Inference of Utilities and Time Preference in Sequential Decision-Making

Haoyang Cao [*] Zhengqi Wu [†] Renyuan Xu [†]

June 3, 2024

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 of a time-varying rate, tailored to each client's risk tolerance, valuation of daily consumption, and significant life goals. We address the resulting time inconsistency issue through state augmentation and the establishment of the 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 log-likelihood function is locally concave, facilitating the 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.

1 Introduction

Automated investment managers, commonly known as robo-advisors, have emerged as a modern alternative to traditional financial advisors in recent years [14, 23, 53]. The effectiveness and viability of robo-advisors depend significantly on their ability to provide customized financial guidance tailored to the unique needs of each client. To provide impactful personalized advice, two critical steps must be undertaken: first, accurately estimate the clients 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. On the other hand, inferring a client's investment preferences is typically challenging, as it involves several complex

[*]Department of Applied Mathematics and Statistics, Johns Hopkins University. **Email:** hycao@jhu.edu. H. C. is partially supported by the departmental startup fund.

[†]Epstein Department of Industrial and Systems Engineering, University of Southern California. **Email:** {zhengqiw,renyuanx}@usc.edu. R. X. is partially supported by the NSF CAREER award DMS-2339240 and a JP Morgan Faculty Research Award.

aspects that vary from individual to individual. For example, clients may have short-term or long-term investment objectives [37]. Additionally, they might exhibit varying utility functions [45, 63], reflecting distinct risk tolerance related to profit-and-loss (PnL) outcomes and valuation of daily consumption. Furthermore, individuals often demonstrate diverse time preferences in terms of the trade-off between immediate and deferred outcomes [6]. Finally, clients may have specific life goals [13], 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 [19, 58]. Healthcare professionals use inferred utility functions to evaluate patient preferences regarding different treatment options, which is essential for effective healthcare management and policy-making [15, 51]. 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 [10, 16].

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 the client to define their risk tolerance related to the PnL outcomes and their valuation of daily consumption. Additionally, it allows for a generic discounting scheme under a time-varying rate, enabling the clients to balance immediate and deferred outcomes. This time-varying discounting scheme 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. The control problem is time inconsistent under the generic time-varying discounting scheme. We address this issue by state augmentation to account for the cumulative discount rate. 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, 6, 5 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. To demonstrate the effectiveness of this procedure, along with tractable theoretical guarantees, we focus on a specific case within the discrete-time Markov Decision Process (MDP), featuring Shannon's entropy regularization over an infinite-time horizon. The discrete-time MDP is especially relevant in the context of statistical inference and machine learning. The entropy term encourages full exploration of the state-action space and simultaneously introduces smoothness into the analysis [32]. 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 log-likelihood function is locally concave near $(\bar{\rho}, \bar{\theta})$; 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 [38] 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. [11]. In fact, economists have long been interested in such questions within the context of determining utility functions from observations such as Samuelson [55] and Richter [52]. For instance, Keeney and Raiffa [40] studied the proper rank of actions based on some deterministic evaluations under a static setting. Sargent [56] later extended this question into a dynamic setting where the actions were specified as labor demand and evaluations as wages. Dybvig and Rogers [21] 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.

Cox et al. [17] studied the utility inference problem for the optimal consumption and allocation of wealth in continuous time by observing the actions of the client. The authors observed that there are infinitely many utility functions generating a given consumption pattern when the dynamic is deterministic and the consumption and investment strategies have to satisfy a consistency condition in the stochastic setting. El Karoui and Mrad [25] took a "forward-looking" perspective of the connection between the observable process $\{\mathscr{X}_t(x) > 0 : \mathscr{X}_0(x) = x > 0\}$ (i.e., the characteristic process) and the corresponding utility process $\{U(t, z) : z > 0, U(0, z) = u(z)\}$ (i.e., the dynamic utility); see the concept forward utility proposed by Musiela and Zariphopoulou [46]. Different than the backward-looking perspective where the connection between the observable and utility is governed by some Markovian decision-making rule, the authors interpreted such a connection as the martingale property of the process $\{U(t, \mathscr{X}_t(x)\}$, since Markov property no longer existed under the forward-looking viewpoint. To fully explore the concavity of utilities, the authors introduced an adjoint process of $\mathscr{X}$, $\{Y_t(y) : Y_0(y) = y\}$, representing the decreasing marginal utility $\{U_z(t, z)\}$ so that $\{Y_t(u_z(x)) = U_z(t, \mathscr{X}_t(x))\}$. Given the initial utility u, the observable process $\mathscr{X}$ and its adjoint process Y, the authors fully characterized the martingale dynamic utility and its dual form via the Itô-Ventzel formula and showed that they are solutions to some Hamilton-Jacobi-Bellman-type stochastic partial differential equations; this set of analytical tools was introduced in [48] and [26]. In [27], the authors also extended the result of [25] to allow an exogenous default time τ.

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. [47] 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 GAIL (generative adversarial imitation learning) algorithm [34]. 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. [66] studied the maximum entropy IRL based on known features. They assumed that the reward is a linear function of such features. Ziebart [65] extended this approach to a selected set of non-linear rewards; see also [42] and [10] for similar settings. Wulfmeier et al. [61] followed this approach but with rewards represented by neural networks. Finn et al. [28] combined the idea of adversarial training and IRL. They trained a discriminator to recover the reward function. Reddy et al. [50] proposed a soft Q imitation learning algorithm to imitate the expert's policy by learning her Q function. Garg et al. [31] proposed an algorithm to learn the soft Q function which implicitly represents both the reward function and the policy. Zeng et al. [64] 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 [54] 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 and assumptions are required. Under an entropy regularized MDP setting, Cao et al. [12] 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 [12] 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 [30]. Both Cao et al. [12] and Kim et al. [41] provided sufficient structural conditions for the MDP environment that guarantee identifiability.

However, as pointed out by Schlaginhaufen and Kamgarpour [57], 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 [20] 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.

Time inconsistency in stochastic control. Unlike assuming an exponential discounting scheme for the client, a general discounting scheme may lead to a *time-inconsistent* policy. In economics, one of the earliest studies on the inconsistency in dynamic utility maximization is [59], where the optimality of the problem derived today is different from that of tomorrow due to some non-exponential discounting mechanism. Later Pollak [49] proposed a game-theoretic consistent planning approach 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, [7, 8, 24, 35, 36, 62], and more recently, [18, 33]. Apart from this game-theoretic approach, Karnam et al. [39] introduced the idea of "dynamic utility" to a family of time-inconsistent 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 different way to revive DPP; see Propositions 1 and 4 in Section 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 [14, 22, 23]. Here, we mainly review some papers that explore the machine learning and inference aspects of this subject. 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 [60] 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. Her 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*. This could possibly lead to time-inconsistent decision-making. First, for the *optimal control* problem, we analyze the time-inconsistent dynamical decision-making problem for such a client, assuming the client's utility functions of consumption and wealth as well as her 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 her optimal joint consumption-allocation plan is disclosed. Such an identifiability result provides inspirations for the algorithm proposed in Section 3.

2.1 Finite-time Horizon

We first focus on a finite-time horizon setting, with a decision horizon T, to address scenarios where the client has a short-term investment plan.

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, 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

$$\mathrm{d}S_t^0 = r\mathrm{d}t, \tag{1}$$

and the price of the stock follows

$$\mathrm{d}S_t = S_t(\mu \mathrm{d}t + \sigma \mathrm{d}W_t). \tag{2}$$

Assume the client choose 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 making consumption c_t to achieve certain satisfaction in life.

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

$$\begin{aligned}\mathcal{A} := \Big\{ (\boldsymbol{\alpha}, \boldsymbol{c})\{(\alpha_t, c_t)\}_{t\geq 0} \,\Big|\, & (\alpha_t, c_t) \in \mathcal{K},\, (\alpha_t, c_t) \in \bar{\mathcal{F}}_t := \sigma\left(\sigma(\beta_s, 0 \leq s \leq t) \times \mathcal{F}_t^W\right), \\ & \mathbb{E}\left[|X_t^{\boldsymbol{\alpha},\boldsymbol{c}}|^2\right] < \infty,\, \forall t \geq 0 \Big\}\end{aligned} \tag{3}$$

as the admissible set of all possible joint consumption-allocation processes. Hence the wealth process follows:

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

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

$$\begin{aligned}\mathcal{U} := \Big\{ U : \mathbb{R} \to [-\infty, +\infty) \,\Big|\, & U \text{ is strictly positive, increasing and concave on } (0, +\infty), \\ & \text{there exists a sufficiently large constant } C \in \mathbb{R}^+ \text{ such that} \\ & |U(x)| \leq C\left(1 + x^2\right) \text{ for all } x \in (0, \infty), \\ & U \in \mathcal{C}^2((0, +\infty)),\, U(x) = U(0) \text{ for } x \leq 0 \Big\}.\end{aligned} \tag{5}$$

Here U_1 quantifies the the client's evaluation regarding the consumption whereas U_2 quantifies her evaluation regarding the terminal wealth at the end of the investment plan.

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 bounded and 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.

Then for any $(t, x, z) \in [0, T] \times \mathbb{R} \times [0, 1]$, define the total reward as

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

subject to the wealth process (4) and

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

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 also aim to recover the DPP to the above time-inconsistent utility optimization problem (8), where the time-inconsistency is particularly due to the general discounting scheme. We take a different approach than the BSDE characterization of dynamic utility in [39]; instead, we extend the state space to incorporate the discounting process (similar to [5]) and then re-establish DPP accordingly.

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.

Lemma 1. *Assume that $U_1, U_2 \in \mathcal{U}$. Moreover, assume that $U_1(0) = 0$ and $U_2(0) = -\infty$. For any $(t, x, z) \in [0, T] \times (0, \infty) \times [0, 1]$, if the policy $\boldsymbol{\alpha}^*, \boldsymbol{c}^*$ satisfies that $J(t, x, z, \boldsymbol{\alpha}^*, \boldsymbol{c}^*) = V(t, x, z)$, then it holds almost surely that*

$$X_s^{\boldsymbol{\alpha}^*, \boldsymbol{c}^*} \in (0, \infty) \quad \textit{for all} \quad s \in [t, T], \tag{9}$$

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

$$X_T^{\boldsymbol{\alpha}, \boldsymbol{c}} \leq X_T^{\boldsymbol{\alpha}, \boldsymbol{0}} \leq 0,$$

and hence $J(t, x, z, \boldsymbol{\alpha}, \boldsymbol{c}) = -\infty$. On the other hand,

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

Then if $J(t, x, z, \boldsymbol{\alpha}^*, \boldsymbol{c}^*) = V(t, x, z)$, then $X_s^{\boldsymbol{\alpha}^*, \boldsymbol{c}^*} \in (0, \infty)$ for all $s \in [t, T]$ almost surely. $\square$

Lemma 2. *Assume that $U_1, U_2 \in \mathcal{U}$. Moreover, assume that $U_1(0) = 0$ and $U_2(0) = -\infty$. Then it holds that the value function $V : [0,T] \times \mathbb{R} \times [0,1] \to [-\infty, +\infty)$ defined in (8) is strictly concave and strictly increasing in $x \in (0,\infty)$ given any $(t,z) \in [0,T] \times (0,1]$.*

Proof. Fix any $(t,x,z) \in [0,T] \times (0,\infty) \times (0,1]$.

1. *Strictly concave and positive.* Take $y \in (0,\infty)\backslash\{x\}$ and $\lambda \in (0,1)$. Define $u = \lambda x + (1-\lambda)y$ and $u \in (0,\infty)$. Take any $(\boldsymbol{\alpha}^x, \boldsymbol{c}^x), (\boldsymbol{\alpha}^y, \boldsymbol{c}^y) \in \mathcal{A}$ and define $(\boldsymbol{\alpha}^u, \boldsymbol{c}^u) = \lambda(\boldsymbol{\alpha}^x, \boldsymbol{c}^x) + (1-\lambda)(\boldsymbol{\alpha}^y, \boldsymbol{c}^y)$. Then it immediately follows that $(\boldsymbol{\alpha}^u, \boldsymbol{c}^u) \in \mathcal{A}$. Let $X^{\boldsymbol{\alpha}^u, \boldsymbol{c}^u}$ (resp. $X^{\boldsymbol{\alpha}^x, \boldsymbol{c}^x}$ or $X^{\boldsymbol{\alpha}^y, \boldsymbol{c}^y}$) be the solution to the SDE (4) over $[t,T]$ given $(\boldsymbol{\alpha}, \boldsymbol{c}) = (\boldsymbol{\alpha}^u, \boldsymbol{c}^u)$ (resp. $(\boldsymbol{\alpha}, \boldsymbol{c}) = (\boldsymbol{\alpha}^x, \boldsymbol{c}^x)$ or $(\boldsymbol{\alpha}, \boldsymbol{c}) = (\boldsymbol{\alpha}^y, \boldsymbol{c}^y)$) and $X_t^{\boldsymbol{\alpha}, \boldsymbol{c}} = u$ (resp. $X_t^{\boldsymbol{\alpha}, \boldsymbol{c}} = x$ or $X_t^{\boldsymbol{\alpha}, \boldsymbol{c}} = y$). Then we have

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

   By Lemma 1, we can assume that both $X_s^{\boldsymbol{\alpha}^x, \boldsymbol{c}^x}$ and $X_s^{\boldsymbol{\alpha}^y, \boldsymbol{c}^y}$ are strictly positive for $s \in [t,T]$ almost surely. Since $U_1, U_2 \in \mathcal{U}$, then

$$J(t,u,z,\boldsymbol{\alpha}^u, \boldsymbol{c}^u) > \lambda J(t,x,z,\boldsymbol{\alpha}^x, \boldsymbol{c}^x) + (1-\lambda) J(t,y,z,\boldsymbol{\alpha}^y, \boldsymbol{c}^y) > 0.$$

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

$$V(t,u,z) > \lambda V(t,x,z) + (1-\lambda) V(t,y,z) > 0.$$

2. *Strictly increasing.* 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,

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

   and therefore

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

   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 on U_1, U_2 as in Lemma 1. 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*

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

with $V(T,x,z) = zU_2(x)$.

