Inverse Risk-Sensitive Reinforcement Learning

Lillian J. Ratliff, *Member, IEEE*, and Eric Mazumdar, *Student Member, IEEE*

***Abstract*—This work addresses the problem of *inverse reinforcement learning* in Markov decision processes where the decision-making agent is *risk-sensitive*. In particular, a risk-sensitive reinforcement learning algorithm with convergence guarantees that makes use of coherent risk metrics and models of human decision-making which have their origins in behavioral psychology and economics is presented. The risk-sensitive reinforcement learning algorithm provides the theoretical underpinning for a gradient-based inverse reinforcement learning algorithm that seeks to minimize a loss function defined on the observed behavior. It is shown that the gradient of the loss function with respect to the model parameters is well defined and computable via a contraction map argument. Evaluation of the proposed technique is performed on a *Grid World* example, a canonical benchmark problem.**

***Index Terms*—Autonomous systems, Markov processes, optimization, reinforcement learning.**

I. Introduction

Complex risk-sensitive behavior arising from human interaction with automation has attracted research efforts from a variety of communities including psychology, economics, engineering, and computer science. The adoption of diverse behavioral models in engineering—in particular, in learning and control—is growing due to the fact that humans are increasingly playing an integral role in automation both at the individual and societal scale. Learning accurate models of human decision-making is important for both *prediction* and *description*. For instance, control/incentive schemes need to predict human behavior as a function of external stimuli including not only potential disturbances but also the control/incentive mechanism itself. On the other hand, policy makers are interested in interpreting and describing human reactions to implemented regulations and policies.

There are many challenges to capturing representative, salient features of human decision-making, not the least of which is the fact that humans are known to behave in ways that are not completely rational. For instance, there is mounting evidence to support the fact that humans often use *reference points*—e.g., the *status quo*, former experiences, or recent expectations about the future that are otherwise perceived to be related to the decision the human is making [1], [2]. Empirical evidence also suggests that human decision-making is impacted by perceptions of the external world (exogenous factors) and their present state of mind (endogenous factors) as well as how the decision is *framed* or presented [3]. Furthermore, humans are *risk-sensitive*: they are risk-averse when close to a desired state and risk-seeking otherwise.

Approaches for integrating risk-sensitivity into algorithms for control synthesis and reinforcement learning via behavioral models have recently emerged [4]–[7]. These approaches largely assume a risk-sensitive Markov decision process (MDP) formulated based on a model that captures behavioral aspects of the human's decision-making process. We refer the problem of learning the optimal policy in this setting as the *forward* problem. Our primary interest is the so-called *inverse* problem which seeks to estimate the decision-making process given a set of demonstrations. Inverse reinforcement learning in the context of recovering policies directly (or indirectly via first learning a representation for the reward) has long been studied in the context expected utility maximization and MDPs [8], [9]. There are typically two approaches. 1) producing the value and reward functions (or at least, characterizing the space of these functions) that mimic behaviors matching that which is observed; 2) directly extracting the optimal policy from a set of demonstrations. In order to do so, a well formulated forward problem with convergence guarantees is required.

We model human decision-makers as *risk-sensitive Q-learning agents*. To capture both risk-sensitivity as well as other empirically observed behavioral decision-making traits such as loss aversion and reference point dependence, within a reinforcement learning framework, we combine behavioral psychology models of decision-making such as those from prospect theory [10] with appropriate—and computationally tractable—risk metrics that take into account such models. We construct a forward reinforcement learning framework for which we provide convergence guarantees in support of the development of an inverse reinforcement learning algorithm. We leverage the developed forward algorithm in to derive an *inverse risk-sensitive reinforcement learning* algorithm with theoretical guarantees. We show that the gradient of the loss function with respect to the model parameters is well defined and computable via a contraction map argument. We demonstrate the efficacy of the learning scheme on the canonical Grid World example.

The remainder of the paper is organized as follows. In Section II, the contributions are detailed. In Section III, the model for risk-sensitive agents is presented; we show that behavioral decision-theoretic value functions can be integrated into the decision-making framework and present a risk-sensitive Q-learning convergence result. In Section IV, we formulate the inverse reinforcement learning problem and propose a gradient–based algorithm to solve it. Illustrative examples are presented in Section V, and we conclude in Section VI.

II. Contributions and Related Work

The goal of this work is to provide a theoretical and algorithmic framework for recovering interpretable behavioral models of human decision-makers. Toward this end, the main contribution of this work is the development of a gradient-based inverse risk-sensitive reinforcement learning algorithm that enables recovery of prospect theoretic value functions and parameters of the class of coherent risk metrics—*utility-based shortfall*—that we consider.

Manuscript received July 31, 2018; revised August 1, 2018, August 2, 2018, June 6, 2019, and June 9, 2019; accepted June 22, 2019. Date of publication July 3, 2019; date of current version February 27, 2020. This work was supported by National Science Foundation Award CNS-1656873. Recommended by Associate Editor Prof. Samer S. Saab. *(Corresponding author: Lillian J. Ratliff.)*

L. Ratliff is with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195 USA (e-mail: ratliffl@uw.edu).

E. Mazumdar is with the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley, CA 94720 USA (e-mail: mazumdar@berkeley.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2019.2926674

The forward risk-sensitive reinforcement learning framework we adopt was first introduced in [11] and later refined in [5], [6], [12]. In preliminary work [12], we examined a similar risk-sensitive reinforcement learning framework to [5] and leveraged it to develop a gradient-based inverse reinforcement learning algorithm. Building on these works, we construct a new value function—ℓ–prospect value function—which is Lipschitz on the domain of interest and retains the convex–concave shape of a prospect theoretic value function. Similar to [5], we provide a convergence theorem, though with high probability due to the fact that the ℓ-prospect function leads to a reinforcement learning scheme that is a contraction on a finite radius ball. We show that the ℓ–prospect value function—along with other value functions considered in [5]—satisfies the assumptions. The assumptions of the theorem are also stated explicitly in terms of MDP parameters. Given the forward risk-sensitive reinforcement learning algorithm, we propose a gradient-based *inverse risk-sensitive reinforcement learning* algorithm for inferring the decision-making model parameters from demonstrations. We show that the gradient of the loss function with respect to the model parameters is well defined and computable via a contraction map argument.

The primary motivation for most other work on inverse risk-sensitive reinforcement learning is to recover a prescriptive model or algorithm for humans amidst autonomy so that the human can be accounted for in the design of control policies. For example, in [5], in order to learn the decision-making model the approach is to parameterize unknown quantities of interest, sample the parameter space, and use a model selection criteria (specifically, the Bayesian information criteria) to select parameters that best fit the observed behavior. We, on the other hand, derive a well-formulated gradient-based procedure for finding the value function and policy best matching the observed behavior.

In other promising work [7], the authors leverage a more expansive set of coherent risk metrics to capture risk sensitivity, yet without the focus on prospect theoretic value functions. In comparison, our approach focuses on estimating the agent's behavior and the value function which also induces the risk metric via an acceptance level set. In addition, the parameters of the value function are interpretable in terms of the degree of risk sensitivity and loss aversion. Thus, our technique supports prescriptive and descriptive analysis, both of which are important for the design of incentives and policies that takes into consideration the nuances of human decision-making behavior.

III. Risk-Sensitive Reinforcement Learning

Consider a class of finite MDPs consisting of a state space X, an admissible action space $U(x) \subset U$ for each $x \in X$, a transition kernel $P(x'|x,u)$ that denotes the probability of moving from state x to x' given action u, and a reward function $r : X \times U \times W \to \mathbb{R}$ where W is the space of bounded disturbances and has distribution $P_r(\cdot|x,u)$. Including disturbances allows us to model random rewards; we use the notation $R(x',u)$ to denote the random reward having distribution $P_r(\cdot|x,u)$.

In the classical expected utility maximization framework, the agent seeks to maximize the sum of their expected discounted reward over time by selecting a Markov policy π which is a distribution across actions for each state $x \in X$, i.e., $\pi(x) \in \Delta(U)$. For instance, given an infinite horizon MDP, the optimal policy is obtained by maximizing

$$J(x_0,\pi) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t R(x_t,u_t)\right] \tag{1}$$

with respect to π where x_0 is the initial state and $\gamma \in (0,1)$ is the discount factor.

The risk-sensitive reinforcement learning problem transforms the above problem to account for salient features of the human decision-making process such as loss aversion, reference point dependence, and risk-sensitivity. In this work, like others [5], [6], we introduce prospect theoretic value functions [10] and coherent risk metrics [13] to capture such features. Specifically, we introduce two key components, *value functions* and *valuation functions*, that capture these features. The former captures risk-sensitivity, loss-aversion, and reference point dependence in its transformation of outcome values to their value as perceived by the agent and the latter generalizes the expectation operator to more general measures of risk.

