What Matters when Modeling Human Behavior using Imitation Learning?

Aneri Muni [1 2] **Esther Derman** [1 2] **Vincent Taboga** [1 2] **Pierre-Luc Bacon** [1 2] **Erick Delage** [3 2]

Abstract

As AI systems become increasingly embedded in human decision-making process, aligning their behavior with human values is critical to ensuring safe and trustworthy deployment. A central approach to AI Alignment called Imitation Learning (IL), trains a learner to directly mimic desirable human behaviors from expert demonstrations. However, standard IL methods assume that (1) experts act to optimize expected returns; (2) expert policies are Markovian. Both assumptions are inconsistent with empirical findings from behavioral economics, according to which humans are (1) risk-sensitive; and (2) make decisions based on past experience. In this work, we examine the implications of risk sensitivity for IL and show that standard approaches do not capture all optimal policies under risk-sensitive decision criteria. By characterizing these expert policies, we identify key limitations of existing IL algorithms in replicating expert performance in risk-sensitive settings. Our findings underscore the need for new IL frameworks that account for both risk-aware preferences and temporal dependencies to faithfully align AI behavior with human experts.

1. Introduction

With the rapid development and widespread deployment of AI-driven technologies, people are increasingly relying on AI systems for making decisions in their daily lives. From personalized recommendations to critical applications in healthcare, finance and transportation, AI is shaping how individuals and societies make choices, creating an imminent need for AI systems that can quickly adapt to nuanced human preferences from limited behavioral data. There is a plethora of literature that highlights the benefits of aligning AI systems with human intentions (Sadigh et al., 2018; Carroll et al., 2019; Mandlekar et al., 2021; Kwon et al., 2020; Ethayarajh et al., 2024). However, formalizing reward objectives to train such aligned AI systems is far from trivial, as it requires incorporating diverse human preferences, values, and user beliefs.

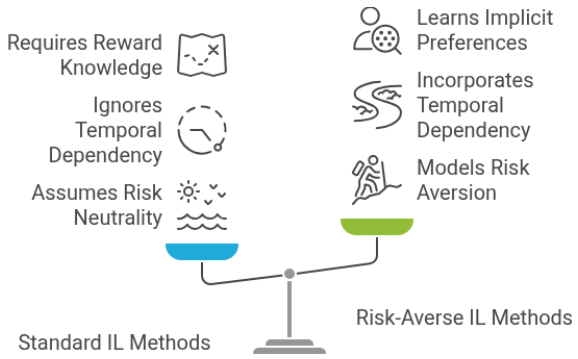


Figure 1. Key characteristics for designing IL algorithms that better reflect human preferences and decision-making. (Image generated using Napkin AI).

Instead of hand-crafting reward functions that encode human intentions, a common approach to modeling human decision-making is through *expert* demonstrations. These demonstrations can be used to either (1) directly learn a policy to match the expert's performance, commonly known as *Imitation Learning (IL)* (Pomerleau, 1991; Ross et al., 2011) or (2) indirectly imitate the policy by learning the expert's reward function, an approach known as *Inverse Reinforcement Learning (IRL)* (Abbeel & Ng, 2004; Ng & Russell, 2000). The term "expert" highlights the implicit assumption that the demonstrator makes decisions optimally w.r.t. its (unknown) objective.

Existing IL and IRL methods commonly assume that the human (expert) demonstrator is optimizing the expected return, implying that the underlying objective is known and the expert's policy is *risk neutral*. This modeling assumption corresponds to the expected utility theory (EUT) proposed by (Neumann & Morgenstern, 1944), and contradicts empirical studies from behavioral economics and decision theory (Kahneman & Tversky, 1979; Ellsberg, 1961; Tversky, 1975). To better understand why, consider an experiment in which a subject is asked to pick between two bets in each of the following scenarios:

[1]Université de Montréal, Montréal, Canada [2]Mila - Québec AI Institute, Montréal, Canada [3]Department of Decision Sciences & GERAD, HEC Montréal, Canada. Correspondence to: Aneri Muni <aneri.muni@mila.quebec>.

Proceedings of the ICML 2025 Workshop on Models of Human Feedback for AI Alignment, Vancouver, Canada.

Scenario I	A = $(\$1000, 0.5, \$0)$	B = $(\$400)$
Scenario II	C = $(\$1000, 0.1, \$0)$	D = $(\$400, 0.2, \$0)$

For instance, in Scenario I the bet A implies that the subject receives either $\$1000$ or $\$0$ with probability 0.5 whereas in bet B the subject receives $\$400$ with probability 1. The expected utility of each bet can be given as $\{A : 500, B : 400, C : 100, D : 200\}$. Experiments by (Kahneman & Tversky, 1979; Tversky, 1975) showed that subjects consistently choose bet B over A, and C over D. This behavior contradicts the *objective* EU optimum and highlights the limitation of the EU theory in modeling *subjective* human preferences.

The key insight of this paper is to model humans as risk-averse decision-making agents. We provide theoretical arguments for rethinking IL methods when the expert data is human-generated. We identify three key limitations of common IL methods: (1) They assume that the expert demonstrations are optimal for the expected return; (2) They require prior knowledge of the expert's risk measure and/or the reward model input to the risk measure; and (3) They rely on occupancy matching and Markov policies, which are insufficient for capturing history-dependent behavior, as is often the case for risk-sensitive agents. Our findings highlight the need for IL frameworks that accommodate temporal dependencies and risk-aware decision-making.

2. Background

2.1. Risk-Neutral Reinforcement Learning

A discounted Markov decision process (MDP) is a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, P, r, p_0, \gamma)$ where $\mathcal{S}$ is the state-space, $\mathcal{A}$ is the action space, $P(\cdot|s,a)$ is the transition probability from state s by taking action a, $r(s,a)$ is the immediate reward, p_0 is the initial state-distribution, and $\gamma \in (0,1)$ is a discount factor. A trajectory up to time t is denoted by $h_t := (s_0, a_0, \cdots, s_{t-1}, a_{t-1}, s_t)$, which belongs to the set of length-t histories $\mathcal{H}_t$. A policy is a sequence of decision rules $\boldsymbol{\pi} := (\pi_t)_{t\in\mathbb{N}}$ such that $\pi_t : \mathcal{H}_t \to \Delta_{\mathcal{A}}$, in which case $\boldsymbol{\pi} \in \Pi^{\text{H}}$. We call Markov policy any sequence $\boldsymbol{\pi}$ of decision rules $\pi_t : \mathcal{S} \to \Delta_{\mathcal{A}}$ that only depend on the current state, and denote by Π^{M} the set of Markov policies. The *risk-neutral* objective is to maximize $J(\boldsymbol{\pi}) := \lim_{T\to\infty} J_T(\boldsymbol{\pi})$ over policies $\boldsymbol{\pi}$ where

$$\begin{aligned} J_T(\boldsymbol{\pi}) :&= \mathbb{E}^{\pi_0}_{s_0}[r(s_0,a_0) + \gamma \mathbb{E}^{\pi_1}_{s_1}[r(s_1,a_1) + \cdots \\ &\quad + \gamma \mathbb{E}^{\pi_T}_{s_T}[r(s_{T-1}, a_{T-1})] \cdots]] \\ &= \mathbb{E}^{\boldsymbol{\pi}}\left[\sum_{t=0}^{T-1} \gamma^t r(s_t,a_t)\right]. \end{aligned} \tag{1}$$

It is known that there exists a stationary policy $\boldsymbol{\pi} = (\pi_0, \pi_0, \cdots)$ that is optimal for the above objective (1) (Puterman, 1994). Therefore, in a risk-neutral setting, restricting policy search to the class of stationary policies does not impair performance. In the sequel, we will denote by Π^{S} the set of stationary policies π, with a slight abuse of notation.