Proof. Fix any $(t, x, z) \in [0, T) \times \mathbb{R} \times [0, 1]$, $(\boldsymbol{\alpha}, \boldsymbol{c}) \in \mathcal{A}$ and $\tau \in \mathbb{T}_t$. We have

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

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

$$J(t+\Delta t, x, z, \boldsymbol{\alpha}^{\epsilon, t+\Delta t, x, z}, \boldsymbol{c}^{\epsilon, t+\Delta t, x, z}) > \sup_{(\boldsymbol{\alpha}, \boldsymbol{c}) \in \mathcal{A}} J(t+\Delta t, x, z, \boldsymbol{\alpha}, \boldsymbol{c}) - \epsilon = V(t+\Delta t, x, z) - \epsilon, \tag{10}$$

and

$$J(t+\Delta t, x, z, \boldsymbol{\alpha}, \boldsymbol{c}) \leq J(t+\Delta t, x, z, \boldsymbol{\alpha}^{\epsilon, t+\Delta t, x, z}, \boldsymbol{c}^{\epsilon, t+\Delta t, x, z}) \leq V(t+\Delta t, x, z). \tag{11}$$

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

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

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 (10) and (11), for any $(\boldsymbol{\alpha}, \boldsymbol{c}) \in \mathcal{A}$, we have

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

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 $(t, x, z) \in [0, T) \times \mathbb{R} \times [0, 1]$ and $\tau \in \mathbb{T}_t$,

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

$\square$

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 $\phi \in \mathcal{C}_b^\infty([0,T)\times\mathbb{R}\times\mathbb{R}^+)\bigcap\mathcal{C}_b^0([0,T)\times\mathbb{R}\times\mathbb{R}^+)$. Following the DPP under a generic discounting scheme (DPP), we have the following Hamilton-Jacobi-Bellman (HJB) equation

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

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

Proposition 2. *Take the same assumptions on U_1, U_2 as in Lemma 1. Let $w : [0,T]\times\mathbb{R}\times[0,1] \to \mathbb{R}$ be a function such that*

$$w \in \mathcal{C}^{1,2,1}\left([0,T]\times\mathbb{R}\times[0,1]\right)\bigcap\mathcal{C}^0\left([0,T]\times\mathbb{R}\times[0,1]\right),$$

and there exists a constant $C > 0$ with

$$w(t,x,z) \le C(1+|x|^2), \quad \forall(t,x,z)\in[0,T]\times\mathbb{R}\times[0,1].$$

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

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

   Then $w \ge V$ on $[0,T]\times\mathbb{R}\times[0,1]$.

2. *Assume further that there exists $\hat{\alpha} : [0,T]\times\mathbb{R}\times[0,1] \to [-M,M]$ and $\hat{c} : [0,T]\times\mathbb{R}\times[0,1] \to [0,M]$ such that*

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

   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,

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

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

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

   Then $w = V$ on $[0,T]\times\mathbb{R}\times[0,1]$, with $(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}})$ being an optimal joint allocation-consumption process.

Proof. The assumptions on U_1 and U_2 guarantee a quadratic growth rate in x. Consider arbitrary $(t,x,z)\in[0,T)\times\mathbb{R}\times[0,1]$ and $(\boldsymbol{\alpha},\boldsymbol{c})\in\mathcal{A}$.

1. Define
$$\tau_n := \inf\left\{s\geq t\middle|\int_t^s|\partial_x w(u,X_u^{\boldsymbol{\alpha},\boldsymbol{c}},\beta_u)|^2du\geq n\right\},\quad \forall n\in\mathbb{N}^+.$$
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
$$\begin{aligned}&\mathbb{E}\left[w(s\wedge\tau_n,X_{s\wedge\tau_n}^{\boldsymbol{\alpha},\boldsymbol{c}},\beta_{s\wedge\tau_n})\middle|X_t^{\boldsymbol{\alpha},\boldsymbol{c}}=x,\beta_t=z\right]=w(t,x,z)+\mathbb{E}\left[\int_t^{s\wedge\tau_n}\partial_t w(u,X_u^{\boldsymbol{\alpha},\boldsymbol{c}},\beta_u)\right.\\ &\quad\left.+\dot\beta_u\partial_z w(u,X_u^{\boldsymbol{\alpha},\boldsymbol{c}},\beta_u)+rX_u^{\boldsymbol{\alpha},\boldsymbol{c}}\partial_x w(u,X_u^{\boldsymbol{\alpha},\boldsymbol{c}},\beta_u)+\mathcal{L}^{\alpha_u,c_u}w(u,X_u^{\boldsymbol{\alpha},\boldsymbol{c}},\beta_u)du\middle|X_t^{\boldsymbol{\alpha},\boldsymbol{c}}=x,\beta_t=z\right]\\ &\quad\leq w(t,x,z)-\mathbb{E}\left[\int_t^{s\wedge\tau_n}\beta_u U_1(c_u)du\middle|X_t^{\boldsymbol{\alpha},\boldsymbol{c}}=x,\beta_t=z\right],\end{aligned}$$
where the well-posedness of $\mathbb{E}\left[\int_t^{s\wedge\tau_n}\beta_u U_1(c_u)du\middle|X_t^{\boldsymbol{\alpha},\boldsymbol{c}}=x,\beta_t=z\right]$ 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
$$\begin{aligned}&\mathbb{E}\left[\beta_T U_2(X_T^{\boldsymbol{\alpha},\boldsymbol{c}})\middle|X_t^{\boldsymbol{\alpha},\boldsymbol{c}}=x,\beta_t=z\right]\leq\mathbb{E}\left[w(T,X_T^{\boldsymbol{\alpha},\boldsymbol{c}},\beta_T)\middle|X_t^{\boldsymbol{\alpha},\boldsymbol{c}}=x,\beta_t=z\right]\\ &\leq w(t,x,z)-\mathbb{E}\left[\int_t^T\beta_u U_1(c_u)du\middle|X_t^{\boldsymbol{\alpha},\boldsymbol{c}}=x,\beta_t=z\right]\\ &\Longrightarrow w(t,x,z)\geq\mathbb{E}\left[\int_t^T\beta_u U_1(c_u)du+\beta_T U_2(X_T^{\boldsymbol{\alpha},\boldsymbol{c}})\middle|X_t^{\boldsymbol{\alpha},\boldsymbol{c}}=x,\beta_t=z\right]=J(t,x,z,\boldsymbol{\alpha},\boldsymbol{c}).\end{aligned}$$
Hence, $w(t,x,z)\geq V(t,x,z)$ 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 $s\in[t,T]$
$$\begin{aligned}&\mathbb{E}\left[w(s,X_s^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}},\beta_s)\middle|X_t=x,\beta_t=z\right]=w(t,x,z)+\mathbb{E}\left[\int_t^s\partial_t w(u,X_u^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}},\beta_u)\right.\\ &\quad\left.+\dot\beta_u\partial_z w(u,X_u^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}},\beta_u)+rX_u^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}}\partial_x w(u,X_u^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}},\beta_u)+\mathcal{L}^{\hat\alpha_u,\hat c_u}w(u,X_u^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}},\beta_u)du\middle|X_t^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}}=x,\beta_t=z\right]\\ &\quad=w(t,x,z)-\mathbb{E}\left[\int_t^s\beta_u U_1(\hat c_u)du\middle|X_t^{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}}=x,\beta_t=z\right].\end{aligned}$$

In particular, when $s = T$,

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

Then we have $w(t,x,z) = J(t,x,z,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}) \leq V(t,x,z)$. Combined with the result from the previous part, we have $w = V$ on $[0,T] \times \mathbb{R} \times \mathbb{R}^+$, with $(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{c}}) \in \mathcal{A}$ being a corresponding optimal joint allocation-consumption process.

□

Without assuming the existence of a classical solution to (HJB), we could instead consider its viscosity solution.

Definition 1 (Viscosity solution). *Denote $\mathcal{D} = [0,T) \times (0,\infty) \times [0,1]$.*

1. *A lower semi-continuous function $\underline{v} : \mathcal{D} \to \mathbb{R}$ is a viscosity subsolution to (HJB) if for any $(t_0, x_0, z_0) \in \mathcal{D}$ and any $\phi \in \mathcal{C}^{1,2,1}(\mathcal{D})$ 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 neighborhood $B(t_0,x_0,z_0) \subset \mathcal{D}$,

$$-\partial_t\phi(t_0,x_0,z_0) - \dot{\beta}_t\partial_z\phi(t_0,x_0,z_0) - rx_0\partial_x\phi(t_0,x_0,z_0) - \sup_{(\alpha,c)\in\mathcal{K}}\left\{z_0U_1(c) + \mathcal{L}^{\alpha,c}\phi(t_0,x_0,z_0)\right\} \leq 0. \tag{12}$$

2. *An upper semi-continuous function $\overline{v} : \mathcal{D} \to \mathbb{R}$ is a viscosity supersolution to (HJB) if for any $(t_0, x_0, z_0) \in \mathcal{D}$ and any $\psi \in \mathcal{C}^{1,2,1}(\mathcal{D})$ 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 neighborhood $B(t_0,x_0,z_0) \subset \mathcal{D}$,

$$-\partial_t\psi(t_0,x_0,z_0) - \dot{\beta}_t\partial_z\psi(t_0,x_0,z_0) - rx_0\partial_x\psi(t_0,x_0,z_0) - \sup_{(\alpha,c)\in\mathcal{K}}\left\{z_0U_1(c) + \mathcal{L}^{\alpha,c}\psi(t_0,x_0,z_0)\right\} \geq 0. \tag{13}$$

3. *A continuous function $v : \mathcal{D} \to \mathbb{R}$ 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 on U_1, U_2 as in Lemma 1. 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, $i = 1, 2$.*

Proof. First, notice that under the assumptions on admissible control specified in (3) as well as those on the utility functions specified in (5), the continuity of the value function V in (8) over the domain $\mathcal{D}$ can be established following the classical results of [43, 44]. Therefore, V is bounded and uniformly continuous on $\bar{\mathcal{D}}$. 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 a classical comparison principal [29, Theorem V9.1]. □

Given V being a $\mathcal{C}^{1,2,1}([0,T)\times(0,\infty)\times[0,1])\bigcap\mathcal{C}([0,T]\times(0,\infty)\times[0,1])$, define the Hamiltonian as

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

Then, we have that for any $(t,x,z)\in[0,T)\times(0,\infty)\in[0,1]$,

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

where

$$\alpha_t^*=-\frac{\partial_xV(t,x,z)}{\partial_x^2V(t,x,z)}\cdot\frac{(\mu-r)}{\sigma^2x}\wedge M, \tag{14}$$

$$c_t^*=\operatorname{argmax}_{c\in[0,M]}\Big\{-\partial_xV(t,x,z)c+zU_1(c)\Big\}; \tag{15}$$

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 provides her 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

$$(\bar{\alpha},\bar{c}):[0,T]\times(0,\infty)\times[0,1]\to\mathcal{K} \tag{16}$$

be some allocation and consumption policies of a client such that

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

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

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

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

$$(\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{c}})\in\operatorname{argmax}_{(\boldsymbol{\alpha},\boldsymbol{c})\in\mathcal{A}}J(t,x,z,\boldsymbol{\alpha},\boldsymbol{c}),\quad\forall(t,x,z)\in[0,T]\times(0,\infty)\times[0,1]$$

subject to (4) and (7) . Write

$$\bar{V}(t,x,z)=J(t,x,z,\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{c}}),\quad\forall(t,x,z)\in[0,T]\times(0,\infty)\times[0,1]. \tag{17}$$

Note that the inference agent has full access to (16).

Theorem 1 (Identifiability)**.** *Assume that*

1. $\bar{\alpha}, \bar{c} \in \mathcal{C}^{1,1,1}([0,T) \times (0,\infty) \times [0,1]) \bigcap \mathcal{C}^0([0,T] \times (0,\infty) \times [0,1])$;
2. $\bar{\alpha}(t,x,z) \in (0,M)$ *for all* $(t,x,z) \in [0,T] \times (0,\infty) \times [0,1]$;
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*
$$\bar{\alpha}(T,x,z) \equiv \bar{\alpha}_T(x), \quad \bar{c}(T,x,z) \equiv \bar{c}_T(x), \quad \forall (x,z) \in (0,\infty) \times [0,1];$$
5. *for* $x \in (0,\infty)$, $\bar{\alpha}_T(x) > 0$, $\bar{c}_T$ *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.*

Remark 1. *This result is also consistent with the finding in [12] 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. Assumptions 2 and 5 enunciate the precise dependency of value function* $\bar{V}$ *in* (17) *and the observed policy* $(\bar{\alpha}, \bar{c})$.