A. Value Functions

Much like the standard expected utility framework, an agent makes choices based on the value of outcomes as defined by a *value function* $v : \mathbb{R} \to \mathbb{R}$. There are a number of existing approaches to defining value functions that capture risk-sensitivity and loss aversion. These approaches derive from a variety of fields including behavioral psychology/economics, mathematical finance, and even neuroscience. One of the principal features of human decision-making is that losses are perceived more significant than a gain of equal true value—*losses loom larger than gains*. Empirically validated models that capture this affect are convex and concave in different regions of the outcome space. Prospect theory [10] is built on one such model. The prospect theoretic value function is given by

$$v(y) = \begin{cases} k_+(y-y_0)^{\zeta_+}, & y > y_0 \\ -k_-(y_0-y)^{\zeta_-}, & y \le y_0 \end{cases} \tag{2}$$

where y_0 is the *reference point* that the decision-maker compares outcomes against in determining if the decision is a loss or gain. The parameters $(k_+, k_-, \zeta_+, \zeta_-)$ control the degree of loss-aversion and risk-sensitivity; e.g., the following are risk preferences for different parameter values: 1) $0 < \zeta_+, \zeta_- < 1$ correspond to risk-averse preferences on gains and risk-seeking preferences on losses (concave in gains, convex in losses); 2) $\zeta_+ = \zeta_- = 1$ correspond to risk-neutral preferences; 3) $\zeta_+, \zeta_- > 1$ correspond to risk-averse preferences on losses and risk-seeking preferences on gains (convex in gains, concave in losses). Experimental results for a series of one-off decisions show that typically both ζ_+ and ζ_- are less than one thereby indicating that humans are risk-averse on gains and risk-seeking on losses—that is, v is concave for $y > y_0$ and convex otherwise [10], [14].

In addition to the nonlinear transformation of outcome values, the effect of under/over-weighting the likelihood of events that has been commonly observed in human behavior is modeled via *warping* of event probabilities [15], [16]. Other concepts such as framing effects, reference dependence, and loss aversion—captured, e.g., in the (k_+, k_-) parameters in (2)—have also been widely observed in experimental studies on human decision-making (see, e.g., [17]–[19]).

Motivated by the empirical evidence supporting the prospect theoretic value function and numerical considerations, which are discussed in greater detail in subsequent sections, we introduce a new value function that retains the shape of the prospect theory value function over the whole domain—convex–concave structure—while improving the performance (in terms of convergence speed) of the gradient-based inverse reinforcement learning algorithm we propose in Section IV. In particular, we define the locally Lipschitz-prospect (ℓ-prospect) value function given by

$$v(y) = \begin{cases} k_+(y-y_0+\epsilon)^{\zeta_+} - k_+\epsilon^{\zeta_+}, & y > y_0 \\ -k_-(y_0-y+\epsilon)^{\zeta_-} + k_-\epsilon^{\zeta_-}, & y \le y_0 \end{cases} \tag{3}$$

with $k_+, k_-, \zeta_+, \zeta_- > 0$ and $\epsilon > 0$, a small constant. This value function is Lipschitz continuous on a bounded domain. Moreover, the derivative of the ℓ-prospect function is bounded away from zero at the reference point. Hence, in practice it has better numerical properties. Moreover, for given parameters $(k_+, k_-, \zeta_+, \zeta_-)$, the ℓ-prospect function has the same risk-sensitivity as the prospect value function with those same parameters; as $\epsilon \to 0$ the ℓ-prospect value function approaches the prospect value function.

There are, of course, other behaviorally motivated value functions that appear in the literature beyond those from prospect theory. For example, in [6] a piecewise linear value function is considered in a risk-sensitive reinforcement learning context, and another very common example is the *entropic map*— $v(y) = \exp(\lambda y)$.

The fact that each of these value functions is defined by a small number of parameters that are highly interpretable in terms of risk-sensitivity and loss-aversion is one of the motivating factors for integrating them into a reinforcement learning framework. It is our aim to design learning algorithms that will ultimately provide the theoretical underpinnings for designing incentives and control policies taking into consideration salient features of human decision-making behavior.

B. Valuation Functions

Given environment and reward uncertainties, we model the outcome of each action as a real-valued random variable $Y(i) \in \mathbb{R}$, $i \in I$ where I denotes a finite event space and Y is the outcome of ith event with probability $\mu(i)$, where $\mu \in \Delta(I)$, the space of probability distributions on I.

Definition 1 (Valuation Function): A mapping $\mathcal{V} : \mathbb{R}^{|I|} \times \Delta(I) \to \mathbb{R}$ is called a *valuation function* if for each $\mu \in \Delta(I)$, 1) $\mathcal{V}(Y, \mu) \leq \mathcal{V}(Z, \mu)$ whenever $Y \leq Z$ (monotonic) and 2) $\mathcal{V}(Y + y\mathbf{1}, \mu) = \mathcal{V}(Y, \mu) + y$ for any $y \in \mathbb{R}$ (translation invariant).

Typically the valuation function used in MDPs is defined in terms of the expectation operation. For each state–action pair, we define $\mathcal{V}(Y|x, a) : \mathbb{R}^{|I|} \times X \times A \to \mathbb{R}$ a *valuation map* such that $\mathcal{V}_{x,a} \equiv \mathcal{V}(\cdot|x, a)$ is a valuation function where we drop the dependence on μ for simplicity of notation. If we let $\mathcal{V}^{\pi}_{x}(Y) = \sum_{a \in A(x)} \pi(a|x)\mathcal{V}_{x,a}(Y)$, (1) generalizes to

$$\tilde{J}_T(\pi, x_0) = \mathcal{V}^{\pi_0}_{x_0}\left[R(x_0, u_0) + \gamma \mathcal{V}^{\pi_1}_{x_1}\left[R(x_1, u_1) + \cdots + \gamma \mathcal{V}^{\pi_T}_{x_T}\left[R(x_T, u_T)\right]\right]\right]$$

where we define $\tilde{J}(\pi, x_0) = \lim_{T\to\infty} \tilde{J}_T(\pi, x_0)$.

Given that we intend to integrate empirically validated value functions that capture decision-making features of humans, the most appropriate class of coherent risk metrics are those induced by an acceptance level set defined in terms of a value function. Hence, we focus our attention on this particular class, members of which are often referred to as *utility-based shortfall risk metrics*. We are not the first to leverage this class of risk metrics in a similar framework; the authors of [5] take a similar approach.

To define this class of metrics, we first recall the following definition. A *monetary measure of risk* [13] is a functional $\rho : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ on the space $\mathcal{X}$ of measurable functions defined on a probability space $(\Omega, \mathcal{F}, P)$ such that $\rho(0)$ is finite, and for all $X, X' \in \mathcal{X}$, ρ satisfies the following:

1) (monotone) $X \leq X' \Longrightarrow \rho(X') \leq \rho(X)$, and
2) (translation invariant) $m \in \mathbb{R} \Longrightarrow \rho(X + m) = \rho(X) - m$.

If ρ additionally satisfies

$$\rho(\lambda X + (1-\lambda)X') \leq \lambda\rho(X) + (1-\lambda)\rho(X')$$

for $\lambda \in [0, 1]$, then it is a *convex risk measure*. A monetary measure of risk ρ induces an acceptance level set $\mathcal{A}_\rho = \{X \in \mathcal{X} |\ \rho(X) \leq 0\}$ [13, Prop. 4.6] and, conversely, an acceptance level set $\mathcal{A}$ induces a monetary measure of risk $\rho_{\mathcal{A}}(X) = \inf\{m \in \mathbb{R} |\ X + m \in \mathcal{A}\}$ [13, Prop. 4.7].

Utility-based shortfall risk is defined with respect to an *acceptance level set*. The acceptance level set $\mathcal{A} = \{X \in \mathcal{X} |\ \mathbb{E}_\mu[v(X)] \geq v_0\}$ defined in terms of a utility or value function v, where v_0 is the acceptance level induces $\rho(X) = \inf\{m \in \mathbb{R} |\ \mathbb{E}_\mu[v(X + m)] \geq v_0\}$. Given a value function v and acceptance level v_0, we use the utility-based shortfall risk metric to induce a state–action valuation function given by $\mathcal{V}_{x,u}(Y) = \sup\{z \in \mathbb{R} |\ \mathbb{E}[v(Y - z)] \geq v_0\}$, where the expectation is taken with respect to $\mu = P(x'|x, u)P_r(w|x, u)$; $\mathcal{V}_{x,u}(Y)$ has the properties outlined in Definition 1.

