+1}(s'_{t+1})\Big|s'_t=s,a'_t=a\Big]\\ &\quad+\rho\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla_\theta\nabla_\rho V_{\rho,\theta}^{t+1}(s'_{t+1})\Big|s'_t=s,a'_t=a\Big]\\ &=\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla_\theta V_{\rho,\theta}^{t+1}(s'_{t+1})\Big|s'_t=s,a'_t=a\Big]\\ &\quad+\rho\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla_\theta\nabla_\rho Q_{\rho,\theta}^{t+1}(s'_{t+1},a'_{t+1})\Big|s'_t=s,a'_t=a\Big]\\ &\quad+\rho\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\Big[\mathrm{Cov}_{\xi'\sim\pi_{\rho,\theta}}\Big[\nabla_\theta Q_{\rho,\theta}^{t+1}(s'_{t+1},a'_{t+1}),\\ &\quad\nabla_\rho Q_{\rho,\theta}^{t+1}(s'_{t+1},a'_{t+1})\Big|s'_{t+1}\Big]\Big|s'_t=s,a'_t=a\Big].\end{aligned}$$

Applying the last equation recursively yields:

$$\nabla_\theta \nabla_\rho Q^t_{\rho,\theta}(s,a) = \mathbb{E}_{\xi' \sim \pi_{\rho,\theta}} \left[\sum_{k=t+1}^{T} \rho^{k-t-1} \nabla_\theta V^k_{\rho,\theta}(s'_k) \Big| s'_t = s, a'_t = a \right] + \mathbb{E}_{\xi' \sim \pi_{\rho,\theta}} \left[\sum_{k=t+1}^{T} \rho^{k-t} \mathrm{Cov}_{\xi' \sim \pi_{\rho,\theta}} \left[\nabla_\theta Q^k_{\rho,\theta}(s'_k, a'_k), \nabla_\rho Q^k_{\rho,\theta}(s'_k, a'_k) \Big| s'_k \right] \Big| s'_t = s, a'_t = a \right]. \tag{69}$$

In summary, by combining (61), (63), (65), (67), (68), and (69), we have the following formulas of the second-order gradients of the value function: for any $(\rho, \theta) \in \Theta$, for any $(s_t, a_t) \in \mathcal{S} \times \mathcal{A}$,

$$\nabla^2_\theta V^t_{\rho,\theta}(s) = \mathbb{E}_{\xi' \sim \pi_{\rho,\theta}} \left[\sum_{k=t}^{T} \rho^{k-t} \nabla^2_\theta U_\theta(R(s'_k, a'_k)) \Big| s'_t = s \right] + \mathbb{E}_{\xi' \sim \pi_{\rho,\theta}} \left[\sum_{k=t}^{T} \rho^{k-t} \mathbb{V}_{\xi' \sim \pi_{\rho,\theta}} \left[\nabla_\theta Q^k_{\rho,\theta}(s'_k, a'_k) \Big| s'_k \right] \Big| s'_t = s \right], \tag{70}$$

$$\nabla^2_\rho V^t_{\rho,\theta}(s) = 2\mathbb{E}_{\xi' \sim \pi_{\rho,\theta}} \left[\sum_{k=t+1}^{T} \rho^{k-t-1} \nabla_\rho Q^k_{\rho,\theta}(s'_k, a'_k) \Big| s'_t = s \right] + \mathbb{E}_{\xi' \sim \pi_{\rho,\theta}} \left[\sum_{k=t}^{T} \rho^{k-t} \mathbb{V}_{\xi' \sim \pi_{\rho,\theta}} \left[\nabla_\rho Q^k_{\rho,\theta}(s'_k, a'_k) \Big| s'_k \right] \Big| s'_t = s \right], \tag{71}$$

$$\nabla_\theta \nabla_\rho V^t_{\rho,\theta}(s) = \mathbb{E}_{\xi' \sim \pi_{\rho,\theta}} \left[\sum_{k=t+1}^{T} \rho^{k-t-1} \nabla_\theta V^k_{\rho,\theta}(s'_k) \Big| s'_t = s \right] + \mathbb{E}_{\xi' \sim \pi_{\rho,\theta}} \left[\sum_{k=t}^{T} \rho^{k-t} \mathrm{Cov}_{\xi' \sim \pi_{\rho,\theta}} \left[\nabla_\theta Q^k_{\rho,\theta}(s'_k, a'_k), \nabla_\rho Q^k_{\rho,\theta}(s'_k, a'_k) \Big| s'_k \right] \Big| s'_t = s \right]. \tag{72}$$

Step 2 Next, we calculate the derivatives of the log-likelihood function. By straight-forward calculations using (50) and combining with (70), the second-order derivative of the log-likelihood function to θ satisfies:

$$\begin{aligned}\nabla^2_\theta \mathcal{L}(\rho,\theta) &= \mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=0}^{T}\gamma^t\nabla^2_\theta U_\theta(R(s_t,a_t))\right] - \mathbb{E}_{s_0\sim\mu(\cdot)}\left[\nabla^2_\theta V^0_{\rho,\theta}(s_0)\right]\\ &\quad + (\rho-\gamma)\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=1}^{T}\gamma^{t-1}\nabla^2_\theta V^t_{\rho,\theta}(s_t)\right]\\ &= \mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=0}^{T}\gamma^t\nabla^2_\theta U_\theta(R(s_t,a_t))\right] - \mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\sum_{t=0}^{T}\rho^t\nabla^2_\theta U_\theta(R(s'_t,a'_t))\right]\\ &\quad - \mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\sum_{t=0}^{T}\rho^t\mathbb{V}_{\xi'\sim\pi_{\rho,\theta}}\left[\nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t)\middle|s'_t\right]\right]\\ &\quad + (\rho-\gamma)\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=1}^{T}\gamma^{t-1}\nabla^2_\theta V^t_{\rho,\theta}(s_t)\right].\end{aligned}$$

Note that when $(\rho,\theta)=(\bar\rho,\bar\theta)$,

$$\begin{aligned}&\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=1}^{T}\gamma^{t-1}\nabla^2_\theta V^t_{\rho,\theta}(s_t)|_{(\rho,\theta)=(\bar\rho,\bar\theta)}\right]\\ &\quad= \mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=1}^{T}\gamma^{t-1}\sum_{k=t}^{T}\bar\rho^{k-t}\nabla^2_\theta U_\theta(R(s_k,a_k))|_{\theta=\bar\theta}\right]\\ &\qquad+ \mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=1}^{T}\gamma^{t-1}\sum_{k=t}^{T}\bar\rho^{k-t}\mathbb{V}_{\xi'\sim\pi_{\bar\rho,\bar\theta}}\left[\nabla_\theta Q^k_{\rho,\theta}(s'_k,a'_k)|_{(\rho,\theta)=(\bar\rho,\bar\theta)}\middle|s'_k=s_k\right]\right]\\ &\quad= \mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{k=1}^{T}\frac{\gamma^k-\bar\rho^k}{\gamma-\bar\rho}\nabla^2_\theta U_\theta(R(s_k,a_k))|_{\theta=\bar\theta}\right]\\ &\qquad+ \mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{k=1}^{T}\frac{\gamma^k-\bar\rho^k}{\gamma-\bar\rho}\mathbb{V}_{\xi'\sim\pi_{\bar\rho,\bar\theta}}\left[\nabla_\theta Q^k_{\rho,\theta}(s'_k,a'_k)|_{(\rho,\theta)=(\bar\rho,\bar\theta)}\middle|s'_k=s_k\right]\right],\end{aligned}$$

where the first equality holds by (70) and the last equality holds by changing the order of summations. Plugging the above result to $\nabla^2_\theta\mathcal{L}(\rho,\theta)$, we obtain that,