Proof. First, by Assumption 2, for all $(t,x,z) \in [0,T] \times (0,\infty) \times [0,1]$, (HJB) is equivalent to

$$\begin{cases} \partial_t V(t,x,z) + \dot{\beta}_t \partial_z V(t,x,z) + rx\partial_x V(t,x,z) \\ \qquad\qquad - \dfrac{(r-\mu)^2}{2\sigma^2} \dfrac{[\partial_x V(t,x,z)]^2}{\partial_x^2 V(t,x,z)} = zU_1^*\left(\dfrac{\partial_x V(t,x,z)}{z}\right), \quad t \in [0,T); \\ V(T,x,z) = zU_2(x), \end{cases} \tag{18}$$

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 (17) from (18). By (14) and Assumption 2,

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

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 (15), 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 (14)–(15),

$$\begin{cases} \bar{\alpha}(t, x, z) = -\dfrac{\partial_x \bar{V}(t, x, z)}{\partial_x[\partial_x \bar{V}(t, x, z)]} \dfrac{\mu - r}{\sigma^2 x}, \\ \partial_x \bar{V}(t, x, z) = z U_1'(\bar{c}(t, x, z)); \end{cases} \implies \partial_x \bar{V}(t, x, z) = K_1(t, z) \exp\left\{ \int_x^1 \frac{\mu - r}{\sigma^2 y \bar{\alpha}(t, y, z)} dy \right\},$$

where

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

Rewrite (18) as

$$\partial_t \bar{V}(t, x, z) + \dot{\beta}_t \partial_z \bar{V}(t, x, z) = -\left\{ \left[r + \frac{\bar{\alpha}(t, x, z)(\mu - r)}{2} \right] x - \bar{c}(t, x, z) \right\} \partial_x \bar{V}(t, x, z) - z U_1(\bar{c}(t, x, z)).$$

Differentiating with respect to x on both sides, we have

$$\dot{\beta}_t = -\frac{\partial_t[\partial_x \bar{V}(t, x, z)] + \partial_x \left\{ \left\{ \left[r + \frac{\bar{\alpha}(t, x, z)(\mu - r)}{2} \right] x - \bar{c}(t, x, z) \right\} \partial_x \bar{V}(t, x, z) + z U_1(\bar{c}(t, x, z)) \right\}}{\partial_z[\partial_x \bar{V}(t, x, z)]}.$$

□

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 $U_2(x) = \dfrac{x^\theta}{\theta}$ with $0 < \theta < 1$. Also set the constant M such that $M > \dfrac{\mu - r}{\sigma^2}$. In this case, we face a classic control problem that is time-consistent. 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:

$$V(t, x) = \sup_{\alpha \in \mathcal{A}} \mathbb{E}\Big[U_2(X_T) \;\Big|\; X_t = x \Big]. \tag{19}$$

The value function satisfies the following HJB equation:

$$-\partial_t V - \sup_{\alpha \in \mathcal{A}} \Big[\mathcal{L}^\alpha V(t, x) \Big] = 0, \tag{20}$$

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

$$\bar{\alpha}(t, x) = -M \vee \left(-\frac{(\mu - r)\partial_x V}{x \sigma^2 \partial_x^2 V} \right) \wedge M. \tag{21}$$

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*

$$-\frac{\partial V}{\partial t} = xr\partial_x V - \frac{1}{2}\frac{(\mu-r)(\partial_x V)^2}{\sigma^2\partial_x^2 V} \tag{22}$$

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:*

$$\phi'(t) = \rho\phi(t) = 0, \quad \phi(T) = 1, \tag{23}$$

where $\rho = p \times \sup_{\alpha\in[-M,M]}\left[\alpha(\mu-r) + r - \frac{1}{2}a^2(1-p)\sigma^2\right]$. *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 $t > 0$.

For any $(t,x,z) \in [0,\infty)\times\mathbb{R}\times[0,1]$, define the total reward function as

$$J_\infty(t,x,z,\boldsymbol{\alpha},\boldsymbol{c}) := \mathbb{E}\left[\int_t^\infty \beta_s U_1(c_s)ds \middle| X_t = x, \beta_t = z\right] \tag{24}$$

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

$$V_\infty(t,x,z) = \sup_{(\boldsymbol{\alpha},\boldsymbol{c})\in\mathcal{A}} J_\infty(t,x,z,\boldsymbol{\alpha},\boldsymbol{c}),\ t\in[0,t);\quad \lim_{t\to\infty} V_\infty(t,x,z) = 0, \tag{25}$$

subject to (4) and (7).

It is easy to show that the value function V_∞ in (25) will have similar results as specified in Section 2.1 therefore here we state these results without proofs. First, we have the necessary condition for V_∞.

Proposition 4. *For any* $(t,x,z) \in [0,\infty)\times\mathbb{R}\times[0,1]$ *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 (25) satisfies*

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

The corresponding HJB equation is given by

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

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

Proposition 5. *If the value function V_∞ in (25) is jointly continuous on $\mathcal{D}' = [0,\infty)\times(0,\infty)\times[0,1]$, 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 $U_1 : [0,\infty) \to \mathbb{R}^+ \in \mathcal{U}$ continuous at 0. Let $w : [0,\infty)\times\mathbb{R}\times[0,1] \to \mathbb{R}$ be a function such that*

$$w \in \mathcal{C}^{1,2,1}\left([0,\infty)\times\mathbb{R}\times[0,1]\right),$$

and there exists a constant $C>0$ with

$$w(t,x,z) \leq C(1+|x|^2), \quad \forall (t,x,z)\in[0,\infty)\times\mathbb{R}\times[0,1].$$

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

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

   Then $w \geq V_\infty$ on $[0,\infty)\times\mathbb{R}\times[0,1]$.

2. *Assume further that there exists $\hat{\alpha} : [0,\infty)\times\mathbb{R}\times[0,1] \to [-M,M]$ and $\hat{c} : [0,\infty)\times\mathbb{R}\times[0,1] \to [0,M]$ such that*

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

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

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

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

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

   Then $w = V_\infty$ on $[0,\infty)\times\mathbb{R}\times[0,1]$, 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

$$\bar{\alpha}^*(t,x,z) = -M \vee -\frac{\partial_x V_\infty(t,x,z)}{\partial_x^2 V_\infty(t,x,z)} \cdot \frac{(\mu - r)}{\sigma^2 x} \wedge M, \tag{26}$$

$$\bar{c}^*(t,x,z) = \operatorname{argmax}_{c\in[0,M]} \Big\{ -\partial_x V_\infty(t,x,z)c + zU_1(c) \Big\}. \tag{27}$$

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

Theorem 2. *Assume that*

1. $\bar{\alpha}, \bar{c} \in \mathcal{C}^{1,1,1}([0,\infty)\times(0,\infty)\times[0,1]) \bigcap \mathcal{C}^0([0,\infty)\times(0,\infty)\times[0,1])$;
2. $\bar{\alpha}(t,x,z) \in (-M, M)$ *for all* $(t,x,z) \in [0,T]\times(0,\infty)\times[0,1]$;
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* $\bar{c}_0(\cdot) \coloneqq \bar{c}(t_0,\cdot,z_0)$ *is invertible, and the following difference for any* $(t,x,z) \in [0,T)\times(0,\infty)\times(0,1]$,
$$\Delta(t,x,z) \coloneqq \int_x^1 \frac{dy}{y\bar{\alpha}(t,y,z)} - \int_{\bar{c}_0^{-1}(\bar{c}(t,x,z))}^1 \frac{dy}{y\bar{\alpha}_T(y)}$$
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 Discrete-time MDP with Entropy Regularization

The continuous-time framework in Section 2 emphasizes the well-definedness of the mathematical framework when the client is subject to a generic discounting scheme with a time-varying rate, and it outlines conditions necessary for ensuring identifiability for both the utility functions and the discounting scheme. Building on these insights, this section explores a practical scenario focusing on the inference procedure. We adopt a parametric framework in which the client utilizes an exponential discounting scheme, parameterized by $\bar{\rho}$, alongside a utility function parameterized by $\bar{\theta} \in \mathbb{R}^d$. 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 $(\bar{\rho}, \bar{\theta})$. This analysis is conducted within a discrete-time MDP setting under the regularization of a Shannon entropy type. It encourages the client to fully explore the state-action space and introduces smoothness to the analysis at the same time; see [32], for instance.

Mathematically, let us consider the entropy regularized MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$, which can be finite or infinite. The state process follows $s_{t+1} \sim P(\cdot|s_t, a_t)$, with $P : \mathcal{S}\times\mathcal{A} \to \mathcal{P}(\mathcal{A})$ 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 will use the notation $\sum\!\!\!\!\!\!\int$ to denote the summation or integration over the action space, emphasizing that our framework accommodates both finite and infinite action spaces.

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

$$Q^*_{\rho,\theta}(s,a) = \max_{\pi} \mathbb{E}^{\pi}\left[\sum_{t=1}^{\infty} \rho^t \Big(U_\theta\big(R(s_t,a_t)\big) + \mathcal{H}\big(\pi(\cdot|s_t)\big)\Big) \middle| s_0 = s, a_0 = a, a_t \sim \pi(s_t)\right], \tag{28}$$

where $\mathcal{H}\big(\pi(\cdot|s)\big) := -\sum_{a\in\mathcal{A}} \pi(a|s)\log\big(\pi(a|s)\big)$ is the Shannon's entropy. The optimal policy is then given by:

$$\pi_{\rho,\theta}(a|s) = \frac{e^{Q^*_{\rho,\theta}(s,a)}}{\sum_{a\in\mathcal{A}} e^{Q^*_{\rho,\theta}(s,a')}}, \tag{29}$$

and the soft Bellman equation holds:

$$V^*_{\rho,\theta}(s) = \log\Big(\sum_{a\in\mathcal{A}} e^{Q^*_{\rho,\theta}(s,a)}\Big). \tag{30}$$

3.1 Maximum Likelihood Estimation

With a trajectory $\tau = \{(s_t,a_t)\}_{t=0}^{\infty}$ following the client's policy $\pi_{\bar\rho,\bar\theta}$, 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 $\tau = \{(s_t,a_t)\}_{t=0}^{\infty}$ following the client's policy $\pi_{\bar\rho,\bar\theta}$ is defined as

$$\begin{aligned}
&\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\left[\log\left(\prod_{t=0}^{\infty}\big(P(s_{t+1}|s_t,a_t)\pi_{\rho,\theta}(a_t|s_t)\big)^{\gamma^t}\right)\right] \\
&\quad = \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{\infty}\gamma^t\log\pi_{\rho,\theta}(a_t|s_t)\Big] + \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{\infty}\gamma^t\log P(s_{t+1}|s_t,a_t)\Big],
\end{aligned} \tag{31}$$

where the notation $\tau \sim \pi_{\bar\rho,\bar\theta}$ represents that the trajectory τ is sampled from applying 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 2. *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 [9]. For example, Zeng et al. [64] studied the IRL problem using a maximum likelihood estimator by setting $\gamma = \rho$ in (31) 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:

$$\max_{(\rho,\theta)\in\Theta} \quad \mathcal{L}(\rho,\theta) := \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{\infty}\gamma^t\log\pi_{\rho,\theta}(a_t|s_t)\Big], \tag{32}$$

where $\pi_{\rho,\theta}$ is the optimal policy under the preference parameter (ρ,θ) defined in (29). Here for simplicity we set $\Theta := (0,1)\times\mathbb{R}^d$. The maximum likelihood 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 $(\bar\rho,\bar\theta)$, which is also referred to as the inverse problem. 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 likelihood function is concave near $(\bar\rho,\bar\theta)$ (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*

$$\nabla_\theta \mathcal{L}(\rho,\theta)\Big|_{(\rho,\theta)=(\bar\rho,\bar\theta)} = 0, \quad \nabla_\rho \mathcal{L}(\rho,\theta)\Big|_{(\rho,\theta)=(\bar\rho,\bar\theta)} = 0. \tag{33}$$

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 provide some useful formulas regarding the first-order and second-order derivatives of Q, V with respect to (ρ,θ). With such formulas, we next derive the derivatives of the log-likelihood function. Finally, we show that (33) holds.