Occupancy measure. For any $\pi \in \Pi^{\text{S}}$, we define its occupancy measure $\mu^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as:

$$\mu^\pi(s,a) = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}^\pi(s_t = s, a_t = a). \tag{2}$$

One can formulate problem (1) as a linear program and get the duality:

$$J(\pi) = \sum_{s,a} \mu^\pi(s,a) r(s,a), \qquad \forall \pi \in \Pi^{\text{S}}. \tag{3}$$

For any $\pi \in \Pi^{\text{S}}$, its occupancy measure also satisfies a form of Bellman recursion:

$$p_0 = \sum_a \mu^\pi(\cdot, a) - \gamma \sum_{s',a} P(s'|\cdot, a)\mu^\pi(\cdot, a)$$

so that the set of occupancies $\mathcal{V} := \{\mu^\pi : \pi \in \Pi^{\text{S}}\}$ can be written as a feasible set of affine constraints, i.e.,

$$\begin{aligned} \mathcal{V} = \{\mu \in &\mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \mu \geq 0, \\ &p_0 = \sum_a \mu(\cdot,a) - \gamma \sum_{s',a} P(s'|\cdot,a)\mu(\cdot,a)\}. \end{aligned}$$

Finally, there is a one-to-one correspondence between Π^{S} and $\mathcal{V}$, as stated below.

Lemma 1 ((Puterman, 1994)[Thm. 6.9.1]). There exists a bijection $g : \Pi^{\text{S}} \to \mathcal{V}$ given by

$$\begin{aligned} g(\pi) &= \mu^\pi, \qquad \forall \pi \in \Pi^{\text{S}} \\ g^{-1}(\mu) &= \pi_\mu := \frac{\mu}{\langle \mu, \mathbf{1}_{\mathcal{A}} \rangle}, \qquad \forall \mu \in \mathcal{V}. \end{aligned}$$

2.2. Risk-Sensitive Reinforcement Learning

The risk-sensitive objective generalizes the risk-neutral objective by replacing the expectation by a possibly non-linear functional $\rho : \mathcal{Z} \to \mathbb{R}$ called a risk measure. The risk measure maps each random variable $Z \in \mathcal{Z}$ to a real number reflecting the agent's sensitivity to risk. Examples of risk measures include conditional value at risk (CVaR), entropic risk, or expectiles. Risk can be incorporated into an agent's decision criteria in different ways:

Static risk objective. Apply a risk measure ρ to the entire return, thus leading to the objective:

$$J_{\text{static}}(\boldsymbol{\pi}) = \rho\left(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right) \tag{4}$$

Optimal policies under static risks are in general non-Markovian (Bäuerle & Ott, 2011; Chow et al., 2015).

Risk-constrained objective. Optimize the expected return subject to a risk constraint with threshold β (Geibel & Wysotzki, 2005), which yields the objective:

$$J_{\text{constrained}}(\boldsymbol{\pi}) = \mathbb{E}^{\boldsymbol{\pi}}\left[\sum_{t=0}^{T-1}\gamma^t r(s_t,a_t)\right] \quad \text{s.t. } \rho\left(\sum_{t=0}^{T-1}\gamma^t r(s_t,a_t)\right) \leq \beta. \tag{5}$$

Optimal policies under the risk-constrained expected return criterion generally lie in the class of history-dependent policies (Chow et al., 2017; Greenberg et al., 2022).

Dynamic Risk Objective. Nested risk measures (also called dynamic risk measures) capture the risk of the rewards-to-go at each time step (Tamar et al., 2015; Majumdar et al., 2017; Coache & Jaimungal, 2021). It is written through the following recursion:

$$J_{\text{nested}}(\boldsymbol{\pi}) = \rho\left(r_0 + \gamma\rho\left(r_1 + \gamma\rho(r_2 + \cdots)\right)\right) \tag{6}$$

The nested risk formulation conveniently satisfies Bellman equations, and an optimal stationary deterministic policy exists as in risk-neutral RL (Ruszczyński, 2010). If the risk measure ρ satisfies the *tower property* (Lin Hau et al., 2023), then the dynamic risk objective is equivalent to the static risk objective. The expectation is one of them, see Eq. (1).

3. Imitation Learning: Problem Statement

Consider a set of expert demonstrations where each demonstration ζ is a sequence of state-action pairs,

$$\zeta = \{(s_0,a_0),(s_1,a_1),\dots\}$$

drawn from an expert policy $\boldsymbol{\pi}^{\text{E}}$. The demonstrations implicitly describe the expert's performance criterion. The IL problem aims to leverage a set of demonstrations, $\mathcal{D} = \{\zeta_1,\dots,\zeta_n\}$ from an expert policy $\boldsymbol{\pi}^{\text{E}}$ to learn a policy $\hat{\boldsymbol{\pi}}^{\text{E}}$ that imitates the expert policy $\boldsymbol{\pi}^{\text{E}}$. Most IL methods rely on at least one of the following assumptions, which we will further analyze in the realm of imitating risk-sensitive human behavior.

Assumption 1: The *expert acts optimally* w.r.t some unknown performance objective.

Assumption 2: The *expert is risk-neutral*, i.e., it is optimal for the objective (1).

Assumption 3: The *expert policy is stationary*, meaning that its action choices follow the same distribution over time. This implies that the *expert follows a Markov policy*, so decisions are independent of history.

Combining these three assumptions (or sufficiently, Assumption 2-3) implies that the expert policy $\boldsymbol{\pi}^{\text{E}} \in \Pi^{\text{S}}$. Therefore, most IL methods limit their search to the set Π^{S}, e.g., by parameterizing the learner policy with a feedforward network.

Behavioral Cloning (BC) is a type of IL method where an agent learns to mimic expert behavior by treating the problem as a supervised learning one (Pomerleau, 1991). Instead of learning from trial and error, the agent is trained on an offline dataset of state-action pairs collected from an expert, and the goal is to learn a policy that maximizes likelihood of the demonstration data:

$$\max_{\pi} \mathbb{E}_{\tau\sim\pi^{\text{E}}}\left[\sum_{(s,a)\in\tau}\log(\pi(a|s))\right]$$

Although simple and effective in some settings, BC ignores environmental dynamics, and therefore suffers from distributional shift. This leads the agent to perform poorly in unfamiliar states that are not seen in the expert demonstrations.

Distribution Matching methods learn a policy by minimizing the divergence between the γ-discounted state-action distribution under the learner's policy π and the discounted state-action distribution of the expert policy π^{E}. This approach is often combined with the MaxEnt IRL framework (Ziebart et al., 2008; Ho & Ermon, 2016; Fu et al., 2017) resulting in the minimization problem:

$$J(\pi) = \mathcal{D}_f(\mu^{\pi}||\mu^{\pi^{\text{E}}}) - \lambda H(\pi),$$

where $H(\pi) = \mathbb{E}_{\mu^{\pi}}\log(\pi(a|s))$ is the causal entropy of the policy, and $\mathcal{D}_f = \mathbb{E}_{\mu^{\pi}}\left[f\left(\frac{\mu^{\pi}(s,a)}{\mu^{\pi^{\text{E}}}(s,a)}\right)\right]$ is an f-divergence (Ghasemipour et al., 2019). Distributional matching approaches include generative adversarial IL (GAIL) (Ho & Ermon, 2016), AIRL (Fu et al., 2017), RS-GAIL (Lacotte et al., 2018), f-MAX (Ghasemipour et al., 2019), PWIL (Dadashi et al., 2020), SIL (Papagiannis & Li, 2022), and ValueDice (Kostrikov et al., 2019).