C. Risk-Sensitive Q-Learning Convergence

In the classical reinforcement learning framework, the Bellman equation is used to derive a Q-learning procedure. Generalizations of the Bellman equation for risk-sensitive reinforcement learning—derived, e.g., in [5], [6], [20]—have been used to formulate Q-learning procedures for the risk-sensitive reinforcement learning problem. In particular, as shown in [20], if V^* satisfies

$$V^*(x_0) = \max_{u \in U(x)} \mathcal{V}_{x,u}(R(x, u) + \gamma V^*) \tag{4}$$

then $V^* = \max_\pi \tilde{J}(\pi, x_0)$ holds for all $x_0 \in X$; moreover, a deterministic policy is optimal if $\pi^*(x) = \arg\max_{u \in U(x)} \mathcal{V}_{x,u}(R + \gamma V^*)$ [20, Th. 5.5]. Defining $Q^*(x, u) = \mathcal{V}_{x,u}(R + \gamma V^*)$ for each $(x, u) \in X \times U$, (4) becomes $Q^*(x, u) = \mathcal{V}_{x,u}(R + \gamma \max_{u \in U(x')} Q^*(x', u))$. As shown in [5, Prop. 3.1], by letting $Y = R + \gamma V^*$ and directly applying Proposition 4.104 of [13] with $z^* \equiv Q^*$, we have that

$$\mathbb{E}[v(r(x, u, w) + \gamma \max_{u' \in U(x')} Q^*(x', u') - Q^*(x, u))] = v_0$$

where the expectation is with respect to $\mu = P(x'|x, u)P_r(w|x, u)$. This leads naturally the Q-learning procedure

$$Q(x_t, u_t) \leftarrow Q(x_t, u_t) + \alpha_t(x_t, u_t)[v(y_t) - v_0] \tag{5}$$

where the nonlinear transformation v is applied to the temporal difference $y_t = r_t + \gamma \max_u Q(x_{t+1}, u) - Q(x_t, u_t)$. Transforming temporal differences avoids certain pitfalls of the reward transformation approach such as poor convergence performance.

It has been shown that under some assumptions on v and the sequence α_t, that the above Q-learning procedure converges with probability one [5, Th. 3.2]. Indeed, suppose that 1) v is strictly increasing in y, 2) there exist constants $\varepsilon, L > 0$ such that $\varepsilon \leq \frac{v(y) - v(y')}{y - y'} \leq L$ for all $y \neq y'$, and 3) there exists a $\bar{y}$ such that $v(\bar{y}) = v_0$. Then, if the nonnegative learning rates $\alpha_t(x, u)$ are such that $\sum_{t=0}^{\infty} \alpha_t(x, u) = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2(x, u) < \infty, \forall (x, u) \in X \times U$, then the procedure in (5) converges to $Q^*(x, u)$ for all $(x, u) \in X \times U$ with probability one.

The assumptions on α_t are fairly standard and the core of the convergence proof is based on the Robbins–Siegmund Theorem appearing in the seminal work [21]. On the other hand, the assumptions on the value function v are fairly restrictive, excluding many of the value functions presented in Section III-A; e.g., value functions of the form e^x and x^ζ do not satisfy the global Lipschitz condition. To address this, the Lipschitz assumption can be relaxed to a local condition assuming the rewards are bounded [5, Th. A.1]. However the result still requires the derivative to be bounded away from zero. We provide a slightly modified result that introduces conditions on v—in terms the bound on the rewards and the size of the ball on which v is Lipschitz—under which the derivative is bounded away from zero and show that the functions considered in [5] as well as the ℓ-prospect function we introduce satisfy these conditions. In addition, we provide a more streamlined proof technique that leverages a well-known fixed point theorem.

Assumption 1: The value function $v \in C^1(Y, \mathbb{R})$ satisfies the following: 1) it is strictly increasing in y and there exists a $\bar{y}$ such that $v(\bar{y}) = v_0$ and, 2) it is Lipschitz on any ball of finite radius centered at the origin.

Let $\mathcal{X}$ be a complete metric space endowed with the L_∞ norm and let $\mathcal{Q} \subset \mathcal{X}$ be the space of maps $Q : X \times U \to \mathbb{R}$. Further, define $\tilde{v} \equiv v - v_0$. We then rewrite the Q–update equation in the form

$$Q_{t+1}(x, u) = \left(1 - \frac{\alpha_t}{\alpha}\right) Q_t(x, u) + \frac{\alpha_t}{\alpha}(\alpha(v(y_t) - v_0) + Q_t(x, u))$$

where $\alpha \in (0, \min\{L^{-1}, 1\}]$ and we have suppressed the dependence of α_t on (x, u). This is a standard update equation form in, e.g., the stochastic approximation algorithm literature [22], [23]. In addition,

we define the map

$$(TQ)(x,u) = \alpha \mathbb{E}_{x',w}[\tilde{v}(r(x,u,w) + \gamma \max_{u' \in U(x')} Q(x',u') - Q(x,u))] + Q(x,u). \quad (6)$$

For any given $K > 0$ and $M > 0$, we use the notation I_K for the interval $[-M-2K, M+2K]$. Moreover, for any given K such that $0 < K < \infty$, let $\alpha \in (0, \min\{1, L^{-1}\}]$, where L is the Lipschitz constant of v on I_K.

Theorem 1: Suppose v satisfies Assumption 1 and for each $(x,u) \in X \times U$ the reward $r(x,u,w)$ is bounded almost surely—there exists $0 < M < \infty$ such that $|r| < M$ almost surely.

a) Consider any given $K \in (0,\infty)$ and let $B_K(0) \subset \mathcal{Q}$ be a ball of radius K centered at zero. Then, $T : \mathcal{Q} \to \mathcal{X}$ is a contraction on $B_K(0)$.
b) Suppose $\bar{K}$ is chosen such that

$$\frac{\max\{|\tilde{v}(M)|, |\tilde{v}(-M)|\}}{(1-\gamma)} < \bar{K} \min_{y \in I_{\bar{K}}} D\tilde{v}(y). \quad (7)$$

Then, T has a unique fixed point in $B_K(0)$ for any $K \in [\bar{K}, \infty)$.

The proof is provided in Appendix A.

The following proposition shows that the ℓ-prospect as well as the class of functions considered in [5] satisfy (7). Moreover, it shows that the value functions which satisfy Assumption 1 also satisfy (7).

Proposition 1: Suppose $r : X \times U \times W \to \mathbb{R}$ is bounded almost surely by M and $\gamma \in (0,1)$. Consider the condition

$$\frac{\max\{|\tilde{v}(M)|, |\tilde{v}(-M)|\}}{(1-\gamma)} < K \min_{y \in I_K} D\tilde{v}(y). \quad (8)$$

a) Suppose v satisfies Assumption 1 and that for some $\varepsilon > 0$, $\varepsilon < \frac{v(y)-v(y')}{y-y'}$ for all $y \neq y'$. Then (8) holds.
b) Suppose v is an ℓ-prospect value function with arbitrary parameters $(k_-, k_+, \zeta_-, \zeta_+)$ satisfying Assumption 1. Then there exists a K such that the ℓ-prospect value function satisfies (8).

With Theorem 1 and Proposition 1, we can prove convergence of Q-learning for risk-sensitive reinforcement learning.

Theorem 2 (Q-learning Convergence on $B_K(0)$): Suppose that v satisfies Assumption 1 and that for each $(x,u) \in X \times U$ the reward $r(x,u,w)$ is bounded almost surely—that is, there exists $0 < M < \infty$ such that $|r| < M$ almost surely. Moreover, suppose the ball $B_K(0)$ is chosen such that (7) holds and $Q_0 \in B_K(0)$. If the nonnegative learning rates $\alpha_t(x,u)$ are such that $\sum_{t=0}^{\infty} \alpha_t(x,u) = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2(x,u) < \infty$, $\forall (x,u) \in X \times U$. Let $\varepsilon > 0$. Then, if $T \geq g_1(\varepsilon)$ and $1/\gamma_k \geq g_2(\varepsilon)$ for all $k \geq 0$ and for some functions $g_1(\varepsilon) = O(\log(1/\varepsilon))$ and $g_2(\varepsilon) = O(1/\varepsilon)$, then the procedure in (5) converges to $Q^* \in B_K(0)$ with high probability— $\Pr(\|Q_t - Q^*\| \leq \varepsilon, \forall t \geq T+1) \geq 1 - \delta(\varepsilon)$ for some constant $\delta(\varepsilon)$.

The proof of the above theorem is provided in Appendix B. It replies on a standard argument which combines the fixed point result of Theorem 1 with and the ordinary differential equation (ODE) method for analyzing stochastic approximation algorithms [24, Ch. 1–4][25]. Since Theorem 1 holds on any $B_K(0)$ with $\bar{K} < K < \infty$, so does Theorem 2.

IV. Inverse Risk-Sensitive Reinforcement Learning

Given a set of *demonstrations* $\mathcal{D} = \{(x_k, u_k)\}_{k=1}^N$, our goal is to recover an estimate of the policy and value function used to generate the demonstrations. Let $\Pi = \{\pi_\theta\}_\theta$ be a class of parameterized policies and $\mathcal{F}$ be a class of parameterized value functions where $\theta \in \Theta \subset \mathbb{R}^d$ and $v \in \mathcal{F}$ is such that $v : Y \times \Theta \to \mathbb{R} : (y(\theta), \theta) \mapsto v(y(\theta), \theta)$. We use the notation v_θ where convenient. We also indicate the dependence of Q on θ using the notation $Q(x,u,\theta)$. We seek to minimize some loss $\ell(\pi_\theta)$ which is a function of the parameterized policy π_θ. By an abuse of notation, we introduce the shorthand $\ell(\theta) = \ell(\pi_\theta)$. The optimization problem is specified by

$$\min_{\theta \in \Theta} \{\ell(\theta) |\ \pi_\theta = H_\theta(Q^*), v_\theta \in \mathcal{F}\} \quad (9)$$

where H_θ belongs to a parameterized policy class. There are several possible loss functions that may be employed.