Step 1. To begin with, for any $(\rho,\theta)\in\Theta$, for any $(s_t,a_t)\in\mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
\nabla_\theta Q_{\rho,\theta}(s_t,a_t) &= \nabla_\theta U_\theta(R(s_t,a_t)) + \rho\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}\Big[\nabla_\theta V_{\rho,\theta}(s_{t+1})\Big] \\
&= \nabla_\theta U_\theta(R(s_t,a_t)) + \rho\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}\Big[\nabla_\theta \log(\sum_a e^{Q_{\rho,\theta}(s_{t+1},a)})\Big] \qquad (34)\\
&= \nabla_\theta U_\theta(R(s_t,a_t)) + \rho\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}\Big[\sum_a \pi_{\rho,\theta}(a|s_{t+1})\nabla_\theta Q_{\rho,\theta}(s_{t+1},a)\Big] \qquad (35)\\
&= \nabla_\theta U_\theta(R(s_t,a_t)) + \rho\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t),a_{t+1}\sim\pi_{\rho,\theta}(\cdot|s_{t+1})}\Big[\nabla_\theta Q_{\rho,\theta}(s_{t+1},a_{t+1})\Big], \qquad (36)
\end{aligned}$$

where (34) holds by the soft Bellman equation and (35) holds because $\pi_{\rho,\theta}$ is the optimal policy. Applying (36) recursively yields:

$$\nabla_\theta Q_{\rho,\theta}(s_t,a_t) = \mathbb{E}_{\tau'\sim\pi_{\rho,\theta}}\Big[\sum_{k=t}^{\infty}\rho^{k-t}\nabla_\theta U_\theta(R(s'_k,a'_k))\Big|s_t,a_t\Big], \tag{37}$$

where $\tau' = \{s'_k,a'_k\}_{k=0}^{\infty}$ denotes a trajectory following $\pi_{\rho,\theta}$. Similarly,

$$\begin{aligned}
\nabla_\rho Q_{\rho,\theta}(s_t,a_t) &= \nabla_\rho U_\theta(R(s_t,a_t)) + \nabla_\rho(\rho\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}\Big[V_{\rho,\theta}(s_{t+1})\Big]) \qquad (38)\\
&= \mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}\Big[V_{\rho,\theta}(s_{t+1})\Big] + \rho\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}\Big[\nabla_\rho \log(\sum_a e^{Q_{\rho,\theta}(s_{t+1},a)})\Big] \qquad (39)\\
&= \mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}\Big[V_{\rho,\theta}(s_{t+1})\Big] + \rho\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}\Big[\sum_a \pi_{\rho,\theta}(a|s_{t+1})\nabla_\rho Q_{\rho,\theta}(s_{t+1},a)\Big] \qquad (40)\\
&= \mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}\Big[V_{\rho,\theta}(s_{t+1})\Big] + \rho\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t),\ a_{t+1}\sim\pi_{\rho,\theta}(\cdot|s_{t+1})}\Big[\nabla_\rho Q_{\rho,\theta}(s_{t+1},a_{t+1})\Big], \qquad (41)
\end{aligned}$$

where (39) holds by the Bellman equation and (40) holds because $\pi_{\rho,\theta}$ is the optimal policy. Applying (41) recursively yields:

$$\nabla_\rho Q_{\rho,\theta}(s_t,a_t) = \mathbb{E}_{\tau'\sim\pi_{\rho,\theta}}\Big[\sum_{k=t+1}^{\infty}\rho^{k-t-1}V_{\rho,\theta}(s'_k)\Big|s_t,a_t\Big]. \tag{42}$$

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

$$\nabla_\theta V_{\rho,\theta}(s_t) = \nabla_\theta (\log \sum_a e^{Q_{\rho,\theta}(s_t,a)})$$

$$= \mathbb{E}_{a_t \sim \pi_{\rho,\theta}(\cdot|s_t)} \Big[\nabla_\theta Q_{\rho,\theta}(s_t, a_t) \Big] = \mathbb{E}_{\tau' \sim \pi_{\rho,\theta}} \Big[\sum_{k=t}^{\infty} \rho^{k-t} \nabla_\theta U_\theta(R(s'_k, a'_k)) \Big| s_t \Big], \tag{43}$$

where the last equation holds by (37). In addition,

$$\nabla_\rho V_{\rho,\theta}(s_t) = \nabla_\rho (\log \sum_a e^{Q_{\rho,\theta}(s_t,a)})$$

$$= \mathbb{E}_{a_t \sim \pi_{\rho,\theta}(\cdot|s_t)} \Big[\nabla_\rho Q_{\rho,\theta}(s_t, a_t) \Big] = \mathbb{E}_{\tau' \sim \pi_{\rho,\theta}} \Big[\sum_{k=t+1}^{\infty} \rho^{k-t-1} V_{\rho,\theta}(s'_k) \Big| s_t \Big], \tag{44}$$

where the last equation holds by (42).

In summary, it holds that for any $(\rho, \theta) \in \Theta, s_t \in \mathcal{S}, a_t \in \mathcal{A}$,

$$\nabla_\theta Q_{\rho,\theta}(s_t, a_t) = E_{\tau' \sim \pi_{\rho,\theta}} \Big[\sum_{k=t}^{\infty} \rho^{k-t} \nabla_\theta U_\theta(R(s'_k, a'_k)) \Big| s_t, a_t \Big], \tag{45}$$

$$\nabla_\rho Q_{\rho,\theta}(s_t, a_t) = E_{\tau' \sim \pi_{\rho,\theta}} \Big[\sum_{k=t+1}^{\infty} \rho^{k-t-1} V_{\rho,\theta}(s'_k) \Big| s_t, a_t \Big], \tag{46}$$

$$\nabla_\theta V_{\rho,\theta}(s_t) = \mathbb{E}_{\tau' \sim \pi_{\rho,\theta}} \Big[\sum_{k=t}^{\infty} \rho^{k-t} \nabla_\theta U_\theta(R(s'_k, a'_k)) \Big| s_t \Big], \tag{47}$$

$$\nabla_\rho V_{\rho,\theta}(s_t) = \mathbb{E}_{\tau' \sim \pi_{\rho,\theta}} \Big[\sum_{k=t+1}^{\infty} \rho^{k-t-1} V_{\rho,\theta}(s'_k) \Big| s_t \Big]. \tag{48}$$

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

$$\mathcal{L}(\rho, \theta) = \mathbb{E}_{\tau \sim \pi_{\bar\rho,\bar\theta}} \Big[\sum_{t=0}^{\infty} \gamma^t \log \pi_{\rho,\theta}(a_t|s_t) \Big] = \mathbb{E}_{\tau \sim \pi_{\bar\rho,\bar\theta}} \Big[\sum_{t=0}^{\infty} \gamma^t \log \frac{e^{Q_{\rho,\theta}(s_t,a_t)}}{\sum_a e^{Q_{\rho,\theta}(s_t,a)}} \Big] \tag{49}$$

$$= \mathbb{E}_{\tau \sim \pi_{\bar\rho,\bar\theta}} \Big[\sum_{t=0}^{\infty} \gamma^t \Big(Q_{\rho,\theta}(s_t, a_t) - V_{\rho,\theta}(s_t) \Big) \Big]$$

$$= \mathbb{E}_{\tau \sim \pi_{\bar\rho,\bar\theta}} \Big[\sum_{t=0}^{\infty} \gamma^t U_\theta(R(s_t, a_t)) \Big] - \mathbb{E}_{s_0 \sim \mu(\cdot)} \Big[V_{\rho,\theta}(s_0) \Big] + (\rho - \gamma) \mathbb{E}_{\tau \sim \pi_{\bar\rho,\bar\theta}} \Big[\sum_{t=1}^{\infty} \gamma^{t-1} V_{\rho,\theta}(s_t) \Big], \tag{50}$$

where μ is the distribution of the initial state s_0. (49) holds by the optimality of the policy, and (50) holds by the soft Bellman equation. Taking the gradient of (50) with respect to θ gives

$$\nabla_\theta \mathcal{L}(\rho, \theta) = \mathbb{E}_{\tau \sim \pi_{\bar\rho,\bar\theta}} \Big[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta U_\theta(R(s_t, a_t)) \Big] - \mathbb{E}_{s_0 \sim \mu(\cdot)} \Big[\nabla_\theta V_{\rho,\theta}(s_0) \Big]$$

$$+ (\rho - \gamma) \mathbb{E}_{\tau \sim \pi_{\bar\rho,\bar\theta}} \Big[\sum_{t=1}^{\infty} \gamma^{t-1} \nabla_\theta V_{\rho,\theta}(s_t) \Big]. \tag{51}$$

Combining (51) with (47) gives:

$$\nabla_\theta \mathcal{L}(\rho,\theta) = \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta U_\theta(R(s_t,a_t))\Big] - \mathbb{E}_{\tau\sim\pi_{\rho,\theta}}\Big[\sum_{t=0}^{\infty}\rho^t \nabla_\theta U_\theta(R(s_t,a_t))\Big]$$
$$+ (\rho-\gamma)\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\mathbb{E}_{\tau'\sim\pi_{\rho,\theta}}\Big[\sum_{k=t}^{\infty}\rho^{k-t}\nabla_\theta U_\theta(R(s'_k,a'_k))\,\Big|\, s_t\Big]\Big]. \tag{52}$$

Similarly, taking the gradient of (50) with respect to ρ gives

$$\nabla_\rho \mathcal{L}(\rho,\theta) = -\mathbb{E}_{s_0\sim\mu(\cdot)}\Big[\nabla_\rho V_{\rho,\theta}(s_0)\Big] + \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}V_{\rho,\theta}(s_t)\Big]$$
$$+ (\rho-\gamma)\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\nabla_\rho V_{\rho,\theta}(s_t)\Big]. \tag{53}$$

Combining (53) with (48) yields:

$$\nabla_\rho \mathcal{L}(\rho,\theta) = \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}V_{\rho,\theta}(s_t)\Big] - \mathbb{E}_{\tau\sim\pi_{\rho,\theta}}\Big[\sum_{t=1}^{\infty}\rho^{t-1}V_{\rho,\theta}(s_t)\Big]$$
$$+ (\rho-\gamma)\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\mathbb{E}_{\tau'\sim\pi_{\rho,\theta}}\Big[\sum_{k=t+1}^{\infty}\rho^{k-t-1}V_{\rho,\theta}(s'_k)\,\Big|\, s_t\Big]\Big]. \tag{54}$$

<u>Step 3.</u> Finally, when $(\rho,\theta)=(\bar\rho,\bar\theta)$, by (52) we have

$$\nabla_\theta \mathcal{L}(\rho,\theta)|_{(\rho,\theta)=(\bar\rho,\bar\theta)} = \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta U_\theta(R(s_t,a_t))\Big] - \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{\infty}\rho^t \nabla_\theta U_\theta(R(s_t,a_t))\Big]$$
$$+(\rho-\gamma)\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\sum_{k=t}^{\infty}\rho^{k-t}\nabla_\theta U_\theta(R(s_k,a_k))\Big]. \tag{55}$$

Note that for the last line of the above equation,

$$\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\sum_{k=t}^{\infty}\rho^{k-t}\nabla_\theta U_\theta(R(s_k,a_k))\Big] = \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\Big(\frac{\gamma}{\rho}\Big)^{t-1}\sum_{k=t}^{\infty}\rho^{k-1}\nabla_\theta U_\theta(R(s_k,a_k))\Big]$$
$$= \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{k=1}^{\infty}\sum_{t=1}^{k}\Big(\frac{\gamma}{\rho}\Big)^{t-1}\rho^{k-1}\nabla_\theta U_\theta(R(s_k,a_k))\Big] \tag{56}$$
$$= \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{k=1}^{\infty}\frac{(\gamma/\rho)^k-1}{\gamma/\rho-1}\rho^{k-1}\nabla_\theta U_\theta(R(s_k,a_k))\Big] = \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{k=1}^{\infty}\frac{\gamma^k-\rho^k}{\gamma-\rho}\nabla_\theta U_\theta(R(s_k,a_k))\Big], \tag{57}$$

where (56) holds by changing the order of summations. Plugging (57) back into (55), 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*

$$\nabla^2_\theta \mathcal{L}(\rho,\theta)|_{(\rho,\theta)=(\bar\rho,\bar\theta)} = -\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^\infty \gamma^t \mathbb{V}_{a\sim\pi_{\bar\rho,\bar\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a)\Big]\Big],$$

$$\nabla^2_\rho \mathcal{L}(\rho,\theta)|_{(\rho,\theta)=(\bar\rho,\bar\theta)} = -\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^\infty \gamma^t \mathbb{V}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\rho Q_{\rho,\theta}(s_t,a)\Big]\Big],$$

$$\nabla_\theta\nabla_\rho \mathcal{L}(\rho,\theta)|_{(\rho,\theta)=(\bar\rho,\bar\theta)} = -\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^\infty \gamma^t \mathbf{Cov}_{a\sim\pi_{\bar\rho,\bar\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a),\nabla_\rho Q_{\rho,\theta}(s_t,a)\Big]\Big],$$

in which we define

$$\begin{aligned}\mathbb{V}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a)\Big] &:= \mathbb{E}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a)\nabla_\theta Q_{\rho,\theta}(s_t,a)^\top\Big] \\ &\quad - \mathbb{E}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a)\Big]\mathbb{E}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a)\Big]^\top, \end{aligned} \tag{58}$$

and

$$\begin{aligned}&\mathbf{Cov}_{a\sim\pi_{\rho,\theta}(\cdot|s_{t+1})}\Big[\nabla_\theta Q_{\rho,\theta}(s_{t+1},a),\nabla_\rho Q_{\rho,\theta}(s_{t+1},a)\Big] \\ &:= \mathbb{E}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a)\nabla_\rho Q_{\rho,\theta}(s_t,a)\Big] \\ &\qquad - \mathbb{E}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a)\Big]\mathbb{E}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\rho Q_{\rho,\theta}(s_t,a)\Big] \in \mathbb{R}^d. \end{aligned} \tag{59}$$

In addition,

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

is negative semi-definite.