$$
\begin{aligned}
\nabla_\theta^2 \mathcal{L}(\rho,\theta)|_{(\rho,\theta)=(\bar\rho,\bar\theta)} &= \mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=0}^{T}\gamma^t \nabla_\theta^2 U_\theta(R(s_t,a_t))|_{\theta=\bar\theta}\right] \\
&- \mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=0}^{T}\bar\rho^t \nabla_\theta^2 U_\theta(R(s_t,a_t))|_{\theta=\bar\theta}\right] \\
&- \mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=0}^{T}\bar\rho^t \mathbb{V}_{\xi'\sim\pi_{\bar\rho,\bar\theta}}\left[\nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t)|_{(\rho,\theta)=(\bar\rho,\bar\theta)}\middle| s'_t=s_t\right]\right] \\
&+ \mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=1}^{T}\bar\rho^t \nabla_\theta^2 U_\theta(R(s_t,a_t))|_{\theta=\bar\theta}\right] - \mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=1}^{T}\gamma^t \nabla_\theta^2 U_\theta(R(s_t,a_t))|_{\theta=\bar\theta}\right] \\
&+ \mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=1}^{T}\bar\rho^t \mathbb{V}_{\xi'\sim\pi_{\bar\rho,\bar\theta}}\left[\nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t)|_{(\rho,\theta)=(\bar\rho,\bar\theta)}\middle| s'_t=s_t\right]\right] \\
&- \mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=1}^{T}\gamma^t \mathbb{V}_{\xi'\sim\pi_{\bar\rho,\bar\theta}}\left[\nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t)|_{(\rho,\theta)=(\bar\rho,\bar\theta)}\middle| s'_t=s_t\right]\right] \\
&= -\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=0}^{T}\gamma^t \mathbb{V}_{\xi'\sim\pi_{\bar\rho,\bar\theta}}\left[\nabla_\theta Q^t_{\rho,\theta}(s'_t,a'_t)|_{(\rho,\theta)=(\bar\rho,\bar\theta)}\middle| s'_t=s_t\right]\right].
\end{aligned}
$$

Similarly, using the result in (71) and (52),

$$
\begin{aligned}
\nabla_\rho^2 \mathcal{L}(\rho,\theta) = &- 2\mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\sum_{t=0}^{T}\rho^t \nabla_\rho Q^t_{\rho,\theta}(s'_{t+1},a'_{t+1})\right] \\
&- \mathbb{E}_{\xi'\sim\pi_{\rho,\theta}}\left[\sum_{t=0}^{T}\rho^t \mathbb{V}_{\xi'\sim\pi_{\rho,\theta}}\left[\nabla_\theta Q^t_{\rho,\theta}(s'_t,a')\middle| s'_t=s_t\right]\right] \\
&+ 2\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=1}^{T}\gamma^{t-1} \nabla_\rho V^t_{\rho,\theta}(s_t)\right] \\
&+ (\rho-\gamma)\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=1}^{T}\gamma^{t-1} \nabla_\rho^2 V^t_{\rho,\theta}(s_t)\right].
\end{aligned} \tag{73}
$$

Note that when $(\rho,\theta)=(\bar\rho,\bar\theta)$,

$$
\begin{aligned}
&\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=1}^{T}\gamma^{t-1}\nabla_\rho^2 V_{\rho,\theta}^t(s_t)|_{(\rho,\theta)=(\bar\rho,\bar\theta)}\right]\\
&\quad=2\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=1}^{T}\gamma^{t-1}\sum_{k=t+1}^{T}\bar\rho^{k-t-1}\nabla_\rho Q_{\rho,\theta}^k(s_k,a_k)|_{(\rho,\theta)=(\bar\rho,\bar\theta)}\right]\\
&\qquad+\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=1}^{T}\gamma^{t-1}\sum_{k=t}^{T}\bar\rho^{k-t}\mathbb{V}_{\xi'\sim\pi_{\bar\rho,\bar\theta}}\left[\nabla_\rho Q_{\rho,\theta}^k(s_k',a_k')|_{(\rho,\theta)=(\bar\rho,\bar\theta)}\middle|s_k'=s_k\right]\right]\\
&\quad=2\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{k=2}^{T}\frac{\gamma^{k-1}-\bar\rho^{k-1}}{\gamma-\bar\rho}\nabla_\rho Q_{\rho,\theta}^k(s_k,a_k)|_{(\rho,\theta)=(\bar\rho,\bar\theta)}\right]\\
&\qquad+\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{k=1}^{T}\frac{\gamma^{k}-\bar\rho^{k}}{\gamma-\bar\rho}\mathbb{V}_{\xi'\sim\pi_{\bar\rho,\bar\theta}}\left[\nabla_\rho Q_{\rho,\theta}^k(s_k',a_k')|_{(\rho,\theta)=(\bar\rho,\bar\theta)}\middle|s_k'=s_k\right]\right],
\end{aligned}
$$

where the last equality holds by changing the order of summations. Plugging the above result to $\nabla_\rho^2\mathcal{L}(\rho,\theta)$ and combining with (42) and (43), we have

$$
\begin{aligned}
&\nabla_\rho^2\mathcal{L}(\rho,\theta)|_{(\rho,\theta)=(\bar\rho,\bar\theta)}\\
&\quad=-\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\left[\sum_{t=0}^{T}\gamma^t\mathbb{V}_{\xi'\sim\pi_{\bar\rho,\bar\theta}}\left[\nabla_\rho Q_{\rho,\theta}^t(s_t',a_k')|_{(\rho,\theta)=(\bar\rho,\bar\theta)}\middle|s_t'=s_t\right]\right].
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
\nabla_\theta\nabla_\rho\mathcal{L}(\rho,\theta)|_{(\rho,\theta)=(\bar\rho,\bar\theta)}&=-\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\\
\Bigg[\sum_{t=0}^{T}\gamma^t\mathrm{Cov}_{\xi'\sim\pi_{\bar\rho,\bar\theta}}&\left[\nabla_\theta Q_{\rho,\theta}^t(s_t',a_t')|_{(\rho,\theta)=(\bar\rho,\bar\theta)},\right.\\
&\left.\nabla_\rho Q_{\rho,\theta}^t(s_t',a_t')|_{(\rho,\theta)=(\bar\rho,\bar\theta)}\middle|s_t'=s_t\right]\Bigg].
\end{aligned}
$$

To summarize,

$$
\begin{aligned}
\mathcal{H}&:=\begin{pmatrix}\nabla_\theta^2\mathcal{L}(\rho,\theta) & \nabla_\theta\nabla_\rho\mathcal{L}(\rho,\theta)\\ \nabla_\theta\nabla_\rho\mathcal{L}(\rho,\theta)^\top & \nabla_\rho^2\mathcal{L}(\rho,\theta)\end{pmatrix}\Bigg|_{(\rho,\theta)=(\bar\rho,\bar\theta)}\\
&=-\mathbb{E}_{\xi\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{T}\gamma^t\mathbb{V}_{\xi'\sim\pi_{\bar\rho,\bar\theta}}\Big[\begin{pmatrix}\nabla_\theta Q_{\rho,\theta}^t(s_t',a_t')\\ \nabla_\rho Q_{\rho,\theta}^t(s_t',a_t')\end{pmatrix}\Big|s_t'=s_t\Big]\Big].
\end{aligned}
$$

Therefore, $\mathcal{H}$ is negative-semi definite by the definition of the covariance notation $\mathbb{V}$ in (57). □

3.2 Algorithm Design and Implementation

Motivated by the landscape analysis in Sect. 3.1, we design an algorithm that iteratively updates ρ and θ to maximize the likelihood function; see Algorithm 1. At each iteration k, the value function $V_{\rho^{(k)},\theta^{(k)}}$ is first computed by the soft Q iteration (see e.g. [64]) in lines 3–7, and the parameters $\rho^{(k)}, \theta^{(k)}$ are then updated in line 10 using the gradient computed in line 9.

