Models of human preference for learning reward functions

Anonymous Author(s)
Affiliation
Address
email

Abstract

The utility of reinforcement learning is limited by the alignment of reward functions with the interests of human stakeholders. One promising method for alignment is to learn the reward function from human-generated preferences between pairs of trajectory segments. These human preferences are typically assumed to be informed solely by partial return, the sum of rewards along each segment. We find this assumption to be flawed and propose modeling preferences instead as arising from a different statistic: each segment's regret, a measure of a segment's deviation from optimal decision-making. Given infinitely many preferences generated according to regret, we prove that we can identify a reward function equivalent to the reward function that generated those preferences. We also prove that the previous partial return model lacks this identifiability property without preference noise that reveals rewards' relative proportions, and we empirically show that our proposed regret preference model outperforms it with finite training data in otherwise the same setting. Additionally, our proposed regret preference model better predicts real *human* preferences and also learns reward functions from these preferences that lead to policies that are better human-aligned. Overall, this work establishes that the choice of preference model is impactful, and our proposed regret preference model provides an improvement upon a core assumption of recent research.

1 Introduction

Improvements in reinforcement learning (RL) have led to notable recent achievements [1–6], increasing its applicability to real-world problems. Yet, like all optimization algorithms, even *perfect* RL optimization is limited by the objective it optimizes. For RL, this objective is created in large part by the reward function. Poor alignment between reward functions and the interests of human stakeholders limits the utility of RL and may even pose catastrophic risks [7, 8].

Influential recent research has focused on reward learning from preferences over pairs of fixed-length trajectory segments. Nearly all of this recent work assumes that human preferences arise probabilistically from *only* the sum of rewards over a segment, i.e., the segment's **partial return** [9–16]. That is, these works assume that people tend to prefer trajectory segments that yield greater rewards *during the segment*. However, this preference model ignores seemingly important information about the segment's desirability, including the state values of the segment's start and end states. Separately, this partial return preference model can prefer suboptimal actions with lucky outcomes, like buying a lottery ticket.

This paper proposes an alternative preference model based on the **regret** of each segment, which is equivalent to the negated sum of an optimal policy's advantage of each transition in the segment (Section ~~2~~2.2).

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Figure 1 shows an intuitive example of when these two models disagree. Other classes of domains that the models will differ on are those with constant reward until the end, including competitive games like chess, go, and soccer as well as tasks for which the objective is to minimize time until reaching a goal.

For these two preference models, we first focus theoretically on a normative analysis (Section 3)—i.e., what preference model would we *want* humans to use if we could choose—proving that reward learning on infinite, exhaustive preferences with our proposed regret preference model identifies a reward function with the same set of optimal policies as the reward function with which the preferences are generated. We also prove that the partial return preference model is not guaranteed to identify such a reward function without preference noise. We follow up with a descriptive analysis of how well each of these proposed models align with *actual* human preferences by collecting a human-labeled dataset of preferences in a rich grid world domain (Section 4) and showing that the regret preference model better predicts these human preferences (Section 5). Finally, we find that the policies ultimately created through the regret preference model tend to outperform those from the partial return model learning—both when assessed with collected human preferences or when assessed with synthetic preferences (Section 6).

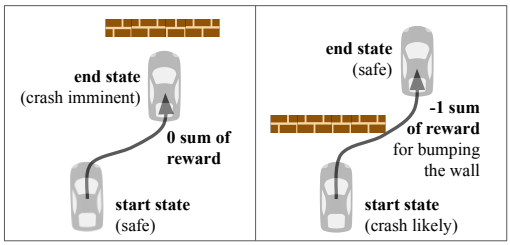


Figure 1: Two segments of a car moving at high speed near a brick wall. Assume the right segment is optimal and the left segment is suboptimal (as defined in Sec. 2.1). The left segment has a higher sum of reward, so the partial return preference model tends to prefer it. The regret preference model instead tends to prefer the right segment because optimal segments have minimal regret. If we also assume deterministic transitions, then the regret model includes the difference in values between the start state and the end state (Eq. 3), and the right segment would tend to be preferred because it greatly improves its state values from start to end, whereas the left segment's state values greatly worsen. We suspect our human readers will also tend to prefer the right segment.

2 Preference models for learning reward functions

We assume that the task environment is a Markov decision process (MDP) specified by the tuple $(S, A, T, \gamma, D_0, r)$. S and A are the sets of possible states and actions, respectively. T is a transition function, ~~$T: S \times A \times \rightarrow S$~~ $T: S \times A \rightarrow S$. γ is the discount factor and D_0 is the distribution of start states. Unless otherwise stated, we assume undiscounted tasks (i.e., $\gamma = 1$) that have terminal states, after which only 0 reward can be received. r is a reward function, $r: S \times A \times S \rightarrow \mathbb{R}$, where the reward r_t at time t is a function of s_t, a_t, and s_{t+1}. An MDP$\backslash r$ is an MDP without a reward function.

Throughout this paper, r refers to the ground-truth reward function for some MDP; $\hat{r}$ refers to a learned approximation of r; and $\tilde{r}$ refers to any reward function (including r or $\hat{r}$). A policy ($\pi: S \times A \rightarrow [0,1]$) specifies the probability of an action given a state. $Q^*_{\tilde{r}}$ and $V^*_{\tilde{r}}$ refer respectively to the state-action value function and state value function for an optimal policy, π^*, under $\tilde{r}$. The optimal advantage function is defined as $A^*_{\tilde{r}}(s,a) \triangleq Q^*_{\tilde{r}}(s,a) - V^*_{\tilde{r}}(s)$. Throughout this paper, the ground-truth reward function r is used to algorithmically generate preferences when they are not human-generated, is hidden during reward learning, and is used to evaluate the performance of optimal policies under a learned $\hat{r}$.

2.1 Reward learning from pairwise preferences

A reward function can be learned by minimizing the cross-entropy loss—i.e., maximizing the likelihood—of observed human preferences, a common approach in recent literature [9–11, 14, 16].

Segments Let σ denote a segment starting at state $s_{\sigma,0}$. Its length $|\sigma|$ is the number of transitions within the segment. A segment includes $|\sigma|+1$ states and $|\sigma|$ actions: $(s_{\sigma,0}, a_{\sigma,0}, s_{\sigma,1}, a_{\sigma,1}, \ldots, s_{\sigma,|\sigma|})$. In this problem setting, segments lack any reward information. As shorthand, we define $\sigma_t \triangleq (s_{\sigma,t}, a_{\sigma,t}, s_{\sigma,t+1})$. A segment σ is **optimal** with respect to $\tilde{r}$ if, for every $i \in \{1, \ldots, |\sigma|\text{-}1\}$, $Q^*_{\tilde{r}}(s_{\sigma,i}, a_{\sigma,i}) = V^*_{\tilde{r}}(s_{\sigma,i})$. A segment that is not optimal is **suboptimal**. Given some $\tilde{r}$ and a segment σ, $\tilde{r}_t \triangleq \tilde{r}(s_{\sigma,t}, a_{\sigma,t}, s_{\sigma,t+1})$, and the **partial return** of a segment σ is $\sum_{t=0}^{|\sigma|-1} \gamma^t \tilde{r}_t$, denoted in shorthand as $\Sigma_\sigma\, r$.

Preference datasets Each preference over a pair of segments creates a sample $(\sigma_1, \sigma_2, \mu)$ in a preference dataset $D_{\succ}$. Vector $\mu = \langle \mu_1, \mu_2 \rangle$ represents the preference; specifically, if σ_1 is preferred over σ_2, denoted $\sigma_1 \succ \sigma_2$, $\mu = \langle 1,0 \rangle$. μ is $\langle 0,1 \rangle$ if $\sigma_1 \prec \sigma_2$ and is $\langle 0.5,0.5 \rangle$ for $\sigma_1 \sim \sigma_2$ (no preference).

Loss function To learn a reward function from a preference dataset, $D_{\succ}$, a common assumption is that these preferences were generated by a preference model P that arises from an unobservable *ground-truth* reward function r. We approximate r by minimizing cross-entropy loss to learn $\hat{r}$:

$$loss(\hat{r}, D_{\succ}) = -\sum_{(\sigma_1,\sigma_2,\mu) \in D_{\succ}} \mu_1 \log P(\sigma_1 \succ \sigma_2 | \hat{r}) + \mu_2 \log P(\sigma_1 \prec \sigma_2 | \hat{r}) \tag{1}$$

This loss is under-specified until $P(\sigma_1 \succ \sigma_2 | \hat{r})$ is defined, which is the focus of this paper. We show that the common model of preference probabilities is flawed and introduce an improved preference model.

Preference models A preference model determines the probability of one trajectory segment being preferred over another, $P(\sigma_1 \succ \sigma_2 | \tilde{r})$. Preference models could be applied to model preferences provided by humans or other systems. Preference models can also directly generate preferences, and in such cases we refer to them as **preference generators**.

2.2 Choice of preference model: partial return and regret

Partial return Recent work assumes human preferences are generated by a Boltzmann distribution over the two segments' partial returns [9–16], expressed here as a logistic function[1]:

$$P_{\Sigma_r}(\sigma_1 \succ \sigma_2 | \tilde{r}) = logistic\Big(\Sigma_{\sigma_1} \tilde{r} - \Sigma_{\sigma_2} \tilde{r}\Big). \tag{2}$$

Regret We introduce an alternative preference model based on the regret of each transition in a segment. We first focus on segments with deterministic transitions. For a transition (s_t, a_t, s_{t+1}) in a deterministic segment, $regret_{\mathrm{d}}(\sigma_t | \tilde{r}) \triangleq V^*_{\tilde{r}}(s_{\sigma,t}) - [\tilde{r}_t + V^*_{\tilde{r}}(s_{\sigma,t+1})]$. For a full deterministic segment,

$$regret_d(\sigma | \tilde{r}) \triangleq \sum_{t=0}^{|\sigma|-1} regret_d(\sigma_t | \tilde{r}) = V^*_{\tilde{r}}(s_{\sigma,0}) - (\Sigma_\sigma \tilde{r} + V^*_{\tilde{r}}(s_{\sigma,|\sigma|})), \tag{3}$$

with the right-hand expression arising from cancelling out intermediate state values. Therefore, deterministic regret measures how much the segment reduces expected return from $V^*_{\tilde{r}}(s_{\sigma,0})$. An optimal segment, σ^*, always has 0 regret, and a suboptimal segment, $\sigma^{\neg *}$, will always have positive regret, a intuitively appealing property that also plays a role in the identifiability proof of Theorem 3.1.

Stochastic transitions, however, can result in $regret_d(\sigma^* | \hat{r}) > regret_d(\sigma^{\neg *} | \tilde{r})$, losing the property above. To retain it, we note that the effect on expected return of transition stochasticity from a transition (s_t, a_t, s_{t+1}) is $[\tilde{r}_t + V^*_{\tilde{r}}(s_{t+1})] - Q^*_{\tilde{r}}(s_t, a_t)$ and add this expression once per transition to get $regret(\sigma)$, removing the subscript d that refers to determinism. The regret for a single transition becomes $regret(\sigma_t | \tilde{r}) = [V^*_{\tilde{r}}(s_{\sigma,t}) - [\tilde{r}_t + V^*_{\tilde{r}}(s_{\sigma,t+1})]] + [[\tilde{r}_t + V^*_{\tilde{r}}(s_{\sigma,t+1})] - Q^*_{\tilde{r}}(s_{\sigma,t}, a_{\sigma,t})] = V^*_{\tilde{r}}(s_{\sigma,t}) - Q^*_{\tilde{r}}(s_{\sigma,t}, a_{\sigma,t}) = -A^*_{\tilde{r}}(s_{\sigma,t}, a_{\sigma,t})$. Regret for a full segment is

$$regret(\sigma | \tilde{r}) = \sum_{t=0}^{|\sigma|-1} regret(\sigma_t | \tilde{r}) = \sum_{t=0}^{|\sigma|-1} \Big[V^*_{\tilde{r}}(s_{\sigma,t}) - Q^*_{\tilde{r}}(s_{\sigma,t}, a_{\sigma,t}) \Big] = \sum_{t=0}^{|\sigma|-1} -A^*_{\tilde{r}}(s_{\sigma,t}, a_{\sigma,t}). \tag{4}$$

The regret preference model is the Boltzmann distribution over negated regret:

$$P_{regret}(\sigma_1 \succ \sigma_2 | \tilde{r}) \triangleq logistic\Big(regret(\sigma_2 | \tilde{r}) - regret(\sigma_1 | \tilde{r})\Big). \tag{5}$$

Lastly, we note that if two segments have deterministic transitions, end in terminal states, and have the same starting state, this regret model reduces to the partial return model: $P_{regret}(\cdot | \tilde{r}) = P_{\Sigma_r}(\cdot | \tilde{r})$.

Algorithms in this paper All algorithms in the body of this paper are defined simply as "minimize Equation 1". They differ only in how the preference probabilities are calculated. All reward function learning via partial return uses Equation 2. We use two algorithms for reward function learning

[1] See Appendix B for a derivation of this logistic expression from a Boltzmann distribution with a temperature of 1. Unless otherwise stated, we ignore the temperature because scaling reward has the same effect.

via regret. The theory in Section 3 assumes exact measurement of regret, using Equation 5. Our experimental results in Section 6 use Equation 6 to approximate regret. Appendix B introduces other algorithms that use Equation 1, as well as one in Appendix B.2 that generalizes Equation 1.