Since we seek a probability distribution π_θ, it is natural to formulate the loss in terms of the principle of maximum *entropy*, a tool for building probability distributions to match observations. It has been shown in the classical inverse reinforcement learning approach that specifying the problem in terms of *maximum casual entropy* [26]–[28] avoids certain pitfalls—e.g., nonconvexity and learning from suboptimal demonstrations. Motivated by this, we consider two related cost functions: the negative weighted log-likelihood of the demonstrated behavior and the *relative entropy* or *Kullback–Leibler (KL) divergence* between the empirical distribution of the state–action trajectories and their distribution under the learned policy. The former is given by

$$\ell(\theta) = \sum_{(x,u) \in \mathcal{D}} w(x,u) \log(\pi_\theta(u|x))$$

where $w(x,u)$ may, e.g., be the normalized empirical frequency of observing (x,u) pairs in $\mathcal{D}$— $n(x,u)/N$ with $n(x,u)$ denoting the frequency of (x,u) and the latter is given by

$$\ell(\theta) = \sum_{x \in \mathcal{D}_x} D_{\mathrm{KL}}(\hat{\pi}(\cdot|x) || \pi_\theta(\cdot|x))$$

where $D_{\mathrm{KL}}(\pi||\pi') = \sum_i \pi(i) \log(\pi(i)/\pi'(i))$ is the KL divergence, $\mathcal{D}_x \subset \mathcal{D}$ is the sequence of observed states, and $\hat{\pi}$ is the empirical distribution on the trajectories of $\mathcal{D}$. These losses are essentially the same under a reweighting: the weighted log-likelihood can be rewritten as $\ell(\theta) = \sum_{x \in \mathcal{D}_x} w(x) D_{KL}(\hat{\pi}_n(\cdot|x) || \pi_\theta(\cdot|x))$, where $w(x)$ is the frequency of state x normalized by $|\mathcal{D}| = N$. This approach has the added benefit that it is independent of θ and thus, is not affected by scaling of the value functions.

It is also common to adopt a smooth map H that operates on the action-value function space for defining the parametric policy space—e.g., *soft-max* or *Boltzmann* policies [27]–[29] of the form

$$H_\theta(Q)(u|x) = \frac{\exp(\beta Q(x,u,\theta))}{\sum_{u' \in U(x)} \exp(\beta Q(x,u',\theta))} \quad (10)$$

to the action-value functions Q where $\beta > 0$ controls how close $H_\theta(Q)$ is to a *greedy policy* which we define to be any policy π such that $\sum_{u \in U(x)} \pi(u|x) Q(x,u,\theta) = \max_{u \in U(x)} Q(x,u,\theta)$ at all states $x \in X$. This is one class of smooth policies dependent on θ through Q; we use this class in the examples in Section V. We use value functions such as those described in Section III-A; e.g., if v is the prospect theory value function defined in (2), then the parameter vector is $\theta = (k_-, k_+, \zeta_-, \zeta_+, \beta)$.

A. Gradient–Based Approach

In this subsection, we show (Theorem 3) that gradient descent is well-defined in the sense that 1) the derivative is computable via a contraction map and 2) the update step is in the direction of steepest descent. This result requires computing the derivative of $Q^*(x,u,\theta)$ with respect to θ. In particular, our result applies to any smooth policy class Π dependent on θ through Q. For instance, given policies of the form (10), the derivative with respect to an element θ_j of θ of the loss ℓ depends on the policy π_θ which, in turn, depends on $Q^*(\cdot,\cdot,\theta)$. Further, considering the log-based loss functions described above, $\log(\pi_\theta(u|x)) = \beta(Q^*(x,u,\theta) - \sum_{u' \in U(x)} Q^*(x,u',\theta))$ so that we simply need to show that $D_{\theta_j} Q^*$ can be computed. We do this by showing it can be calculated almost everywhere on Θ by solving fixed-point equations similar to the Bellman-optimality equations. We require some assumptions on the value function v.

Algorithm 1: Gradient-Based IRSRL.

procedure IRSRL $\mathcal{D}$
  Initialize: $\theta \leftarrow \theta_0$
  while $k <$ MAXITER & $\|\ell(\theta) - \ell(\theta_-)\| \geq \delta$ **do**
    $\theta_- \leftarrow \theta$
    $\eta_k \leftarrow$ LINESEARCH$(\ell(\theta_-), D_\theta \ell(\theta_-))$
    $\theta \leftarrow \theta_- - \eta_k D_\theta \ell(\theta_-)^T$
    $k \leftarrow k + 1$
  return θ

Assumption 2: The value function $v \in C^1(Y \times \Theta, \mathbb{R})$ satisfies the following conditions: (i) v is strictly increasing in y and for each $\theta \in \Theta$, there exists a $\bar{y}$ such that $v(\bar{y}, \theta) = v_0$; (ii) for each $\theta \in \Theta$, on any ball centered around the origin of finite radius, v is locally Lipschitz in y with constant $L_y(\theta)$ and Lipschitz in θ on Θ with constant L_θ.

Define $L_y = \max_\theta L_y(\theta)$ and $L = \max_\theta \{L_y(\theta), L_\theta\}$. As before, let $\tilde{v} \equiv v - v_0$. The Q–update equation can be re-written as

$$Q_{t+1}(x_t, u_t, \theta) = \left(1 - \frac{\alpha_t}{\alpha}\right) Q_t(x_t, u_t, \theta) + \frac{\alpha_t}{\alpha}(\alpha(v(y_t(\theta), \theta) - v_0) + Q_t(x_t, u_t, \theta)) \quad (11)$$

where

$$y_t(\theta) = r_t + \gamma \max_u Q_t(x_{t+1}, u, \theta) - Q_t(x_t, u_t, \theta)$$

is the temporal difference, $\alpha \in (0, \min\{L^{-1}, 1\}]$ and we have suppressed the dependence of α_t on (x_t, u_t). In addition, define the map T such that

$$(TQ)(x, u, \theta) = \alpha \mathbb{E}_{x',w} \tilde{v}(y(\theta), \theta) + Q(x, u, \theta)$$

where

$$y(\theta) = r(x, u, w) + \gamma \max_{u' \in U(x')} Q(x', u', \theta) - Q(x, u, \theta).$$

By the results of the proceeding section, this map is a contraction in Q for each fixed θ. Let $D_i \tilde{v}(\cdot, \cdot)$ be the derivative of $\tilde{v}$ with respect to the ith argument where $i = 1, 2$.

Theorem 3: Assume that $v \in C^1(Y \times \Theta, \mathbb{R})$ satisfies Assumption 2 and that the reward $r : X \times U \times W \to \mathbb{R}$ is bounded almost surely, i.e, $|r| < M$ for $M > 0$. Let $B_K(0)$ be any ball with radius K satisfying

$$\frac{\max\{|\tilde{v}(M, \theta)|, |\tilde{v}(-M, \theta)|\}}{1 - \gamma} < K \min_{(\theta, y(\theta)) \in \Theta \times I_K} D_1 \tilde{v}(y(\theta), \theta). \quad (12)$$

Then the following statements hold:

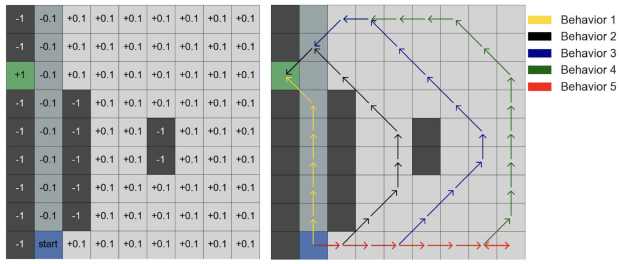
a) Q^* is locally Lipschitz-continuous on $B_K(0)$ as a function of θ—that is, for any $(x, u) \in X \times U$, $\theta, \theta' \in \Theta$, $|Q^*(x, u, \theta) - Q^*(x, u, \theta')| \leq C\|\theta - \theta'\|$ for some $C > 0$.

b) Except on a set of measure zero, the gradient $D_\theta Q^*(x, u, \theta) \in B_K(0)$ is given by the solution of the fixed–point equation

$$\phi_\theta(x, u) = \alpha \mathbb{E}_{x',w}[D_2 \tilde{v}(y(\theta), \theta) + D_1 \tilde{v}(y(\theta), \theta) \cdot (\gamma \phi_\theta(x', u^*_{x'}) - \phi_\theta(x, u))] + \phi_\theta(x, u) \quad (13)$$

where $\phi_\theta : X \times U \to \mathbb{R}^d$ and $u^*_{x'}$ is an action that maximizes $Q(x', u, \theta)$.