4. Limitations of Imitation Learning

The limitations of existing IL methods in modeling human behavior can be cast into three categories, each of them being problematic for risk-sensitive experts.

Risk-neutral expert. Starting from the seminal work of (Ng & Russell, 2000; Abbeel & Ng, 2004), several authors have proposed IL algorithms, including (Ratliff et al., 2006; Ziebart et al., 2008; Syed et al., 2008; Ho et al., 2016). All of these approaches were developed under the *Assumption 2* of a risk-neutral expert.

(Ho & Ermon, 2016; Fu et al., 2017) presented a distribution matching perspective on IL. Leveraging the dual formulation of RL (Lemma 1), they reformulate IL as an occupancy matching problem. The authors then proposed an adversarial IL method, GAIL, to learn a policy that matches the

occupancy measure of the expert by training a discriminator (i.e. a classifier) between expert and learner data. The discriminator serves as a proxy reward signal used for policy updates. GAIL was designed under the assumption that the expert is maximizing an entropy-regularized expected return criterion.

Non-adversarial IL methods such as IQ-learn (Garg et al., 2021) and SQIL (Reddy et al., 2020) leverage the closed-form optimal policy of the maximum entropy RL objective, which also relies on a risk-neutral Bellman equation. Off-policy IL methods such as ValueDice (Kostrikov et al., 2019) also follow *Assumption 2*. Recently, (Ghasemipour et al., 2019) showed that many well-known IL methods can be viewed as special instances of a general divergence minimization problem between the state-action distribution induced by the learner and the expert policies. However, these occupancy matching algorithms all rely on the hypothesis that the expert is maximizing the expected return.

Interestingly, we note that behavioral cloning, and its online version, DAgger (Ross et al., 2011), both solve IL as a supervised learning problem, without making any assumptions on the expert's underlying decision-criteria. These methods directly aim to maximize the probability of expert actions under the learned policy. Recently, IL with optimal transport (Dadashi et al., 2020; Papagiannis & Li, 2022) has emerged as an approach to minimize the distance between the state-action distribution of the expert and the learner, in its primal form, rather than the dual (as done in discriminator-based IL methods), without any assumptions on the objective function of the expert. These methods are promising in avoiding the limitations of *Assumption 2*. However, in the next sections we show that merely matching state-action occupancy distributions is insufficient to guarantee that the learned policy will match the risk-sensitive expert's performance.

Known risk or reward function. Several works have incorporated risk aversion into the IRL or IL frameworks. (Ratliff & Mazumdar, 2017) proposed a gradient-based inverse risk-sensitive RL formulation leveraging the risk measure proposed by prospect theory (Kahneman & Tversky, 1979). To this end, the authors assume access to the expert's nominal reward, and learn the parameters of the risk function while learning a policy under which the learner's behavior matches that of the demonstration dataset. (Majumdar et al., 2017; Singh et al., 2018) proposed an IRL approach to learn both the reward and the risk preference of an expert acting according to a coherent risk-aware (Shapiro et al., 2014) objective. However, their approach involves solving a linear program for every (s, a) data point in the demonstration dataset, making it difficult to scale. Recently, (Lazzati & Metelli, 2024) proposed an IRL method under the expected utility framework (*Assumption 2*) that aims to learn the utility of a risk-averse agent. While their approach considers a class of non-Markov policies (thus avoiding *Assumption 3*), they assume access to the agent's reward function while learning the utility function.

Another line of work builds on GAIL to accommodate learning from a risk-averse expert. (Santara et al., 2017) proposed Risk-Averse IL (RAIL), which modifies GAIL's objective to now learn a policy that achieves a maximum expected sum of discounted rewards with a conditional value-at-risk (CVaR_α) at least as good as the expert's. (Lacotte et al., 2018) later showed that RAIL does not accurately take the expert's risk into account, and proposed Risk Sensitive GAIL (RS-GAIL) to fix this limitation. Both methods correctly identified the limitation of *Assumption 2* ((Lacotte et al., 2018)[Theorem 3]), and modified GAIL's risk-neutral objective with a risk-constrained performance metric. However, they fail to account for the fact that this change in objective leads to an expert that is non-Markovian. By only searching for a Markov expert policy, RAIL and RS-GAIL, still follow *Assumption 3* and are therefore insufficient to effectively match the expert's performance.

Moreover, both approaches assume that the risk level α of the expert is known apriori, a strong assumption in practice. Additionally, the RS-GAIL algorithm relies on the dual representation of coherent risk measures (Ang et al., 2018), and cannot be used to model non-coherent risks like an entropic risk (Föllmer & Schied, 2016).

Markov expert policy. In addition to assuming access to the expert's risk measure, both RAIL and RS-GAIL restrict their search for policies to the set Π^{M}. However, as illustrated in Section 2.2, it has been shown that for the case of static risk objectives eq. (4) and risk-constrained objectives eq. (5), all optimal policies can be history-dependent (Bäuerle & Ott, 2011; Bäuerle & Rieder, 2014). This implies that to accurately imitate a risk-averse agent, one would need to search over a class of non-Markovian policies, Π^{H}, for which the class of Markov policies is only a subset. In the following section, we present a simple example that highlights the importance of considering history-dependent policies when modeling risk-averse experts.

5. Imitating Risk-Sensitive Experts

5.1. Nested Risk-Averse Imitation Learning

When the expert dataset is the result of an optimal policy for a nested risk measure, namely Eq. (6), we show that occupancy matching is sufficient to recover an optimal risk-sensitive return, regardless of whether the expert policy was Markov or history-dependent.

Theorem 5.1. *Denote by $\boldsymbol{\pi}^{\text{E}}$ the expert policy and $\mu^{\boldsymbol{\pi}^{\text{E}}}$ the associated occupancy measure. Given a Markov policy $\boldsymbol{\pi} \in \Pi^{\text{M}}$ and its occupancy measure $\mu^{\boldsymbol{\pi}}$, if $\mu^{\boldsymbol{\pi}} = \mu^{\boldsymbol{\pi}^{\text{E}}}$, then it holds that $J_{nested}(\boldsymbol{\pi}) \geq J_{nested}(\boldsymbol{\pi}^{\text{E}})$.*

Proof. Under mild assumptions on the risk measure, for any Markov policy, the risk-sensitive Bellman operator is a contraction (Shen et al., 2013)[Prop. 5.2]. Using monotonicity of risk measures, the risk-sensitive value for policy $\boldsymbol{\pi}$ is also the optimal solution of:

$$\max\langle v, \mu_0\rangle \text{ s.t. } v \leq r^{\boldsymbol{\pi}} + \gamma\rho_{\text{nested}}(v)$$

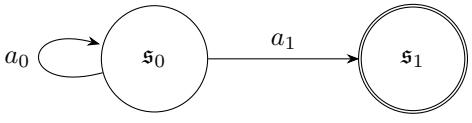
The solution of the dual gives the occupancy measure of $\boldsymbol{\pi}$, which is the same as that of $\boldsymbol{\pi}^{\text{E}}$, by assumption. Passing again to the primal yields the risk-sensitive value of $\boldsymbol{\pi}$, which is the same as that of $\boldsymbol{\pi}$ by strong duality. □

A direct consequence of this result is that distribution matching is sufficient to imitate a risk-averse expert when the risk measure is nested. Therefore, existing IL methods that minimize statistical divergences between occupancy measures can be used to imitate risk-averse experts. This includes methods like PWIL (Dadashi et al., 2020) and SIL (Papagiannis & Li, 2022) that directly learn a policy matching the occupancy measure induced by the expert's policy.