Regret as a model for human preference P_{regret} makes at least three assumptions worth noting. First, it keeps the assumption that human preferences follow a Boltzmann distribution over some statistic, which is a common model of choice behavior in economics and psychology, where it is called the Luce-Shepard choice rule [17, 18]. Second, P_{regret} implicitly assumes humans can identify optimal and suboptimal segments when they see them, which will less true in domains where the human has less expertise. Lastly, P_{regret} assumes that in stochastic settings where the best *outcome* may only result from suboptimal decisions (e.g., buying a lottery ticket), humans instead prefer optimal *decisions*. We suspect humans are capable of expressing either type of preference—based on decision quality or desirability of outcomes—and can be influenced by training or the preference elicitation interface. In practice we determine that the regret model produces improvements over the partial-return model (Section 6), and its assumptions represent an opportunity for follow-up research.

Alternative methods for learning reward functions Other methods for learning reward functions include inverse reinforcement learning from demonstrations [19, 20] (discussed in Appendix B.5) and inverse reward design from trial-and-error reward design in multiple instances of a task domain [21].

3 Theoretical comparisons

In this section, we consider how different ways of generating preferences affect reward inference, setting aside whether humans can be influenced to give preferences in accordance with a specific preference method. In economic terms, this analysis—and all of our analyses with synthetic preferences—could be considered a normative analysis. In artificial intelligence, this analysis might be cast as a step towards defining criteria for a rational preference model.

Definition 3.1 (An identifiable preference model)**.** *For a preference model P, assume an infinite dataset $D_{\succ}$ of n-length pairs of segments is constructed by repeatedly choosing (σ_1,σ_2) and sampling a label $\mu \sim P(\sigma_1 \succ \sigma_2|r)$, using P as a preference generator. Further assume that in this dataset, all possible n-length segment pairs appear infinitely many times. For some MDP$\backslash r$ M, let $M_{\tilde{r}}$ be M with the reward function $\tilde{r}$. Let $\Pi^*_{\tilde{r}}$ be the set of optimal policies in $M_{\tilde{r}}$. Let reward-equivalence class $\mathfrak{R}$ be the set of all reward functions such that if $r_1,r_2 \in \mathfrak{R}$ then $\Pi^*_{r_1} = \Pi^*_{r_2}$. Preference model P is* ***identifiable*** *if, for any choice of n and M_r , any $\hat{r} = argmin_{\tilde{r},D_{\succ}}[loss(\tilde{r})]$—for the cross-entropy loss (Eqn. 1), with P as the preference model—is in the same reward equivalence class as r. I.e., $\Pi^*_r = \Pi^*_{\hat{r}}$.*

Theorem 3.1 (P_{regret} is identifiable)**.** *Let P_{regret} be any function such that if $regret(\sigma_1|\tilde{r}) < regret(\sigma_2|\tilde{r})$, $P_{regret}(\sigma_1 \succ \sigma_2|\tilde{r}) > 0.5$, and if $regret(\sigma_1|\tilde{r}) = regret(\sigma_2|\tilde{r})$, $P_{regret}(\sigma_1 \succ \sigma_2|\tilde{r}) = 0.5$. P_{regret} is identifiable.*

This class of regret preference models includes but is not limited to the Boltzmann distribution of Eqn. 5 and the narrower class that Theorem 3.1 focuses upon.

~~**Proof sketch** — Make all assumptions in Definition 3.1. Since $\hat{r}$ minimizes cross-entropy loss, $P_{regret}(\cdot|r) = P_{regret}(\cdot|\hat{r})$ for all possible segment pairs. Also, by Equation 4 $regret(\sigma|\tilde{r}) = 0$ if and only if σ is optimal with respect to $\tilde{r}$. And $regret(\sigma|\tilde{r}) > 0$ if and only if σ is suboptimal with $\tilde{r}$.~~

~~With respect to some $\tilde{r}$, let σ^* be any optimal segment and $\sigma^{\neg *}$ be any suboptimal segment. $regret(\sigma^*|\tilde{r}) < regret(\sigma^{\neg *}|\tilde{r})$. $P_{regret}(\sigma^*,\sigma^{\neg *}|\tilde{r}) > 0.5$, which we refer to as being *preferred by* $P_{regret}(\cdot|\tilde{r})$. In the total ordering created by $P_{regret}(\cdot|\tilde{r})$, let $\Sigma^*_{\tilde{r}}$ be the set of maximal segments in $D_{\succ}$ (i.e., the segments over which no segment is preferred by $P_{regret}(\cdot|\tilde{r})$). A policy $\pi \in \Pi^*_{\tilde{r}}$ if and only if all state-action pairs producible by π are within one or more of the segments in $\Sigma^*_{\tilde{r}}$.~~

~~Since $P_{regret}(\cdot|r) = P_{regret}(\cdot|\hat{r})$, $\Sigma^*_r = \Sigma^*_{\tilde{r}}$. Therefore $\Pi^*_r = \Pi^*_{\tilde{r}}$.~~

Theorem 3.2 (Noiseless P_{Σ_r} is not identifiable)**.** *Let P_{Σ_r} be any function such that if $\Sigma_{\sigma_1}\tilde{r} > \Sigma_{\sigma_2}\tilde{r}$, $P_{\Sigma_r}(\sigma_1 \succ \sigma_2|\tilde{r}) = 1$, and if $\Sigma_{\sigma_1}\tilde{r} = \Sigma_{\sigma_2}\tilde{r}$, $P_{\Sigma_r}(\sigma_1 \succ \sigma_2|\tilde{r}) = 0.5$. There exists an MDP in which P_{Σ_r} is not identifiable.*

Appendix C contains a ~~detailed~~ proof of Theorem 3.1 and two proofs by example for Theorem 3.2, each focusing on a different weakness of P_{Σ_r}.The first proof by example reveals issues when learning reward functions with stochastic transitions with either P_{Σ_r} or *deterministic* P_{regret_d}. These issues directly correspond to the need for preferences over distributions over outcomes (i.e., lotteries) to construct a cardinal utility function (see Russell and Norvig [22, Ch. 16]). Note that the noiseless version of P_{Σ_r} in Theorem 3.2 is achieved in the limit as reward values are scaled higher; equivalently, one could include a Boltzmann temperature parameter in Equation 2 and scale it towards 0. Intuitively, Theorem 3.2 says that P_{Σ_r} is not identifiable without the distribution over preferences providing information about the proportions of rewards with respect to each other. In contrast, to be identifiable, the regret preference model does not require this preference error (though it can presumably benefit from it in certain contexts).

4 Creating a human-labeled preference dataset

To empirically investigate the consequences of each preference model when learning reward from *human* preferences, we created a preference dataset labeled by human subjects via Amazon Mechanical Turk. This data collection was IRB-approved. Appendix D adds detail to the content below.

4.1 The general delivery domain

The delivery domain consists of a grid of cells, each of a specific road surface type. The delivery agent's state is its location. The agent's action space is moving one cell in one of the four cardinal directions. The episode can terminate either at the destination for $+50$ reward or in failure at a sheep for -50 reward. The reward for a non-terminal transition is the sum of any reward components. Cells with a white road surface have a -1 reward component, and cells with brick surface have a -2 component. Additionally, each cell may contain a coin ($+1$) or a roadblock (-1). Coins do not disappear and at best cancel out the road surface cost. Actions that would move the agent into a house or beyond the grid's perimeter result in no motion and receive reward that includes the current cell's surface reward component but not any coin or roadblock components. In this work, the start state distribution, D_0, is always uniformly random over non-terminal states. This domain was designed to permit subjects to easily identify bad behavior yet also to be difficult for them to determine *optimal* behavior from most states, which is representative of many common tasks.

4.1.1 The delivery task

We chose one instantiation of the delivery domain for gathering our dataset of human preferences. This specific MDP has a 10×10 grid. From every state, the highest return possible involves reaching the goal, rather than hitting a sheep or perpetually avoiding termination. Figure 2 shows this task.

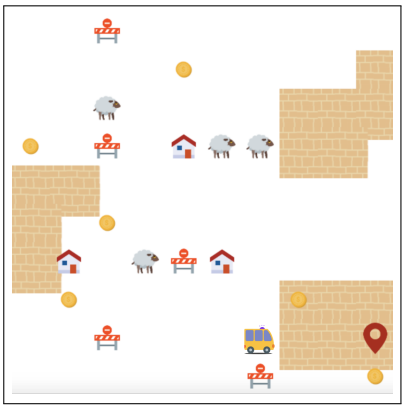


Figure 2: The delivery task used to gather human preferences. The yellow van is the agent and the red inverted teardrop is the destination.

4.2 The user interface and survey

This subsection describes the three main stages of the experimental session. A video showing the full experimental protocol can be seen at ....

Teaching subjects about the task Subjects first view instructions describing the general domain. To avoid the jargon of "return" and "reward," these terms are mapped to equivalent values in US dollars, and the instructions describe the goal of the task as maximizing the delivery vehicle's financial outcome, where the reward components are specific financial impacts. This information is shared amongst interspersed interactive episodes, in which the subject controls the agent in domain maps that are each designed to teach one or two concepts. Our intention during this stage is to inform the later preferences of the subject by teaching them about the domain's dynamics and its reward function, as well as to develop the subject's sense of

how desirable various behaviors are. At the end of this stage, the subject controls the agent for two episodes in the specific delivery task shown in Figure 2.

Preference elicitation After each subject is trained to understand the task, they indicate their preferences between 40–50 randomly-ordered pairs of segments, using the interface shown in Figure 3. The users select a preference, no preference ("same"), or "can't tell". In this work, we exclude responses labeled "can't tell", though one might alternatively try to extract information from these responses.

Users' task comprehension Subjects then answered questions testing their understanding of the task, and we removed their data if they scored poorly. We also removed a subject's data if they preferred colliding the vehicle into a sheep over not doing so, which we interpreted as poor task understanding or inattentiveness. This filtered dataset contains 1812 preferences from 50 subjects.

4.3 Selection of segment pairs for labeling

We collected human preferences in two stages, each with different methods for selecting which segment pairs to present for labeling. The second stage's sole purpose was to improve the reward-learning performance of P_{Σ_r}. ~~Before this second stage~~Without second-stage data, P_{Σ_r} compared even worse to P_{regret} than in the results described in Section 6 ~~.~~(see Appendix **??**). Both stages' data are combined and used as a single dataset. These methods and their justification are described in Appendix D.3.

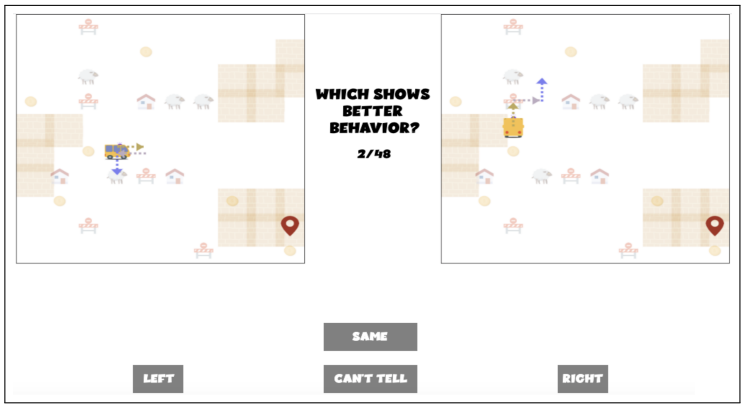


Figure 3: Interface shown to subjects during preference elicitation.

5 Descriptive results

This section considers how well different preference models explain our dataset of human preferences.

5.1 Correlations between preferences and segment statistics

We hypothesize that the values of segments' start and end states—which are included in P_{regret} but not in P_{Σ}—affect human preferences, independent of partial return. To simplify analysis, we combine the two parts of $regret_{\mathrm{d}}(\sigma|r)$ that are additional to $\Sigma_\sigma \tilde{r}$ and introduce the following shorthand: $\Delta_\sigma V_{\tilde{r}} \triangleq V^*_{\tilde{r}}(s_{\sigma,|\sigma|}) - V^*_{\tilde{r}}(s_{\sigma,0})$. Note that with an algebraic manipulation (see Appendix E.1), $regret_{\mathrm{d}}(\sigma_2|\tilde{r}) - regret_{\mathrm{d}}(\sigma_1|\tilde{r}) = (\Delta_{\sigma_1} V_{\tilde{r}} - \Delta_{\sigma_2} V_{\tilde{r}}) + (\Sigma_{\sigma_1}\tilde{r} - \Sigma_{\sigma_2}\tilde{r})$. Therefore, on the diagonal line in Figure 4, $regret_{\mathrm{d}}(\sigma_2|r) = regret_{\mathrm{d}}(\sigma_1|r)$, making the P_{regret_d} preference model indifferent.

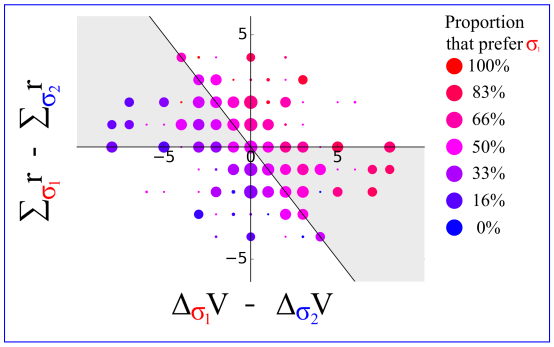


Figure 4: Proportions at which subjects preferred each segment in a pair, plotted by the difference in the segments' changes in state values (x-axis) and partial returns (y-axis). The diagonal line shows points of preference indifference for P_{regret}. Points of indifference for P_{Σ} lie on the x-axis. The shaded gray area indicates where the two models disagree, each giving a different segment a preference probability greater than 0.5. Each circle's area is proportional to the number of samples it describes.