1: Initialize $\rho^{(0)}, \theta^{(0)}$.
2: **for** $k = 1, 2, \cdots, K$ **do**
3: Set $Q^T_{\rho^{(k)},\theta^{(k)}}(s,a) = U_{\theta^{(k)}}(R(s,a))$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, and $V^T_{\rho^{(k)},\theta^{(k)}}(s) = \log\Big(\sum_{a\in\mathcal{A}} e^{Q^T_{\rho^{(k)},\theta^{(k)}}(s,a)}\Big)$ for all $s \in \mathcal{S}$.
4: **for** $t = T-1, T-2, \ldots, 0$ **do**
5: **for** $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**

$$Q^t_{\rho^{(k)},\theta^{(k)}}(s,a) = U_{\theta^{(k)}}(R(s,a)) + \rho^{(k)}\, \mathbb{E}_{s'\sim P(\cdot|s,a)}\Big[V^{t+1}_{\rho^{(k)},\theta^{(k)}}(s')\Big].$$

6: **end for**
7: Compute $V^t_{\rho^{(k)},\theta^{(k)}}$ using the soft Bellman equation (29).
8: **end for**
9: With the value of $V^0_{\rho^{(k)},\theta^{(k)}}$, compute $\nabla\mathcal{L}(\rho^{(k)},\theta^{(k)})$ using (51) and (53).
10: Update $(\rho^{(k+1)},\theta^{(k+1)}) = (\rho^{(k)},\theta^{(k)}) + \zeta^{(k)}\nabla\mathcal{L}(\rho^{(k)},\theta^{(k)})$.
11: **end for**

Algorithm 1 Maximum likelihood update

Numerical Example One: Merton's Problem We implement the discrete-time version of Merton's problem introduced in Sect. 2.2. The price of the bond follows $S^0_{t+1} = S^0_t + r\,\Delta$ and the price of the stock follows $S_{t+1} - S_t = S_t(\nu\Delta + \sigma\sqrt{\Delta}B_t)$, where B_t are iid sampled from $\mathcal{N}(0,1)$. Denote $(\alpha_t, c_t) \in \mathcal{A} := [0,1] \times [0,2]$ as the pair of the consumption-allocation policy at time t, then the wealth process follows:

$$X_{t+1} - X_t = \Big[X_t(\alpha_t\nu + (1-\alpha_t)r) - c_t\Big]\Delta + X_t\alpha_t\sigma\sqrt{\Delta}B_t. \tag{74}$$

The client provides a policy $\pi_{\bar\rho,\bar\theta} = \{\pi^t_{\bar\rho,\bar\theta}\}_{t=0}^T$ with $\pi^t_{\bar\rho,\bar\theta}(X_t) \in \mathcal{P}(\mathcal{A})$ to the inference agent, which solves

$$\sup_{\pi=\{\pi^t\}_{t=0}^T} \mathbb{E}^\pi\left[\sum_{t=0}^T(\bar\rho)^t\Big(1 - \exp(-\bar\theta\, c_t)\Big) + \mathcal{H}\big(\pi^t(\cdot|X_t)\big)\Big)\right] \tag{75}$$

with $c_t = c(X_t)$ and $\alpha_t = \alpha(X_t)$. Here both $\bar\rho$ and $\bar\theta$ are unknown.

In the experiment, we set $\bar\rho = 0.3$, $\bar\theta = 2$, $T = 100$, $\gamma = 0.6$, $r = 1.05$, $\Delta = 1$, $\nu = 1.06$, and $\sigma = 0.05$. We discretize and truncate the state space of the wealth process as $\mathcal{S} = \{0.13, 0.39, \ldots, 2.23, 2.5\}$, with evenly distanced values such that $|\mathcal{S}| = 10$. In addition, we discretize the joint space of the allocation and consumption processes as $\mathcal{A} = \{0.1, 0.11, \ldots, 0.98, 1\} \times \{0, 0.22, \ldots, 1.77, 2\}$, with evenly distanced values such that $|\mathcal{A}| = 50$.

We visualize the log-likelihood function and its gradient in Fig. 1. One can see that the likelihood function is locally concave in θ and ρ around $(\bar{\theta}, \bar{\rho})$ in a sufficiently large area, enabling us to find the true parameters by Algorithm 1 under fast convergence rate.

When implementing Algorithm 1, we initialize the parameters randomly with $\theta^{(0)}$ sampled uniformly from [0, 1] and $\rho^{(0)}$ sampled uniformly from [0.1, 0.2]. We set the learning rate as $\zeta^{(k)} = \frac{1000}{k}$. As shown in Fig. 2, both θ and ρ converge to the ground-truth value within 100 iterations.

Additionally, we analyze the behaviors of the client at time step $t = 0$ under different $\bar{\rho}$ values. Figure 3 suggests that the client opts for an overall higher consumption when $\bar{\rho} = 0.1$ and an overall lower consumption when $\bar{\rho} = 0.75$, indicating a bigger emphasis on deferred outcomes for the latter case.

Numerical Example Two: Investment under Unhedgeable Risk We consider a more complex investment problem, where the price of the primitive asset is modeled as a diffusion process whose coefficients evolve according to a correlated diffusive factor [82]. The price of the bond follows the same dynamics as in Example One:

$$S^0_{t+1} = S^0_t + r\,\Delta.$$

On the other hand, the price of the stock follows

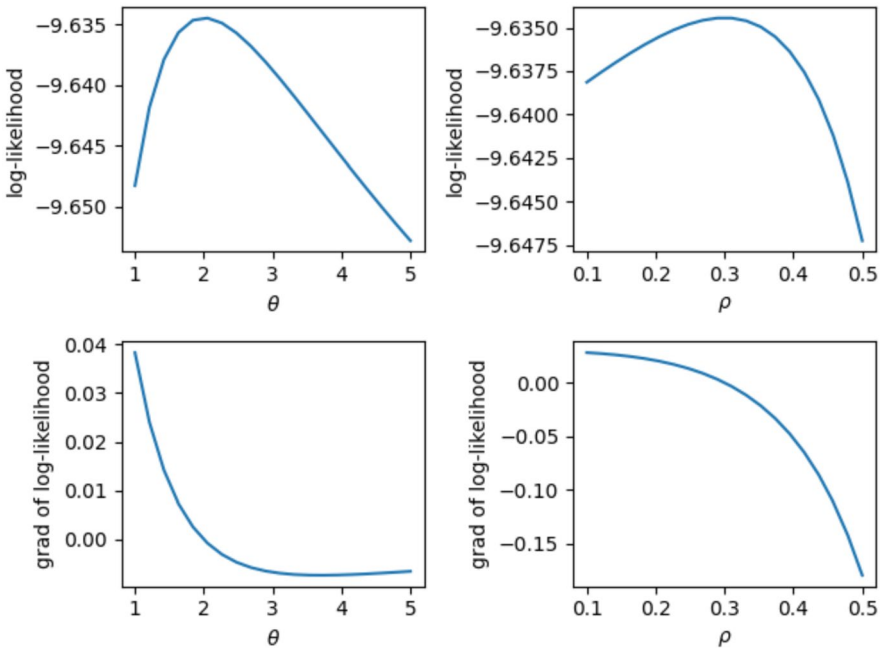

Fig. 1 Visualization of the log-likelihood function and its gradients (**Left columns** visualization with respect to θ (under $\rho = \bar{\rho}$). **Right columns** visualization with respect to ρ (under $\theta = \bar{\theta}$))

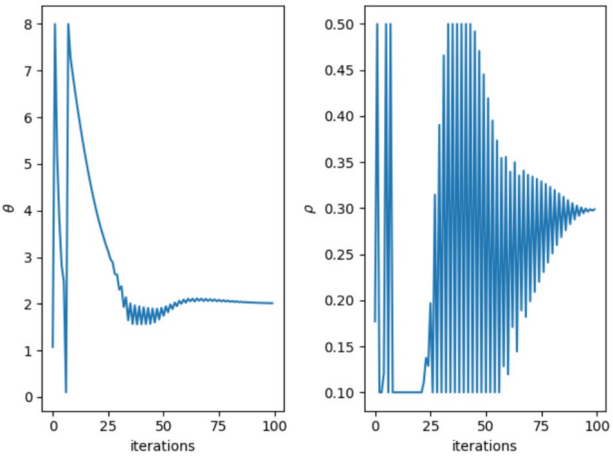

Fig. 2 The convergence result of Algorithm 1. The left plot shows the value of θ at each iteration, while the right plot displays the values for ρ