5.2. Static Risk-Averse Imitation Learning

All optimal policies of a static risk-sensitive objective may lie in the class of history-dependent policies (Ruszczyński, 2010; Bäuerle & Ott, 2011). We may apply (Laroche & Tachet Des Combes, 2023)[Thm. 4], first proposed by (Szepesvári, 2010), stating under mild assumptions that the occupancy measure of any history-dependent policy, i.e., the distribution of transition samples collected with it, can be equivalently generated by a Markov one. Then, a naive idea would be to find a Markov policy with the same occupancy measure as a static risk-averse agent, despite optimal polices being non-Markovian. However, as we establish next, this is insufficient and can be arbitrarily suboptimal. Through the counterexample below, we indeed show that two policies inducing the same occupancy measure may not perform equivalently for a static risk measure. Therefore, IL methods explicitly need to account for the history dependence when mimicking the expert's policy.

Consider the simple MDP below adapted from (Laroche & Tachet Des Combes, 2023). It has two states and two actions: $\mathcal{S} = \{\mathfrak{s}_0, \mathfrak{s}_1\}$, $\mathcal{A} = \{a_0, a_1\}$. Transition dynamics are given by $p(\mathfrak{s}_0|\mathfrak{s}_0, a_0) = 1$ and $p(\mathfrak{s}_1|\mathfrak{s}_0, a_1) = 1$, $\mathfrak{s}_0$ is the starting state and $\mathfrak{s}_1$ is an absorbing state where the episode effectively terminates.



Consider a reward function of the form:

$$r(\mathfrak{s}_0, a_0) = -1, \quad r(\mathfrak{s}_0, a_1) \sim \mathcal{N}(0, \sigma^2)$$
$$r(\mathfrak{s}_1, a_0) = 0, \quad r(\mathfrak{s}_1, a_1) = 0.$$

The **Entropic Risk Measure (ERM)** with parameter $\alpha \in \mathbb{R}_+ \cup \{\infty\}$ is a concave risk measure defined for a random variable $X \in \mathbb{X}$ as (Föllmer & Schied, 2016):

$$\text{ERM}_\alpha[X] = -\frac{1}{\alpha} \cdot \log\left(\mathbb{E}\left[e^{-\alpha X}\right]\right).$$

The ERM of a Gaussian variable $X \sim \mathcal{N}(\mu, \sigma^2)$ can be conveniently written as:

$$\text{ERM}_\alpha[X \sim \mathcal{N}(\mu, \sigma^2)] = \mu - \frac{\alpha}{2}\sigma^2.$$

We are interested in maximizing the ERM of the infinite-horizon discounted return so an optimal policy would be:

$$\boldsymbol{\pi}^* \in \arg\max_{\boldsymbol{\pi}} \text{ERM}_\alpha\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]. \tag{7}$$

History-dependent policy. Given some timestep T, consider the following deterministic, non-Markov policy $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_t)_{t\in\mathbb{N}}$ of the form:

$$\begin{aligned} &\tilde{\pi}_t(a = a_0|s = \mathfrak{s}_0) = 1 \text{ if } t < T, \\ &\tilde{\pi}_t(a = a_1|s = \mathfrak{s}_0) = 1 \text{ otherwise.} \end{aligned} \tag{8}$$

This policy takes action a_0 for T steps, then takes action a_1. The occupancy measure induced by the above policy is:

$$\mu^{\tilde{\boldsymbol{\pi}}}(s = \mathfrak{s}_0, a = a_0) = \frac{1-\gamma^T}{1-\gamma}$$
$$\mu^{\tilde{\boldsymbol{\pi}}}(s = \mathfrak{s}_0, a = a_1) = \gamma^T.$$

The static entropic risk of policy $\tilde{\boldsymbol{\pi}}$ from Eq. (8) is:

$$\begin{aligned} \text{ERM}_\alpha^{\tilde{\boldsymbol{\pi}}} &:= \text{ERM}_\alpha\left(\sum_{t=0}^{T-1} \gamma^t r(a_0) + \gamma^T r(a_1)\right) \\ &= -\frac{1-\gamma^T}{1-\gamma} + \text{ERM}_\alpha(\gamma^T r(a_1)) \\ &= -\frac{1-\gamma^T}{1-\gamma} - \frac{\alpha\gamma^{2T}\sigma^2}{2} \\ &= \left(-\frac{\alpha\sigma^2}{2}\right)(\gamma^T)^2 + \left(\frac{1}{1-\gamma}\right)(\gamma^T) - \frac{1}{1-\gamma} \\ &= -\frac{a}{2}x^2 + bx - c. \end{aligned} \tag{9}$$

where $a := \alpha\sigma^2$, $b = c := \frac{1}{1-\gamma}$. To find an optimal policy, we must find a stationary point $\nabla_x J = 0$. Thus, $x^* = b/a$. Substituting yields the closed-form solution:

$$(\gamma^T)^* = \frac{\frac{1}{1-\gamma}}{\alpha\sigma^2} = \frac{1}{(1-\gamma)\alpha\sigma^2}.$$

Since $\gamma \in (0,1)$, we are interested in cases where $\lceil (1-\gamma)\alpha\sigma^2 \rceil \geq 1$. Thus, we have:

$$T = \left\lfloor -\log_\gamma \left((1-\gamma)\alpha\sigma^2\right) \right\rfloor \text{ or } \left\lceil -\log_\gamma \left((1-\gamma)\alpha\sigma^2\right) \right\rceil.$$

In the case where $-\log_\gamma((1-\gamma)\alpha\sigma^2)$ is an integer, we can substitute T in Eq. (9) to get the return:

$$\begin{aligned}\text{ERM}_\alpha^{\tilde{\boldsymbol{\pi}}} &= -\frac{1}{2(1-\gamma)^2\alpha\sigma^2} + \frac{1}{(1-\gamma)^2\alpha\sigma^2} - \frac{1}{(1-\gamma)} \\ &= \frac{1}{(1-\gamma)}\left[-1 + \frac{1}{2(1-\gamma)\alpha\sigma^2}\right]. \end{aligned} \tag{10}$$