The dataset of preferences is visualized in Figure 4. This plot shows how $\Delta_\sigma V_r$ has influence independent of partial return by focusing only on points at a chosen y-axis value; if the colors

Preference model	Loss
$P(\cdot) = 0.5$ (uninformed)	0.69
P_{Σ_r} (partial return)	0.62
P_{regret}	**0.57**

Table 1: Mean cross-entropy test loss over 10-fold cross validation (n=1812) from predicting human preferences. Lower is better.

along the corresponding horizontal line reddens as the x-axis value increases, then $\Delta_\sigma V_r$ appears to have independent influence. To statistically test for independent influence of $\Delta_\sigma V_r$ on preferences, we consider subsets of data where $\Sigma_{\sigma_1} r - \Sigma_{\sigma_2} r$ is constant. For $\Sigma_{\sigma_1} r - \Sigma_{\sigma_2} r = -1$ and $\Sigma_{\sigma_1} r - \Sigma_{\sigma_2} r = -2$, the only values with more than 30 samples that also include informative samples with both negative and positive values of $regret(\sigma_1|r) - regret(\sigma_2|r)$, the Spearman's rank correlations between $\Delta_\sigma V_r$ and the preferences are significant ($r >= 0.3$, $p < 0.0001$). This result indicates that $\Delta_\sigma V_r$ *influences human preferences independent of partial return*, validating our hypothesis that humans form preferences based on information about segments' start states and end states, not only partial returns.

5.2 Likelihood of human preferences under different preference models

To examine how well each preference model predicts human preferences, we calculate the cross-entropy loss for each model (Eqn. 1)—i.e., the negative log likelihood—of the preferences in our dataset. Scaling reward by a constant factor does not affect the set of optimal policies. Therefore, throughout this work we ensure that our analyses of preference models are insensitive to reward scaling. To do so for this specific analysis, we conduct 10-fold cross validation to learn a reward scaling factor for each of P_{regret} and P_{Σ_r}. Table 1 shows that the loss of P_{regret} is lower than that of P_{Σ_r}, indicating that it is more reflective of how people actually express preferences.

6 Results from learning reward functions

Analysis of a preference model's predictions of human preferences is informative, but such predictions are a means to the ends of learning human-aligned reward functions and policies. We now examine each preference model's performance on these ends. In all cases, we learn a reward function $\hat{r}$ according to Eqn. 1 and apply value iteration [23] to find the approximately optimal $Q^*_{\hat{r}}$ function. For this $Q^*_{\hat{r}}$, we then evaluate the mean return of the maximum-entropy optimal policy—which chooses uniformly randomly among all *optimal* actions—with respect to the ground-truth reward function r, over D_0. To compare performance across different MDPs, the mean return of a policy π, V^π_r, is normalized to $(V^\pi_r - V^U_r)/V^*_r$, where V^*_r is the optimal expected return and V^U_r is the expected return of the uniformly random policy (both given D_0). Normalized mean return above 0 is better than V^U_r. Optimal policies have a normalized mean return of 1, and we consider above 0.9 to be *near optimal*.

6.1 An algorithm to learn reward functions with $regret(\sigma_\sigma|\hat{r})$

Algorithm 1 is a general algorithm for learning a *linear* reward function according to P_{regret}. ~~Adding to our assumptions listed in Section 2, this algorithm introduces two more approximations. First, to avoid expensive evaluation of $V^*_{\hat{r}}(s_{\sigma,t})$ and $Q^*_{\hat{r}}(s_{\sigma,t},a_{\sigma,t})$.~~ This regret-specific algorithm only changes the regret-based algorithm from Section 2.2 by replacing Equation 5 with a tractable approximation of regret, avoiding expensive repeated evaluation of $V^*_{\hat{r}}(\cdot)$ and $Q^*_{\hat{r}}(\cdot,\cdot)$ to compute $P_{regret}(\cdot|\hat{r})$ during reward learning. Specifically, successor features for a set of policies are used to approximate the optimal state values and state-action values for *any* reward function. ~~Second, softmax weighting is used, as explained below.~~

Approximating P_{regret} with successor features Following the notation of Barreto et al. [24], assume the ground-truth reward is linear with respect to a feature vector extracted by $\boldsymbol{\phi}: S \times A \times S \to \mathbb{R}^d$ and a weight vector $\boldsymbol{w_r} \in \mathbb{R}^d$: $r(s,a,s') = \boldsymbol{\phi}(s,a,s')^\top \boldsymbol{w_r}$. During learning, $\boldsymbol{w_{\hat{r}}}$ similarly expresses $\hat{r}$ as $\hat{r}(s,a,s') = \boldsymbol{\phi}(s,a,s')^\top \boldsymbol{w_{\hat{r}}}$.

Given a policy π, the successor features for (s,a) are the expectation of discounted reward features from that state-action pair when following π: $\boldsymbol{\psi}^\pi_{\boldsymbol{Q}}(s,a) = E^\pi[\sum_{i=t}^{\infty} \gamma^{i-t} \boldsymbol{\phi}(s_t,a_t,s_{t+1}) | s_t = s, a_t = a]$. Therefore, $Q^\pi_{\hat{r}}(s,a) = \boldsymbol{\psi}^\pi_{\boldsymbol{Q}}(s,a)^\top \boldsymbol{w}_{\hat{r}}$. Additionally, state-based successor features can be calculated from the $\boldsymbol{\psi}^\pi_{\boldsymbol{Q}}$ above as $\boldsymbol{\psi}^\pi_{\boldsymbol{V}}(s) = \sum_{a \in A} \pi(a|s) \boldsymbol{\psi}^\pi_{\boldsymbol{Q}}(s,a)$, making $V^\pi_{\hat{r}}(s) = \boldsymbol{\psi}^\pi_{\boldsymbol{V}}(s)^\top \boldsymbol{w}_{\hat{r}}$.

Algorithm 1 Linear reward learning with regret preference model (P_{regret}), using successor features

1: Input: a set of reward functions and a set of policies (where one set can be $\varnothing$)
2: $\Psi \leftarrow \varnothing$
3: **for** *each reward function* r_{SF} *or policy* π_{SF} *in the input sets* **do**
4: **if** r_{SF} **then** $\pi_{SF} \leftarrow$ estimate of optimal maximum-entropy policy for r_{SF}
5: estimate $\psi_{Q}^{\pi_{SF}}$ and $\psi_{V}^{\pi_{SF}}$ (if not estimated already during step 4)
6: add $\psi_{Q}^{\pi_{SF}}$ to Ψ_{Q}
7: add $\psi_{V}^{\pi_{SF}}$ to Ψ_{V}
8: **end for**
9: **repeat**
10: optimize $\boldsymbol{w}_{\hat{r}}$ by loss of Eqn. 1, calculating $\tilde{P}_{regret}(\sigma_1 \succ \sigma_2 | \hat{r})$ via Eqn. 6, using Ψ_{Q} and Ψ_{V}
11: **until** *stopping criteria are met*
12: **return** $\boldsymbol{w}_{\hat{r}}$

Given a set Ψ_{Q} of state-action successor feature functions and a set Ψ_{V} of state successor feature functions for various policies and given a reward function via $\boldsymbol{w}_{\hat{r}}$, $Q_{\hat{r}}^{\pi^*}(s,a) \geq max_{\boldsymbol{\psi_Q} \in \Psi_{Q}} [\boldsymbol{\psi}_{\boldsymbol{Q}}^{\pi}(s,a)^{\top} \boldsymbol{w}_{\hat{r}}]$ and $V_{\hat{r}}^{\pi^*}(s) \geq max_{\boldsymbol{\psi_V} \in \Psi_{V}} [\boldsymbol{\psi}_{\boldsymbol{V}}^{\pi}(s)^{\top} \boldsymbol{w}_{\hat{r}}]$ [24], so we use these two maximizations as approximations of $Q_{\hat{r}}^{*}(s,a)$ and $V_{\hat{r}}^{*}(s)$, respectively. In practice, to enable gradient-based optimization with current tools, the maximization in this expression is replaced with the softmax-weighted average, making the loss function linear. Focusing first on the approximation of $V_{\hat{r}}^{*}(s)$, for each $\boldsymbol{\psi_V} \in \Psi_{V}$, a softmax weight is calculated for $\boldsymbol{\psi}_{\boldsymbol{V}}^{\pi}(s)$: $softmax_{\Psi_{V}}(\boldsymbol{\psi}_{\boldsymbol{V}}^{\pi}(s)^{\top} \boldsymbol{w}_{\hat{r}}) \triangleq [(\boldsymbol{\psi}_{\boldsymbol{V}}^{\pi}(s)^{\top} \boldsymbol{w}_{\hat{r}})^{1/T}] / [(\sum_{\boldsymbol{\psi'_V} \in \Psi_{V}} \boldsymbol{\psi'}_{\boldsymbol{V}}^{\pi}(s)^{\top} \boldsymbol{w}_{\hat{r}})^{1/T}]$, where temperature T is a constant hyperparameter. The resulting approximation of $V_{\hat{r}}^{*}(s)$ is therefore defined as $\tilde{V}_{\hat{r}}^{*}(s) \triangleq \sum_{\boldsymbol{\psi_V} \in \Psi_{V}} softmax_{\Psi_{V}}(\boldsymbol{\psi}_{\boldsymbol{V}}^{\pi}(s)^{\top} \boldsymbol{w}_{\hat{r}}) [\boldsymbol{\psi}_{\boldsymbol{V}}^{\pi}(s)^{\top} \boldsymbol{w}_{\hat{r}}]$. Similarly, to approximate $Q_{\hat{r}}^{*}(s,a)$, $softmax_{\Psi_{Q}}(\boldsymbol{\psi}_{\boldsymbol{Q}}^{\pi}(s,a)^{\top} \boldsymbol{w}_{\hat{r}}) \triangleq [(\boldsymbol{\psi}_{\boldsymbol{Q}}^{\pi}(s,a)^{\top} \boldsymbol{w}_{\hat{r}})^{1/T}] / [(\sum_{\boldsymbol{\psi'_Q} \in \Psi} \boldsymbol{\psi'}_{\boldsymbol{Q}}^{\pi}(s,a)^{\top} \boldsymbol{w}_{\hat{r}})^{1/T}]$ and $\tilde{Q}_{\hat{r}}^{*}(s,a) \triangleq \sum_{\boldsymbol{\psi_Q} \in \Psi_{Q}} softmax_{\Psi_{Q}}(\boldsymbol{\psi}_{\boldsymbol{Q}}^{\pi}(s,a)^{\top} \boldsymbol{w}_{\hat{r}}) [\boldsymbol{\psi}_{\boldsymbol{Q}}^{\pi}(s,a)^{\top} \boldsymbol{w}_{\hat{r}}]$. Consequently, from Eqns. 4 and 5, the corresponding approximation $\tilde{P}_{regret}$ of the regret preference model is:

$$\tilde{P}_{regret}(\sigma_1 \succ \sigma_2 | \hat{r}) = logistic\left(\sum_{t=0}^{|\sigma_2|-1} \left[\tilde{V}_{\hat{r}}^{*}(s_{\sigma_2,t}) - \tilde{Q}_{\hat{r}}^{*}(s_{\sigma_2,t}, a_{\sigma_2,t}) \right] - \sum_{t=0}^{|\sigma_1|-1} \left[\tilde{V}_{\hat{r}}^{*}(s_{\sigma_1,t}) - \tilde{Q}_{\hat{r}}^{*}(s_{\sigma_1,t}, a_{\sigma_1,t}) \right] \right) \quad (6)$$

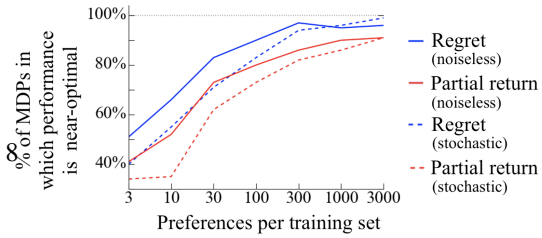
The algorithm In Algorithm 1, lines 9–12 describe the supervised-learning optimization using the approximation $\tilde{P}_{regret}$, and the prior lines create Ψ_{Q} and Ψ_{V}. Specifically, given a set of reward functions, a corresponding set of policies is created (line 4), where each policy is an estimate of the maximum entropy policy for a reward function. Standard policy improvement methods can be used to create each such policy. Alternatively, some or all of the set of policies can be given as input directly, not derived from input reward functions. For each such policy π_{SF}, successor feature functions $\Psi_{Q}^{\pi_{SF}}$ and $\Psi_{V}^{\pi_{SF}}$ are estimated (line 5), which by default would be performed by a minor extension of a standard policy evaluation algorithm as detailed by Barreto et al. [24]. Note that the reward function that is ultimately learned is not restricted to be in the input set of reward functions, which is used only to create an approximation of regret.

The details of our instantiation of Algorithm 1 for the delivery domain can be found in Appendix F.1, along with guidance for extending it to reward functions that might be non-linear.

6.2 Results from synthetic preferences

Before considering human preferences, we first ask how each preference model performs when it is correct. In other words, we investigate empirically how well the preference model could perform if humans perfectly adhered to it. Recall that the ground-truth reward function, r, is used to create these preferences but is inaccessible to the reward-learning algorithms.

For these evaluations, either a stochastic or noiseless preference model acts a preference generator to create a preference dataset, and



then the stochastic version of the same model is used for reward learning. For the noiseless case, the deterministic preference generator compares a segment pair's $\Sigma_\sigma r$ values for P_{Σ_r} or their $regret(\sigma|r)$ values for P_{regret}. Note that through reward scaling the preference generators approach determinism in the limit, so this noiseless analysis examines minimal-entropy versions of the two preference-generating models. (The opposite extreme, uniformly random preferences, would remove all information from preferences and therefore is not examined.) In the stochastic case, for each preference model, each segment pair is labeled by sampling from that preference generator's output distribution (Eqs 2 or 5), using the unscaled ground-truth reward function.