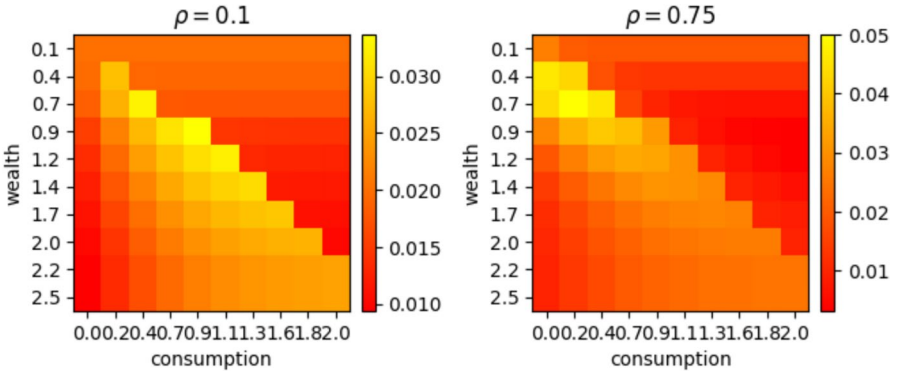

Fig. 3 Visualization of the client's consumption policy at timestamp $t = 0$. The left plot illustrates consumption at various wealth levels under $\bar{\rho} = 0.1$, while the right plot corresponds to $\bar{\rho} = 0.75$

$$S_{t+1} - S_t = S_t(\nu(Y_t, t)\Delta + \sigma(Y_t, t)\sqrt{\Delta}B_t^1),$$

with Y_t the "stochastic factor model" and it is assumed to satisfy

$$Y_{t+1} - Y_t = b(Y_t, t)\Delta + d(Y_t, t)\sqrt{\Delta}B_t^1.$$

Here B_t^1 and B_t^2 are iid sampled from $\mathcal{N}(0, 1)$. We assume the correlation between B_t^1 and B_t^2 is $\eta \in (0, 1)$.

Consider a problem with only investment and no consumption. Then the wealth process follows:

$$X_{t+1} - X_t = \Big[X_t(\alpha_t\, \nu(t, Y_t) + (1 - \alpha_t)r)\Big]\Delta + X_t\alpha_t\sigma(t, Y_t)\,\sqrt{\Delta}B_t^1, \quad (76)$$

under the investment strategy $\alpha_t \in \mathcal{A} = [0, 1]$. The client provides a policy $\pi_{\bar{\rho},\bar{\theta}} = \{\pi^t_{\bar{\rho},\bar{\theta}}\}_{t=0}^T$ with $\pi^t_{\bar{\rho},\bar{\theta}}(x, y) \in \mathcal{P}(\mathcal{A})$ to the inference agent, which solves

$$\sup_{\pi=\{\pi^t\}_{t=0}^T} \mathbb{E}^\pi \left[\sum_{t=0}^{T} (\bar{\rho})^t \frac{1}{\bar{\theta}_1 \bar{\theta}_2} (X_t)^{\bar{\theta}_1} (Y_t)^{\bar{\theta}_2} + \mathcal{H}\Big(\pi^t(\cdot|(X_t, Y_t))\Big) \right] \tag{77}$$

for some $\bar{\rho} > 0$ and $\bar{\theta}_1, \bar{\theta}_2 \in (0, 1)$ that are unknown to the inference agent.

In the experiment, we set $T = 100$, discretize and truncate the state space for the wealth process and the stochastic factor model as $\mathcal{S} = \{0.1, 0.7, 1.3, 1.9, 2.5\} \times \{0.1, 0.32, 0.55, 0.77, 1\}$, with evenly distanced values such that $|\mathcal{S}| = 25$. In addition, we discretize the action space of the allocation process as $\mathcal{A} = \{0.1, 0.32, 0.55, 0.77, 1\}$, with evenly distanced values such that $|\mathcal{A}| = 5$. We set $r = 1.05$, $\Delta = 1$, $\bar{\theta}_1 = 3$, $\bar{\theta}_2 = 2$, $\bar{\rho} = 0.3$, and $\gamma = 0.6$. For the drift and diffusion terms, we set $b(y, t) = -0.6y + 0.2$, $d(y, t) = 0.3y + 0.3$, $\nu(t, y) = y$, and $\sigma(t, y) = 0.5y + 0.3$.

As shown in Fig. 4, we visualize the log-likelihood function and its gradient. One can see that the likelihood function is locally concave in θ and ρ in an area around $(\bar{\theta}, \bar{\rho})$.

When implementing Algorithm 1, we initialize the parameters randomly with $\theta_1^{(0)}$, $\theta_2^{(0)}$ sampled uniformly from [1, 2] and $\rho^{(0)}$ sampled uniformly from [0.1, 0.2]. We

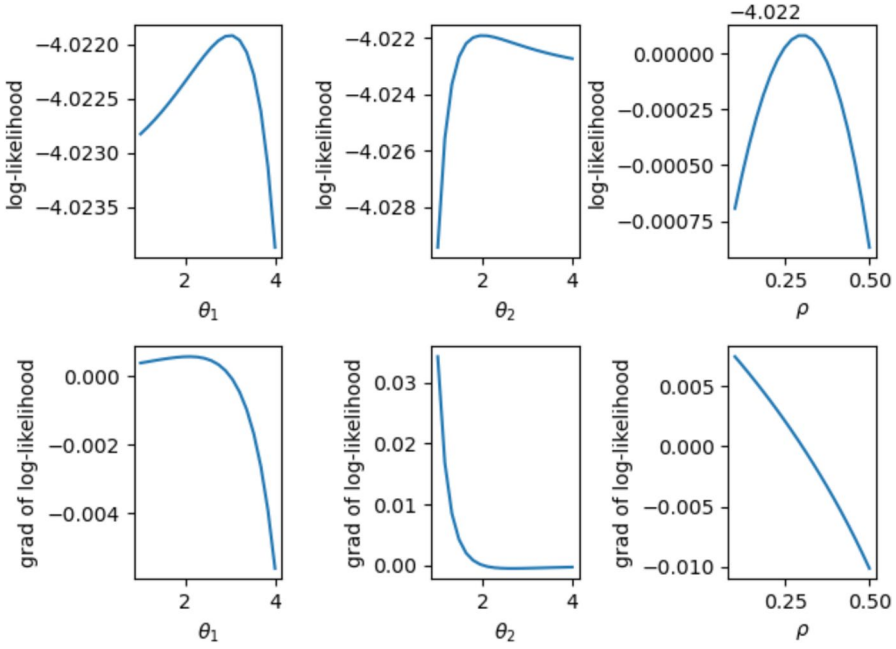

Fig. 4 Visualization of the log-likelihood function and its gradients (**Left columns** visualization with respect to θ_1 (under $\theta_2 = \bar{\theta}_2$ and $\rho = \bar{\rho}$). **Middle columns** visualization with respect to θ_2 (under $\theta_1 = \bar{\theta}_1$ and $\rho = \bar{\rho}$). **Right columns** visualization with respect to ρ (under $\theta_1 = \bar{\theta}_1$ and $\theta_2 = \bar{\theta}_2$)

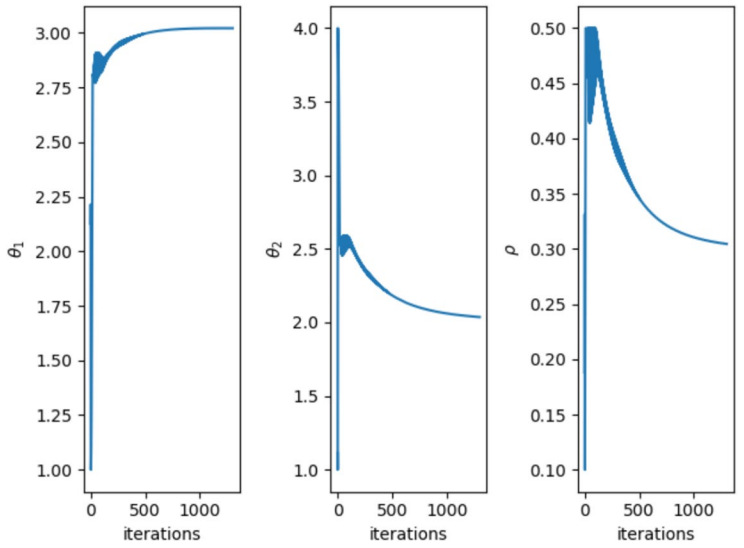

Fig. 5 The convergence result of Algorithm 1. The left plot shows the value of θ_1 at each iteration, the middle plot is for θ_2, and the right plot is for ρ