Assume the optimal time step is $T = n \geq 1$. We then compute the occupancy measure induced by $\tilde{\boldsymbol{\pi}}$:

$$\begin{aligned}\mu^{\tilde{\boldsymbol{\pi}}}(s=\mathfrak{s}_0, a=a_0) &= 1 \cdot \sum_{t=0}^{T-1} \gamma^t \mathbf{1}(s=\mathfrak{s}_0, a=a_0) \\ &= \frac{1-\gamma^T}{1-\gamma} = \frac{1-\gamma^n}{1-\gamma}, \\ \mu^{\tilde{\boldsymbol{\pi}}}(s=\mathfrak{s}_0, a=a_1) &= 1 \cdot \gamma^T \mathbf{1}(s=\mathfrak{s}_0, a=a_0) \\ &= \gamma^T = \gamma^n, \\ \mu^{\tilde{\boldsymbol{\pi}}}(s=\mathfrak{s}_1, a=\emptyset) &= 1 \cdot \sum_{t=T}^{\infty} \gamma^t \mathbf{1}(s=\mathfrak{s}_1, a=\emptyset) \\ &= \gamma^T \cdot \frac{1}{1-\gamma} = \frac{\gamma^n}{1-\gamma}. \end{aligned} \tag{11}$$

Equivalent Markov policy. Since our goal is to compare the performance of a history-dependent policy with that of an "equivalent" (in terms of the induced occupancy measure) Markov policy, consider the following stationary policy π:

$$\begin{aligned}\pi(a_0|\mathfrak{s}_0) &= \frac{\mu^{\tilde{\boldsymbol{\pi}}}(\mathfrak{s}_0, a_0)}{\mu^{\tilde{\boldsymbol{\pi}}}(\mathfrak{s}_0)} = \frac{1-\gamma^T}{1-\gamma^{T+1}}, \\ \pi(a_1|\mathfrak{s}_0) &= \frac{\mu^{\tilde{\boldsymbol{\pi}}}(\mathfrak{s}_0, a_1)}{\mu^{\tilde{\boldsymbol{\pi}}}(\mathfrak{s}_0)} = \frac{\gamma^T(1-\gamma)}{1-\gamma^{T+1}}. \end{aligned} \tag{12}$$

We want to analyze the occupancy measure induced by π. We first substitute $T = n$ in Eq. (12), and use $\tilde{T}$ to denote the time step at which the agent acting under the Markov policy takes action a_1 in state $\mathfrak{s}_0$:

$$\begin{aligned}\mu^\pi(s=s_0) &= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{1}(s_t = \mathfrak{s}_0)\right] = \mathbb{E}\left[\sum_{t=0}^{\tilde{T}} \gamma^t\right] \\ &= \sum_{T=0}^{\infty}\sum_{t=0}^{T} \gamma^t \Pr(\tilde{T}=T) \\ &= \sum_{T=0}^{\infty}\sum_{t=0}^{T} \gamma^t \left(\frac{1-\gamma^n}{1-\gamma^{n+1}}\right)^T \left(\frac{\gamma^n(1-\gamma)}{1-\gamma^{n+1}}\right) \\ &= \sum_{T=0}^{\infty} \left(\frac{1-\gamma^n}{1-\gamma^{n+1}}\right)^T \left(\frac{\gamma^n(1-\gamma)}{1-\gamma^{n+1}}\right) \sum_{t=0}^{T} \gamma^t \\ &= \sum_{T=0}^{\infty} \left(\frac{1-\gamma^n}{1-\gamma^{n+1}}\right)^T \left(\frac{\gamma^n(1-\gamma)}{1-\gamma^{n+1}}\right) \left(\frac{1-\gamma^{T+1}}{1-\gamma}\right) \\ &= \left(\frac{\gamma^n}{1-\gamma^{n+1}}\right) \sum_{T=0}^{\infty} \left(\frac{1-\gamma^n}{1-\gamma^{n+1}}\right)^T (1-\gamma^{T+1}) \\ &= \left(\frac{\gamma^n}{1-\gamma^{n+1}}\right) \left[\sum_{T=0}^{\infty} \left(\frac{1-\gamma^n}{1-\gamma^{n+1}}\right)^T \right. \\ &\qquad \left. - \gamma \sum_{T=0}^{\infty} \gamma^T \left(\frac{1-\gamma^n}{1-\gamma^{n+1}}\right)^T\right] \\ &= \left(\frac{\gamma^n}{1-\gamma}\right)\left(\frac{1}{\gamma^n} - \gamma\right) = \left(\frac{1-\gamma^{n+1}}{1-\gamma}\right), \end{aligned}$$

where by construction, $\tilde{T}$ is a geometric distribution of success parameter $\frac{1-\gamma^n}{1-\gamma^{n+1}}$. We then have:

$$\begin{aligned}\mu^\pi(s=\mathfrak{s}_0, a=a_0) &= \mu^\pi(s=\mathfrak{s}_0)\pi(\mathfrak{s}_0, a=a_0) \\ &= \frac{1-\gamma^{n+1}}{1-\gamma}\frac{1-\gamma^n}{1-\gamma^{n+1}} = \frac{1-\gamma^n}{1-\gamma}, \\ \mu^\pi(s=\mathfrak{s}_0, a=a_0) &= \mu^\pi(s=\mathfrak{s}_0)\pi(\mathfrak{s}_0, a=a_1) \\ &= \frac{1-\gamma^{n+1}}{1-\gamma}\frac{\gamma^n(1-\gamma)}{1-\gamma^{n+1}} = \gamma^n. \end{aligned} \tag{13}$$

Comparing the occupancy measure induced by $\tilde{\boldsymbol{\pi}}$ in equation (11) with the occupancy measure under π in equation (13), we have shown that $\mu^\pi = \mu^{\tilde{\boldsymbol{\pi}}}$. We are interested in comparing the performance of these two policies. Define:

$$P_T := \mathbb{P}(\tilde{T}=T) = \left(\frac{1-\gamma^n}{1-\gamma^{n+1}}\right)^T \left(\frac{\gamma^n(1-\gamma)}{1-\gamma^{n+1}}\right).$$

The static entropic risk performance of policy π is:

$$\begin{aligned}\text{ERM}_\alpha^\pi &= \text{ERM}_\alpha\left(-\frac{1-\gamma^{\tilde{T}}}{1-\gamma} + \gamma^{\tilde{T}} r(a_1)\right) \\ &= -\frac{1}{\alpha}\log\left(\sum_{T=0}^{\infty} P_T \exp\left(-\alpha\left[-\left(\frac{1-\gamma^T}{1-\gamma}\right)\right.\right.\right. \\ &\qquad \left.\left.\left. -\frac{\alpha}{2}\gamma^{2T}\sigma^2\right]\right)\right). \end{aligned} \tag{14}$$

In particular, take $\alpha = \frac{10}{0.9}, \sigma^2 = 1, \gamma = 0.9$. Then, substituting these values in Eqs. (14) and (10) and numerically solving for the performance of both policies, we get:

$$T = 1, \quad \text{ERM}_\alpha^{\tilde{\pi}} = -5.5, \quad \text{ERM}_\alpha^{\pi} = -8.02.$$

This indicates that the return under the optimal, deterministic, non-Markov policy $\tilde{\pi}$ is strictly greater than that of the equivalent Markov policy π, despite both policies inducing the same state-action occupancy measure.

Consider now a case where we hold $\alpha = \frac{10}{0.9}$ and $\gamma = 0.9$ constant, and vary σ values such that $\sigma^2 = (0.9)^{-k}$ where $k \in \mathbb{N}$. Then, T will result in a positive integer value, ensuring that the occupancy induced by π and $\tilde{\pi}$ is the same. We plot the corresponding performance of each policy in Fig. 2 and provide numerical values in Tab. 1. The results show that as σ increases, the performance gap between Markov and non-Markov policies diverges, highlighting the importance of the choice of policy class when mimicking risk-averse experts.