We created 100 deterministic MDPs that instantiate variants of our delivery domain (see Section 4.1). To create each MDP, we sampled from sets of possible widths, heights, and reward component values, and the resultant grid cells were randomly populated with a destination, objects, and road surface types (see Appendix F.2 for details). Each segment in the preference datasets for each MDP was generated by choosing a start state and three actions, all uniformly randomly. For a set number of preferences, each method had the same set of segment pairs in its preference dataset. Figure 5 shows the percentage of MDPs in which each preference model results in near-optimal performance. The regret preference model outperforms the partial return model at every dataset size, both with and without noise. By a Wilcoxon paired signed-rank test on normalized mean returns, $p < 0.05$ for 86% of these comparisons and $p < 0.01$ for 57% of them, as reported in Appendix F.2.

Further analyses can be found in Appendix F.2, including with stochastic transitions, with different segment lengths, and while artificially lowering the discount factor (as is common in deep RL and recent work on deep reward learning from preferences).

6.3 Results from human preferences

We randomly assign human preferences from our gathered dataset to different numbers of same-sized partitions, resulting in different training set sizes, and test each preference model on each partition. Figure 6 shows the results. With smaller training sets (20–100 partitions), the regret preference model results in near-optimal performance more often. With larger training sets (1–10 partitions), both preference models always reach near-optimal return, but the mean return from the regret preference model is higher for all of these partitions except for 3 partitions in the 10-partition test. Applying a Wilcoxon paired signed-rank test on normalized mean return to each group with 5 or more partitions, $p < 0.05$ for all numbers of partitions except 100 and $p < 0.01$ for 20 and 50 partitions.

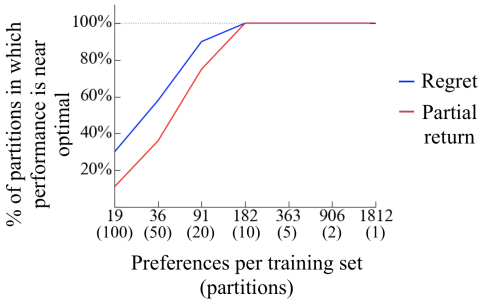


Figure 6: Performance comparison over various amounts of human preferences. Each partition has the number of preferences shown or one less.

7 Conclusion

Over numerous evaluations with human preferences, our proposed regret preference model (P_{regret}) shows improvements summarized below over the previous partial return preference model (P_{Σ_r}). When each preference model generates the preferences for its own infinite and exhaustive training set, we prove that P_{regret} identifies the set of optimal policies, whereas P_{Σ_r} is not guaranteed to do so without preference noise that reveals the proportions of rewards with respect to each other. With finite training data of synthetic preferences, P_{regret} also empirically results in learned policies that tend to outperform those resulting from P_{Σ_r}. This superior performance of P_{regret} is also seen with human preferences. In summary, our analyses suggest that regret preference models are more effective both descriptively with respect to human preferences and also normatively, as the model we want humans to follow if we had the choice.

Independent of P_{regret}, this paper also reveals that segments' changes in state values provide information about human preferences that is not fully provided by partial return. More generally, we show that the choice of preference model impacts the performance of learned reward functions.

This study motivates several new directions for research. Future work could address any of the limitations detailed in Appendix A.1. Specifically, future work could further test the general superiority of P_{regret} or apply it to deep learning settings. Additionally, *prescriptive* methods could be developed via the user interface or elsewhere to nudge humans to conform more to P_{regret} or to other normatively appealing preference models. Lastly, subsequent efforts could seek preference models that are even more effective with preferences from actual humans, now that this work has provided conclusive evidence that the choice of preference model is impactful.

References

[1] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[2] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[3] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[4] Marc G Bellemare, Salvatore Candido, Pablo Samuel Castro, Jun Gong, Marlos C Machado, Subhodeep Moitra, Sameera S Ponda, and Ziyu Wang. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588(7836):77–82, 2020.

[5] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

[6] Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, 2022.

[7] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[8] W Bradley Knox, Alessandro Allievi, Holger Banzhaf, Felix Schmitt, and Peter Stone. Reward (mis)design for autonomous driving. *arXiv preprint arXiv:2104.13906*, 2021.

[9] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4299–4307, 2017.

[10] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in atari. *arXiv preprint arXiv:1811.06521*, 2018.

[11] Xiaofei Wang, Kimin Lee, Kourosh Hakhamaneshi, Pieter Abbeel, and Michael Laskin. Skill preferences: Learning to extract and execute robotic skills from human feedback. In *Conference on Robot Learning*, pages 1259–1268. PMLR, 2022.

[12] Erdem Bıyık, Dylan P Losey, Malayandi Palan, Nicholas C Landolfi, Gleb Shevchuk, and Dorsa Sadigh. Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences. *The International Journal of Robotics Research*, page 02783649211041652, 2021.

[13] Dorsa Sadigh, Anca D Dragan, Shankar Sastry, and Sanjit A Seshia. Active preference-based learning of reward functions. *Robotics: Science and Systems*, 2017.

[14] Kimin Lee, Laura Smith, and Pieter Abbeel. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091*, 2021.

[15] Kimin Lee, Laura Smith, Anca Dragan, and Pieter Abbeel. B-pref: Benchmarking preference-based reinforcement learning. *arXiv preprint arXiv:2111.03026*, 2021.

[16] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

[17] R Duncan Luce. *Individual choice behavior: A theoretical analysis*. John Wiley, 1959.

[18] Roger N Shepard. Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, 22(4):325–345, 1957.

[19] A.Y. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *Seventeenth International Conference on Machine Learning (ICML)*, 2000.

[20] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *Twenty-third AAAI Conference on Artificial Intelligence*, volume 8, pages 1433–1438, 2008.

[21] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse reward design. In *Advances in Neural Information Processing Systems (NIPS)*, pages 6765–6774, 2017.

[22] Stuart Russell and Peter Norvig. *Artificial intelligence: a modern approach*. 2020.

[23] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[24] André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado Van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. *arXiv preprint arXiv:1606.05312*, 2016.

[25] Riad Akrour, Marc Schoenauer, and Michele Sebag. Preference-based policy learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 12–27. Springer, 2011.

[26] Daniel Brown, Russell Coleman, Ravi Srinivasan, and Scott Niekum. Safe imitation learning via fast bayesian reward inference from preferences. In *International Conference on Machine Learning*, pages 1165–1177. PMLR, 2020.

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

[28] Kuno Kim, Shivam Garg, Kirankumar Shiragur, and Stefano Ermon. Reward identification in inverse reinforcement learning. In *International Conference on Machine Learning*, pages 5496–5505. PMLR, 2021.

[29] Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2021.

[30] John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. Princeton university press, 1944.

[31] Yuchen Cui, Qiping Zhang, Alessandro Allievi, Peter Stone, Scott Niekum, and W Bradley Knox. The empathic framework for task learning from implicit human feedback. *arXiv preprint arXiv:2009.13649*, 2020.

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Checklist

~~https://neurips.cc/public/guides/PaperChecklist~~

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Appendix A.1.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Appendix A.2.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Sections 3 and C include all assumptions.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section 3 and Appendix C.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] However, the learning code, the code for running experiments, the code and UI elements for gathering human preferences on Mechanical Turk, and the anonymized human preferences data will be opened. We are particularly excited to provide the first open dataset of human preferences over pairs of trajectory segments.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Appendix F.1

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Error bars do not seem applicable to our plots, which do not show the exact data that we do statistical testing on. However, statistical significance testing was reported, in Sections 5.1 and 6.2 (with a pointer to the appendix for details).

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix F.1.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] Appendix D does so for visual assets used to visualize the delivery task.

(b) Did you mention the license of the assets? [Yes] Appendix D mentions the license for visual assets used to visualize the delivery task.

(c) Did you include any new assets either in the supplemental material or as a URL? [No]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Appendix D.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Appendix D

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] Section 4.1.1 includes a link to a video of a full experimental session (with an author acting as the subject).

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] We discuss participant risks from our crowdsourced study and provide a link to the IRB approval in Appendix D.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See Appendix D.

After Appendix A, the appendix is organized according to the major sections and subsections of the main content. We consider Appendix C to be its most important addition to the main content.

A Limitations and ethics

A.1 Limitations

Some limitations of the regret preference model are discussed in the paragraph "Regret as a model for human preference" in Section ~~2~~2.2, including assumptions that a person giving preferences can distinguish between optimal and suboptimal segments, that they follow a Boltzmann distribution (i.e., a Luce Shepard choice rule), and that they base their preferences on the desirability of decisions even when transition stochasticity results in segment pairs for which the worse decision has a better outcome.

Our proposed algorithm (Section 6.1) has a few additional limitations. Generating candidate successor features for the approximations $\tilde{Q}^*_{\hat{r}}$ and $\tilde{V}^*_{\hat{r}}$ may be difficult in complex domains. Specifically, challenges include choosing the set of policies or reward functions for which to compute successor features (line 3 of Algorithm 1) and creating a reward feature vector ϕ for non-linear reward functions (discussed in Appendix F.1). Additionally, although learning with P_{regret} is more sample efficient in our experiments, it is computationally slower than learning with P_{Σ_r} because of the additional need to compute successor features and the use of the softmax function to approximate $Q^*_{\hat{r}}$ and $V^*_{\hat{r}}$. Nonetheless, we may accept the tradeoff of an increase in computational time that reduces the number of human samples needed or that improves the reward function's alignment with human stakeholders' interests. Lastly, the loss during optimization with P_{regret} was unstable, which we addressed by taking the minimum loss over all epochs during training. Therefore, for more complex reward feature vectors (ϕ) than our 6-element vector for the delivery task, extra care might be needed to avoid overfitting $\hat{r}$, for example by withholding some preference data to serve as a test set.

We also generally assume that the RL algorithm and reward learning algorithm use the same discount factor as in the MDP$\setminus r$ specification. One weakness of contemporary deep RL is that RL algorithms require artificially lower discount factors than the true discount factor of the task. The interaction of this discounting with preference models is considered in Appendix F.2. Our expectation though is that this weakness of deep RL algorithms is likely a temporary one, and so we focused our analysis on simple tasks in which we do not need to artificially lower the RL algorithm's discount factor. However, further investigation of the interaction between preference models and discount factors would aid near-term application of P_{regret} to deep RL domains.

This work also does not consider which segment pairs should be presented for labeling with preferences used for reward learning. However, other research has addressed this problem through active learning [14][9][25], and it may be possible to simply swap our Algorithm 1 into these active learning methods, combining the improved sample efficiency of P_{regret} with that of these active learning methods.

Regarding the human side of the problem of reward learning from preferences, further research could provide several improvements. First, we are confident that humans can be influenced by their training and by the preference elicitation interface, which is a particularly rich direction for follow-up study. We also do not consider how to handle learning reward functions from multiple human stakeholders who have different preferences, a topic we revisit in Appendix A.2. Lastly, we expect humans to deviate from any simple model, including P_{regret}, and a fine-grained characterization of how humans generate preferences could produce preference models that further improve the alignment of the reward functions that are ultimately learned from human preferences.

A.2 Ethical statement

This work is meant to address ethical issues that arise when autonomous systems are deployed without properly aligning their objectives with those of human stakeholders. It is merely a step in that direction, and overly trusting in our methods—even though they improve on previous methods for alignment—could result in harm caused by poorly aligned autonomous systems.

When considering the objectives for such systems, a critical ethical question is *which* human stakeholders' interests the objectives should be aligned with and how multiple stakeholders' interests should be combined into a single objective for an autonomous system. We do not address these important questions, instead making the convenient-but-flawed assumption that many different humans' preferences can simply be combined. In particular, care should be taken that vulnerable and marginalized communities are adequately represented in any technique or deployment to learn a reward function from human preferences in high-impact settings. The stakes are high: for example, a reward function that is only aligned with a corporation's financial interests could lead to exploitation of such communities or more broadly to exploitation of or harm to users.

In this specific work, our filter for which Mechanical Turk Workers could join our study is described in Appendix D. We did not gather demographic information and therefore we cannot assess how representative our subjects are of any specific population.

A.3 On the challenge of using regret preference models in practice

We have provided evidence—theoretically and with experimentation—that the regret preference model is more effective when precisely measured or effectively approximated. The challenge of efficiently creating such approximations presents one clear path for future research and does not justify staying within the local maximum of the partial return preference model.

Like the regret preference model, inverse reinforcement learning (IRL) was founded on an algorithm that requires solving an MDP in an inner loop of learning a reward function. For example, see the seminal work on IRL by Ng and Russell [19]. This challenge has not stopped IRL from being an impactful problem, and handling this inner-loop computational demand is the focus of much IRL research.

Future work on the application of the regret preference model can face the challenge of scaling to more complex problems. Given that IRL has made tremendous progress in this direction and Brown et al. [26] have scaled an algorithm with similar needs to those of Algorithm 1, we are optimistic that the methods to scale can be developed, likely with light adaptation from existing methods (e.g., in Brown et al. or in Appendix F.1, under "Instantiating Algorithm 1 for reward functions that may be non-linear").