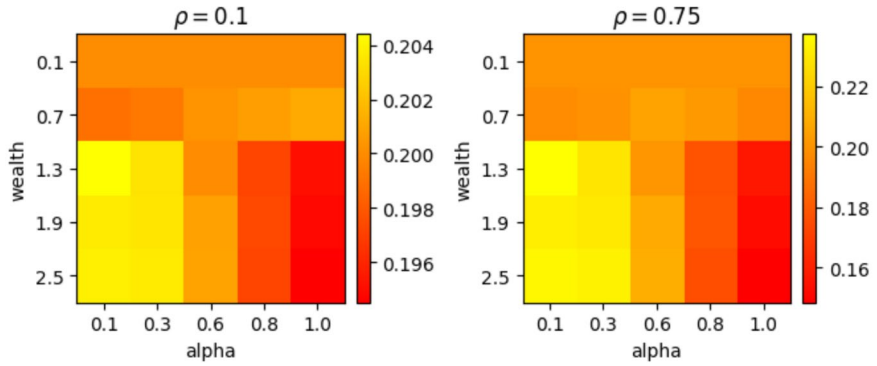

Fig. 6 Visualization of the client's allocation policy at timestamp $t = 0$ (under fixed factor value 1). The left plot illustrates the allocation at various wealth levels with $\bar{\rho} = 0.1$, while the right plot is for $\bar{\rho} = 0.75$

set the learning rate as $\zeta^{(k)} = \frac{1000}{\sqrt{k}}$. As shown in Fig. 5, both θ_1, θ_2 and ρ converge to the ground-truth values within 1500 iterations.

Furthermore, Fig. 6 illustrates the client's investment allocation policy α across various wealth levels (under fixed factor value 1) at time step $t = 0$, considering $\bar{\rho} = 0.1$ and $\bar{\rho} = 0.75$. The influence of $\bar{\rho}$ on the investment decisions in this example is less pronounced compared to Merton's problem. This difference arises because, in Merton's problem, the client confronts a trade-off between higher consumption for instantaneous rewards and lower consumption for better future rewards. Conversely, the client addressing (77) strives for a higher X_t regardless of the value of $\bar{\rho}$. Our

algorithm consistently finds the optimal parameters, although the convergence speed here is slower compared to that for Merton's problem due to the above-mentioned reasons.

A Proofs in Section 2.1.1

A.1 Proof to Proposition 1

Fix any $(t,x,z)\in\mathcal{D}$, $(\boldsymbol{\alpha},\boldsymbol{c})\in\mathcal{A}$ and $\tau\in\mathbb{T}_t$. We have

$$\begin{aligned}
J(t,x,z,\boldsymbol{\alpha},\boldsymbol{c}) &= \mathbb{E}\left[\int_t^\tau \beta_s U_1(c_s)ds + \int_\tau^T \beta_s U_1(c_s)ds + \beta_T U_2\big(X_T^{\boldsymbol{\alpha},\boldsymbol{c}}\big)\Big| X_t^{\boldsymbol{\alpha},\boldsymbol{c}}=x,\beta_t=z\right]\\
&= \mathbb{E}\left[\int_t^\tau \beta_s U_1(c_s)ds\Big| X_t^{\boldsymbol{\alpha},\boldsymbol{c}}=x,\beta_t=z\right]\\
&\quad+\mathbb{E}\left[\mathbb{E}\left[\int_\tau^T \beta_s U_1(c_s)ds + \beta_T U_2(X_T^{\boldsymbol{\alpha},\boldsymbol{c}})\Big| X_\tau^{\boldsymbol{\alpha},\boldsymbol{c}},\beta_\tau\right]\Big| X_t^{\boldsymbol{\alpha},\boldsymbol{c}}=x,\beta_t=z\right]\\
&= \mathbb{E}\left[\int_t^\tau \beta_s U_1(c_s)ds\Big| X_t^{\boldsymbol{\alpha},\boldsymbol{c}}=x,\beta_t=z\right]\\
&\quad+\mathbb{E}\left[\mathbb{E}\left[J(\tau,X,Z,\boldsymbol{\alpha},\boldsymbol{c})\Big|\tau,X=X_\tau^{\boldsymbol{\alpha},\boldsymbol{c}},Z=\beta_\tau\right]\Big| X_t^{\boldsymbol{\alpha},\boldsymbol{c}}=x,\beta_t=z\right].
\end{aligned}$$

By the definition given by (8), for any $\epsilon>0$ and $\Delta t\in[0,T-t]$, there exists $(\boldsymbol{\alpha}^{\epsilon,t+\Delta t,x,z},\boldsymbol{c}^{\epsilon,t+\Delta t,x,z})\in\mathcal{A}$ such that

$$\begin{aligned}
&J(t+\Delta t,x,z,\boldsymbol{\alpha}^{\epsilon,t+\Delta t,x,z},\boldsymbol{c}^{\epsilon,t+\Delta t,x,z})\\
&>\sup_{(\boldsymbol{\alpha},\boldsymbol{c})\in\mathcal{A}} J(t+\Delta t,x,z,\boldsymbol{\alpha},\boldsymbol{c})-\epsilon = V(t+\Delta t,x,z)-\epsilon,
\end{aligned}\tag{78}$$

and

$$\begin{aligned}
J(t+\Delta t,x,z,\boldsymbol{\alpha},\boldsymbol{c}) &\le J(t+\Delta t,x,z,\boldsymbol{\alpha}^{\epsilon,t+\Delta t,x,z},\boldsymbol{c}^{\epsilon,t+\Delta t,x,z})\\
&\le V(t+\Delta t,x,z).
\end{aligned}\tag{79}$$

Then consider $(\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{c}})=\{(\bar\alpha_s,\bar c_s)\}_{s\in[t,T]}$ such that

$$(\bar\alpha_s,\bar c_s)=(\alpha_s,c_s)\mathbf{1}\{\tau>s\}+\left(\alpha_s^{\epsilon,\tau,X_\tau^{\boldsymbol{\alpha},\boldsymbol{c}},\beta_\tau},c_s^{\epsilon,\tau,X_\tau^{\boldsymbol{\alpha},\boldsymbol{c}},\beta_\tau}\right)\mathbf{1}\{\tau\le s\},$$

where

$$\beta_\tau=z+\int_t^\tau\dot\beta_s ds,\quad X_\tau^{\boldsymbol{\alpha},\boldsymbol{c}}=x+\int_t^\tau dX_s^{\boldsymbol{\alpha},\boldsymbol{c}},$$

according to (4) and (7). Notice that $(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{c}}) \in \mathcal{A}$. By (78) and (79), for any $(\boldsymbol{\alpha}, \boldsymbol{c}) \in \mathcal{A}$, we have

$$V(t,x,z) \geq J(t,x,z,\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{c}}) > \mathbb{E}\left[\int_t^\tau \beta_s U_1(c_s)ds \middle| X_t^{\boldsymbol{\alpha},\boldsymbol{c}} = x, \beta_t = z\right] + \mathbb{E}\left[V(\tau, X_\tau^{\boldsymbol{\alpha},\boldsymbol{c}}, \beta_\tau)\middle| X_t^{\boldsymbol{\alpha},\boldsymbol{c}} = x, \beta_t = z\right] - \epsilon$$

for any $\epsilon > 0$, and

$$J(t,x,z,\boldsymbol{\alpha},\boldsymbol{c}) \leq \mathbb{E}\left[\int_t^\tau \beta_s U_1(c_s)ds + V(\tau, X_\tau^{\boldsymbol{\alpha},\boldsymbol{c}}, \beta_\tau)\middle| X_t^{\boldsymbol{\alpha},\boldsymbol{c}} = x, \beta_t = z\right].$$

It follows that for any $(t,x,z) \in \mathcal{D}$ and $\tau \in \mathbb{T}_t$,

$$V(t,x,z) = \sup_{(\boldsymbol{\alpha},\boldsymbol{c})\in\mathcal{A}} \mathbb{E}\left[\int_t^\tau \beta_s U_1(c_s)ds + V(\tau, X_\tau^{\boldsymbol{\alpha},\boldsymbol{c}}, \beta_\tau)\middle| X_t^{\boldsymbol{\alpha},\boldsymbol{c}} = x, \beta_t = z\right].$$

□

A.2 Proof to Proposition 2

The assumptions on U_1 guarantee a quadratic growth of U_1 rate in x. Consider arbitrary $(t,x,z) \in \mathcal{D}$ and $(\boldsymbol{\alpha},\boldsymbol{c}) \in \mathcal{A}$.