Table 1. Comparing return of policies π and $\tilde{\pi}$ with varying values of σ such that $\sigma(k) := (0.9)^{\frac{-k}{2}}$ where $k = 0, 1, 2 \ldots$.

σ^2	T	$\text{ERM}_\alpha^{\tilde{\pi}}$	ERM_α^{π}
$(0.9)^{0}$	1	-5.5	-8.02
$(0.9)^{-1}$	2	-5.95	-8.77
$(0.9)^{-2}$	3	-6.35	-9.12
$(0.9)^{-5}$	6	-7.34	-9.57
$(0.9)^{-10}$	11	-8.43	-15.65
$(0.9)^{-20}$	21	-9.45	-45.29
$(0.9)^{-50}$	51	-9.98	-1077.27

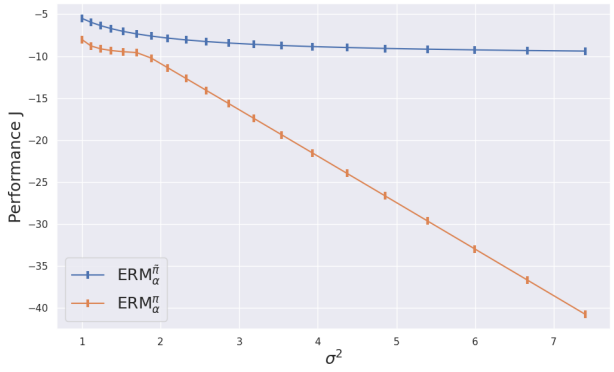


Figure 2. Comparing the performance of π and $\tilde{\pi}$ as σ^2 increases. Modeling a risk-averse expert using Markov policies can result in an arbitrarily large performance gap with the expert's performance.

This counterexample questions the modeling assumptions behind IL methods used for mimicking human behavior. Specifically, parameterizing the expert's objective with a static risk measure or using a risk-constrained formulation must rely on a careful choice of policy class as in both cases, optimal policies may be history-dependent. An IL algorithm that only searches for Markov policies may thus underperform the expert policy. Specifically, the state-action distribution matching objective commonly used in IL (Ho & Ermon, 2016; Dadashi et al., 2020; Ghasemipour et al., 2019) can be problematic for risk-sensitive data . Indeed, as shown above, occupancy matching is not a sufficient statistic for capturing the behavior of a risk-averse expert. Since standard IL algorithms focus on Markov or even stationary policies, they may converge to a Markov policy that matches the occupancy measure of the expert but fails to capture its risk-sensitive preferences. This highlights the need for IL methods that address such non-identifiability.

Appx. A presents a counterexample using the static CVaR objective, leading to similar conclusions as the ERM example. This demonstrates that the observed performance mismatch is not specific to entropic risk, but rather holds for general risk measures.

5.3. Constrained Risk-Averse Imitation Learning

Although most of our analysis focused on *static* and *nested* risk-averse experts, any IL method developed under *Assumptions 2-3* also fails in modeling a risk-constrained expert. This is because an optimal policy of the risk-constrained decision criteria (Eq. (5)) can be non-Markov and therefore requires appropriate policy parameterization to mimic the corresponding risk-averse expert.

6. Conclusion and Future Directions

To faithfully model human behavior and achieve effective alignment, IL methods must move beyond the assumptions of risk-neutrality and Markovianity. Risk-averse behavior, especially under static risk criteria, necessitates modeling temporal dependencies that current methods overlook.

We call for the development of IL algorithms that: (1) Infer risk-sensitive policies from demonstration data without assuming access to reward or risk models; (2) Support non-Markov policy classes, for example, parameterizing policies with RNNs or LSTMs; (3) Utilize richer statistics than occupancy measures to capture risk-sensitive behavior, for example, trajectory matching.

Aligning AI behavior with human decision-making requires confronting the complexity of risk preferences. Modeling humans as risk-averse agents is fundamental for building intelligent systems that can interact with users effectively, align with their preferences, and contribute to the development of user-centric AI applications. We believe this work highlights important and promising future directions towards a better understanding of human feedback models and better AI alignment.

References

Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, pp. 1, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138385. doi: 10.1145/1015330.1015430. URL https://doi.org/10.1145/1015330.1015430.

Ang, M., Sun, J., and Yao, Q. On the dual representation of coherent risk measures. *Annals of Operations Research*, 262, 03 2018. doi: 10.1007/s10479-017-2441-3.

Bäuerle, N. and Rieder, U. More risk-sensitive markov decision processes. *Math. Oper. Res.*, 39(1):105–120, February 2014. ISSN 0364-765X. doi: 10.1287/moor.2013.0601. URL https://doi.org/10.1287/moor.2013.0601.

Bäuerle, N. and Ott, J. Markov decision processes with average-value-at-risk criteria. *Mathematical Methods of Operations Research*, 74:361–379, 12 2011. doi: 10.1007/s00186-011-0367-0.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T. L., Seshia, S. A., Abbeel, P., and Dragan, A. *On the utility of learning about humans for human-AI coordination*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Chow, Y., Tamar, A., Mannor, S., and Pavone, M. Risk-sensitive and robust decision-making: a cvar optimization approach. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pp. 1522–1530, Cambridge, MA, USA, 2015. MIT Press.

Chow, Y., Ghavamzadeh, M., Janson, L., and Pavone, M. Risk-constrained reinforcement learning with percentile risk criteria. *J. Mach. Learn. Res.*, 18(1):6070–6120, January 2017. ISSN 1532-4435.

Coache, A. and Jaimungal, S. Reinforcement learning with dynamic convex risk measures. *Mathematical Finance*, 34:557 – 587, 2021. URL https://api.semanticscholar.org/CorpusID:245501888.

Dadashi, R., Hussenot, L., Geist, M., and Pietquin, O. Primal wasserstein imitation learning. *ArXiv*, abs/2006.04678, 2020. URL https://api.semanticscholar.org/CorpusID:219531578.

Ellsberg, D. Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*, 75(4):643–669, 1961. ISSN 00335533, 15314650. URL http://www.jstor.org/stable/1884324.

Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., and Kiela, D. Kto: Model alignment as prospect theoretic optimization. *ArXiv*, abs/2402.01306, 2024. URL https://api.semanticscholar.org/CorpusID:267406810.

Fu, J., Luo, K., and Levine, S. Learning robust rewards with adversarial inverse reinforcement learning. 10 2017. doi: 10.48550/arXiv.1710.11248.

Föllmer, H. and Schied, A. *Stochastic Finance*. De Gruyter, Berlin, Boston, 2016. ISBN 9783110463453. doi: doi: 10.1515/9783110463453. URL https://doi.org/10.1515/9783110463453.

Garg, D., Chakraborty, S., Cundy, C., Song, J., and Ermon, S. Iq-learn: inverse soft-q learning for imitation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Geibel, P. and Wysotzki, F. Risk-sensitive reinforcement learning applied to control under constraints. *J. Artif. Int. Res.*, 24(1):81–108, July 2005. ISSN 1076-9757.

Ghasemipour, S. K. S., Zemel, R. S., and Gu, S. S. A divergence minimization perspective on imitation learning methods. In *Conference on Robot Learning*, 2019. URL https://api.semanticscholar.org/CorpusID:207880680.