B Preference models for learning reward functions

For the reader's convenience, below we derive the logistic expression of a function that is based on two subtracted values from the Boltzmann distribution (i.e., softmax) representation that is more common in past work. These values are specifically the same function f applied to each segment, which is a general expression of both of the preference models considered here.

$$\begin{aligned} P(\sigma_1 \succ \sigma_2) &= \frac{\exp\left[f(\sigma_1)\right]}{\exp\left[f(\sigma_1)\right]+\exp\left[f(\sigma_2)\right]} \\ &= \frac{1}{1+\frac{\exp\left[f(\sigma_2)\right]}{\exp\left[f(\sigma_1)\right]}} \\ &= \frac{1}{1+\exp\left[f(\sigma_2)-f(\sigma_1)\right]} \\ &= logistic(f(\sigma_1)-f(\sigma_2)). \end{aligned} \tag{7}$$

B.1 Logistic-linear preference model

In Appendix E.2, we also consider preference models that arise by making the noiseless preference model a linear function over the 3 components of P_{regret_d}. Building upon Eqn. 7 above, we set $f(\sigma) = \vec{w} \cdot \langle V^*_{\tilde{r}}(s_{\sigma,0}), \Sigma_\sigma, V^*_{\tilde{r}}(s_{\sigma,|\sigma|})\rangle$. This preference model, $P_{log-lin}$, can be expressed after algebraic manipulation as

$$P_{log-lin}(\sigma_1 \succ \sigma_2 | \tilde{r}) =$$
$$logistic([\vec{w} \cdot \langle V^*_{\tilde{r}}(s_{\sigma_1,0}) - V^*_{\tilde{r}}(s_{\sigma_2,0}), \Sigma_{\sigma_1} - \Sigma_{\sigma_2}, V^*_{\tilde{r}}(s_{\sigma_1,|\sigma_1|}) - V^*_{\tilde{r}}(s_{\sigma_2,|\sigma_2|}) \rangle]). \quad (8)$$

This logistic-linear preference model is a generalization of P_{Σ_r} and also of P_{regret_d}, the regret preference model for deterministic transitions. Specifically, if $\vec{w} = \langle 0,1,0 \rangle$, then $P_{log-lin}(\cdot|\tilde{r}) = P_{\Sigma_r}(\cdot|\tilde{r})$. And if $\vec{w} = \langle -1,1,1 \rangle$, then $P_{log-lin}(\cdot|\tilde{r}) = P_{regret_d}(\cdot|\tilde{r})$.

B.2 Adding a constant probability of uniformly distributed preference

~~Some of our analysis in the appendix—but not in the body of this paper—includes~~ Appendix E.2 also considers adaptations of P_{Σ_r}, P_{regret_d}, and $P_{log-lin}$ that add a constant probability of uniformly distributed preference, as was done by Christiano et al. [9]. The body of the paper does not consider these adaptations.

We create this adaptation, which we will call P' here, from another preference model P by $P'(\sigma_1 \succ \sigma_2) = [(1 - logistic(c)) * P(\sigma_1 \succ \sigma_2)] + [logistic(c)/2]$, where c is a constant that in practice we fit to data and $logistic(c)$ is the constant probability of uniformly random preference. The $logistic(c)$ allows any constant c to result in a the constant probability of uniformly distributed preference to be in $(0,1)$. The term $logistic(c)/2$ gives half of the constant probability to σ_1 and half to σ_2. The term $[1 - logistic(c)]$ scales the $P(\sigma_1 \succ \sigma_2)$ probability—which could be P_{Σ_r}, P_{regret_d}, or $P_{log-lin}$—to a proportion of the remaining probability. The only difference in this adaptation and Christiano et al.'s 0.1 probability of uniformly distributed preference is that we learn the value of c from training data (in a k-fold cross-validation setting), as we see in Appendix E, whereas Christiano et al. do not share how 0.1 was chosen.

B.3 Expected return preference model

In Appendix F.3, we test reward learning on a third preference model. This expected return preference model is derived by making $f(\sigma) = -(\Sigma_\sigma \tilde{r} + V^*_{\tilde{r}}(s_{\sigma,|\sigma|}))$, in Equation 7. This segment statistic $f(\sigma)$ can be considered be in between deterministic regret (Equation 3) and partial return, differing from each by one term.

We include this preference model because judging by expected return is intuitively appealing in that it considers the partial return along the segment and the end state value of the segment, and we found it plausible that human preference providers might tend to ignore start state value, as this preference model does. However, reward learning with the regret model outperforms or matches that by this expected return preference model, as we show in Appendix F.3.

B.4 Policy learning from regret-based preferences without learning a reward function

When preferences are based upon regret, an entirely different method of policy learning appears reasonable. This method is motivated by regret itself differentiating which of two segments is more desirable, including differentiating an optimal segment from a suboptimal segment.

When learning a reward function is not desired, one can instead learn a segment scoring function that estimates the regret of a segment directly. Such a regret estimator can be learned by following the same algorithm for reward learning with the *partial return* preference model that is defined in Section ~~2~~ 2.2 and used throughout this paper, with two changes. First, each preference sample's segment ordering needs to be reversed before training. Second, the learned "reward function" is instead interpreted as a regret function, outputting a regret estimate for any segment. (Without reversing the order, the learned function would be *negated* regret.) If we assume the regret estimator perfectly models regret, any action that minimizes expected regret is an optimal action. Therefore, the policy is defined by choosing such a regret-minimizing action.

Though theoretical analysis of this approach is beyond the scope of this paper, we suspect that it would be possible to show that the above algorithm will result in an optimal policy for noiseless or stochastic

preferences, given the infinite and exhaustive dataset described in Definition 3.1. In Appendix F.2.3, we provide limited empirical evidence supporting this approach.

Even when learning a reward function, this approach could be used to help create reward features for reward functions that are non-linear or have unknown reward features (as discussed in Appendix F.1) and to identify a policy for which to learn successor features.

B.5 Relationship to inverse reinforcement learning

The inputs to inverse reinforcement learning (IRL) and learning reward functions from pairwise preferences are different: IRL requires demonstrations, not preferences over segment pairs. However, because a a regret-based preference model always prefers optimal segments over suboptimal segments, at least one connection can be made. If one assumes that a demonstrated trajectory segment is noiselessly optimal—as in the foundational IRL paper on apprenticeship learning [27])—then such a demonstration is equivalent to expressing preference or indifference for the demonstrated segment over all other segments (or, equivalently, that no other segment is preferred over the demonstrated segment). However, IRL has its own identifiability issues in noiseless settings (e.g., see Kim et al. [28]) that, viewed from the lens of preferences, come in part from the "indifference" part of the above statement: since there can be multiple optimal actions from a single state, it is not generally correct to assume that a demonstration of one such action shows a preference over all others, and therefore it remains unclear in IRL what other actions are optimal. Note that since partial-return-based preferences can prefer suboptimal segments over optimal segments, the common assumption in IRL that demonstrations are optimal does not map as cleanly to partial-return-based preferences.

The regret preference model also relates to IRL in that the most basic version of IRL requires solving an MDP in the inner loop, as appears necessary for a perfect measure of regret while learning a reward function [29, Algorithm 1].

C Theoretical comparisons

For convenience, Theorems 3.1 and 3.2 from Section 3 are reprinted below. Consider reviewing the definitions of optimal segments and suboptimal segments in Section 2.1 and Definition 3.1 before proceeding.

Theorem 3.1 (P_{regret} is identifiable)**.** *Let P_{regret} be any function such that if $regret(\sigma_1|\tilde{r}) < regret(\sigma_2|\tilde{r})$, $P_{regret}(\sigma_1 \succ \sigma_2|\tilde{r}) > 0.5$, and if $regret(\sigma_1|\tilde{r}) = regret(\sigma_2|\tilde{r})$, $P_{regret}(\sigma_1 \succ \sigma_2|\tilde{r}) = 0.5$. P_{regret} is identifiable.*

Proof Make all assumptions in Definition 3.1. Since $\hat{r}$ minimizes cross-entropy loss, $P_{regret}(\cdot|r) = P_{regret}(\cdot|\hat{r})$ for all possible segment pairs. Also, by Equation 4 $regret(\sigma|\tilde{r}) = 0$ if and only if σ is optimal with respect to $\tilde{r}$. And $regret(\sigma|\tilde{r}) > 0$ if and only if σ is suboptimal with $\tilde{r}$.

With respect to some $\tilde{r}$, let σ^* be any optimal segment and $\sigma^{\neg *}$ be any suboptimal segment. $regret(\sigma^*|\tilde{r}) < regret(\sigma^{\neg *}|\tilde{r})$. $P_{regret}(\sigma^* \succ \sigma^{\neg *}|\tilde{r}) > 0.5$, which we refer to as being *preferred by* $P_{regret}(\cdot|\tilde{r})$. $P_{regret}(\cdot|\tilde{r})$ induces a total ordering over segments, defined by $regret(\sigma_1|\tilde{r}) < regret(\sigma_2|\tilde{r}) \Longleftrightarrow P_{regret}(\sigma_1 \succ \sigma_2|\tilde{r}) > 0.5 \Longleftrightarrow \sigma_1 > \sigma_2$ and $regret(\sigma_1|\tilde{r}) = regret(\sigma_2|\tilde{r}) \Longleftrightarrow P_{regret}(\sigma_1 \succ \sigma_2|\tilde{r}) = 0.5 \Longleftrightarrow \sigma_1 = \sigma_2$. Because regret has a minimum (0), there must be a set of segments which are ranked highest under this ordering, denoted $\Sigma^*_{\tilde{r}}$. These segments in $\Sigma^*_{\tilde{r}}$ are exactly those that achieve the minimum regret (0) and so are optimal with respect to $\tilde{r}$.

Since the dataset ($D_{\succ}$) contains all segments by assumption, $\Sigma^*_{\tilde{r}}$ contains all optimal segments with respect to $\tilde{r}$. If a state-action pair (s,a) is in an optimal segment, then by the definition of an optimal segment $Q^*_{\tilde{r}}(s,a) = V^*_{\tilde{r}}(s)$. The set of optimal policies $\Pi^*_{\tilde{r}}$ for $\tilde{r}$ is all π such that, for all (s,a), if $\pi(s,a) > 0$, then $Q^*_{\tilde{r}}(s,a) = V^*_{\tilde{r}}(s)$. In short, $\Sigma^*_{\tilde{r}}$ determines the set of every state-action pair (s,a) such that $Q^*_{\tilde{r}}(s,a) = V^*_{\tilde{r}}(s)$, and that set determines $\Pi^*_{\tilde{r}}$. Therefore $\Sigma^*_{\tilde{r}}$ determines $\Pi^*_{\tilde{r}}$, and we will refer to this determination as the function g.

We now focus on the reward function used to generate preferences, r, and on the learned reward function, $\hat{r}$. Since $P_{regret}(\cdot|r) = P_{regret}(\cdot|\hat{r})$, r and $\hat{r}$ induce the same total ordering over segments, and so $\Sigma_r^* = \Sigma_{\hat{r}}^*$. Therefore $g(\Sigma_r^*) = g(\Sigma_{\hat{r}}^*)$. Since $g(\Sigma_r^*) = \Pi_r^*$ and $g(\Sigma_{\hat{r}}^*) = \Pi_{\hat{r}}^*$, $\Pi_r^* = \Pi_{\hat{r}}^*$. □

Theorem 3.2 (Noiseless P_{Σ_r} is not identifiable)**.** *Let P_{Σ_r} be any function such that if $\Sigma_{\sigma_1}\tilde{r} > \Sigma_{\sigma_2}\tilde{r}$, $P_{\Sigma_r}(\sigma_1 \succ \sigma_2|\tilde{r}) = 1$, and if $\Sigma_{\sigma_1}\tilde{r} = \Sigma_{\sigma_2}\tilde{r}$, $P_{\Sigma_r}(\sigma_1 \succ \sigma_2|\tilde{r}) = 0.5$. There exists an MDP in which P_{Σ_r} is not identifiable.*

Below we present two proofs of Theorem 3.2. Each are proofs by counterexample. Though only one proof is needed, we present two because each counterexample demonstrates a qualitatively different category of how the partial return preference model can fail to identify the set of optimal policies.

Proof based on stochastic transitions: Assume the following class of MDPs, illustrated in Figure 7. The agent always begins at start state s_0. From s_0, action a_{safe} always transitions to s_{safe}, getting a reward of 0. From s_0, action a_{risk} transitions to s_{win} with probability 0.5, getting a reward of r_{win}, and transitions to s_{lose} with with probability 0.5, getting a reward of -10. In all MDPs in this class, $r_{win} > 0$. All 3 possible resulting states (s_{safe}, s_{win}, and s_{lose}) are absorbing states, from which all further reward is 0.

If $r_{win} \geq 10$, a_{risk} is optimal in s_0. If $r_{win} \leq 10$, a_{safe} is optimal in s_0. Three single-transition segments exist: $(s_0, a_{safe}, s_{safe})$, (s_0, a_{risk}, s_{win}), and $(s_0, a_{risk}, s_{lose})$. By noiseless P_{Σ_r}, $(s_0, a_{risk}, s_{win}) \succ (s_0, a_{safe}, s_{safe}) \succ (s_0, a_{risk}, s_{lose})$, regardless of the value of r_{win}. In other words, P_{Σ_r} is insensitive the what the optimal action is from s_0 in this class of MDPs.

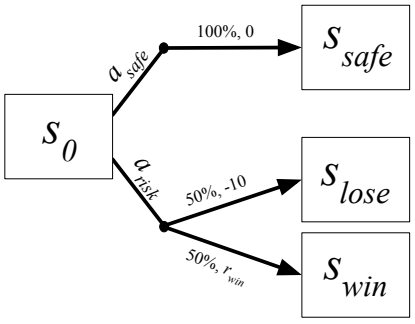


Figure 7: A class of MDPs in which, if $r_{win} > 0$, the partial return preference model fails the test for identifiability.