1. Define

$$\tau_n := \inf\left\{s \geq t \middle| \int_t^s |\partial_x w(u, X_u^{\boldsymbol{\alpha},\boldsymbol{c}}, \beta_u)|^2 du \geq n\right\}, \quad \forall n \in \mathbb{N}^+.$$

Then we have $\lim_{n\uparrow\infty} \tau_n \overset{a.s.}{=} \infty$ and the stopped process $\left\{\int_t^{s\wedge\tau_n} \partial_x w(u, X_u^{\boldsymbol{\alpha},\boldsymbol{c}}, \beta_u) dW_u\right\}_{s\in[t,T]}$ is a martingale for all $n \in \mathbb{N}^+$. The for any $s \in [t,T]$, by Itô's formula, we have

$$\begin{aligned} w(s\wedge\tau_n, X_{s\wedge\tau_n}^{\boldsymbol{\alpha},\boldsymbol{c}}, \beta_{s\wedge\tau_m}) = w(t,x,z) + \int_t^{s\wedge\tau_n} \big\{&\partial_t w(u, X_u^{\boldsymbol{\alpha},\boldsymbol{c}}, \beta_u) + \dot{\beta}_u \partial_z w(u, X_u^{\boldsymbol{\alpha},\boldsymbol{c}}, \beta_u) \\ &. + rX_u^{\boldsymbol{\alpha},\boldsymbol{c}} \partial_x w(u, X_u^{\boldsymbol{\alpha},\boldsymbol{c}}, \beta_u) + \mathcal{L}^{\alpha_u, c_u} w(u, X_u^{\boldsymbol{\alpha},\boldsymbol{c}}, \beta_u)\big\} du \\ &+ \int_t^{s\wedge\tau_n} \partial_x w(u, X_u^{\boldsymbol{\alpha},\boldsymbol{c}}, \beta_u) dW_u. \end{aligned}$$

Therefore, taking expectations on both sides we have

$$\begin{aligned}
&\mathbb{E}\left[w(s\wedge\tau_n, X^{\boldsymbol{\alpha},\boldsymbol{c}}_{s\wedge\tau_n}, \beta_{s\wedge\tau_n})\middle| X^{\boldsymbol{\alpha},\boldsymbol{c}}_t = x, \beta_t = z\right] = w(t,x,z) + \mathbb{E}\left[\int_t^{s\wedge\tau_n} \partial_t w(u, X^{\boldsymbol{\alpha},\boldsymbol{c}}_u, \beta_u)\right.\\
&\quad + \dot{\beta}_u \partial_z w(u, X^{\boldsymbol{\alpha},\boldsymbol{c}}_u, \beta_u) + r X^{\boldsymbol{\alpha},\boldsymbol{c}}_u \partial_x w(u, X^{\boldsymbol{\alpha},\boldsymbol{c}}_u, \beta_u)\\
&\quad \left.+\mathcal{L}^{\alpha_u, c_u} w(u, X^{\boldsymbol{\alpha},\boldsymbol{c}}_u, \beta_u) du\middle| X^{\boldsymbol{\alpha},\boldsymbol{c}}_t = x, \beta_t = z\right]\\
&\quad \le w(t,x,z) - \mathbb{E}\left[\int_t^{s\wedge\tau_n} \beta_u U_1(c_u) du\middle| X^{\boldsymbol{\alpha},\boldsymbol{c}}_t = x, \beta_t = z\right],
\end{aligned}$$

where the well-posedness of $\mathbb{E}\left[\int_t^{s\wedge\tau_n} \beta_u U_1(c_u) du\middle| X^{\boldsymbol{\alpha},\boldsymbol{c}}_t = x, \beta_t = z\right]$ is guaranteed by the quadratic growth rate condition on U_1 and the fact that $(\boldsymbol{\alpha}, \boldsymbol{c}) \in \mathcal{A}$. The quadratic growth rate assumption on w together with $(\boldsymbol{\alpha}, \boldsymbol{c}) \in \mathcal{A}$ allows us to apply dominated convergence theorem and get

$$\begin{aligned}
&\mathbb{E}\left[\beta_T U_2\big(X^{\boldsymbol{\alpha},\boldsymbol{c}}_T\big)\middle| X^{\boldsymbol{\alpha},\boldsymbol{c}}_t = x, \beta_t = z\right] \le \mathbb{E}\left[w(T, X^{\boldsymbol{\alpha},\boldsymbol{c}}_T, \beta_T)\middle| X^{\boldsymbol{\alpha},\boldsymbol{c}}_t = x, \beta_t = z\right]\\
&\le w(t,x,z) - \mathbb{E}\left[\int_t^T \beta_u U_1(c_u) du\middle| X^{\boldsymbol{\alpha},\boldsymbol{c}}_t = x, \beta_t = z\right]\\
&\implies w(t,x,z) \ge \mathbb{E}\left[\int_t^T \beta_u U_1(c_u) du + \beta_T U_2\big(X^{\boldsymbol{\alpha},\boldsymbol{c}}_T\big)\middle| X^{\boldsymbol{\alpha},\boldsymbol{c}}_t = x, \beta_t = z\right]\\
&= J(t,x,z,\boldsymbol{\alpha},\boldsymbol{c}).
\end{aligned}$$

Hence, $w(t,x,z) \ge V(t,x,z)$ by taking the supreme of $(\boldsymbol{\alpha}, \boldsymbol{c})$ over $\mathcal{A}$.

2. Applying a similar localization-and-Itô argument as in the previous part, we have that for any $s \in [t,T]$

$$\begin{aligned}
&\mathbb{E}\left[w\big(s, X^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}}_s, \beta_s\big)\middle| X_t = x, \beta_t = z\right] = w(t,x,z) + \mathbb{E}\left[\int_t^s \partial_t w\big(u, X^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}}_u, \beta_u\big)\right.\\
&\quad + \dot{\beta}_u \partial_z w\big(u, X^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}}_u, \beta_u\big) + r X^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}}_u \partial_x w\big(u, X^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}}_u, \beta_u\big) +\\
&\quad \left.\mathcal{L}^{\hat{\alpha}_u, \hat{c}_u} w\big(u, X^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}}_u, \beta_u\big) du\middle| X^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}}_t = x, \beta_t = z\right]\\
&\quad = w(t,x,z) - \mathbb{E}\left[\int_t^s \beta_u U_1(\hat{c}_u) du\middle| X^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}}_t = x, \beta_t = z\right].
\end{aligned}$$

In particular, when $s = T$,

$$\mathbb{E}\left[\beta_T U_2\big(X_T^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}}\big)\Big|X_t^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}}=x,\beta_t=z\right]=\mathbb{E}\left[w\big(T,X_T^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}},\beta_T\big)\Big|X_t^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}}=x,\beta_t=z\right]$$
$$=w(t,x,z)-\mathbb{E}\left[\int_t^T\beta_u U_1(\hat{c}_u)du\Big|X_t^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}}=x,\beta_t=z\right]$$
$$\Longrightarrow\ w(t,x,z)=\mathbb{E}\left[\int_t^T\beta_u U_1(\hat{c}_u)du+\beta_T U_2\big(X_T^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}}\big)\Big|X_t^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}}=x,\beta_t=z\right]$$
$$=J(t,x,z,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}).$$

Then we have $w(t,x,z)=J(t,x,z,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}})\le V(t,x,z)$. Combined with the result from the previous part, we have $w=V$ on $\mathcal{D}$, with $(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}})\in\mathcal{A}$ being a corresponding optimal joint allocation-consumption process. □

A.3 Proof to Proposition 3

First, notice that under the assumptions on admissible control specified in (3) as well as those on the utility functions specified in Proposition 2, the continuity of the value function V given by (8) over domain $\mathcal{D}$ can be established following the classical results of [52, 53], and therefore continuous on $\mathcal{D}$. Following the proof of Lemma 2, for any $z\in(0,1]$, there exists constants $C_1\ge 0$ and $C_2\in\mathbb{R}^+$ such that

$$C_1U_2(x)\le V(t,x,z)\le C_2(1+U_2(xe^{T(M|\mu-r|+r)})),\quad\forall(t,x,z)\in\mathcal{D}.$$

The polynomial growth of the value function V in x follows from the polynomial growth assumption of U_2. Combining Proposition 1 and similar arguments of Itô's formula in its proof, the viscosity solution property in Definition 1 can be established. The uniqueness result follows comparison principal; see [62, Theorem 4.4.5]. The conclusion then follows for any domain $\mathcal{D}$. □

Funding Author H. C. is partially supported by the DSAI and Department startup fund for early career faculty. Author R. X. is partially supported by the NSF CAREER award DMS-2524465 and a JP Morgan Faculty Research Award.