Theorem 3 suggests that the log-liklihood function $\mathcal{L}(\rho,\theta)$ is concave near the client's preference parameter $(\bar\rho,\bar\theta)$. As mentioned earlier, 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 derive formulas for the second-order derivatives of Q, V 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 $(\rho,\theta)\in\Theta$ by taking the derivative of (43), we have:

$$\begin{aligned}\nabla^2_\theta V_{\rho,\theta}(s_t) &= \nabla_\theta\Big(\sum\!\!\!\!\!\!\int_a \pi_{\rho,\theta}(a|s_t)\nabla_\theta Q_{\rho,\theta}(s_t,a)^\top\Big) \\ &= \sum\!\!\!\!\!\!\int_a \nabla_\theta\Big(e^{Q_{\rho,\theta}(s_t,a)-V_{\rho,\theta}(s_t)}\Big)\nabla_\theta Q_{\rho,\theta}(s_t,a)^\top + \sum\!\!\!\!\!\!\int_a \pi_{\rho,\theta}(a|s_t)\nabla^2_\theta Q_{\rho,\theta}(s_t,a) \qquad (60) \\ &= \mathbb{E}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a)\nabla_\theta Q_{\rho,\theta}(s_t,a)^\top\Big] \\ &\quad - \mathbb{E}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a)\Big]\mathbb{E}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a)\Big]^\top + \mathbb{E}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla^2_\theta Q_{\rho,\theta}(s_t,a)\Big] \qquad (61) \\ &= \mathbb{V}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a)\Big] + \mathbb{E}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla^2_\theta Q_{\rho,\theta}(s_t,a)\Big], \qquad (62)\end{aligned}$$

where the covariance matrix $\mathbb{V}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a)\Big]$ is defined in (58). In particular, (60) holds because:

$$\pi_{\rho,\theta}(a|s_t) = \frac{e^{Q_{\rho,\theta}(s_t,a)}}{\sum\!\!\!\!\!\!\int_{a'} e^{Q_{\rho,\theta}(s,a')}} = e^{Q_{\rho,\theta}(s_t,a)-V_{\rho,\theta}(s_t)},$$

and (61) holds by (43). In addition, we have:

$$\begin{aligned}
\nabla_\theta^2 Q_{\rho,\theta}(s_t,a_t) &= \nabla_\theta^2 U_\theta(R(s_t,a_t)) + \rho\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}\Big[\nabla_\theta^2 V_{\rho,\theta}(s_{t+1})\Big] \\
&= \nabla_\theta^2 U_\theta(R(s_t,a_t)) + \rho\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}\Big[\mathbb{V}_{a_{t+1}\sim\pi_{\rho,\theta}(\cdot|s_{t+1})}\Big[\nabla_\theta Q_{\rho,\theta}(s_{t+1},a_{t+1})\Big]\Big] \\
&\quad + \rho\mathbb{E}_{\tau\sim\pi_{\rho,\theta}}\Big[\nabla_\theta^2 Q_{\rho,\theta}(s_{t+1},a_{t+1})\Big|s_t,a_t\Big],
\end{aligned} \tag{63}$$

where (63) holds by (62). Applying (63) recursively yields:

$$\begin{aligned}
\nabla_\theta^2 Q_{\rho,\theta}(s_t,a_t) &= \mathbb{E}_{\tau'\sim\pi_{\rho,\theta}}\Big[\sum_{k=t}^{\infty}\rho^{k-t}\nabla_\theta^2 U_\theta(R(s_k',a_k'))\Big|s_t,a_t\Big] \\
&\quad + \mathbb{E}_{\tau'\sim\pi_{\rho,\theta}}\Big[\sum_{k=t}^{\infty}\rho^{k-t+1}\mathbb{V}_{a\sim\pi_{\rho,\theta}(\cdot|s_{k+1}')}\Big[\nabla_\theta Q_{\rho,\theta}(s_{k+1}',a)\Big]\Big|s_t,a_t\Big].
\end{aligned} \tag{64}$$

Similarly, for any $\rho\in(0,1)$, $\theta\in\Theta$, by taking the gradient of (44) with respect to ρ, we have

$$\begin{aligned}
\nabla_\rho^2 V_{\rho,\theta}(s_t) &= \nabla_\rho\Big(\sum\!\!\!\!\!\!\int_a \pi_{\rho,\theta}(a|s_t)\nabla_\rho Q_{\rho,\theta}(s_t,a)\Big) \\
&= \sum\!\!\!\!\!\!\int_a \nabla_\rho\Big(e^{Q_{\rho,\theta}(s_t,a)-V_{\rho,\theta}(s_t)}\Big)\nabla_\rho Q_{\rho,\theta}(s_t,a) + \sum\!\!\!\!\!\!\int_a \pi_{\rho,\theta}(a|s_t)\nabla_\rho^2 Q_{\rho,\theta}(s_t,a) \\
&= \mathbb{E}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[(\nabla_\rho Q_{\rho,\theta}(s_t,a))^2\Big] - \mathbb{E}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\rho Q_{\rho,\theta}(s_t,a)\Big]^2 + \mathbb{E}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\rho^2 Q_{\rho,\theta}(s_t,a)\Big] \\
&= \mathbb{V}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\rho Q_{\rho,\theta}(s_t,a)\Big] + \mathbb{E}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\rho^2 Q_{\rho,\theta}(s_t,a)\Big].
\end{aligned} \tag{65}$$

Similarly, by taking the gradient of (38),

$$\begin{aligned}
\nabla_\rho^2 Q_{\rho,\theta}(s_t,a_t) &= \nabla_\rho\Big(\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}\Big[V_{\rho,\theta}(s_{t+1})\Big] + \rho\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}\Big[\nabla_\rho V_{\rho,\theta}(s_{t+1})\Big]\Big) \\
&= 2\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}\Big[\nabla_\rho V_{\rho,\theta}(s_{t+1})\Big] + \rho\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}\Big[\nabla_\rho^2 V_{\rho,\theta}(s_{t+1})\Big] \\
&= 2\mathbb{E}_{\tau\sim\pi_{\rho,\theta}}\Big[\nabla_\rho Q_{\rho,\theta}(s_{t+1},a_{t+1})\Big|s_t,a_t\Big] + \rho\mathbb{E}_{s_{t+1}}\Big[\mathbb{V}_{a\sim\pi_{\rho,\theta}(\cdot|s_{t+1})}\Big[\nabla_\rho Q_{\rho,\theta}(s_{t+1},a)\Big]\Big] \\
&\quad + \rho\mathbb{E}_{a_{t+1}\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\rho^2 Q_{\rho,\theta}(s_{t+1},a_{t+1})\Big|s_t,a_t\Big].
\end{aligned} \tag{66}$$

Applying (66) recursively yields:

$$\begin{aligned}
\nabla_\rho^2 Q_{\rho,\theta}(s_t,a_t) &= 2\mathbb{E}_{\tau'\sim\pi_{\rho,\theta}}\Big[\sum_{k=t}^{\infty}\rho^{k-t}\nabla_\rho Q_{\rho,\theta}(s_{k+1}',a_{k+1}')\Big|s_t,a_t\Big] \\
&\quad + \mathbb{E}_{\tau'\sim\pi_{\rho,\theta}}\Big[\sum_{k=t}^{\infty}\rho^{k-t+1}\mathbb{V}_{a\sim\pi_{\rho,\theta}(\cdot|s_{k+1}')}\Big[\nabla_\rho Q_{\rho,\theta}(s_{k+1}',a)\Big]\Big|s_t,a_t\Big].
\end{aligned} \tag{67}$$

Furthermore, for any $(\rho, \theta) \in \Theta$, by taking the gradient of (44) with respect to θ , we have

$$
\begin{aligned}
&\nabla_\theta \nabla_\rho V_{\rho,\theta}(s_t) = \nabla_\theta \Big(\sum_a \pi_{\rho,\theta}(a|s_t) \nabla_\rho Q_{\rho,\theta}(s_t, a) \Big) \\
&= \sum_a \nabla_\theta \Big(e^{Q_{\rho,\theta}(s_t,a) - V_{\rho,\theta}(s_t)} \Big) \nabla_\rho Q_{\rho,\theta}(s_t, a) + \sum_a \pi_{\rho,\theta}(a|s_t) \nabla_\theta \nabla_\rho Q_{\rho,\theta}(s_t, a) \\
&= \mathbb{E}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)} \Big[\nabla_\theta Q_{\rho,\theta}(s_t, a) \nabla_\rho Q_{\rho,\theta}(s_t, a) \Big] - \mathbb{E}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)} \Big[\nabla_\theta Q_{\rho,\theta}(s_t, a) \Big] \mathbb{E}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)} \Big[\nabla_\rho Q_{\rho,\theta}(s_t, a) \Big] \\
&\quad + \mathbb{E}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)} \Big[\nabla_\theta \nabla_\rho Q_{\rho,\theta}(s_t, a) \Big] \\
&= \mathbf{Cov}_{a\sim\pi_{\rho,\theta}(\cdot|s_{t+1})} \Big[\nabla_\theta Q_{\rho,\theta}(s_{t+1}, a), \nabla_\rho Q_{\rho,\theta}(s_{t+1}, a) \Big] \\
&\quad + \mathbb{E}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)} \Big[\nabla_\theta \nabla_\rho Q_{\rho,\theta}(s_t, a) \Big], \qquad (68)
\end{aligned}
$$

where the "covariance" between $\nabla_\theta Q$ and $\nabla_\rho Q$ is defined in (59). Note that $\mathbf{Cov}_{a\sim\pi_{\rho,\theta}(\cdot|s_{t+1})} \Big[\nabla_\theta Q_{\rho,\theta}(s_{t+1}, a), \nabla_\rho Q_{\rho,\theta}(s_{t+1}, a) \Big] \in \mathbb{R}^d$, as we have $\theta \in \mathbb{R}^d$ and $\rho \in \mathbb{R}$.

Lastly, by taking the gradient of (38), we obtain

$$
\begin{aligned}
&\nabla_\theta \nabla_\rho Q_{\rho,\theta}(s_t, a_t) \\
&= \nabla_\theta \Big(\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)} \Big[V_{\rho,\theta}(s_{t+1}) \Big] + \rho \mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)} \Big[\nabla_\rho V_{\rho,\theta}(s_{t+1}) \Big] \Big) \\
&= \mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)} \Big[\nabla_\theta V_{\rho,\theta}(s_{t+1}) \Big] + \rho \mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)} \Big[\nabla_\theta \nabla_\rho V_{\rho,\theta}(s_{t+1}) \Big] \\
&= \mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)} \Big[\nabla_\theta V_{\rho,\theta}(s_{t+1}) \Big] + \rho \mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)} \Big[\mathbf{Cov}_{a\sim\pi_{\rho,\theta}(\cdot|s_{t+1})} \Big[\nabla_\theta Q_{\rho,\theta}(s_{t+1}, a), \nabla_\rho Q_{\rho,\theta}(s_{t+1}, a) \Big] \Big] \\
&\quad + \rho \mathbb{E}_{\tau\sim\pi_{\rho,\theta}} \Big[\nabla_\theta \nabla_\rho Q_{\rho,\theta}(s_{t+1}, a) \Big| s_t, a_t \Big].
\end{aligned}
$$

Applying the last equation recursively yields:

$$
\begin{aligned}
\nabla_\theta \nabla_\rho Q_{\rho,\theta}(s_t, a_t) &= \mathbb{E}_{\tau'\sim\pi_{\rho,\theta}} \Big[\sum_{k=t}^{\infty} \rho^{k-t} \nabla_\theta V_{\rho,\theta}(s'_{k+1}) \Big| s_t, a_t \Big] \\
&\quad + \mathbb{E}_{\tau'\sim\pi_{\rho,\theta}} \Big[\sum_{k=t}^{\infty} \rho^{k-t+1} \mathbf{Cov}_{a\sim\pi_{\rho,\theta}(\cdot|s_{k+1})} \Big[\nabla_\theta Q_{\rho,\theta}(s'_{k+1}, a), \nabla_\rho Q_{\rho,\theta}(s'_{k+1}, a) \Big] \Big| s_t, a_t \Big]. \qquad (69)
\end{aligned}
$$

In summary, by combining (62), (64), (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 $(s_t,a_t)\in\mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}\nabla_\theta^2 V_{\rho,\theta}(s_t) &= \mathbb{E}_{\tau'\sim\pi_{\rho,\theta}}\Big[\sum_{k=t}^{\infty}\rho^{k-t}\nabla_\theta^2 U_\theta(R(s'_k,a'_k))\Big|s_t\Big]\\ &\quad+\mathbb{E}_{\tau'\sim\pi_{\rho,\theta}}\Big[\sum_{k=t}^{\infty}\rho^{k-t}\mathbb{V}_{a\sim\pi_{\rho,\theta}(\cdot|s'_k)}\Big[\nabla_\theta Q_{\rho,\theta}(s'_k,a)\Big]\Big|s_t\Big], \end{aligned}\tag{70}$$

$$\begin{aligned}\nabla_\rho^2 V_{\rho,\theta}(s_t) &= 2\mathbb{E}_{\tau'\sim\pi_{\rho,\theta}}\Big[\sum_{k=t}^{\infty}\rho^{k-t}\nabla_\rho Q_{\rho,\theta}(s'_{k+1},a_{k+1})\Big|s_t\Big]\\ &\quad+\mathbb{E}_{\tau'\sim\pi_{\rho,\theta}}\Big[\sum_{k=t}^{\infty}\rho^{k-t}\mathbb{V}_{a\sim\pi_{\rho,\theta}(\cdot|s'_k)}\Big[\nabla_\rho Q_{\rho,\theta}(s'_k,a)\Big]\Big|s_t\Big], \end{aligned}\tag{71}$$

$$\begin{aligned}\nabla_\theta\nabla_\rho V_{\rho,\theta}(s_t) &= \mathbb{E}_{\tau'\sim\pi_{\rho,\theta}}\Big[\sum_{k=t}^{\infty}\rho^{k-t}\nabla_\theta V_{\rho,\theta}(s'_{k+1})\Big|s_t\Big]\\ &\quad+\mathbb{E}_{\tau'\sim\pi_{\rho,\theta}}\Big[\sum_{k=t}^{\infty}\rho^{k-t}\mathbf{Cov}_{a\sim\pi_{\rho,\theta}(\cdot|s'_k)}\Big[\nabla_\theta Q_{\rho,\theta}(s'_k,a),\nabla_\rho Q_{\rho,\theta}(s'_k,a)\Big]\Big|s_t\Big]. \end{aligned}\tag{72}$$