The proof is provided in Appendix D. Theorem 3 gives us a procedure—namely, a fixed–point equation which is a contraction—to compute the derivative $D_{\theta_j} Q^*$ so that, in turn, we can compute the derivative of $\ell(\theta)$ with respect to θ. Hence, the gradient method provided in Algorithm 1 for solving the inverse risk-sensitive reinforcement learning problem is well formulated. Note that for each fixed θ, condition (12) is the same as condition (7). Moreover, Proposition 1 shows that for the ℓ-prospect value functions and functions v such that $\varepsilon < \frac{v(y) - v(y')}{y - y'}$, such a K must exist for any choice of parameters and, hence, the result of Theorem 3 holds for these functions.



(a) (b)

Fig. 1. (a) Grid World layout. (b) Maximum likelihood paths corresponding to the five behavior profiles of risk-sensitive policies with various parameter combinations ($\{k_-, k_+, \zeta_-, \zeta_+\}$) for the prospect and ℓ-prospect value functions: *Behavior 1*: $\{0.1, 1.0, 0.5, 1.5\}$; *Behavior 2*: $\{1.0, 1.0, 1.0, 1.0\}$; *Behavior 3*: $\{1.0, 1.0, 1.1, 0.9\}$; *Behavior 4*: $\{5.0, 1.0, 1.1, 0.8\}$; *Behavior 5*: $\{5.0, 1.0, 1.5, 0.7\}$.

Given that the gradient of Q^* with respect to θ is computable, the gradient-based approach in Algorithm 1 simply implements an update scheme of the form $\theta_{k+1} = \theta_k - \eta_k g_k$, where $-g_k(\theta_k) = -D_\theta \ell(\theta_k)^T$ points in the direction of steepest descent. We also note that, following [30] this method is amenable to letting g_k be the *natural gradient*. Indeed, let $h(\theta) = H_\theta(Q)$ be a mapping from the parameter space to the policy space. Then, $g_k = G_\theta^\dagger D_\theta \ell(\theta_k)^T$ be the *natural gradient*, where $G_\theta = Dh(\theta) Dh(\theta)^T$ is a pseudo-Riemannian metric at θ induced by (d, Π, h) with d a metric on Π [31, Th. 1]. Since our intention is to find the best policy π_θ matching the empirical policy, this approach is beneficial as it allows us to update θ by taking a step in the direction of steepest ascent on the surface $(\pi, \ell(\pi))$.

V. Examples

We demonstrate the proposed approach on Grid World. While the formulation of inverse risk-sensitive reinforcement learning is amenable to learning β, we assume it is known for the purpose of explicitly exploring the effects of changing the value function parameters on the resulting policy. In all experiments, $\gamma = 0.95$, $\beta = 4$, the objective is the negative log-likelihood of the data, and the valuation function is induced by an acceptance level set defined by a parameterized value function and acceptance level of zero. For the prospect and ℓ-prospect value functions, the reference point is zero. These choices are aimed at further deconflating observations of behavior—in terms of risk-sensitivity and loss-aversion—that result from different choices of the value function parameters from characteristics of the MDP or learning algorithm.

The Grid World instance is shown in Fig. 1(a). The agent starts in the blue box and aims to maximize their reward via the risk-sensitive reinforcement learning procedure described in Section III over an infinite time horizon. Every square in the grid represents a state, and the action space is $U = \{N, NE, E, SE, S, SW, W, NW\}$. Each action corresponds to a movement in the specified direction. The black and green states are absorbing. In all the other states, the agent moves in the direction specified by their action with probability 0.93 and in any of the other seven directions with probability 0.01. To make the grid finite, any action taking the agent out of the grid has probability zero, and the other actions are re-weighted accordingly. Rewards in the black and green states are -1 and $+1$, respectively. In the darker gray states, the agent gets a reward of -0.1. In all other states, the agent gets a reward of $+0.1$.

We conduct two types of experiments: 1) learning the value function of an agent with the correct model for the value function (e.g., learning

TABLE I
THE MEAN AND VARIANCE OF THE TV DISTANCE BETWEEN THE TRUE POLICY AND THE POLICY UNDER THE LEARNED VALUE FUNCTION

Value Function	Prospect		ℓ-prospect	
Behavior	**Mean**	**Variance**	**Mean**	**Variance**
Behavior 1	1.9e-2	6.3e-4	1.3e-2	2.3e-4
Behavior 2	1.5e-2	2.0e-4	1.0e-2	9.6e-5
Behavior 3	2.0e-2	3.6e-4	1.1e-2	1.3e-4
Behavior 4	1.6e-2	2.0e-4	1.2e-2	1.4e-4
Behavior 5	4.7e-2	3.0e-3	1.0e-2	3.4e-4

(a) Experiment A: Learning with the correct model

Value Function	Mean	Variance
Prospect	1.5e-2	1.6e-4
ℓ-prospect	1.5e-2	1.6e-4

(b) Experiment B: Learning with an incorrect model

a prospect value function when the agent also has a prospect value function); 2) learning the value function of an agent with the wrong model for the value function (e.g., learning an ℓ–prospect value function when the agent has a prospect value function). Performance is measured via the total variation (TV) distance.

In Experiment A, we trained agents with various parameter combinations of the prospect and ℓ-prospect value functions. The resulting policies of these agents are classified into five behavior profiles via their maximum likelihood path: (*Behavior 1*) profile that is risk-seeking on gains, (Behavior 2) profile that is risk neutral on gains and losses (this is also the behavior corresponding to the nonrisk-sensitive reinforcement learning approach), and (*Behaviors 3–5*) profiles that are increasingly risk averse on losses and increasingly weigh losses more than gains. The parameters are given in Fig. 1(b). We sampled 1,000 trajectories from the policies of these agents and used the data in the inverse risk-sensitive reinforcement learning framework. The learned value function is of the same type as that of the agent. Due to the nonconvexity of the loss, we use five randomly generated initial parameter choices.

The results we report are associated with the value function that achieves the minimum value of the objective. In Table I a, we report the mean and variance of the TV distance between the two policies across all states. In all the cases the learned value functions produce policies that correctly match the maximum likelihood path of the true agent. The performance for learning a prospect value function is consistently worse than learning an ℓ-prospect function and requires significantly more computation time. This is most likely due to the fact that the prospect value function is not Lipschitz around the reference point. Thus, we have no guarantees of differentiability of Q^* with respect to θ for the prospect value function.

Experiment B consists of learning different types of value functions from the same dataset. The motivation for this experiment is to ensure that the results and risk-profiles learned were consistent across the choice of model. We generated a data set with 10 000 samples from an agent with a prospect value function, and used it to learn prospect and ℓ-prospect value functions. The mean TV distance between the policy of the true agent and the policies under the learned value functions are shown in Table I b. The true agent's value function has parameters $\{k_-, k_+, \zeta_-, \zeta_+\} = \{2.0, 1.0, 0.9, 0.7\}$, i.e., it is risk-seeking in losses, risk-averse in gains, and loss averse. Again, the learned value functions all have policies that replicated the maximum likelihood behavior of the true agent. We note that the ℓ-prospect and prospect functions perform as well as each other on this data (likely due to the fact that they have the same underlying shape), but the ℓ-prospect function showed none of the numerical issues that we encountered with the prospect function. Further, learning with the ℓ-prospect function is markedly faster than with the prospect function. Again, this is most likely due to the fact that the prospect function is not locally Lipschitz continuous around the reference point.

VI. DISCUSSION

We present a gradient-based technique for learning risk-sensitive decision-making models of agents operating in uncertain environments. Moreover, we introduce a Lipschitz variation of the prospect value function, which retains the convex–concave structure of the prospect theory value function while satisfying the assumptions of the theorems we present on a bounded domain and possessing better numerical properties. We demonstrate the algorithm's performance for agents based on several types of behavioral models on the Grid World benchmark problem. Looking forward, there are a number of interesting open questions regarding convergence of the gradient-based procedure (perhaps using a multitimescale stochastic approximation technique), expanding the theory to handle multiple value functions to tradeoff between different outcomes, and estimating the reference point and acceptance level.

APPENDIX

A. Proof of Theorem 1

The proof of Theorem 1 relies on a fixed point theorem.