Declarations

Competing Interests The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

1. Abbeel, P., Ng, A.Y.: Apprenticeship learning via inverse reinforcement learning. In: Proceedings of the Twenty-First International Conference on Machine Learning, p. 1 (2004)
2. Alsabah, H., Capponi, A., Ruiz Lacedelli, O., Stern, M.: Robo-advising: learning investors' risk preferences via portfolio choices. J. Financ. Econom. **19**(2), 369–392 (2021)
3. Amin, K., Singh, S.: Towards resolving unidentifiability in inverse reinforcement learning. arXiv Preprint arXiv:1601.06569 (2016)

4. Amin, K., Jiang, N., Singh, S.: Repeated inverse reinforcement learning. Adv. Neural Inf. Process. Syst. **30**, 1815–1824 (2017)
5. Angoshtari, B., Zariphopoulou, T., Zhou, X.Y.: Predictable forward performance processes: the binomial case. SIAM J. Control Optim. **58**(1), 327–347 (2020)
6. Bäuerle, N., Rieder, U.: More risk-sensitive Markov decision processes. Math. Oper. Res. **39**(1), 105–120 (2014)
7. Bjork, T., Murgoci, A.: A general theory of Markovian time inconsistent stochastic control problems. SSRN 1694759 (2010)
8. Björk, T., Murgoci, A.: A theory of Markovian time-inconsistent stochastic control in discrete time. Financ. Stoch. **18**, 545–592 (2014)
9. Björk, T., Khapko, M., Murgoci, A.: On time-inconsistent stochastic control in continuous time. Financ. Stoch. **21**, 331–360 (2017)
10. Black, F.: Investment and consumption through time. Financial Notes 6B (1968)
11. Black, F.: Individual investment and consumption under uncertainty. In: Portfolio Insurance: A Guide to Dynamic Hedging, pp. 207–225 (1988)
12. Bloem, M., Bambos, N.: Infinite time horizon maximum causal entropy inverse reinforcement learning. In: 53rd IEEE Conference on Decision and Control, pp. 4911–4916. IEEE (2014)
13. Boularias, A., Kober, J., Peters, J.: Relative entropy inverse reinforcement learning. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pp. 182–189. JMLR Workshop and Conference Proceedings (2011)
14. Boyd, S., El Ghaoui, L., Feron, E., Balakrishnan, V.: Linear matrix inequalities in system and control theory. In: Society for Industrial and Applied Mathematics, Philadelphia (1994)
15. Cao, H., Cohen, S., Szpruch, L.: Identifiability in inverse reinforcement learning. Adv. Neural Inf. Process. Syst. **34**, 12362–12373 (2021)
16. Capponi, A., Zhang, Y.: A continuous time framework for sequential goal-based wealth management. Manag. Sci. *70*(11), 7664–7691 (2024). https://doi.org/10.1287/mnsc.2022.02047
17. Capponi, A., Olafsson, S., Zariphopoulou, T.: Personalized robo-advising: enhancing investment through client interaction. Manag. Sci. **68**(4), 2485–2512 (2022)
18. Chewning, B., Bylund, C.L., Shah, B., Arora, N.K., Gueguen, J.A., Makoul, G.: Patient preferences for shared decisions: a systematic review. Patient Educ. Couns. **86**(1), 9–18 (2012)
19. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30, pp. 4300–4308. Curran Associates, Inc., New York (2017)
20. Cox, A.M., Hobson, D., Obłój, J.: Utility theory front to back–inferring utility from agents' choices. Int. J. Theor. Appl. Financ. **17**(03), 1450018 (2014)
21. Cox, J.C., Huang, C.-F.: Optimal consumption and portfolio policies when asset prices follow a diffusion process. J. Econ. Theory **49**(1), 33–83 (1989)
22. Crandall, M.G., Ishii, H., Lions, P.-L.: User's guide to viscosity solutions of second order partial differential equations. Bull. Am. Math. Soc. **27**(1), 1–67 (1992)
23. Dai, M., Dong, Y., Jia, Y.: Learning equilibrium mean-variance strategy. Math. Financ. **33**(4), 1166–1212 (2023)
24. Derbaix, C., Abeele, P.V.: Consumer inferences and consumer preferences. the status of cognition and consciousness in consumer behavior theory. Int. J. Res. Mark. **2**(3), 157–174 (1985)
25. Dong, C., Wang, Y.: Towards generalized inverse reinforcement learning. arXiv Preprint arXiv:2402.07246 (2024)
26. Dybvig, P.H., Rogers, L.C.G.: Recovery of preferences from observed wealth in a single realization. Rev. Financ. Stud. **10**(1), 151–174 (1997). (http://www.jstor.org/stable/2962259)
27. D'Acunto, F., Rossi, A.G.: Robo-advising. In: Rau, R., Wardrop, R., Zingales, L. (eds.) The Palgrave Handbook of Technological Finance, pp. 725-749. Palgrave Macmillan, Cham (2021). https://doi.org/10.1007/978-3-030-65117-6_26
28. D'Acunto, F., Prabhala, N., Rossi, A.G.: The promises and pitfalls of robo-advising. Rev. Financ. Stud. **32**(5), 1983–2020 (2019)
29. Ekeland, I., Lazrak, A.: The golden rule when preferences are time inconsistent. Math. Financ. Econ. **4**, 29–55 (2010)
30. Karoui, N., Mrad, M.: Recover dynamic utility from observable process: application to the economic equilibrium. SIAM J. Financ. Math. **12**(1), 189–225 (2021)