Greenberg, I., Chow, Y., Ghavamzadeh, M., and Mannor, S. Efficient risk-averse reinforcement learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Ho, J. and Ermon, S. Generative adversarial imitation learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 4572–4580, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Ho, J., Gupta, J., and Ermon, S. Model-free imitation learning with policy optimization. In *International conference on machine learning*, pp. 2760–2769. PMLR, 2016.

Kahneman, D. and Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1914185.

Kostrikov, I., Nachum, O., and Tompson, J. Imitation learning via off-policy distribution matching, 12 2019.

Kwon, M., Biyik, E., Talati, A., Bhasin, K., Losey, D. P., and Sadigh, D. When humans aren't optimal: Robots that collaborate with risk-aware humans. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, pp. 43–52, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367462. doi: 10.1145/3319502.3374832. URL https://doi.org/10.1145/3319502.3374832.

Lacotte, J., Chow, Y., Ghavamzadeh, M., and Pavone, M. Risk-sensitive generative adversarial imitation learning. *ArXiv*, abs/1808.04468, 2018. URL https://api.semanticscholar.org/CorpusID:52004328.

Laroche, R. and Tachet Des Combes, R. On the occupancy measure of non-Markovian policies in continuous MDPs. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 18548–18562. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/laroche23a.html.

Lazzati, F. and Metelli, A. M. Learning utilities from demonstrations in markov decision processes. *arXiv preprint arXiv:2409.17355*, 2024.

Lin Hau, J., Petrik, M., and Ghavamzadeh, M. Entropic risk optimization in discounted mdps. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 47–76. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/lin-hau23a.html.

Majumdar, A., Singh, S., Mandlekar, A., and Pavone, M. Risk-sensitive inverse reinforcement learning via coherent risk models, 07 2017.

Mandlekar, A., Xu, D., Wong, J., Nasiriany, S., Wang, C., Kulkarni, R., Fei-Fei, L., Savarese, S., Zhu, Y., and Mart'in-Mart'in, R. What matters in learning from offline human demonstrations for robot manipulation. *ArXiv*, abs/2108.03298, 2021. URL https://api.semanticscholar.org/CorpusID:236956615.

Neumann, J. V. and Morgenstern, O. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, USA, 1944.

Ng, A. Y. and Russell, S. J. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pp. 663–670, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

Papagiannis, G. and Li, Y. Imitation learning withsinkhorn distances. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part IV*, pp. 116–131, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-26411-5. doi: 10.1007/978-3-031-26412-2_8. URL https://doi.org/10.1007/978-3-031-26412-2_8.

Pomerleau, D. A. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3(1):88–97, 1991. doi: 10.1162/neco.1991.3.1.88.

Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994. ISBN 0471619779.

Ratliff, L. J. and Mazumdar, E. Inverse risk-sensitive reinforcement learning, 2017. URL https://arxiv.org/abs/1703.09842.

Ratliff, N. D., Bagnell, J. A., and Zinkevich, M. A. Maximum margin planning. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 729–736, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143936. URL https://doi.org/10.1145/1143844.1143936.

Reddy, S., Dragan, A. D., and Levine, S. SQIL: imitation learning via reinforcement learning with sparse rewards. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=S1xKd24twB.

Ross, S., Gordon, G., and Bagnell, D. A reduction of imitation learning and structured prediction to no-regret online learning. In Gordon, G., Dunson, D., and Dudík, M. (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/ross11a.html.

Ruszczyński, A. Risk-averse dynamic programming for markov decision processes. *Math. Program.*, 125(2): 235–261, oct 2010. ISSN 0025-5610.

Sadigh, D., Landolfi, N., Sastry, S. S., Seshia, S. A., and Dragan, A. D. Planning for cars that coordinate with people: leveraging effects on human actions for planning and active information gathering over human internal state. *Auton. Robots*, 42(7):1405–1426, October 2018. ISSN 0929-5593. doi: 10.1007/s10514-018-9746-1. URL https://doi.org/10.1007/s10514-018-9746-1.

Santara, A., Naik, A., Ravindran, B., Das, D., Mudigere, D., Avancha, S., and Kaul, B. Rail: Risk-averse imitation learning. 07 2017. doi: 10.48550/arXiv.1707.06658.

Shapiro, A., Dentcheva, D., and Ruszczyński, A. *Lectures on Stochastic Programming: Modeling and Theory, Second Edition*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2014. doi: 10.1137/1.9781611973433. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611973433.

Shen, Y., Stannat, W., and Obermayer, K. Risk-sensitive markov control processes. *SIAM Journal on Control and Optimization*, 51(5):3652–3672, 2013.

Singh, S., Lacotte, J., Majumdar, A., and Pavone, M. Risk-sensitive inverse reinforcement learning via semi- and non-parametric methods. *Int. J. Rob. Res.*, 37(13–14): 1713–1740, December 2018. ISSN 0278-3649. doi: 10.1177/0278364918772017. URL https://doi.org/10.1177/0278364918772017.

Syed, U., Bowling, M., and Schapire, R. E. Apprenticeship learning using linear programming. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 1032–1039, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390286. URL https://doi.org/10.1145/1390156.1390286.

Szepesvári, C. *Algorithms for Reinforcement Learning*, volume 4. 01 2010. doi: 10.2200/S00268ED1V01Y201005AIM009.

Tamar, A., Chow, Y., Ghavamzadeh, M., and Mannor, S. Policy gradient for coherent risk measures. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/024d7f84fff11dd7e8d9c510137a2381-Paper.pdf.

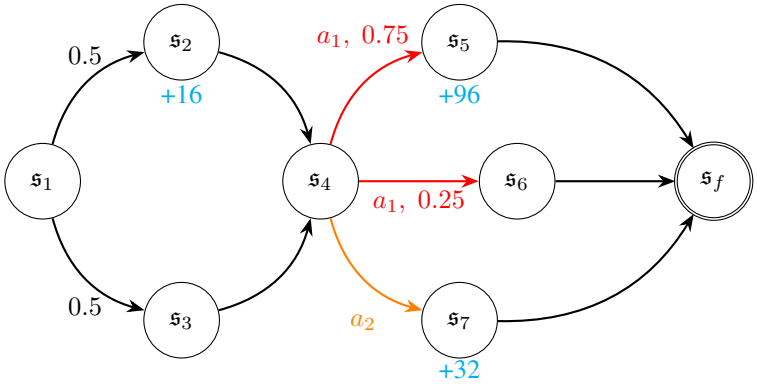
Tversky, A. A critique of expected utility theory: Descriptive and normative considerations. *Erkenntnis (1975-)*, 9(2):163–173, 1975. ISSN 01650106, 15728420. URL http://www.jstor.org/stable/20010465.

Wang, K., Liang, D., Kallus, N., and Sun, W. Risk-sensitive rl with optimized certainty equivalents via reduction to standard rl. *arXiv preprint arXiv:2403.06323*, 2024.

Ziebart, B., Maas, A., Bagnell, J., and Dey, A. Maximum entropy inverse reinforcement learning., 01 2008.

A. Counterexample with *static* Conditional Value-at-Risk (CVaR)

Consider the following synthetic MDP adapted from (Wang et al., 2024). The agent starts in state $\mathfrak{s}_1$. There is only one decision to take in state $\mathfrak{s}_4$ where two actions a_1, a_2 are available. A terminating state $\mathfrak{s}_f$ with no reward simplifies analysis for a discounted, infinite-horizon setting.