Theorem 4 ([32, Th. 2.2]): Let (X, d) be a complete metric space and $B_r(y) = \{x \in X |\, d(x, y) < r\}$ be a ball of radius $r > 0$ centered at $y \in X$. Let $f : B_r(y) \to X$ be a contraction map with contraction constant $h < 1$. Further, assume that $d(y, f(y)) < r(1 - h)$. Then, f has a unique fixed point in $B_r(y)$. ■

Proof: [Proof of Th. 1. (a)] The map T, defined by $(TQ)(x, u) = \alpha \mathbb{E}_{x',w}[\tilde{v}(y(Q(x, u), x'))] + Q(x, u)$, is a contraction with constant $\bar{\alpha}_K = 1 - \alpha(1 - \gamma)\varepsilon_K$ where $\varepsilon_K = \min\{D\tilde{v}(y) |\ y \in I_K\}$, $I_K = [-M - 2K, M + 2K]$ and $\alpha \in (0, \min\{1, L^{-1}\}]$ with L the Lipschitz constant of v on I_K. Indeed, let $y(Q(x, u), x') = r(x, u, w) + \gamma \max_{u'} Q(x', u') - Q(x, u)$ and define $g(x') = \max_{u'} Q(x', u')$. For any $Q \in B_K(0)$ we note that the temporal differences are bounded—in fact, $y(Q(x, u), x') \in I_K = [-M - 2K, M + 2K]$. For any $y', y \in I_K$, $\tilde{v}(y) - \tilde{v}(y') = \xi(y - y')$ for some $\xi \in [\varepsilon_K, L]$ by the monotonicity assumption on v. Then, for any $Q_1, Q_2 \in B_K(0)$,

$$
\begin{aligned}
&(TQ_1 - TQ_2)(x, u)\\
&\quad = \alpha \mathbb{E}_{x',w}[\tilde{v}(y(Q_1(x, u), x')) - \tilde{v}(y(Q_2(x, u), x'))] + Q_1(x, u)\\
&\qquad - Q_2(x, u)\\
&\quad = \alpha \mathbb{E}_{x',w}[\xi_{x',w}(\gamma g_1(x') - \gamma g_2(x') - Q_1(x, u) + Q_2(x, u))]\\
&\qquad + Q_1(x, u) - Q_2(x, u)\\
&\quad = \alpha\gamma \mathbb{E}_{x',w}[\xi_{x',w}(g_1(x') - g_2(x'))] + (1 - \alpha \mathbb{E}_{x',w}[\xi_{x',w}])\\
&\qquad \cdot (Q_1(x, u) - Q_2(x, u)).
\end{aligned}
$$

so that $|(TQ_1 - TQ_2)(x, u)| \le (1 - \alpha(1 - \gamma)\varepsilon_K)\|Q_1 - Q_2\|_\infty$. We claim that the constant $\bar{\alpha}_K = 1 - \alpha(1 - \gamma)\varepsilon_K < 1$. Indeed, recall that $0 < \alpha \le \min\{1, L^{-1}\}$ so that if $\alpha = L^{-1}$, then $\bar{\alpha}_K < 1$ since $L = \max_{y \in I_K} D\tilde{v}(y)$ and $\varepsilon_K = \min_{y \in I_K} D\tilde{v}(y)$. On the other hand, if $\alpha = 1$, then $1 \le L^{-1} \le (\varepsilon_K)^{-1}$ so that $\varepsilon_K \le 1$ which, in turn, implies that $\bar{\alpha}_K < 1$. If $0 < \alpha < \min\{1, L^{-1}\}$, then $\bar{\alpha}_K < 1$ follows trivially from the implications in the above two cases. Thus, T is a contraction on $B_K(0)$.

Proof: [Proof of Th. 1. (b)] Let K be chosen such that

$$
\frac{\max\{|\tilde{v}(M)|, |\tilde{v}(-M)|\}}{1 - \gamma} < K \min_{y \in I_K} D\tilde{v}(y). \tag{14}
$$

The map T applied to the zero map, $0 \in B_K(0)$, is strictly less than $K(1 - \bar{\alpha}_K)$. Indeed, for any $\alpha \in (0, \min\{1, L^{-1}\}]$, $\|T(0)\| \le \alpha \max\{|\tilde{v}(M)|, |\tilde{v}(-M)|\} < (1 - \gamma)K\varepsilon_K\alpha$ since $\tilde{v}$ is increasing by assumption. Since T is a contraction, combining the above with the fact that $(1 - \gamma)K\varepsilon_K\alpha = K(1 - \bar{\alpha}_K)$, the assumptions of Theorem 4 hold and, hence there is a unique fixed point $Q^* \in B_K(0)$. ■

B. Proof of Theorem 2

Proof: Following [33], let (i,a) index the state–action pairs in $X \times U$. Consider

$$Q_{t+1}(i,a) = \left(1 - \frac{\alpha_t(i,a)}{\alpha}\right) Q_t(i,a) + \frac{\alpha_t(i,a)}{\alpha}(TQ_t(i,a) + d_t(i,a))$$

where $d_t(i,a) = \alpha(\tilde{v}(y_t(i,a)) - \mathbb{E}_{x',w}\tilde{v}(y_t(i,a)))$ is a random noise term and α_t is the learning rate such that $\alpha_t(i,a) = 0$ if $Q_t(i,a)$ is not updated, i.e., $\alpha_t(i,a) = 0$ if $1\{x_t = i, u_t = a\} = 0$. Let $\mathcal{F}_t = \{(Q_k(i,a), \alpha_k(i,a))_{k=0}^t, (d_\ell(i,a))_{\ell=0}^{t-1}, (i,a) \in X \times U\}$. Since we have already shown the map T is a contraction, following [25], we simply need to show that $\mathbb{E}_{x',w}[d_t|\mathcal{F}_t] = 0$ and $\mathbb{E}_{x',w}[d_t^2|\mathcal{F}_k] \leq A + B\|Q_t\|_\infty$ for some constants A and B. Clearly, $\mathbb{E}_{x',w}[d_t(i)] = 0$. It is also the case that

$$\begin{aligned}\mathbb{E}_{x',w}[d_t^2|\mathcal{F}_t] &\leq \alpha^2 \mathbb{E}_{x',w}[\tilde{v}(y_t)^2|\mathcal{F}_t] - \alpha^2(\mathbb{E}_{x',w}[\tilde{v}(y_t)|\mathcal{F}_t])^2 \\ &\leq \alpha^2 \mathbb{E}_{x',w}[\tilde{v}(y_t)^2|\mathcal{F}_t].\end{aligned}$$

Since the rewards are bounded by M, $|y_t| \leq M + 2\|Q_t\|_\infty$. Moreover, $\tilde{v}$ is Lipschitz on I_K so that $|\tilde{v}(y_t)| \leq |\tilde{v}(0)| + L(M + 2\|Q_t\|_\infty)$. Applying the triangle inequality, we get that $(|\tilde{v}(0)| + L(M + 2\|Q_t\|_\infty))^2 \leq 2(|\tilde{v}(0)| + LM)^2 + 8L^2\|Q_t\|_\infty^2$ so that

$$\alpha^2 \mathbb{E}_{x',w}[\tilde{v}(y_t)^2|\mathcal{F}_t] \leq 2\alpha^2(|\tilde{v}(0)| + 2M)^2 + 8\alpha^2 L^2\|Q_t\|_\infty.$$

Note this is stronger than [25, Assum. A.3]. Since T is a contraction, we can construct a local Lyapunov function: $V(Q_t(x,u)) = \frac{1}{2}\|Q_t(x,u) - Q^*(x,u)\|_2^2$ with

$$\begin{aligned}\dot{V}(Q_t(x,u)) &= 2(Q_t(x,u) - Q^*(x,u))(TQ_t(x,u) - TQ^*(x,u)) \\ &\quad - 2\|Q_t(x,u) - Q^*(x,u)\|^2 \\ &< 2(\alpha - 1)\|Q_t(x,u) - Q^*(x,u)\|^2 < 0.\end{aligned}$$

Hence, applying [25, Cor. 1.1], we get convergence with high probability given some $\varepsilon > 0$, i.e., suppressing the dependence on $(i,a) \in X \times U$, there exists constants $\lambda, C_1, C_2 > 0$ such that

$$\Pr(\|Q_t - Q^*\| \leq \varepsilon, \quad \forall t \geq T+1) \geq 1 - \delta(\varepsilon)$$

where, by letting $\beta_n = \max_{0\leq k\leq n-1}\{\exp(-\lambda\sum_{i=k+1}^{n-1}\alpha_i)\alpha_k\}$,

$$\delta(\varepsilon) = \begin{cases} \sum_{n=0}^\infty C_1 e^{-C_2\sqrt{\varepsilon}/\sqrt{\alpha}} - \sum_{n=0}^\infty C_1 e^{-C_2\varepsilon^2/\beta_n}, & \varepsilon \leq 1 \\ \sum_{n=0}^\infty C_1 e^{-C_2\sqrt{\varepsilon}/\sqrt{\alpha_t}} - \sum_{n=0}^\infty C_1 e^{-C_2\varepsilon/\beta_n}, & \varepsilon > 1 \end{cases}$$

■

We refer the reader to [25, Cor. 1.1] for a more explicit characterization of the constants.