Now assume MDP M, where r_{win} = 11. In linear form, the weight vector for the reward function r_M can be expressed as $w_{r_{M_1}} = <-10, 0, 11>$. Let $\hat{r}_M$ have $w_{\hat{r}_M} = <-10, 0, 9>$. Both r_M and $\hat{r}_M$ have the same preferences as above, meaning that $\hat{r}_M$ minimizes loss on an infinite preferences dataset $D_{\succ}$ created by P_{Σ_r}, yet it has a different optimal policy. Therefore, noiseless P_{Σ_r} is not identifiable. □

In contrast, note that by noiseless P_{regret}, the preferences are different than those above for P_{Σ_r}. If $r_{win} > 10$, then $(s_0, a_{risk}, s_{win}) \sim (s_0, a_{risk}, s_{lose}) \succ (s_0, a_{safe}, s_{safe})$, If $r_{win} < 10$, then $(s_0, a_{safe}, s_{safe}) \succ (s_0, a_{risk}, s_{win}) \sim (s_0, a_{risk}, s_{lose})$. Intuitively, this difference comes from P_{regret} always giving higher preference probability to optimal actions, even if they result in bad outcomes. Another perspective can be found from the utility theory of Von Neumann and Morgenstern [30]. Specifically, P_{Σ_r} gives preferences over outcomes, which in the terms of utility theory can only learn an ordinal utility function. Ordinal utility functions are merely consistent with the preference ordering over outcomes and do not generally capture preferences over actions when their outcomes are stochastically determined. The deterministic regret preference model, P_{regret_d}, also has this weakness in tasks with stochastic transitions. On the other hand, P_{regret} forms preferences

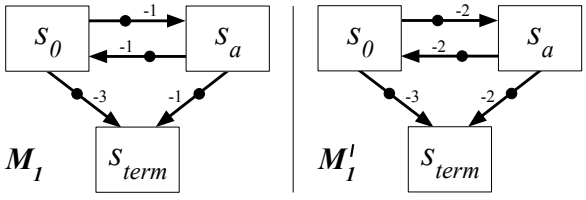


Figure 8: An MDP (M_1) where $\Pi_r^* = \Pi_{\hat{r}}^*$ is not guaranteed for the partial return preference model, failing the test for identifiability with segments of length 1. The ground truth reward function is shown to the left, and an MDP M_1' with an alternative reward function is shown to the right. Under partial return, both create the same set of preferences despite having different optimal actions from s_0.

over so-called lotteries—the distribution over possible outcomes—and can therefore learn a cardinal utility function, which can explain preferences over risky action.

Since the proof above is focused on stochastic settings, we show the lack of identifiability for noiseless P_{Σ_r} can be found for quite different reasons in a deterministic MDP.

Proof based on segments of fixed length: Consider the MDP M_1 in Figure 8 and assume preferences are given over segments with length 1 (i.e., containing one transition). The optimal policy for M_1 is to move rightward from s_0, whereas optimal behavior for M_1' is to move leftward from s_0. In *both* M_1 and M_1', preferences by P_{Σ_r} are as follows, omitting the action for brevity: $(s_a,s_0) \sim (s_a,s_{term}) \sim (s_0,s_a) \succ (s_0,s_{term})$. As in the previous proof, P_{Σ_r} is insensitive to certain changes in the reward function that alter the set of optimal policies. Whenever this characteristic is found, $\Pi_r^* = \Pi_{\hat{r}}^*$ is not guaranteed, failing the test for identifiability. Here specifically, the reward function for M_1' would achieve 0 cross-entropy loss on an exhaustive preference dataset created in M_1 with the noiseless preferences from the partial return preference model, despite the optimal policy in M_1' conflicting with the ground truth optimal policy.

The logic of this proof can be applied for trajectories of length 2 in the MDP M_2 shown in Figure 9. Together, M_1 and M_2 suggest a rule for constructing an MDP where $\Pi_r^* = \Pi_{\hat{r}}^*$ is not guaranteed for P_{Σ_r}, failing the identifiability test for any fixed segment length, $|\sigma|$: set the number of states to the right of s_0 to $|\sigma|$ (not counting s_{term}), set the reward r_{fail} for (s_0, s_{term}) such that $r_{fail} < 0$, and set the reward for each other transition to $c + r_{fail}/(|\sigma|+1)$, where $c > 0$. Given an MDP constructed this way, an alternative reward function that results in the same preferences under P_{Σ_r} yet has a different optimal action from s_0 can then be constructed by changing all reward other than r_{fail} to $c + r_{fail}/(|\sigma|+1)$, where c now is constrained to $c < 0$ and $c \times |\sigma| < r_{fail}$. Note that the set of preferences for each of these MDPs is the same even when including segments that reach terminal state before $|\sigma|$ transitions (which can still be considered to be of length $|\sigma|$ if the terminal state is an absorbing state from which reward is 0).

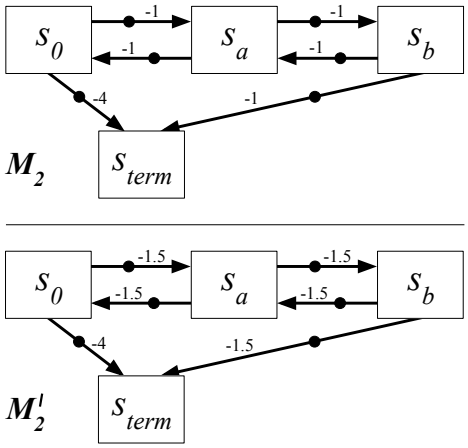


Figure 9: An MDP (M_2) where $\Pi_r^* = \Pi_{\hat{r}}^*$ is not guaranteed for the partial return preference model, failing the test for identifiability with segments of length 2. The ground truth reward function is shown in the top diagram, and an MDP M_2' with an alternative reward function is shown in the bottom diagram. Under partial return, both create the same set of preferences despite having different optimal actions from s_0.

The relevance of noiseless preference generators Because we model preferences as stochastic in Section 2, at this point one might reasonably wonder how the above theoretical analysis of noiseless preference generators are relevant. We offer four arguments below.

First, having structured noise provides information that can help both preference models, but these proofs show that there are cases where the signal behind the noise—either regret or partial return—is not sufficient in the partial return case to identify an equivalent reward function. So, in a rough sense, regret more effectively uses both the signal and the noise, which might explain its superior sample efficiency in our experiments across both human labels and synthetic labels. Relatedly, the noiseless setting can help us understand each preference model's sample efficiency in a low-noise setting.

Second, noiseless preferences are also feasible, even if they are rare. Therefore, understanding what can be learned from them is worthwhile. Theorem 3.2 shows that there are MDPs in which there is *no* class of preference models—stochastic or deterministic—that can identify an equivalent

reward function from partial-return-based preferences if the preference generator noiselessly prefers according to partial return. Specifically, we show that the mapping from two reward functions with different sets of optimal policies to partial-return based preferences is a many-to-one-mapping, and therefore the information simply does not exist to invert that mapping and identify a reward function with the same set of optimal policies. In contrast, Theorem 3.1 shows that preferences generated noiselessly (and in certain stochastic settings) by regret do not have this issue.

Third, noise is often motivated as modeling human error. Having an algorithm rely on noise—structured in a very specific, Boltzmann-rational way—is an undesirable crutch.

Lastly, there is precedent for considering noiseless human input for theory or derivations. For instance, the foundational IRL research by Abbeel and Ng on apprenticeship learning [27] treats demonstrations as noiselessly optimal. Recent work by Kim et al. [28] focuses on reward identifiability with noiseless, optimal demonstrations.

D Additional information for creating a human-labeled preference dataset

D.1 The user interface and study overview

Here we share miscellaneous details about the user interface from which we collected human subjects' preferences. This description builds on Section 4.2.

In selecting preferences, subjects had four options. They could prefer either trajectory (left or right), or they could express their preference to be the same or indistinguishable. To provide these preferences, subjects could either click on each of the buttons labeled "LEFT", "RIGHT", "SAME", or "CAN'T TELL" (shown in Figure 3) or by using the arrow keys to select amongst these choices.

For the interface, all icons used to visualize the task were obtained from icons8.com under their Paid Universal Multimedia Licensing Agreement.

We paid all subjects $5 per experiment (i.e., for each a Mechanical Turk HIT), which was chosen using the median time subjects took during a pilot study and then calculating the payment to result in $15 USD per hour. This hourly rate of $15 was chosen because it is commonly recommended as an improved US federal minimum wage. The human subject experiments cost $2,145 USD in total.

~~The NeurIPS submission checklist triggered a conversation among the authors, since the~~ An experimental error resulted in the IRB-approved consent form ~~was not~~ not being presented to human subjects after Mechanical Turk Workers accepted our study. ~~It is our understanding that due to the nature of the study, which may be considered low-risk data labeling, explicit consent for this use was not needed or is implicit by the nature of the study being accepted on the Mechanical Turk interface. We are confirming with the IRB and will include a statement to this effect in the next version.~~ We reported this error to our IRB and received their approval to use the data.

~~A copy of our IRB letter of approval can be read here. Link inactivated during review to maintain this submission's anonymity.~~

D.2 Filtering subject data

Before someone could join our study via Amazon Mechancial Turk, they had to meet the following criteria. They had to be located in the United States, have an approval rating of at least 99%, and have completed at least 100 other MTurk HITs. We selected these criteria to improve the probability of collecting data from subjects who would attentively engage with our study and who would understand our training protocol.

We assessed each subject's understanding of the delivery domain and filtered out those who did not comprehend the task, as described below. Specifically, subjects completed a task-comprehension survey, through which we assigned them a task-comprehension score. The questions and answer choices are shown in Table 2. Each fully correct answer was worth 1 point and each partially correct answer was worth 0.5 points. Task-comprehension scores were bounded between 0 and 7. We removed

Table 2: The task comprehension survey, designed to test participant's comprehension of the domain for the purpose of filtering data. Each full credit answer earned 1 point; each partial credit answer earned 0.5 points. We discarded the data of participants who scored less than 4.5 points overall.

Question	Full credit answer	Partial credit answer	Other answer choices
What is the goal of this world? (Check all that apply.)	• To maximize profit	• To get to a specific location. • To maximize profit Partial credit was given if both answers were selected.	• To drive as far as possible to explore the world. • To collect as many coins as possible. • To collect as many sheep as possible. • To drive sheep to a specific location.
What happens when you run into a house? (Check all that apply.)	• You pay a gas penalty. • You can't run into a house; the world doesn't let you move into it. Full credit was given if both answers were selected.	• You pay a gas penalty. • You can't run into a house; the world doesn't let you move into it. Partial credit was given if only one answer was selected.	• The episode ends. • You get stuck. • To collect as many sheep as possible.
What happens when you run into a sheep? (Check all that apply.)	• The episode ends. • You are penalized for running into a sheep. Full credit was given if both answers were selected.	• The episode ends. • You are penalized for running into a sheep. Partial credit was given if only one answer was selected.	• You are rewarded for collecting a sheep.
What happens when you run into a roadblock? (Check all that apply.)	• You pay a penalty.		• The episode ends. • You get stuck. • You can't run into a roadblock; the world doesn't let you move into it.
Is running into a roadblock ever a good choice in any town?	• Yes, in certain circumstances.		• No.
What happens when you go into the brick area? (Check all that apply.)	• You pay extra for gas.		• The episode ends. • You get stuck in the brick area. • You can't go into the brick area; the world doesn't let you move into it.
Is entering the brick area ever a good choice?	• Yes, in certain circumstances		• No

the data from subjects who scored below a threshold of 4.5. The threshold of 4.5 was chosen based on visual analysis of a histogram of scores, attempting to balance high standards for comprehension with retaining sufficient data for analysis.

In addition to filtering based off the task comprehension survey, we also removed a subject's data if they ever preferred colliding the vehicle into a sheep over not doing so. Since such collisions are highly undesirable in this task, we interpreted this preference as evidence of either poor task understanding or inattentiveness.

In total, we collected data from 143 subjects. 58 of these subjects were removed based on their responses to the survey, and another 35 were removed for making preference errors. After filtering by both the comprehension survey and subject error, we used the data from 50 subjects. This included 1812 preferences over 1245 unique segment pairs.

Regarding potential risks to subjects, this data collection had limited or no risk. No offensive content was shown to subjects while they completed the HIT. Mechanical Turk collected Worker IDs, which were used only to link preference data with the results from the task-comprehension survey for filtering data (see Appendix D.2) and then were deleted from our data. No other potentially personally identifiable information was collected.

D.3 The two stages of data collection

We collected the human preference dataset in two stages, as mentioned in Section 4.2. Here we provide more detail on each stage. These stages differed largely by their goals for data collection and, following those goals, how we chose which segment pairs were presented to subjects for their preference.

~~Coordinates from which segment pairs were sampled from during the first stage of data collection. The x-axis is state value differences between the two segments and the y-axis is partial return differences between the two segments. The areas of the circles are proportional to the number of samples at that point, and the proportionality is consistent across this plot and the 3 subplots of Figure 11.~~

First stage ~~The first stage of data collection involved choosing segment pairs to present to the subjects.~~ Figure 10 illustrates the coordinates that segment pairs were sampled from in the first stage of data collection, varying by state value differences and by differences in partial returns over the segments. We sought a range of points that would allow a characterization of human preferences that is well distributed across different parts of the plot. To better differentiate the consequences of each preference model, we intentionally chose a large number of points in the gray area of Figure 4, where the regret and partial return preference models would disagree (i.e., each giving a different segment a preference probability greater than 0.5). ~~The segments pairs at these points presented to subjects included~~

We now describe our segment-pair sampling process more specifically. We first we constructed all unique segments of length 3 and then exhaustively paired them, resulting in nearly 30 million segment pairs. Each segment pair's partial returns, start-state values, end-state values place the segment pair on a coordinate in Figure 4, and segment pairs that are not on any of the dots in Figure 4 were discarded. For the segment pairs at each coordinate, we further divided them into 5 bins: the terminal segments with the same start state and different end states, non-terminal segments with different start states and different end states, terminal segments with the same start state and same end state, ~~and~~ terminal segments with a different start states and the same end state~~.~~, and bin of segment pairs that fit in none of the other bins. Segment pairs in the 5th bin were discarded. From each of the 4 bins corresponding to each point in Figure 4, we randomly sampled 20 segment pairs. If the bin did not have at least 20 segment pairs, all segment pairs in the bin were "sampled". All sampled segment pairs from all bins for all points in Figure 4 made up the pool of segment pairs used with Mechanical Turk. For each subject, 50 segment pairs were randomly sampled from this pool. We gathered data until we had roughly 20 labeled segment pairs per bin. After filtering subject data, this first stage contributed 1501 segment pairs out of the 1812 pairs used in our reward learning experiments in Section 6.3 and Appendix F.3.