Step 2. Next, we calculate the derivatives of the log-likelihood function. By straight-forward calculations using (51) and combining with (70), the second-order derivative of the log-likelihood function to θ satisfies:

$$\begin{aligned}\nabla_\theta^2\mathcal{L}(\rho,\theta) &= \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta^2 U_\theta(R(s_t,a_t))\Big]-\mathbb{E}_{s_0\sim\mu(\cdot)}\Big[\nabla_\theta^2 V_{\rho,\theta}(s_0)\Big]\\ &\quad+(\rho-\gamma)\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\nabla_\theta^2 V_{\rho,\theta}(s_t)\Big]\\ &= \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta^2 U_\theta(R(s_t,a_t))\Big]-\mathbb{E}_{\tau\sim\pi_{\rho,\theta}}\Big[\sum_{t=0}^{\infty}\rho^t\nabla_\theta^2 U_\theta(R(s_t,a_t))\Big]\\ &\quad-\mathbb{E}_{\tau\sim\pi_{\rho,\theta}}\Big[\sum_{t=0}^{\infty}\rho^t\mathbb{V}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a)\Big]\Big]\\ &\quad+(\rho-\gamma)\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\nabla_\theta^2 V_{\rho,\theta}(s_t)\Big]. \end{aligned}$$

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

$$\begin{aligned}\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\nabla_\theta^2 V_{\rho,\theta}(s_t)\Big] &= \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\sum_{k=t}^{\infty}\rho^{k-t}\nabla_\theta^2 U_\theta(R(s_k,a_k))\Big]\\ &\quad+\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\sum_{k=t}^{\infty}\rho^{k-t}\mathbb{V}_{a\sim\pi_{\bar\rho,\bar\theta}(\cdot|s_k)}\Big[\nabla_\theta Q_{\rho,\theta}(s_k,a)\Big]\Big]\\ &= \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{k=1}^{\infty}\frac{\gamma^k-\rho^k}{\gamma-\rho}\nabla_\theta^2 U_\theta(R(s_k,a_k))\Big]\\ &\quad+\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{k=1}^{\infty}\frac{\gamma^k-\rho^k}{\gamma-\rho}\mathbb{V}_{a\sim\pi_{\bar\rho,\bar\theta}(\cdot|s_k)}\Big[\nabla_\theta Q_{\rho,\theta}(s_k,a)\Big]\Big], \end{aligned}$$

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,

$$\begin{aligned}
\nabla^2_\theta \mathcal{L}(\rho,\theta)|_{(\rho,\theta)=(\bar\rho,\bar\theta)} &= \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{\infty}\gamma^t\nabla^2_\theta U_\theta(R(s_t,a_t))\Big] - \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{\infty}\rho^t\nabla^2_\theta U_\theta(R(s_t,a_t))\Big] \\
&\quad -\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{\infty}\rho^t\mathbb{V}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a)\Big]\Big] \\
&\quad +\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\rho^t\nabla^2_\theta U_\theta(R(s_t,a_t))\Big] - \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^t\nabla^2_\theta U_\theta(R(s_t,a_t))\Big] \\
&\quad +\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\rho^t\mathbb{V}_{a\sim\pi_{\bar\rho,\bar\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a)\Big]\Big] \\
&\quad -\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^t\mathbb{V}_{a\sim\pi_{\bar\rho,\bar\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a)\Big]\Big] \\
&= -\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{\infty}\gamma^t\mathbb{V}_{a\sim\pi_{\bar\rho,\bar\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a)\Big]\Big].
\end{aligned}$$

Similarly, using the result in (71) and (53),

$$\begin{aligned}
\nabla^2_\rho \mathcal{L}(\rho,\theta) = -2\mathbb{E}_{\tau\sim\pi_{\rho,\theta}}\Big[\sum_{t=0}^{\infty}\rho^t\nabla_\rho Q_{\rho,\theta}(s_{t+1},a_{t+1})\Big] - \mathbb{E}_{\tau\sim\pi_{\rho,\theta}}\Big[\sum_{t=0}^{\infty}\rho^t\mathbb{V}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\rho Q_{\rho,\theta}(s_t,a)\Big]\Big] \\
+2\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\nabla_\rho V_{\rho,\theta}(s_t)\Big] \\
+(\rho-\gamma)\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\nabla^2_\rho V_{\rho,\theta}(s_t)\Big]. \qquad (73)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\nabla^2_\rho V_{\rho,\theta}(s_t)\Big] &= 2\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\sum_{k=t}^{\infty}\rho^{k-t}\nabla_\rho Q_{\rho,\theta}(s_{k+1},a_{k+1})\Big] \\
&\quad +\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\sum_{k=t}^{\infty}\rho^{k-t}\mathbb{V}_{a\sim\pi_{\rho,\theta}(\cdot|s_k)}\Big[\nabla_\rho Q_{\rho,\theta}(s_k,a)\Big]\Big] \\
&= 2\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{k=1}^{\infty}\frac{\gamma^k-\rho^k}{\gamma-\rho}\nabla_\rho Q_{\rho,\theta}(s_{k+1},a_{k+1})\Big] \\
&\quad +\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{k=1}^{\infty}\frac{\gamma^k-\rho^k}{\gamma-\rho}\mathbb{V}_{a\sim\pi_{\rho,\theta}(\cdot|s_k)}\Big[\nabla_\rho Q_{\rho,\theta}(s_k,a)\Big]\Big],
\end{aligned}$$

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

$$
\begin{aligned}
&\nabla^2_\rho\, \mathcal{L}(\rho,\theta)|_{(\rho,\theta)=(\bar\rho,\bar\theta)} \\
&= -2\mathbb{E}_{\tau\sim\pi_{\rho,\theta}}\Big[\sum_{t=0}^{\infty}\rho^t\nabla_\rho Q_{\rho,\theta}(s_{t+1},a_{t+1})\Big] - \mathbb{E}_{\tau\sim\pi_{\rho,\theta}}\Big[\sum_{t=0}^{\infty}\rho^t\mathbb{V}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\rho Q_{\rho,\theta}(s_t,a)\Big]\Big] \\
&\quad +2\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\nabla_\rho V_{\rho,\theta}(s_t)\Big] \\
&\quad +2\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\rho^t\nabla_\rho Q_{\rho,\theta}(s_{t+1},a_{t+1})\Big] - 2\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^t\nabla_\rho Q_{\rho,\theta}(s_{t+1},a_{t+1})\Big] \\
&\quad +\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\rho^t\mathbb{V}_{a\sim\pi_{\bar\rho,\bar\theta}(\cdot|s_t)}\Big[\nabla_\rho Q_{\rho,\theta}(s_t,a)\Big]\Big] - \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^t\mathbb{V}_{a\sim\pi_{\bar\rho,\bar\theta}(\cdot|s_t)}\Big[\nabla_\rho Q_{\rho,\theta}(s_t,a)\Big]\Big] \\
&= 2\,\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\nabla_\rho V_{\rho,\theta}(s_t)\Big] - 2\,\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^t\nabla_\rho Q_{\rho,\theta}(s_{t+1},a_{t+1})\Big] \\
&\quad -\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^t\mathbb{V}_{a\sim\pi_{\bar\rho,\bar\theta}(\cdot|s_t)}\Big[\nabla_\rho Q_{\rho,\theta}(s_t,a)\Big]\Big] \\
&\quad -2\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\nabla_\rho Q_{\rho,\theta}(s_1,a_1)\Big] - \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\mathbb{V}_{a\sim\pi_{\rho,\theta}(\cdot|s_0)}\Big[\nabla_\rho Q_{\rho,\theta}(s_0,a)\Big]\Big] \\
&= -\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{\infty}\gamma^t\mathbb{V}_{a\sim\pi_{\bar\rho,\bar\theta}(\cdot|s_t)}\Big[\nabla_\rho Q_{\rho,\theta}(s_t,a)\Big]\Big],
\end{aligned}
$$

where the last equality holds by the expressions of $\nabla_\rho V_{\rho,\theta}$ and $\nabla_\rho Q_{\rho,\theta}$ in (43) and (44).

Similarly,

$$
\begin{aligned}
\nabla_\theta\nabla_\rho\mathcal{L}(\rho,\theta) = -\mathbb{E}_{s_0\sim\mu(\cdot)}\Big[\nabla_\theta\nabla_\rho V_{\rho,\theta}(s_0)\Big] + \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\nabla_\theta V_{\rho,\theta}(s_t)\Big] \\
+(\rho-\gamma)\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\nabla_\theta\nabla_\rho V_{\rho,\theta}(s_t)\Big].
\end{aligned} \tag{74}
$$

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

$$
\begin{aligned}
&\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\nabla_\theta\nabla_\rho V_{\rho,\theta}(s_t)\Big] \\
&= \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\sum_{k=t}^{\infty}\rho^{k-t}\nabla_\theta V_{\rho,\theta}(s_{k+1})\Big] \\
&\quad +\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\sum_{k=t}^{\infty}\rho^{k-t}\mathbf{Cov}_{a\sim\pi_{\bar\rho,\bar\theta}(\cdot|s_k)}\Big[\nabla_\theta Q_{\rho,\theta}(s_k,a),\nabla_\rho Q_{\rho,\theta}(s_k,a)\Big]\Big] \\
&= \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{k=1}^{\infty}\frac{\gamma^k-\rho^k}{\gamma-\rho}\nabla_\theta V_{\rho,\theta}(s_{k+1})\Big] \\
&\quad +\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{k=1}^{\infty}\frac{\gamma^k-\rho^k}{\gamma-\rho}\mathbf{Cov}_{a\sim\pi_{\bar\rho,\bar\theta}(\cdot|s_k)}\Big[\nabla_\theta Q_{\rho,\theta}(s_k,a),\nabla_\rho Q_{\rho,\theta}(s_k,a)\Big]\Big].
\end{aligned}
$$

The last equality holds by changing the order of summations. Plugging the last equality back to

(74) and applying (72) for $t=0$, we have

$$
\begin{aligned}
& \nabla_\theta \nabla_\rho\, \mathcal{L}(\rho,\theta)|_{(\rho,\theta)=(\bar\rho,\bar\theta)} \\
= \quad & -\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{\infty}\rho^t \nabla_\theta V_{\rho,\theta}(s_{t+1})\Big] - \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{\infty}\rho^t \mathbf{Cov}_{a\sim\pi_{\rho,\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a), \nabla_\rho Q_{\rho,\theta}(s_t,a)\Big]\Big] \\
& +\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\nabla_\theta V_{\rho,\theta}(s_t)\Big] \\
& +\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\rho^t \nabla_\theta V_{\rho,\theta}(s_{t+1})\Big] - \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^t \nabla_\theta V_{\rho,\theta}(s_{t+1})\Big] \\
& +\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\rho^t \mathbf{Cov}_{a\sim\pi_{\bar\rho,\bar\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a), \nabla_\rho Q_{\rho,\theta}(s_t,a)\Big]\Big] \\
& -\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^t \mathbf{Cov}_{a\sim\pi_{\bar\rho,\bar\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a), \nabla_\rho Q_{\rho,\theta}(s_t,a)\Big]\Big] \\
= \quad & -\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^t \mathbf{Cov}_{a\sim\pi_{\bar\rho,\bar\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a), \nabla_\rho Q_{\rho,\theta}(s_t,a)\Big]\Big] \\
& -\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\mathbf{Cov}_{a\sim\pi_{\rho,\theta}(\cdot|s_0)}\Big[\nabla_\theta Q_{\rho,\theta}(s_0,a), \nabla_\rho Q_{\rho,\theta}(s_0,a)\Big]\Big] \qquad (75) \\
= \quad & -\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{\infty}\gamma^t \mathbf{Cov}_{a\sim\pi_{\bar\rho,\bar\theta}(\cdot|s_t)}\Big[\nabla_\theta Q_{\rho,\theta}(s_t,a), \nabla_\rho Q_{\rho,\theta}(s_t,a)\Big]\Big],
\end{aligned}
$$

where (75) holds by the fact that

$$
\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}\nabla_\theta V_{\rho,\theta}(s_t)\Big] - \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=1}^{\infty}\gamma^t \nabla_\theta V_{\rho,\theta}(s_{t+1})\Big] = \mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\nabla_\theta V_{\rho,\theta}(s_1)\Big].
$$

To summarize,

$$
\begin{aligned}
\mathcal{H} \;:=\; & \begin{pmatrix} \nabla_\theta^2 \mathcal{L}(\rho,\theta) & \nabla_\theta\nabla_\rho \mathcal{L}(\rho,\theta) \\ \nabla_\theta\nabla_\rho \mathcal{L}(\rho,\theta)^\top & \nabla_\rho^2 \mathcal{L}(\rho,\theta) \end{pmatrix}\Bigg|_{(\rho,\theta)=(\bar\rho,\bar\theta)} \\
& = -\mathbb{E}_{\tau\sim\pi_{\bar\rho,\bar\theta}}\Big[\sum_{t=0}^{\infty}\gamma^t \mathbb{V}_{a\sim\pi_{\bar\rho,\bar\theta}(\cdot|s_t)}\Big[\begin{pmatrix}\nabla_\theta Q_{\rho,\theta}(s_t,a)\\ \nabla_\rho Q_{\rho,\theta}(s_t,a)\end{pmatrix}\Big]\Big].
\end{aligned}
$$

Therefore, $\mathcal{H}$ is negative-semi definite by the definition of the covariance notation $\mathbb{V}$ in (58).