C. Proof of Proposition 1

Proof: [Proof of Proposition 1.a] Suppose v satisfies Assumption 1 and that for some $\varepsilon > 0$, $\varepsilon < \frac{v(y)-v(y')}{y-y'}$ for all $y \neq y'$. Then there exists a value of K, say $\bar{K}$, such that (8) holds for all $K > \bar{K}$. Indeed since $\min_{K>0}\varepsilon_K > \varepsilon$, for all K satisfying $\frac{\max\{|\tilde{v}(M)|,|\tilde{v}(-M)|\}}{\varepsilon(1-\gamma)} < K$, (8) must hold. ■

Proof: [Proof of Proposition 1.b] For $\zeta_+, \zeta_- \geq 1$ and any choice of k_-, k_+, $\min_{K>0}\varepsilon_K > \varepsilon > 0$ where $\varepsilon = \min\{\lim_{y\uparrow 0} D\tilde{v}(y), \lim_{y\downarrow 0} D\tilde{v}(y)\}$. Therefore, with $\zeta_+, \zeta_- \geq 1$, for any K such that $\frac{\max\{|\tilde{v}(M)|,|\tilde{v}(-M)|\}}{\varepsilon(1-\gamma)} < K$, (8) must hold. For the case when either $\zeta_+ < 1$ or $\zeta_- < 1$ or both, we note that $\min_{y\in I_K} D\tilde{v}(y) = \min\{\min_{y\in I_K} D\tilde{v}(y), \varepsilon\}$. so that we need only show that for $\zeta_+ < 1$, there exists a K such that

$$\frac{\max\{|\tilde{v}(M)|, |\tilde{v}(-M)|\}}{1-\gamma} < KD\tilde{v}(2K+M) \tag{15}$$

and, similarly for $\zeta_- < 1$, there exists a K such that the left-hand side of (15) is less than $KD\tilde{v}(-2K - M)$. Without loss of generality, we show (15) must hold for $\zeta_+ < 1$ and reference point $y_0 = 0$ (the proof for $\zeta_- < 1$ follows an exactly analogous argument). Plugging $D\tilde{v}(2K+M) = k_+\zeta_+(2K + M - y_0 + \epsilon)^{\zeta_+ - 1}$ in and rearranging, we simply need to show that there exists a K such that

$$\frac{\max\{|\tilde{v}(M)|, |\tilde{v}(-M)|\}}{(1-\gamma)\xi_+ k_+} < K(2K + M - y_0 + \epsilon)^{\xi_+ - 1}$$

Since the right-hand side is a function of K that is zero at $K = 0$ and approaches infinity as $K \to \infty$, and the left-hand side is a finite constant, there is some $\bar{K}$ such that for all $K > \bar{K}$, the above holds. Thus, for the ℓ-prospect value function, our assumptions are satisfied and there always exists a value of K to choose in Theorem 1.b. ■

D. Proof of Theorem 3

Let U be a Banach space and U^* its dual. The Fréchet subdifferential of $f : U \to \mathbb{R}$ at $u \in U$, denoted by $\partial f(u)$ is the set of $u^* \in U^*$ such that $\lim_{h\to 0}\inf_{h\neq 0}\|h\|^{-1}(f(u+h) - f(u) - \langle u^*, u\rangle) \geq 0$.

Proposition 2 ([31], [34]): For a finite family $(f_i)_{i\in I}$ of real-valued functions (where I is a finite index set) defined on U, let $f(u) = \max_{i\in I} f_i(u)$. If $u^* \in \partial f_i(u)$ and $f_i(u) = f(u)$, then $u^* \in \partial f(u)$.

Proposition 3 ([31], [35]): Consider $(f_n)_{n\in\mathbb{N}}$, a pointwise convergent sequence to f such that $f_n : U \to \mathbb{R}$. Let $u \in U$, $u_n^* \in \partial f_n(u) \subset U^*$. Suppose that $(u_n^*)_{n\in\mathbb{N}}$ is weak*–convergent to u^* and is bounded, and that at u, for any $\varepsilon > 0$, $\exists N > 0, \delta > 0$ such that for any $n \geq N$, $h \in B_\delta(0)$, a δ–ball around $0 \in U$, $f_n(u+h) \geq f_n(u) + \langle u_n^*, h\rangle - \varepsilon\|h\|$. Then $u^* \in \partial f(u)$. ■

Proof: [Proof of Theorem 3.a.] Let $Q_0(x,u,\theta) \equiv 0$. Then it is trivial that $Q_0(x,u,\theta)$ is locally Lipschitz in θ on Θ. Supposing that $Q_t(x,u,\theta)$ is L_t–locally Lipschitz in θ, then we need to show that $TQ_t(x,u,\theta)$ is locally Lipschitz. Since $\tilde{v} \equiv v - v_0$, it also satisfies Assumption 2. Let $L_y = \max\{L_y(\theta)|\theta \in \Theta\}$ and define $g_t(x,\theta) = \max_{u'} Q_t(x,u',\theta)$. Since Q_t is assumed Lipschitz with constant L_t, so is g_t. Let $\Delta TQ_t(\theta,\theta') = TQ_t(\theta) - TQ_t(\theta')$ and $\Delta Q_t(\theta,\theta') = Q_t(\theta) - Q_t(\theta')$. Suppressing the dependence on (x,u),

$$\begin{aligned}\Delta TQ_t(\theta,\theta') = \alpha\mathbb{E}_{x',w}[&\tilde{v}(y(\theta),\theta) - \tilde{v}(y(\theta'),\theta) + \tilde{v}(y(\theta'),\theta) \\ &- \tilde{v}(y(\theta'),\theta')] + \Delta Q_t(\theta,\theta').\end{aligned}$$

Let $\tilde{\varepsilon}_K = \min_{(\theta,y(\theta))\in\Theta\times I_K} D_1\tilde{v}(y(\theta),\theta)$. Due to the monotonicity of $\tilde{v}$ in y, we know that for all y_1, y_2 there exists $\xi \in [\tilde{\varepsilon}_K, L_y]$ such that $\tilde{v}(y_1,\theta) - \tilde{v}(y_2,\theta) = \xi(y_1 - y_2)$. Hence,

$$\begin{aligned}&\Delta TQ_t(\theta,\theta') \\ &= \alpha\mathbb{E}_{x',w}\left[\xi_{x',w}(y(\theta) - y(\theta')) + \tilde{v}(y(\theta'),\theta) - \tilde{v}(y(\theta'),\theta')\right] \\ &\quad + \Delta Q_t(\theta,\theta') \\ &= \alpha\gamma\mathbb{E}_{x',w}[\xi_{x',w}(g_t(x',\theta) - g_t(x',\theta'))] - \alpha\mathbb{E}_{x',w}[\xi_{x',w}] \\ &\quad \cdot\Delta Q_t(\theta,\theta') + \alpha\mathbb{E}_{x',w}[\tilde{v}(y(\theta'),\theta) - \tilde{v}(y(\theta'),\theta')] + \Delta Q_t(\theta,\theta') \\ &= (1 - \alpha\mathbb{E}_{x',w}[\xi_{x',w}])\Delta Q_t(\theta,\theta') + \alpha\gamma\mathbb{E}_{x',w}[\xi_{x',w}(g_t(x',\theta) \\ &\quad - g_t(x',\theta'))] + \alpha\mathbb{E}_{x',w}[\tilde{v}(y(\theta'),\theta) - \tilde{v}(y(\theta'),\theta')]\end{aligned}$$

so that

$$\|\Delta TQ_t(\theta,\theta')\| \le ((1-\alpha(1-\gamma)\mathbb{E}_{x',w}[\xi_{x',w}])L_t + \alpha L_\theta)\|\theta-\theta'\| \le ((1-\alpha(1-\gamma)\tilde{\varepsilon}_K)L_t + \alpha L_\theta)\|\theta-\theta'\|.$$

Since $\bar{\alpha}_K = 1-\alpha(1-\gamma)\tilde{\varepsilon}_K$, $TQ_t(\cdot,\cdot,\theta)$ is L_{t+1}–locally Lipschitz with $L_{t+1} = \bar{\alpha}L_t + \alpha L_\theta$. With $L_0 = 0$, by iterating, we get that $L_{t+1} = (\bar{\alpha}^t + \cdots + \bar{\alpha} + 1)\alpha L_\theta$. As stated in Section IV-A, T is a contraction so that $T^n Q_0 \to Q_\theta^* = Q^*(\cdot,\cdot,\theta)$ as $n \to \infty$. Hence, by the above, Q_θ^* is $\alpha L_\theta/(1-\bar{\alpha}_K)$–Lipschitz continuous.