31. El Karoui, N., Hillairet, C., Mrad, M.: Construction of an aggregate consistent utility, without pareto optimality. Application to long-term yield curve modeling. In: Frontiers in Stochastic Analysis–BSDEs, SPDEs and Their Applications: Edinburgh, July 2017 Selected, Revised and Extended Contributions 8, pp. 169–199. Springer (2019)
32. Karoui, N., Hillairet, C., Mrad, M.: Bi-revealed utilities in a defaultable universe: a new point of view on consumption. Probab. Uncertainty Quant. Risk **9**(1), 13–34 (2024)
33. Finn, C., Christiano, P., Abbeel, P., Levine, S.: A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. arXiv Preprint arXiv:1611.03852 (2016)
34. Fu, J., Luo, K., Levine, S.: Learning robust rewards with adverserial inverse reinforcement learning. In: International Conference on Learning Representations (2018)
35. Garg, D., Chakraborty, S., Cundy, C., Song, J., Ermon, S.: IQ-learn: Inverse soft-Q learning for imitation. Adv. Neural Inf. Process. Syst. **34**, 4028–4039 (2021)
36. Haarnoja, T., Tang, H., Abbeel, P., Levine, S.: Reinforcement learning with deep energy-based policies. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, pp. 1352–1361. JMLR.org (2017)
37. He, H., Huang, C.-F.: Consumption-portfolio policies: an inverse optimal problem. J. Econ. Theory **62**(2), 257–293 (1994)
38. He, X.D., Strub, M.S., Zariphopoulou, T.: Forward rank-dependent performance criteria: time-consistent investment under probability distortion. Math. Financ. **31**(2), 683–721 (2021)
39. Hernández, C., Possamaï, D.: Me, myself and i: a general theory of non-Markovian time-inconsistent stochastic control for sophisticated agents. Ann. Appl. Probab. **33**(2), 1396–1458 (2023)
40. Ho, J., Ermon, S.: Generative adversarial imitation learning. In: Advances in Neural Information Processing Systems, pp. 4565–4573 (2016)
41. Hu, Y., Jin, H., Zhou, X.Y.: Time-inconsistent stochastic linear-quadratic control. SIAM J. Control Optim. **50**(3), 1548–1572 (2012)
42. Hu, Y., Jin, H., Zhou, X.Y.: Time-inconsistent stochastic linear-quadratic control: characterization and uniqueness of equilibrium. SIAM J. Control Optim. **55**(2), 1261–1279 (2017)
43. Jin, H., Yu Zhou, X.: Behavioral portfolio selection in continuous time. Math. Financ.: Int. J. Math. Stat. Financ. Econ. **18**(3), 385–426 (2008)
44. Källblad, S.: Black's inverse investment problem and forward criteria with consumption. SIAM J. Financ. Math. **11**(2), 494–525 (2020)
45. Källblad, S., Obłój, J., Zariphopoulou, T.: Dynamically consistent investment under model uncertainty: the robust forward criteria. Financ. Stoch. **22**(4), 879–918 (2018)
46. Kalman, R.E.: When is a linear control system optimal? J. Basic Eng. **86**(1), 51–60 (1964)
47. Karnam, C., Ma, J., Zhang, J.: Dynamic approaches for some time-inconsistent optimization problems. Ann. Appl. Probab. **27**(6), 3435–3477 (2017). https://doi.org/10.1214/17-AAP1284
48. Keeney, R.L., Raiffa, H.: Decisions with Multiple Objectives: Preferences and Value Tradeoffs. Wiley, New York(1976)
49. Kerimkulov, B., Leahy, J.-M., Siska, D., Szpruch, L., Zhang, Y.: A Fisher-Rao gradient flow for entropy-regularised Markov decision processes in polish spaces. arXiv Preprint arXiv:2310.02951 (2023)
50. Kim, K., Garg, S., Shiragur, K., Ermon, S.: Reward identification in inverse reinforcement learning. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, Volume 139 of Proceedings of Machine Learning Research, pp. 5496–5505. PMLR, 18–24 (2021). http://proceedings.mlr.press/v139/kim21c.html
51. Levine, S., Popovic, Z., Koltun, V.: Nonlinear inverse reinforcement learning with gaussian processes. Adv. Neural Inf. Process. Syst. **24**, 19–27 (2011)
52. Linos, P.L.: Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations part I: the dynamic programming principle and application. Commun. Partial Differ. Equ. **8**(10), 1101–1174 (1983). https://doi.org/10.1080/03605308308820297
53. Lions, P.L.: Optimal stochastic control of diffusion type processes and Hamilton-Jacobi-Bellman equations. In: Fleming, W.H., Gorostiza, L.G. (eds.) Advances in Filtering and Optimal Stochastic Control, p. 199–215, Berlin, Heidelberg. Springer Berlin Heidelberg (1982). ISBN 978-3-540-39517-1
54. Merton, R.C.: Optimum consumption and portfolio rules in a continuous-time model. J. Econ. Theory **3**(4), 373–413 (1971). https://doi.org/10.1016/0022-0531(71)90038-X
55. Mohammed, A.: Open banking and APIs: research on how open banking frameworks and APIs are reshaping the financial ecosystem. Int. J. Adv. Eng. Manag. **7**, 770–784 (2024)

56. Monin, P.: On a dynamic adaptation of the distribution builder approach to investment decisions. Quant. Financ. **14**(5), 749–760 (2014)
57. Musiela, M., Zariphopoulou, T.: Investments and forward utilities. Preprint (2006)
58. Musiela, M., Zariphopoulou, T.: Investment and Valuation Under Backward and Forward Dynamic Exponential Utilities in a Stochastic Factor Model, pp. 303–334. Birkhäuser, Boston (2007). ISBN 978-0-8176-4545-8. https://doi.org/10.1007/978-0-8176-4545-8_16
59. Musiela, M., Zariphopoulou, T.: Initial investment choice and optimal future allocations under time-monotone performance criteria. Int. J. Theor. Appl. Financ. **14**(01), 61–81 (2011)
60. Ng, A.Y., Russell, S., et al.: Algorithms for inverse reinforcement learning. In: ICML, vol. 1, p. 2 (2000)
61. Nicole, E.K., Mohamed, M.: An exact connection between two solvable SDEs and a nonlinear utility stochastic PDE. SIAM J. Financ. Math. **4**(1), 697–736 (2013). https://doi.org/10.1137/10081143X
62. Pham, H.: Continuous-Time Stochastic Control and Optimization with Financial Applications, vol. 61. Springer Science & Business Media, Berlin (2009)
63. Pollak, R.A.: Consistent planning1. Rev. Econ. Stud. **35**(2), 201–208 (1968). https://doi.org/10.2307/2296548
64. Reddy, S., Dragan, A.D., Levine, S.: SQIL: imitation learning via reinforcement learning with sparse rewards. arXiv Preprint arXiv:1905.11108 (2019)
65. Reher, M., Sokolinski, S.: Robo advisors and access to wealth management. J. Financ. Econ. **155**, 103829 (2024)
66. Richesson, R., Vehik, K.: Patient registries: utility, validity and inference. Adv. Exp. Med. Biol. **686**, 87–104 (2010). https://doi.org/10.1007/978-90-481-9485-8_6
67. Richter, M.K.: Revealed preference theory. Econometrica **34**(3), 635–645 (1966)
68. Rossi, A.G., Utkus, S.P.: Who benefits from robo-advising? Evidence from machine learning (2020)
69. Russell, S.: Learning agents for uncertain environments. In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory, pp. 101–103 (1998)
70. Samuelson, P.A.: Consumption theory in terms of revealed preference. Economica **15**(60), 243–253 (1948)
71. Sargent, T.J.: Estimation of dynamic labor demand schedules under rational expectations. J. Polit. Econ. **86**(6), 1009–1044 (1978)
72. Schlaginhaufen, A., Kamgarpour, M.: Identifiability and generalizability in constrained inverse reinforcement learning. In: International Conference on Machine Learning, pp. 30224–30251. PMLR (2023)
73. Sharpe, W.F., Goldstein, D.G., Blythe, P.W.: The distribution builder: a tool for inferring investor preferences. Preprint (2000)
74. Shin, J., Yu, J.: Targeted advertising and consumer inference. Mark. Sci. **40**(5), 900–922 (2021)
75. Strotz, R.H.: Myopia and inconsistency in dynamic utility maximization. Rev. Econ. Stud. **23**(3), 165–180 (1955)
76. Tertilt, M., Scholz, P.: To advise, or not to advise—how robo-advisors evaluate the risk preferences of private investors. J. Wealth Manag. **21**(2), 70–84 (2018)
77. Vila, J.-L., Zariphopoulou, T.: Optimal consumption and portfolio choice with borrowing constraints. J. Econ. Theory **77**(2), 402–431 (1997)
78. Wang, H., Yu, S.: Robo-advising: enhancing investment with inverse optimization and deep reinforcement learning. In: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 365–372. IEEE (2021)
79. Wulfmeier, M., Ondruska, P., Posner, I.: Maximum entropy deep inverse reinforcement learning. arXiv Preprint arXiv:1507.04888 (2015)
80. Yong, J.: Time-inconsistent optimal control problems and the equilibrium HJB equation. Math. Control Relat. Fields **2**(3), 271–329 (2012). https://doi.org/10.3934/mcrf.2012.2.271
81. Zariphopoulou, T.: Consumption-investment models with constraints. SIAM J. Control Optim. **32**(1), 59–85 (1994)
82. Zariphopoulou, T.: A solution approach to valuation with unhedgeable risks. Financ. Stoch. **5**, 61–82 (2001)
83. Zeng, S., Li, C., Garcia, A., Hong, M.: Maximum-likelihood inverse reinforcement learning with finite-time guarantees. Adv. Neural Inf. Process. Syst. **35**, 10122–10135 (2022)
84. Ziebart, B.D.: Modeling purposeful adaptive behavior with the principle of maximum causal entropy. PhD Thesis, Carnegie Mellon University (2010)

85. Ziebart, B.D., Maas, A.L., Bagnell, J.A., Dey, A.K., et al.: Maximum entropy inverse reinforcement learning. In: AAAI, Chicago, IL, USA, vol. 8, pp. 1433–1438 (2008)

Authors and Affiliations

Haoyang Cao[1] · Zhengqi Wu[2] · Renyuan Xu[3]

✉ Renyuan Xu
ryxu@stanford.edu

Haoyang Cao
hycao@jhu.edu

Zhengqi Wu
zhengqiw@usc.edu

[1] Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, USA

[2] Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, USA

[3] Management Science and Engineering, Stanford University, Stanford, USA