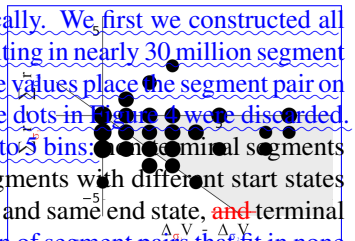


Figure 10: Coordinates from which segment pairs were sampled from during the first stage of data collection. The x-axis is state value differences between the two segments and the y-axis is partial return differences between the two segments. The areas of the circles are proportional to the number of samples at that point, and the proportionality is consistent across this plot and the 3 subplots of Figure 11.

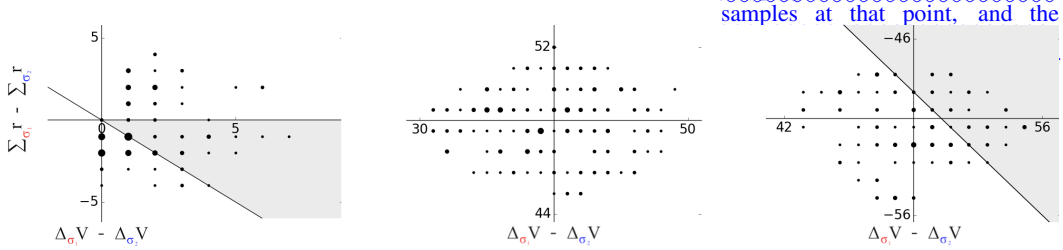


Figure 11: Coordinates from which segment pairs were sampled from during the second stage of data collection. The points are in 3 distant clusters, so they are presented in 3 separate subplots for readability. The areas of the circles are proportional to the number of samples at that point, and the proportionality is consistent across these 3 subplots and Figure 10.

Second stage When we conducted the reward-learning evaluation in Section 6 with only the data from the first stage, P_{Σ_r} performed very poorly, always performing worse than uniformly random. This performance difference is shown in Appendix **??**. In contrast P_{regret} performed well, always achieving near-optimal

performance. To better assess P_{Σ_r}, we investigated its results ~~in detail~~ with synthetic preferences in detail and speculated that two types of additional segment pairs would aid its performance. The first of these two types ~~are pairs of segments between terminal and~~ include one segment that is terminal and one that is non-terminal~~segments, which may~~, which we expected to help differentiate the reward for reaching terminal states from that of reaching non-terminal ones. The second of these two types are ~~pairs of~~ two segments that each terminate at different ~~transitions, which we speculated~~ t values. For example, one segment terminates on its end state, $s_{\sigma,|\sigma|}$, and another terminates after its first transition, at $s_{\sigma,1}$. These early-terminating segments can be viewed either as shorter segments or as segments of the same length as the other segments ($|\sigma|=3$), where they reach absorbing state from which no future reward can be received. We speculated that this second type of segment pairs would help learn the ~~reward~~ negative reward component for each move (i.e., the gas cost). Specifically, in the first stage's data, both segments in a pair always have the same number of non-terminal transitions, seemingly preventing preferences from providing information about whether ~~on extra transition of movement~~ an extra transition (from non-absorbing state) generally resulted in positive or negative reward. These segment pairs were included in all results unless otherwise stated.

~~-~~We now describe our segment-pair sampling process for the second stage more specifically. For the first additional type of segment pair, where one segment is terminal and one is not, we randomly pair terminal and non-terminal segments from the first-stage pool of segment pairs drawn from to present to subjects. In this pairing, each segment is only used once, and pairing stops when one of all terminal segments or all non-terminal segments have been paired. The corresponding coordinates for these pairs are shown in the two right most plots of Figure 11. For the second additional type of segment pair, we utilize all terminal segments from the pool of segment pairs shown to subjects in the first stage. For each of these terminal segments, we construct two additional segments: one that shifts the segment earlier, removing the first state and action and adds a dummy transition within absorbing state at the end, and another that shifts the segment two timesteps earlier and adds two such dummy transitions at the end. These two newly constructed segments are then each paired with the original segment, producing two new pairs for each terminal segment in the data set. The corresponding coordinates for these segment pairs are shown in the left most plot of Figure 11.

All of both types of additional segments pairs are then characterized by the coordinates shown in Figure 11. Then, as with the first stage, we randomly sampled 20 segment pairs from each coordinate to make the experimental pool for the second round of Mechanical Turk data collection. If 20 segment pairs were not available at a coordinate, we used all segment pairs for that coordinate. As in the first stage, 50 segment pairs were randomly sampled from this pool to be presented to each subject during preference elicitation. After filtering subject data, this first stage contributed 311 segment pairs out of the 1812 pairs used in our reward learning experiments in Section 6.3 and Appendix F.3.

D.4 The study design pattern

This work follows an experimental design pattern that is often used for studying methods that take human input for evaluating the desirability of behaviors or outcomes. In this pattern, human subjects are taught to understand a specific task metric and/or are incentivized to align their desires with this metric. The human subjects then provide input to some algorithm that has no knowledge of the performance metric, and this algorithm or learned model is evaluated on how well its output performs with respect to the hidden metric. For another example, see Cui et al. [31].

E Descriptive results

E.1 Derivation of $regret_{\mathbf{d}}(\sigma_2|\tilde{r}) - regret_{\mathbf{d}}(\sigma_1|\tilde{r}) = (\Delta_{\sigma_1} V_{\tilde{r}} - \Delta_{\sigma_2} V_{\tilde{r}}) + (\Sigma_{\sigma_1}\tilde{r} - \Sigma_{\sigma_2}\tilde{r})$

The derivation below supports our assertion in the first paragraph of Section 5.1.

$$
\begin{aligned}
& regret_{\mathrm{d}}(\sigma_2|\tilde{r}) - regret_{\mathrm{d}}(\sigma_1|\tilde{r}) \\
&= \Big([V^*_{\tilde{r}}(s_{\sigma_2,0}) - (\Sigma_{\sigma_2}\tilde{r} + V^*_{\tilde{r}}(s_{\sigma_2,|\sigma_2|})] - [V^*_{\tilde{r}}(s_{\sigma_1,0}) - (\Sigma_{\sigma_1}\tilde{r} + V^*_{\tilde{r}}(s_{\sigma_1,|\sigma_1|})] \Big) \\
&= \Big([V^*_{\tilde{r}}(s_{\sigma_2,0}) - V^*_{\tilde{r}}(s_{\sigma_2,|\sigma_2|})] - [V^*_{\tilde{r}}(s_{\sigma_1,0}) - V^*_{\tilde{r}}(s_{\sigma_1,|\sigma_1|})] \Big) - \Big(\Sigma_{\sigma_2}\tilde{r} - \Sigma_{\sigma_1}\tilde{r} \Big) \\
&= \Big([V^*_{\tilde{r}}(s_{\sigma_1,|\sigma_1|}) - V^*_{\tilde{r}}(s_{\sigma_1,0})] - [V^*_{\tilde{r}}(s_{\sigma_2,|\sigma_2|}) - V^*_{\tilde{r}}(s_{\sigma_2,0})] \Big) + \Big(\Sigma_{\sigma_1}\tilde{r} - \Sigma_{\sigma_2}\tilde{r} \Big) \\
&= (\Delta_{\sigma_1} V_{\tilde{r}} - \Delta_{\sigma_2} V_{\tilde{r}}) + (\Sigma_{\sigma_1}\tilde{r} - \Sigma_{\sigma_2}\tilde{r})
\end{aligned}
\tag{9}
$$

E.2 Losses of an expanded set of preference models on the human preferences dataset

Table 3 shows an expansion of Table 1, including models introduced in Appendix B. The logistic linear preference model provides a lower bound in most cases, given that it can express anything the other preference models can and the rarity of overfitting its 3 parameters. Therefore, the intended comparisons are either between P_{regret} and P_{Σ_r} without the constant probability of a uniformly random response or between them with it. We embolden the result with lower loss between these two preference models for each such comparison.

F Results from learning reward functions

This section provides additional implementation details for Section 6, discussion of potential improvements, and additional analyses that thematically fit in Section 6.

Table 3: Expanding on Table 1, mean cross-entropy test loss over 10-fold cross validation (n=1812) from predicting human preferences. Lower is better.

Preference model	**Loss (n=1,812)**
$P(\cdot) = 0.5$ (uninformed)	0.69
P_{Σ_r} (partial return)	0.62
P_{regret} (regret)	**0.57**
$P_{log\text{-}lin}$ (logistic linear)	0.55
P_{Σ_r} with prob of uniform response	0.63
P_{regret} with prob of uniform response	**0.59**
$P_{log\text{-}lin}$ with prob of uniform response	0.57

F.1 An algorithm to learn reward functions with $regret(\sigma_\sigma|\hat{r})$

We describe below additional details of our instantiation of Algorithm 1.

Because the ordering of preference pairs is arbitrary, for all preference datasets we double the amount of data by duplicating each preference sample with the opposite ordering and the reversed preference. This provides more training data and avoids learning any segment ordering effects.

For this specific instantiation, we compute successor feature functions by first randomly creating a large number of reward functions. Specifically, each reward function is created by sampling with replacement each element of its weight vector, $\boldsymbol{w}_{\tilde{\boldsymbol{r}}}$, from $\{-50,-10,-2,-1,0,1,5,10,50\}$. We also included the ground-truth reward function, r, at this point, resulting in 70 reward functions. For each reward function, we create its maximum entropy optimal policy through value iteration. In practice, we learned the successor feature functions as part of the value iteration process. Finally, we remove any successor feature functions for redundant policies and then also remove the successor features function

for the optimal policy for r. Note that the only effect of including r in the earlier step was to allow us to remove any policies for other reward functions that were also optimal for r, making the regret-based learning problem more difficult. We ensured that the ground-truth reward function was not represented to better approximate real-world reward learning applications, in which one would be unlikely to have the optimal policy for learning a successor features function. (However, the policy from the regret estimator described in Appendix B.4 could be used to learn a successor features function, gaining the benefit of having a successor features function for at least one policy that is likely to perform decently in the task.)

During training, the loss for the P_{regret} model tended to show cyclical fluctuations, reaching low loss and then spiking. To handle this volatility, we used the $\hat{r}$ that achieved the lowest loss over all epochs of training, not the final $\hat{r}$. A better understanding of these cyclical fluctuations could further improve learning with P_{regret}.

Despite the delivery domain being an episodic task, a low-performing policy can endlessly avoid terminal states, resulting in negative-infinity values for both its return and successor features based on the policy. To prevent such negative-infinity values, we apply a discount factor of $\gamma = 0.999$ during value iteration—which is also where successor feature functions are learned—and when assessing the mean returns of policies with respect to the ground-truth reward function, r. We chose this high discount factor to have negligible effect on the returns of high-performing policies (since relatively quick termination is required for high performance) while still allowing value iteration to converge within a reasonable time.

Below we describe the other specific hyperparameters used for learning a reward function with both preference models. These hyperparameters were used across all experiments. For all models, the learning rate, softmax temperature, and number of training iterations were tuned on the noiseless synthetic preference data sets such that each model achieved an accuracy of 100% on our specific delivery task and then were tuned further on stochastic preferences on our specific delivery task.

Reward learning with the partial return preference model learning rate: 2; number of training epochs: 30,000; and optimizer: Adam (with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and eps= $1e-08$).

Reward learning with the regret preference model learning rate: 0.5; number of training epochs: $5{,}000$; optimizer: Adam (with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and eps=$1e-08$); and softmax temperature: 0.001.

Logistic regression with both preference models, for the likelihood analysis in Section 5.2 and Appendix E.2 learning rate: 0.5; number of training iterations: $3{,}000$; optimizer: stochastic gradient descent; and evaluation: 10-fold cross validation.

The computer used to run all experiments had the following specification. Processor: 1x Core™ i9-9980XE (18 cores, 3.00 GHz) & 1x WS X299 SAGE/10G | ASUS | MOBO; GPUs: 4x RTX 2080 Ti; Memory: 128 GB; and operating system drive: 2 TB NVMe (3,500 MB/s read).

Pytorch 1.7.1 [32] was used to implement all reward learning models, and statistical analyses were performed using Scikit-learn 0.23.2 [33].

Instantiating Algorithm ~~F.1~~ 1 for reward functions that may be non-linear Algorithm 1 assumes that the reward function can be expressed as a linear combination of reward features that are provided by a reward-features function ϕ that is input to the algorithm. Here we address situations when that assumption does not hold. If the reward features are unknown or the reward is known to be non-linear, one method is to create a reward features function that permits a linear *approximation* of the reward function. Several methods to derive some or all of these reward features appear promising:

- Reward features can be learned by minimizing several auxiliary losses in a self-supervised fashion, as by Brown et al. [26]. After optimizing for these various objectives using a single neural network, the activations of the penultimate layer of this network can be used as reward features. Such auxiliary tasks may include minimizing the mean squared error of the reconstruction loss for the current state from a lower-dimensional embedding and the original

state, predicting how much time has passed between states by minimizing the mean squared error loss (i.e., learning a temporal difference model), predicting the action taken between two states by minimizing the cross entropy loss (i.e., learning an inverse dynamics model), predicting the next state given the current state and action by minimizing the mean squared error loss(i.e., learning a forward dynamics model), and predicting which of two segments is preferred given a provided ranking by minimizing the t-rex loss.