□

3.2 Algorithm Design and Implementation

Motivated by the landscape analysis in Section 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_{ρ^k,θ^k} is first computed by the soft Q iteration (see e.g. [50]) in lines 3-7, and the parameters ρ^k, θ^k are then updated in line 10 using the gradient computed in line 9.

Algorithm 1 Maximum likelihood update

Initialize ρ^0, θ^0.

for $k = 1, 2, \cdots, K$ **do**

  Set $Q^0_{\rho^k,\theta^k}(s,a) = \dfrac{1}{1-\rho^k}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, and $V^0_{\rho^k,\theta^k}(s) = \dfrac{1}{1-\rho^k}$ for all $s \in \mathcal{S}$.

  for $i = 1, 2, \cdots, I$ **do**

    for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**

$$Q^i_{\rho^k,\theta^k}(s,a) = U_{\theta^k}(R(s,a)) + \rho^k\, \mathbb{E}_{s' \sim P(\cdot|s,a)}\Big[V^i_{\rho^k,\theta^k}(s')\Big].$$

    end for

    Compute $V^I_{\rho^k,\theta^k}$ using the soft Bellman equation (30).

  end for

  With the value of $V^I_{\rho^k,\theta^k}$, compute $\nabla \mathcal{L}(\rho^k, \theta^k)$ using (52) and (54).

  Update $(\rho^{k+1}, \theta^{k+1}) = (\rho^k, \theta^k) + \zeta^k \nabla \mathcal{L}(\rho^k, \theta^k)$.

end for

Numerical example one: Merton's problem We implement the discrete-time version of Merton's problem introduced in Section 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:

$$X_{t+1} - X_t = \Big[X_t(\alpha_t\nu + (1-\alpha_t)r) - c_t\Big]\Delta + X_t\alpha_t\sigma\sqrt{\Delta}B_t. \tag{76}$$

The client provides a time-homogeneous policy $\pi_{\bar{\rho},\bar{\theta}} \in \mathcal{P}(\mathcal{A})$ to the inference agent, which solves

$$\sup_{\pi} \mathbb{E}\Big[\sum_{t=1}^{\infty}(\bar{\rho})^t\Big(1 - \exp(-\bar{\theta}\, c_t)\Big) + \mathcal{H}\big(\pi(\cdot|X_t)\big)\Big)\Big] \tag{77}$$

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

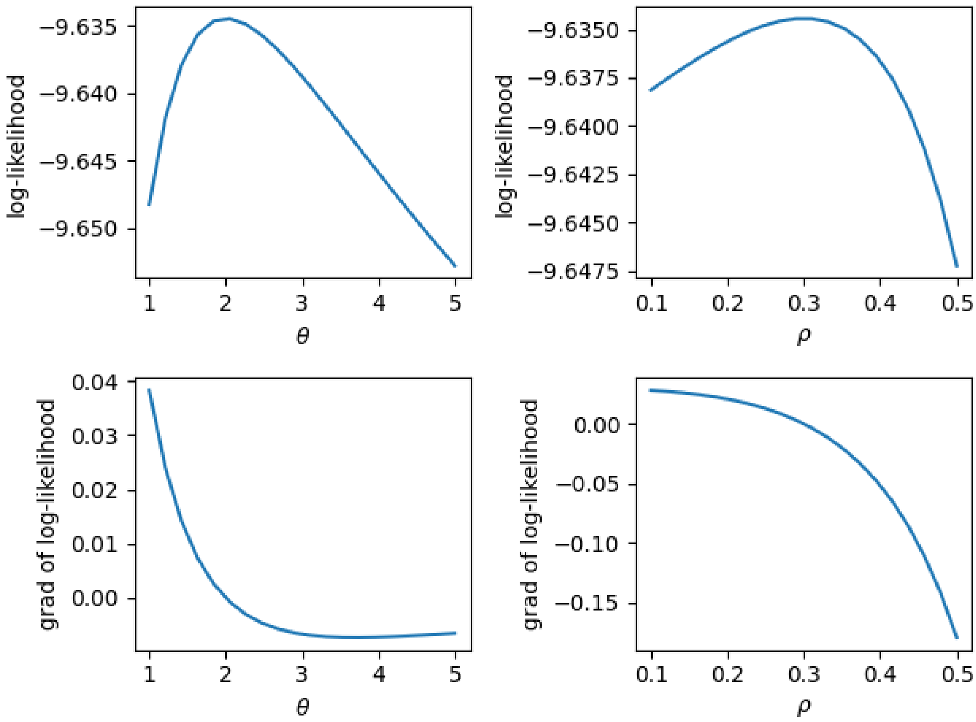

Figure 1: Visualization of the log-likelihood function and its gradients (**Left columns**: visualization with respect to θ (under $\rho = \bar{\rho}$). **Right columns**: visualization with respect to ρ (under $\theta = \bar{\theta}$)).

In the experiment, we set $\bar{\rho} = 0.3$, $\bar{\theta} = 2$, $\gamma = 0.6$, $r = 1.05$, $\Delta = 1$, $\nu = 1.06$, 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 Figure 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 θ^0 sampled uniformly from $[0, 1]$ and ρ^0 sampled uniformly from $[0.1, 0.2]$. We set the learning rate as $\zeta^k = \dfrac{1000}{k}$ and the total steps of the soft Q update as $I = 100$. As shown in Figure 2, both θ and ρ converge to the ground-truth value within 100 iterations.

Additionally, we analyze the behaviors of the client 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.

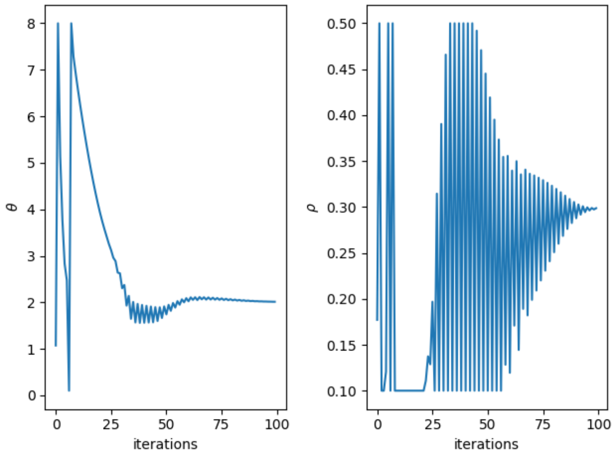

Figure 2: The convergence result of Algorithm 1. The left plot shows the value of θ at each iteration, while the right plot displays the values for ρ.

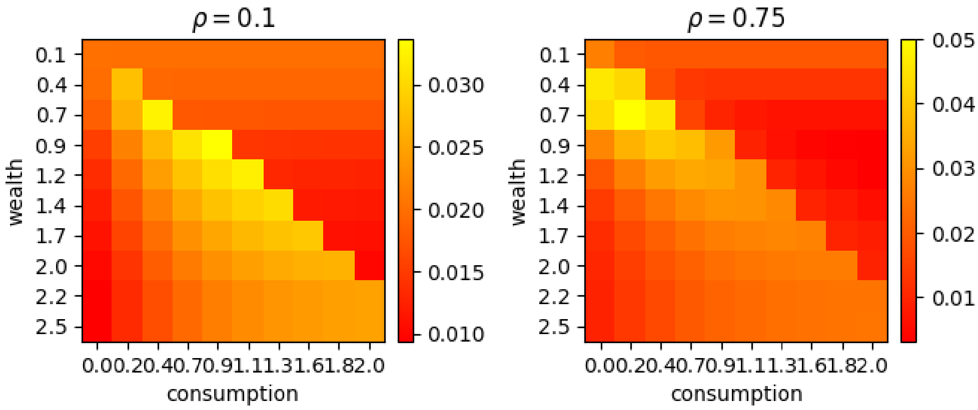

Figure 3: Visualization of the client's consumption policy. The left plot illustrates consumption at various wealth levels under $\bar{\rho} = 0.1$, while the right plot corresponds to $\bar{\rho} = 0.75$.

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 [63]. 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

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

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

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

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

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

$$X_{t+1} - X_t = \Big[X_t(\alpha_t \nu(t, Y_t) + (1-\alpha_t)r)\Big]\Delta + X_t\alpha_t\sigma(t, Y_t)\sqrt{\Delta}B_t^1, \tag{78}$$

under the investment strategy $\alpha_t \in \mathcal{A} = [0, 1]$. The client provides a time-homogeneous policy $\pi_{\bar{\rho},\bar{\theta}}(x, y) \in \mathcal{P}(\mathcal{A})$ to the inference agent, which solves

$$\sup_\pi \mathbb{E}\Big[\sum_{t=1}^{\infty}(\bar{\rho})^t \frac{1}{\bar{\theta}_1\bar{\theta}_2}(X_t)^{\bar{\theta}_1}(Y_t)^{\bar{\theta}_2} + \mathcal{H}\Big(\pi(\cdot|(X_t, Y_t))\Big)\Big] \tag{79}$$

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

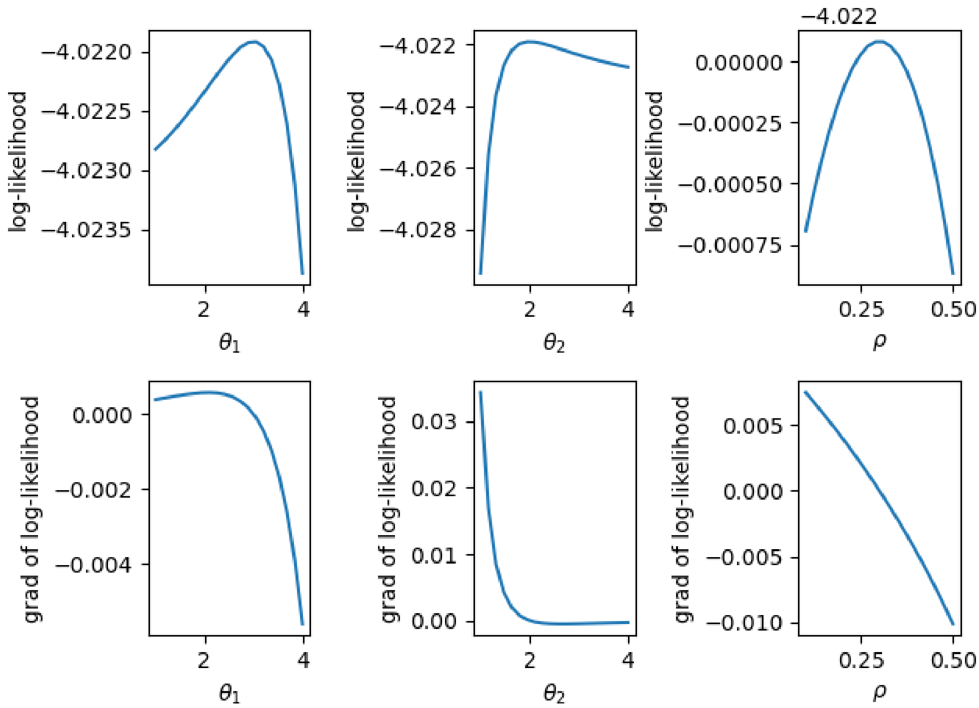

Figure 4: Visualization of the log-likelihood function and its gradients (**Left columns**: visualization with respect to θ_1 (under $\theta_2 = \bar{\theta}_2$ and $\rho = \bar{\rho}$). **Middle columns**: visualization with respect to θ_2 (under $\theta_1 = \bar{\theta}_1$ and $\rho = \bar{\rho}$).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

In the experiment, we 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 Figure 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 θ_1^0, θ_2^0 sampled uniformly from $[1, 2]$ and ρ^0 sampled uniformly from $[0.1, 0.2]$. We set the learning rate as $\zeta^k = \dfrac{1000}{\sqrt{k}}$

and the total steps of the soft Q update as $I = 100$. As shown in Figure 5, both θ_1, θ_2 and ρ converge to the ground-truth values within 1500 iterations.

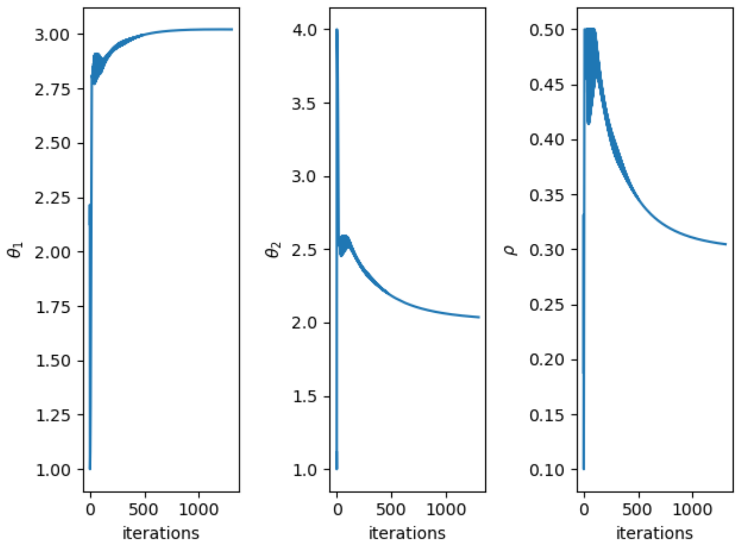

Figure 5: The convergence result of Algorithm 1. The left plot shows the value of θ_1 at each iteration, the middle plot is for θ_2, and the right plot is for ρ.