Stochastic Dynamics:

- $p(\mathfrak{s}_2|\mathfrak{s}_1) = 0.5$ and $p(\mathfrak{s}_3|\mathfrak{s}_1) = 0.5$
- $p(\mathfrak{s}_4|\mathfrak{s}_2) = 1$ and $p(\mathfrak{s}_4|\mathfrak{s}_3) = 1$
- $p(\mathfrak{s}_5|\mathfrak{s}_4, a_1) = 0.75$, $p(\mathfrak{s}_6|\mathfrak{s}_4, a_1) = 0.25$, and $p(\mathfrak{s}_7|\mathfrak{s}_4, a_2) = 1$
- $p(\mathfrak{s}_f|\mathfrak{s}_4) = 1$, $p(\mathfrak{s}_f|\mathfrak{s}_5) = 1$, and $p(\mathfrak{s}_f|\mathfrak{s}_6) = 1$

Reward model:

- $r(\mathfrak{s}_1) = r(\mathfrak{s}_3) = r(\mathfrak{s}_4) = r(\mathfrak{s}_6) = r(\mathfrak{s}_f) = 0$
- $r(\mathfrak{s}_2) = 16, r(\mathfrak{s}_5) = 96, r(\mathfrak{s}_7) = 32$

The agent starts from $\mathfrak{s}_1$ and aims to maximize the CVaR of the discounted cumulative reward:

$$\max_{\pi \in \Pi} \text{CVaR}_\alpha^\pi \left(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right).$$

In this MDP, the policy controls the action implemented when reaching $\mathfrak{s}_4$ at $t = 2$, which may be history-dependent, i.e., output a different action if $(s_0, s_1, s_2) = (\mathfrak{s}_1, \mathfrak{s}_2, \mathfrak{s}_4)$ or $(s_0, s_1, s_2) = (\mathfrak{s}_1, \mathfrak{s}_3, \mathfrak{s}_4)$. Consider the following policy:

$$\bar{\pi}(a_1|(\mathfrak{s}_1, \mathfrak{s}_2, \mathfrak{s}_4)) = 1 - \bar{\pi}(a_2|(\mathfrak{s}_1, \mathfrak{s}_2, \mathfrak{s}_4)) = 0, \qquad \bar{\pi}(a_1|(\mathfrak{s}_1, \mathfrak{s}_3, \mathfrak{s}_4)) = 1 - \bar{\pi}(a_1|(\mathfrak{s}_1, \mathfrak{s}_2, \mathfrak{s}_4)) = 0,$$

which takes action a_2 under trajectory $(\mathfrak{s}_1, \mathfrak{s}_2, \mathfrak{s}_4)$ and action a_1 under $(\mathfrak{s}_1, \mathfrak{s}_3, \mathfrak{s}_4)$. We compare the CVaR performance of this non-Markovian policy with that of three Markov policies: $\pi_1(\mathfrak{s}_4) = a_1$, $\pi_2(\mathfrak{s}_4) = a_2$ are deterministic, while $\pi_3(\mathfrak{s}_4) = Ba_1 + (1 - B)a_2$ with $B \sim \text{Bernoulli}(0.5)$ is stochastic.

Taking as CVaR confidence level $\alpha = 0.25$ and $\gamma = 0.5$ as discount factor, we present the CVaR return of $\bar{\pi}$ (history-dependent), π_1, π_2, π_2 (Markov stationary) in Tab. 2. As we can see, the strictly best-performing policy is the history-dependent one $\bar{\pi}$ with $\text{CVaR}_{0.25}^{\bar{\pi}} = 6 > \text{CVaR}_{0.25}^{\pi_i}, \forall i \in \{1, 2, 3\}$. The Markov randomized policy π_3 shows the worst performance among all. At the same time, $\bar{\pi}$ and π_3 share the same occupancy measure, as established next.

For any policy π, the occupancy measure $\mu_\pi(s, a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is given by:

$$\mu_\pi(s, a) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \mathbf{1}(s_t = s \wedge a_t = a) \right].$$

Table 2. The discounted cumulative reward distribution and $\text{CVaR}_{0.25}$ of three Markovian and one non-Markovian policies. For the distribution, $\{(v_i, p_i)\}_i$ denotes a random variable that takes value v_i w.p. $p_i \geq 0$, s.t. $\sum_i p_i = 1$.

Policy	**Disc. Cumulative Reward Dist.**	$\textbf{CVaR}_{0.25}$
π_1	$\{(0, \frac{1}{2}\cdot\frac{1}{4}), (8, \frac{1}{2}\cdot\frac{1}{4}), (12, \frac{1}{2}\cdot\frac{3}{4}), (20, \frac{1}{2}\cdot\frac{3}{4})\}$	4
π_2	$\{(4, \frac{1}{2}), (12, \frac{1}{2})\}$	4
π_3	$\{(0, \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{4}), (4, \frac{1}{2}\cdot\frac{1}{2}), (8, \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{4}),$ $(12, \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{3}{4} + \frac{1}{2}\cdot\frac{1}{2}), (20, \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{3}{4})\}$	3
$\bar{\pi}$	$\{(0, \frac{1}{2}\cdot\frac{1}{4}), (12, \frac{1}{2}\cdot\frac{3}{4} + \frac{1}{2})\}$	6

We thus compute

$$\mu_{\pi_3}(s) = \begin{cases} 1 & \text{if } s = \mathfrak{s}_1 \\ 1/4 & \text{if } s \in \{\mathfrak{s}_2, \mathfrak{s}_3, \mathfrak{s}_4\} \\ 3/64 & \text{if } s = \mathfrak{s}_5 \\ 1/64 & \text{if } s = \mathfrak{s}_6 \\ 1/16 & \text{if } s = \mathfrak{s}_7 \\ 1/16 & \text{if } s = \mathfrak{s}_f \end{cases}$$

and

$$\mu_{\pi_3}(\mathfrak{s}_4, a_1) = \mu_{\pi_3}(\mathfrak{s}_4)\pi_3(a_1|\mathfrak{s}_4) = 1/4 \cdot 1/2 = 1/8$$
$$\mu_{\pi_3}(\mathfrak{s}_4, a_2) = \mu_{\pi_3}(\mathfrak{s}_4)\pi_3(a_2|\mathfrak{s}_1) = 1/4 \cdot 1/2 = 1/8.$$

On the other hand,

$$\mu_{\bar{\pi}}(s) := \begin{cases} 1 & \text{if } s = \mathfrak{s}_1 \\ 1/4 & \text{if } s \in \{\mathfrak{s}_2, \mathfrak{s}_3, \mathfrak{s}_4\} \\ 3/64 & \text{if } s = \mathfrak{s}_5 \\ 1/64 & \text{if } s = \mathfrak{s}_6 \\ 1/16 & \text{if } s = \mathfrak{s}_7 \\ 1/16 & \text{if } s = \mathfrak{s}_f \end{cases}$$

while

$$\mu_{\bar{\pi}}(\mathfrak{s}_4, a_1) = \mu_{\bar{\pi}}(s_4)\bar{\pi}(a_1|s_4) = 1/4 \cdot 1/2 = 1/8$$
$$\mu_{\bar{\pi}}(\mathfrak{s}_4, a_2) = \mu_{\bar{\pi}}(s_4)\bar{\pi}(a_2|s_4) = 1/4 \cdot 1/2 = 1/8.$$

As a result, the two policies $\bar{\pi}$ and π_3 share the same occupancy measure, but the history-dependent policy $\bar{\pi}$ performs strictly better.