- An additional auxiliary objective that may be promising is to use a neural network to learn a regret estimator as described in Appendix B.4 and then use the activations of its penultimate layer as reward features.
- Reward features could also be learned by first learning a reward function represented as a neural network *using a partial return preference model*, and then using the activations of the penultimate layer of this neural network to provide reward features.

F.2 Results from synthetic preferences

F.2.1 Learning reward functions from 100 randomly generated MDPs

Here we describe how each MDP in the set of 100 MDPs discussed in section 6.2 was generated. We also extend the analysis to illustrate how often each preference model performs better than uniformly random and give further details on our statistical tests.

Design choices The 100 MDPs are all instances of the delivery domain, but they have different reward functions. The height for each MDP is sampled from the set $\{5,6,10\}$, and the width is sampled from $\{3,6,10,15\}$. The proportion of cells that are terminal failure states is sampled from the set $\{0,0.1,0.3\}$. There is always exactly one terminal success state. The proportion of "mildly bad" cells were selected from the set $\{0,0.1,0.5,0.8\}$, and the proportion of "mildly good" cells were selected from $\{0,0.1,0.2\}$. Mildly good cells and mildly bad cells respectively correspond to cells with coins and roadblocks in our specific delivery task, but the semantic meaning of coins and roadblocks is irrelevant here. Each sampled proportion is translated to a number of cells (rounding down to an integer when needed) and then cells are randomly chosen to fill the grid with each of the above types of states until the proportions are satisfied.

Then, the ground truth reward component for each of the above cell types were sampled from the following sets:

- Terminal failure states: $\{0,1,5,10,50\}$
- Terminal success states: $\{-5,-10,-50\}$
- Mildly bad cells: $\{-2,-5,-10\}$

Mildly good cells always have a reward component of 1, and the component for white road surface cells is always -1. There are no cells with a higher road surface penalty (analogous to the bricks in the delivery domain).

Better than random performance Figure 12 complements the results in Figure 5, showing the percentage of MDPs in which each preference model outperforms a policy that chooses actions according to a uniformly random probability distribution. We can see that at this performance threshold, lower than that in Figure 5, the regret preference model outperforms the partial return preference model in most conditions. Even when their performance in this plot—based on outperforming uniformly random actions—is nearly

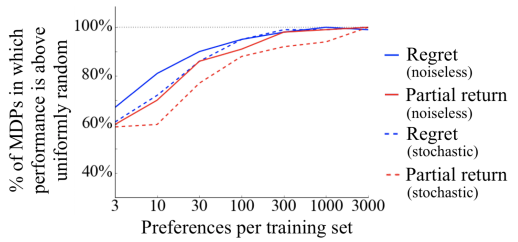


Figure 12: Comparison of performance over 100 randomly generated deterministic MDPs, showing the percentage of MDPs in which each model performed better than an agent taking actions by a uniformly random policy. This plot complements Figure 5, which shows the percentage of MDPs in which the models perform near-optimally.

identical, Figure 5 shows that the regret preference model achieves near optimal performance at a higher proportion.

Details for statistical tests We performed a Wilcoxon signed-rank test on the normalized average returns achieved by each model over the set of 100 randomly generated MDPs. All normalized average returns below -1 were replaced with -1, so that all such returns were in the range $[-1,1]$. This clipping was done because any normalized average return below 0 is worse than uniformly random, so the difference between a normalized return of -1 and -1000 is relatively unimportant compared to the difference between 1 and 0. Results are shown in Table 4.

Table 4: Results of the Wilcoxon signed-rank test on normalized average returns for each preference model.

Preference generator type	$\|D_{\succ}\|=3$	$\|D_{\succ}\|=10$	$\|D_{\succ}\|=30$	$\|D_{\succ}\|=100$	$\|D_{\succ}\|=300$	$\|D_{\succ}\|=1000$	$\|D_{\succ}\|=3000$
Noiseless (P_{regret} vs. P_{Σ_r})	w=1003, p=0.115	w=917, p=0.007	w=739, p=0.012	w=487, p=0.007	w=284, p<0.001	w=301, p=0.002	w=289, p=0.001
Stochastic (P_{regret} vs. P_{Σ_r})	w=979, p=0.541	w=1189.5, p=0.018	w=891, p=0.027	w=710, p=0.018	w=285, p<0.001	w=460, p=0.002	w=199, p<0.001

Additionally, we investigate whether P_{regret} and P_{Σ_r} learn near-optimal policies on the same MDPs within this set of 100 randomly generated MDPs. Results for this analysis are shown below.

Table 5: A table showing the count of the number of MDPs where both, either, or neither of the models achieved near optimal performance.

Model(s)	$\|D_{\succ}\|=3$	$\|D_{\succ}\|=10$	$\|D_{\succ}\|=30$	$\|D_{\succ}\|=100$	$\|D_{\succ}\|=300$	$\|D_{\succ}\|=1000$	$\|D_{\succ}\|=3000$
Both models	31	40	66	72	83	87	88
Only P_{regret}	20	26	17	18	14	8	8
Only P_{Σ_r}	10	12	7	8	3	3	3
Neither	39	22	10	2	0	2	1

F.2.2 Varying segment size

Here we consider the effect of increasing the fixed length of segments in the preference data set. Specifically, we learn a reward function using preference datasets that contained segments of lengths $n \in 6,9,12,15,18,21$ in the specific delivery task, where each dataset contained segments of the same length. Each preference dataset contained 3000 segment pairs. Each segment was generated by choosing a non-terminal start state and n actions, all uniformly randomly. As in Appendix F.2.1, each preference model acts as a preference generator to label these segment pairs, resulting in datasets that differ only in their labels, and then the preference model is used for reward learning on the same dataset it labeled. Unlike in Appendix F.2.1, all segments with termination reach terminal state on their final transition, which we had already observed was problematic for P_{Σ_r} (and the motivation for the second stage of data collection). Therefore, these results are intended to provide comparison between values of n more so than between preference models.

Regardless of the value of n, P_{regret} always achieves the optimal mean return of 41.2. P_{Σ_r} always achieves a mean return of -500.5. This analysis provides limited evidence that segment size does not have a large effect, though further analysis is needed to make this assertion with confidence.

F.2.3 Policy learning without reward learning, using a regret estimator

Here we test the performance of policy learning from regret-based preferences *without learning a reward function*, using a regret estimator, as described in Appendix B.4.

A set of 30 random MDPs is generated as described in Appendix F.2.1, except that the possible MDP widths are instead sampled from the set $\{1,5,6,10\}$ and the heights are sampled from $\{3,5,6,10\}$.

For each MDP, the preference dataset $D_{\succ}$ contains 3000 segment pairs, randomly sampled as in Appendix F.2.1. If 3000 unique segment pairs do not exist, then each possible segment pair is used once. Preference labels are given by P_{regret}.

Anticipating that a linear function over the six-element reward feature vector ϕ would be insufficient to estimate regret, we expanded ϕ to include both the six reward components and an identity function for each possible state-action pair, creating an addition to the feature vector of size $|S| \times |A|$ that contains all zeroes except a single 1 indicating the state and action that initiates a transition.

Table 6 shows results for noiseless and stochastic preferences.

Table 6: Success of learning a regret estimator with noiseless and stochastic preferences.

Preference Model	% of MDPs in which performance was better than uniformly random	% of MDPs in which performance was near optimal
Noiseless P_{regret} preference labels, learning a regret estimator	90%	86.67%
Stochastic P_{regret} preference labels, learning a regret estimator	80%	73.33%

F.2.4 Artificially lowering the discount factor

Almost all deep reinforcement learning algorithms artificially add discounting to tasks that are episodic [9]. Considering that much of the past work that used partial return preference models also involved deep RL, here we re-interpret results above to probe how such discounting affects performance of this preference model if preferences are actually given by P_{regret}. Specifically, the analysis above in Appendix F.2.3 on policy learning without reward learning is applicable to this topic. If the regret estimator is instead considered a negated reward function, then taking a minimal-regret action is equivalent to taking a maximum-value action under fully myopic discounting, $\gamma = 0$.

F.2.5 Reward learning in stochastic MDPs

Although we theoretically consider MDPs with stochastic transitions in Appendix C, we have not yet *empirically* compared P_{Σ_r} and P_{regret} in tasks with stochastic transitions, which we do below.

We randomly generated 20 MDPs, each with a 5×5 grid. Instead of terminal cells that are associated with success or failure, these MDPs have terminal cells that are either risky or safe. A single terminal *safe* cell was randomly placed, and the number of terminal *risk* cells was sampled from the set $\{1,2,7\}$ and then these terminal risk cells were likewise randomly placed. No other special cells were used in this set of MDPs. To add stochastic transitions, the delivery domain was modified such that when an agent moves into a terminal risk cell there is a 50% chance of receiving a lower reward, r_{lose}, and a 50% chance of receiving a higher reward, r_{win}. All other transitions are deterministic. As in the unmodified delivery domain, moving to any non-terminal state results in a reward of -1. Moving to the terminal safe state yields a reward of $+50$, like the terminal success state of the unmodified delivery domain. Therefore, depending on the values of r_{win} and r_{lose}, it may be better to move into a terminal risk state than to avoid it. All segments were generated by choosing a start state and three actions, all uniformly randomly. For each MDP, the preference dataset $D_{\succ}$ contains 3000 segment pairs.

The 10 MDPs of each condition differed from those of the other conditions by their ground-truth reward function r, with different r_{win} and r_{lose} values. The results are shown below, indicating that for both noiseless and stochastic preference datasets, P_{regret} is always able to achieve near-optimal performance, whereas P_{Σ_r} is not.

The results above expand upon and support the first proof of Theorem 3.2 in Appendix C.

F.3 Results from human preferences

Table 7: Stochastic MDPs: % of MDPs in which performance was near optimal, with varied reward functions.

Preference Model	$r_{win}=1$ $r_{lose}=-50$	$r_{win}=10^3$ $r_{lose}=-50$	$r_{win}=100$ $r_{lose}=-1$	$r_{win}=100$ $r_{lose}=-10^3$
Noiseless P_{regret}	100%	100%	100%	100%
Stochastic P_{regret}	100%	100%	100%	100%
Noiseless P_{Σ_r}	100%	0%	100%	0%
Stochastic P_{Σ_r}	100%	0%	100%	100%

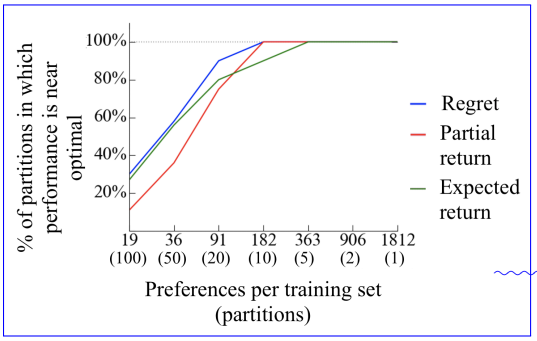


Figure 13: Performance comparison over various amounts of human preferences. Each partition has the number of preferences shown or one less. This plot is identical to Figure 6 except that results for the expected return preference model are included.

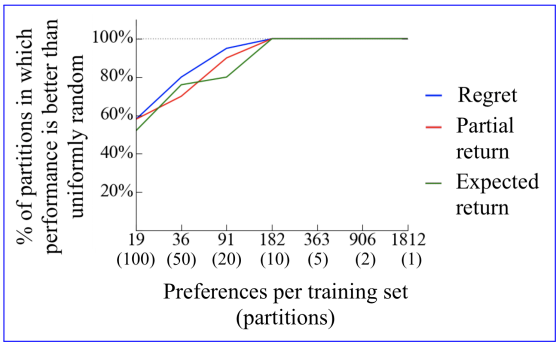


Figure 14: Performance comparison over various amounts of human preferences. Each partition has the number of preferences shown or one less. This plot focuses on outperforming a uniformly random policy, whereas Figure 6 thresholds on near-optimal performance.

In this section we provide further detail regarding the analysis in Section 6.3. For the Wilcoxon paired signed-rank test, normalized mean returns were clipped to $[-1,1]$ as in Appendix F.2.1. The result from each test is shown in Table F.3.

Table 8: Results from Wilcoxon signed-rank tests.

	5 partitions	10 partitions	20 partitions	50 partitions	100 partitions
P_{regret} vs. P_{Σ_r} preference models	w=0 p=0.043	w=6 p=0.028	w=24 p=0.007	w=216 p=0.003	w=939 p=0.076

~~In this section we provide further detail regarding the analysis in Section 6.3. For the Wilcoxon paired signed-rank test, normalized mean returns were clipped to $[-1,1]$ as~~ Figure Figure 13 show the same results as Figure 6, but with additional results from the expected return preference model introduced in Appendix ~~F.2.1. The result from each test is shown above. Figure ?? shows a visualization of the results that complements Figure 6.~~ B.3. Figure 14 shows the same results with a different threshold, that of performing better than uniformly random action selection, which receives a 0 on our normalized mean return metric. The regret preference model matches or outperforms *both* other preference models in all partitionings of the human data, at both thresholds (near optimal and better than random).

As previously mentioned in Section D and Appendix D, when learning reward functions only from the data from the first stage of human data collection, the partial return model does worse. The specific performance of the partial return preference model on the full set of first-stage data (i.e., 1 partition) is a normalized mean return of -3.8, whereas the regret preference model achieves 1.0, close to optimal performance.