Furthermore, Figure 6 illustrates the client's investment allocation policy α across various wealth levels (under fixed factor value 1), 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 (79) strives for a higher X_t regardless of her $\bar{\rho}$. 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.

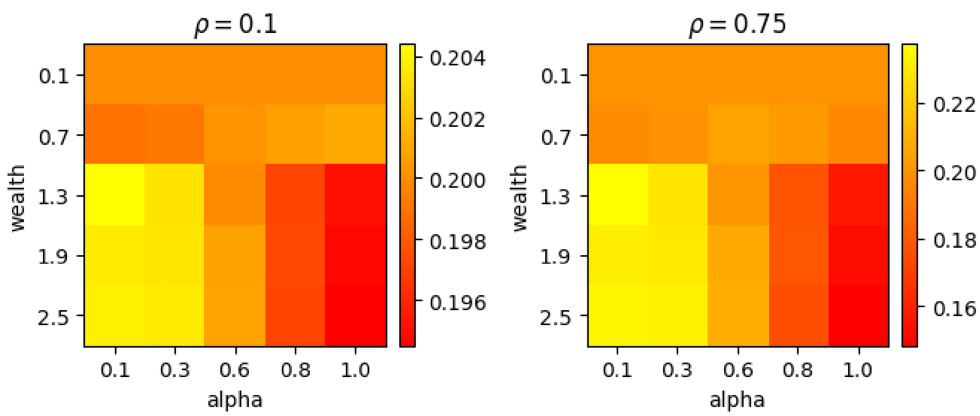

Figure 6: Visualization of the client's allocation policy (under fixed factor value 1). The left plot illustrates her allocation at various wealth levels with $\bar{\rho} = 0.1$, while the right plot is for $\bar{\rho} = 0.75$.

References

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

[2] H. Alsabah, A. Capponi, O. Ruiz Lacedelli, and M. Stern. Robo-advising: Learning investors risk preferences via portfolio choices. *Journal of Financial Econometrics*, 19(2):369–392, 2021.

[3] K. Amin and S. Singh. Towards resolving unidentifiability in inverse reinforcement learning. *arXiv preprint arXiv:1601.06569*, 2016.

[4] K. Amin, N. Jiang, and S. Singh. Repeated inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 30:1815–1824, 2017.

[5] N. Bäuerle and U. Rieder. More risk-sensitive markov decision processes. *Mathematics of Operations Research*, 39(1):105–120, 2014.

[6] T. Bjork and A. Murgoci. A general theory of markovian time inconsistent stochastic control problems. *Available at SSRN 1694759*, 2010.

[7] T. Björk and A. Murgoci. A theory of markovian time-inconsistent stochastic control in discrete time. *Finance and Stochastics*, 18:545–592, 2014.

[8] T. Björk, M. Khapko, and A. Murgoci. On time-inconsistent stochastic control in continuous time. *Finance and Stochastics*, 21:331–360, 2017.

[9] M. Bloem and N. Bambos. Infinite time horizon maximum causal entropy inverse reinforcement learning. In *53rd IEEE conference on decision and control*, pages 4911–4916. IEEE, 2014.

[10] A. Boularias, J. Kober, and J. Peters. Relative entropy inverse reinforcement learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 182–189. JMLR Workshop and Conference Proceedings, 2011.

[11] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear matrix inequalities in system and control theory*. SIAM, 1994.

[12] H. Cao, S. Cohen, and L. Szpruch. Identifiability in inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12362–12373, 2021.

[13] A. Capponi and Y. Zhang. A continuous time framework for sequential goal-based wealth management. *Management Science*, 2024.

[14] A. Capponi, S. Olafsson, and T. Zariphopoulou. Personalized robo-advising: Enhancing investment through client interaction. *Management Science*, 68(4):2485–2512, 2022.

[15] B. Chewning, C. L. Bylund, B. Shah, N. K. Arora, J. A. Gueguen, and G. Makoul. Patient preferences for shared decisions: a systematic review. *Patient education and counseling*, 86(1): 9–18, 2012.

[16] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[17] A. M. Cox, D. Hobson, and J. OBŁÓJ. Utility theory front to backinferring utility from agents'choices. *International Journal of Theoretical and Applied Finance*, 17(03):1450018, 2014.

[18] M. Dai, Y. Dong, and Y. Jia. Learning equilibrium mean-variance strategy. *Mathematical Finance*, 33(4):1166–1212, 2023.

[19] C. Derbaix and P. V. Abeele. Consumer inferences and consumer preferences. the status of cognition and consciousness in consumer behavior theory. *International Journal of Research in Marketing*, 2(3):157–174, 1985.

[20] C. Dong and Y. Wang. Towards generalized inverse reinforcement learning. *arXiv preprint arXiv:2402.07246*, 2024.

[21] P. H. Dybvig and L. C. G. Rogers. Recovery of preferences from observed wealth in a single realization. *The Review of Financial Studies*, 10(1):151–174, 1997. ISSN 08939454, 14657368. URL http://www.jstor.org/stable/2962259.

[22] F. DAcunto and A. G. Rossi. *Robo-advising*. Springer, 2021.

[23] F. DAcunto, N. Prabhala, and A. G. Rossi. The promises and pitfalls of robo-advising. *The Review of Financial Studies*, 32(5):1983–2020, 2019.

[24] I. Ekeland and A. Lazrak. The golden rule when preferences are time inconsistent. *Mathematics and Financial Economics*, 4:29–55, 2010.

[25] N. El Karoui and M. Mrad. Recover dynamic utility from observable process: Application to the economic equilibrium. *SIAM Journal on Financial Mathematics*, 12(1):189–225, 2021.

[26] N. El Karoui, C. Hillairet, and M. Mrad. 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*, pages 169–199. Springer, 2019.

[27] N. El Karoui, C. Hillairet, and M. Mrad. Bi-revealed utilities in a defaultable universe: A new point of view on consumption. *Probability, Uncertainty and Quantitative Risk*, 9(1):13–34, 2024.

[28] C. Finn, P. Christiano, P. Abbeel, and S. Levine. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *arXiv preprint arXiv:1611.03852*, 2016.

[29] W. H. Fleming and H. M. Soner. *Controlled Markov processes and viscosity solutions*, volume 25. Springer Science & Business Media, 2006.

[30] J. Fu, K. Luo, and S. Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *International Conference on Learning Representations*, 2018.

[31] D. Garg, S. Chakraborty, C. Cundy, J. Song, and S. Ermon. Iq-learn: Inverse soft-q learning for imitation. *Advances in Neural Information Processing Systems*, 34:4028–4039, 2021.

[32] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 13521361. JMLR.org, 2017.

[33] C. Hernández and D. Possamaï. Me, myself and i: a general theory of non-markovian time-inconsistent stochastic control for sophisticated agents. *The Annals of Applied Probability*, 33 (2):1396–1458, 2023.

[34] J. Ho and S. Ermon. Generative adversarial imitation learning. In *Advances in neural information processing systems*, pages 4565–4573, 2016.

[35] Y. Hu, H. Jin, and X. Y. Zhou. Time-inconsistent stochastic linear–quadratic control. *SIAM journal on Control and Optimization*, 50(3):1548–1572, 2012.

[36] Y. Hu, H. Jin, and X. Y. Zhou. Time-inconsistent stochastic linear-quadratic control: Characterization and uniqueness of equilibrium. *SIAM Journal on Control and Optimization*, 55 (2):1261–1279, 2017.

[37] H. Jin and X. Yu Zhou. Behavioral portfolio selection in continuous time. *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics*, 18 (3):385–426, 2008.

[38] R. E. Kalman. When Is a Linear Control System Optimal? *Journal of Basic Engineering*, 86 (1):51–60, 03 1964.

[39] C. Karnam, J. Ma, and J. Zhang. Dynamic approaches for some time-inconsistent optimization problems. *The Annals of Applied Probability*, 27(6):3435 – 3477, 2017. doi: 10.1214/17-AAP1284. URL https://doi.org/10.1214/17-AAP1284.

[40] R. L. Keeney and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Wiley, 1976.

[41] K. Kim, S. Garg, K. Shiragur, and S. Ermon. Reward identification in inverse reinforcement learning. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5496–5505. PMLR, 18–24 Jul 2021. URL http://proceedings.mlr.press/v139/kim21c.html.

[42] S. Levine, Z. Popovic, and V. Koltun. Nonlinear inverse reinforcement learning with gaussian processes. *Advances in neural information processing systems*, 24:19–27, 2011.

[43] P. L. Linos. Optimal control of diffustion processes and hamilton-jacobi-bellman equations part i: the dynamic programming principle and application. *Communications in Partial Differential Equations*, 8(10):1101–1174, 1983. doi: 10.1080/03605308308820297. URL https://doi.org/10.1080/03605308308820297.

[44] P. L. Lions. Optimal stochastic control of diffusion type processes and hamilton-jacobi-bellman equations. In W. H. Fleming and L. G. Gorostiza, editors, *Advances in Filtering and Optimal Stochastic Control*, pages 199–215, Berlin, Heidelberg, 1982. Springer Berlin Heidelberg. ISBN 978-3-540-39517-1.

[45] M. Musiela and T. Zariphopoulou. Investments and forward utilities. *preprint*, 2006.

[46] M. Musiela and T. Zariphopoulou. *Investment and Valuation Under Backward and Forward Dynamic Exponential Utilities in a Stochastic Factor Model*, pages 303–334. Birkhäuser Boston, Boston, MA, 2007. ISBN 978-0-8176-4545-8. doi: 10.1007/978-0-8176-4545-8_16. URL https://doi.org/10.1007/978-0-8176-4545-8_16.

[47] A. Y. Ng, S. Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

[48] E. K. Nicole and M. Mohamed. An exact connection between two solvable sdes and a nonlinear utility stochastic pde. *SIAM Journal on Financial Mathematics*, 4(1):697–736, 2013. doi: 10.1137/10081143X. URL https://doi.org/10.1137/10081143X.

[49] R. A. Pollak. Consistent Planning1. *The Review of Economic Studies*, 35(2):201–208, 04 1968. ISSN 0034-6527. doi: 10.2307/2296548. URL https://doi.org/10.2307/2296548.

[50] S. Reddy, A. D. Dragan, and S. Levine. Sqil: Imitation learning via reinforcement learning with sparse rewards. *arXiv preprint arXiv:1905.11108*, 2019.

[51] R. Richesson and K. Vehik. Patient registries: utility, validity and inference. *Rare diseases epidemiology*, pages 87–104, 2010.

[52] M. K. Richter. Revealed preference theory. *Econometrica*, 34(3):635–645, 1966. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1909773.

[53] A. G. Rossi and S. P. Utkus. Who benefits from robo-advising? evidence from machine learning. *Evidence from Machine Learning (March 10, 2020)*, 2020.

[54] S. Russell. Learning agents for uncertain environments. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 101–103, 1998.

[55] P. A. Samuelson. Consumption theory in terms of revealed preference. *Economica*, 15(60): 243–253, 1948. ISSN 00130427, 14680335. URL http://www.jstor.org/stable/2549561.

[56] T. J. Sargent. Estimation of dynamic labor demand schedules under rational expectations. *Journal of Political Economy*, 86(6):1009–1044, 1978.

[57] A. Schlaginhaufen and M. Kamgarpour. Identifiability and generalizability in constrained inverse reinforcement learning. In *International Conference on Machine Learning*, pages 30224–30251. PMLR, 2023.

[58] J. Shin and J. Yu. Targeted advertising and consumer inference. *Marketing Science*, 40(5): 900–922, 2021.

[59] R. H. Strotz. Myopia and inconsistency in dynamic utility maximization. *The Review of Economic Studies*, 23(3):165–180, 1955. ISSN 00346527, 1467937X. URL http://www.jstor.org/stable/2295722.

[60] H. Wang and S. Yu. Robo-advising: Enhancing investment with inverse optimization and deep reinforcement learning. In *2021 20th IEEE international conference on machine learning and applications (ICMLA)*, pages 365–372. IEEE, 2021.

[61] M. Wulfmeier, P. Ondruska, and I. Posner. Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*, 2015.

[62] J. Yong. Time-inconsistent optimal control problems and the equilibrium hjb equation. *Mathematical Control and Related Fields*, 2(3):271–329, 2012. ISSN 2156-8472. doi: 10.3934/mcrf.2012.2.271. URL https://www.aimsciences.org/article/id/0f7d3e02-f039-4ce2-92b6-0328dab062ba.

[63] T. Zariphopoulou. A solution approach to valuation with unhedgeable risks. *Finance and stochastics*, 5:61–82, 2001.

[64] S. Zeng, C. Li, A. Garcia, and M. Hong. Maximum-likelihood inverse reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems*, 35:10122–10135, 2022.

[65] B. D. Ziebart. *Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy*. PhD thesis, Carnegie Mellon University, 2010.

[66] B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.