Proof: [Proof of Theorem 3.b.] Consider a fixed $\theta \in \Theta \subset \mathbb{R}^d$. Since by part (a), Q_θ^* is locally Lipschitz in θ, Rademacher's Theorem (see, e.g., [36, Th. 3.1.6]) tells us it is differentiable almost everywhere (except a set of Lebesgue measure zero). We now show that the operator S acting on the space of functions $\phi_\theta : X \times U \to \mathbb{R}^d$ and defined by $(S\phi_\theta)(x,u) = \alpha\mathbb{E}_{x',w}[D_2\tilde{v}(y(\theta),\theta) + D_1\tilde{v}(y(\theta),\theta)\cdot(\gamma\phi_\theta(x',u^*_{x'}) - \phi_\theta(x,u))] + \phi_\theta(x,u)$ where $u^*_{x'}$ is an action that maximizes $Q(x',u,\theta)$ is a contraction since, by Proposition 2, a subdifferential of the pointwise maximum of functions is equal to the subdifferential of one of the one that achieves the maximum. Indeed,

$$\begin{aligned}&(S\phi_\theta - S\phi'_\theta)(x,u)\\ &\quad= \alpha\mathbb{E}_{x',w}[D_1\tilde{v}(y(\theta),\theta)\big(\gamma(\phi_\theta(x',u^*_{x'}) - \phi'_\theta(x',u^*_{x'}))\\ &\qquad\qquad - (\phi_\theta(x,u) - \phi'_\theta(x,u))\big)] + \phi_\theta(x,u) - \phi'_\theta(x,u)\\ &\quad\le (1-\alpha(1-\gamma)\mathbb{E}_{x',w}[D_1\tilde{v}(y(\theta),\theta)])\|\phi_\theta - \phi'_\theta\|_\infty.\end{aligned}$$

Since we have fixed θ, let $\tilde{\varepsilon}_{K,\theta} = \min_{y\in I_K} D_1\tilde{v}(y,\theta)$. Then, by Assumption 1, $\|(S\phi_\theta - S\phi'_\theta)(x,u)\| \le (1-\alpha(1-\gamma)\tilde{\varepsilon}_{K,\theta})\|\phi_\theta - \phi'_\theta\|_\infty$. Note that $\bar{\alpha}_K = 1-\alpha(1-\gamma)\tilde{\varepsilon}_{K,\theta} < 1$ for the same reasons as given in the proof of Theorem 1 since $\alpha \in (0, \min\{1, L^{-1}\}]$. Note that S operates on each of the d components of θ separately and hence, it is a contraction when restricted to each individual component. Then, for each θ, S has a unique fixed point. In particular, consider the sequence $\phi_{\theta,k}$ such that $\phi_{\theta,0} = 0$ and $\phi_{\theta,k+1} = S\phi_{\theta,k}$. For large enough k, $\phi_{\theta,k+1} = S\phi_{\theta,k}$. Applying the contraction mapping theorem (see, e.g., [37, Th. 3.18]) we get that $\lim_{k\to\infty} S^k\phi_0$ converges to a unique fixed point.

Applying Proposition 2 by induction, $\phi_{\theta,k}(x,u) \in \partial_\theta Q_k(x,u,\theta)$. Indeed, it is obvious for $k=0$. Suppose it holds for k, i.e., $\phi_{\theta,k}(x,u) \in \partial_\theta Q_k(x,u,\theta)$. Then, $\phi_{\theta,k+1}(x,u) = S\phi_{\theta,k}(x,u) \in S(\partial_\theta Q_k(x,u,\theta))$ and $S(\partial_\theta Q_k(x,u,\theta)) \subset \partial_\theta(TQ_k) = \partial_\theta Q_{k+1}(x,u,\theta)$ by the definition of the maps and subdifferentiation. Hence, $\phi_{\theta,k+1}(x,u) \in \partial_\theta Q_{k+1}(x,u,\theta)$. By Proposition 3, the limit is a subdifferential of Q_θ^* since $\tilde{v}$ is Lipschitz on Y and Θ and the derivatives of $\tilde{v}$ are uniformly bounded. By part (a), Q_θ^* is locally Lipschitz in θ so that it is differentiable almost everywhere [36, Th. 3.1]. Since Q_θ^* is differentiable, its subdifferential is its derivative. ■

REFERENCES

[1] B. Köszegi and M. Rabin, "A model of reference-dependent preferences," *Quart. J. Econ.*, vol. 121, no. 4, pp. 1133–1165, 2006.

[2] A. Tversky and D. Kahneman, "Loss aversion in riskless choice: A reference-dependent model," *Quart. J. Econ.*, vol. 106, no. 4, pp. 1039–1061, 1991.

[3] A. Tversky and D. Kahneman, "Rational choice and the framing of decisions," *J. Bus.*, vol. 59, no. 4, pp. S251–S278, 1986.

[4] L. A. Prashanth, C. Jie, M. Fu, S. Marcus, and C. Szepesvári, "Cumulative prospect theory meets reinforcement learning: Prediction and control," in *Proc. 33rd Intern. Conf. Mach. Learn.*, vol. 48, 2016.

[5] Y. Shen, M. J. Tobia, T. Sommer, and K. Obermayer, "Risk-sensitive reinforcement learning," *Neural Comput.*, vol. 26, pp. 1298–1328, 2014.

[6] O. Mihatsch and R. Neuneier, "Risk-sensitive reinforcement learning," *Mach. Learn.*, vol. 49, no. 2, pp. 267–290, 2002.

[7] A. Majumdar, S. Singh, A. Mandlekar, and M. Provone, "Risk-sensitive inverse reinforcement learning via coherent risk models," in *Robot.: Sci. Syst.*, vol. 13, 2017.

[8] A. Y. Ng and S. Russell, "Algorithms for inverse reinforcement learning," in *Proc. 17th Int. Conf. Mach. Learn.*, 2000, pp. 663–670.

[9] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004.

[10] D. Kahneman and A. Tversky, "Prospect theory: An analysis of decision under risk," *Econometrica*, vol. 47, no. 2, pp. 263–291, 1979.

[11] M. Heger, "Consideration of risk in reinforcement learning," in *Proc. 11th Inter. Conf. Mach. Learn.*, 1994, pp. 105–111.

[12] E. Mazumdar, L. J. Ratliff, T. Fiez, and S. S. Sastry, "Gradient–based inverse risk-sensitive reinforcement learning with applications," in *Proc. 56th IEEE Conf. Decis. Control*, 2017, pp. 5796–5801.

[13] H. Föllmer and A. Schied, *Stochastic Finance: An Introduction in Discrete Time*. Berlin, Germany: Walter de Gruyter, 2004.

[14] D. Kahneman and A. Tversky, "Choices, values, and frames," *Amer. Psychologist*, vol. 39, pp. 341–350, 1984.

[15] R. Gonzalez and G. Wu, "On the shape of the probability weighting function," *Cogn. Psychol.*, vol. 38, no. 1, pp. 129–166, 1999.

[16] G. Wu and R. Gonzalez, "Curvature of the probability weighting function," *Manag. Sci.*, vol. 42, no. 12, pp. 1676–1690, 1996.

[17] H. Simon, "Bounded rationality in social science: Today and tomorrow," *Mind Soc.*, vol. 1, no. 1, pp. 25–39, Mar. 2000.

[18] A. Tversky and D. Kahneman, "The framing of decisions and the psychology of choice," *Science*, vol. 211, no. 4481, pp. 453–458, Jan. 1981.

[19] C. F. Camerer, "An experimental test of several generalized utility theories," *J. Risk Uncertainty*, vol. 2, no. 1, pp. 61–104, 1989.

[20] Y. Shen, W. Stannat, and K. Obermayer, "Risk-sensitive Markov control processes," *SIAM J. Control Optim.*, vol. 51, no. 5, pp. 3652–3672, 2013.

[21] H. Robbins and D. Siegmund, "A convergence theorem for non negative almost supermartingales and some applications," in *Optimizing Methods in Statistics*, J. S. Rustagi, Ed. Academic Press, 1971, pp. 233–257.

[22] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, 1951.

[23] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*. New York, NY, USA: Springer, 2003.

[24] V. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*. New York, NY, USA: Springer, 2008.

[25] G. Thoppe and V. Borkar, "A concentration bound for stochastic approximation via Alekseev's formula," 2017, *arXiv:1506.08657v3*.

[26] B. Ziebart, "Modeling purposeful adaptive behavior with the principle of maximum causal entropy," Ph.D. dissertation, Mach. Learn. Dept., Carnegie Mellon Univ., Pittsburg, PA, USA, 2010.

[27] B. D. Ziebart, J. A. Bagnell, and A. K. Dey, "Modeling interaction via the principle of maximum causal entropy," in *Proc. 27th Inter. Conf. Mach. Learn.*, 2010, pp. 1255–1262.

[28] Z. Zhou, M. Bloem, and N. Bambos, "Infinite time horizon maximum causal entropy inverse reinforcement learning," *IEEE Trans. Automat. Contr.*, vol. 63, no. 9, pp. 2787–2802, Sep. 2018.

[29] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning," in *Proc. 23rd AAAI Conf. Artif. Intell.*, 2008, pp. 1433–1438.

[30] S.-I. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, no. 2, pp. 251–276, 1998.

[31] G. Neu and C. Szepesvári, "Apprenticeship learning using inverse reinforcement learning and gradient methods," in *Proc. 23rd Conf. Uncertainty Artif. Intell.*, 2007, pp. 295–302.

[32] A. Latif, *Banach Contraction Principle and Its Generalizations*. New York, NY, USA: Springer, 2014, pp. 33–64.

[33] V. S. Borkar and S. Meyn, "The O.D.E. method for convergence of stochastic approximation and reinforcement learning," *SIAM J. Control Optim.*, vol. 38, no. 2, pp. 447–469, 2000.

[34] A. Y. Kruger, "On Fréchet subdifferentials," *J. Math. Sci.*, vol. 116, no. 3, pp. 3325–3358, 2003.

[35] J. Penot, "On the interchange of subdifferentiation and epi-convergence," *J. Math. Anal. Appl.*, vol. 196, no. 2, pp. 676–698, 1995.

[36] H. Federer, *Geometric Measure Theory*. New York, NY, USA: Springer, 1969.

[37] S. S. Sastry, *Nonlinear Systems*. New York, NY, USA: Springer, 1999.