Towards a Formal Theory of Reward Learning, With Application to Inverse Reinforcement Learning

Joar Skalse

Deducto Limited & King's College London,
formerly Oxford University & the Future of Humanity Institute.

The Iliad Intensive, 2026

Outline

Introduction & Background

What is Reward Learning?

In order to get an AI system to solve a task autonomously, we must first formalise the goal of that task as a *reward function*.

It is often very difficult to directly specify a reward function that never incentivises undesirable behaviour.

Reward learning is an area of machine learning concerned with the problem of *learning* a reward function from data.

In the simplest case, reward learning can be approached as a supervised learning problem.

My Research Agenda

My research has focused on building a formal theory of reward learning, with the aim of developing a deeper theoretical understanding of the reward learning problem.

The aim of this research agenda is partially to rigorously establish under what conditions different reward learning methods can be guaranteed to be reliable.

However, the aim of this work is in part also to increase our *understanding* of reward functions as abstract objects, and thereby give us better intuitions for what kinds of approaches are likely to work and what kinds of failure modes we should anticipate.

Why Study Reward Learning?

Reward optimisation is not the only way to get an AI system to do something; we can also use supervised learning, imitation learning, prompting, or activation engineering, etc.

However, all other methods for directing AI systems are anchored to and limited by current human performance.

If we want an AI system to solve a task that we cannot solve ourselves, then reward optimisation is by far the most promising method.

Theoretical Problems in Reward Learning

1. How 'close' must two reward functions R_1, R_2 be, before it would be bad to maximise R_2 instead of R_1? What is the right way to define 'close'?
2. How 'close' will the learnt reward function be to the true reward function, if we use a given reward learning method?
3. Can all important preference structures be expressed as reward functions?
4. How sensitive are different reward learning methods to misspecified data models?
5. What happens if you optimise a misspecified reward function?
6. Are there reward optimisation methods that are robust to misspecification of the reward function?

What is Inverse Reinforcement Learning?

Inverse Reinforcement Learning (IRL) is a type of reward learning that is concerned with inferring what objective an agent is pursuing based on the actions taken by that agent.

This roughly corresponds to the notion of *revealed preferences* in psychology and economics, since we are attempting to infer *preferences* from *behaviour*.

In its most general form, IRL could be used as a self-supervised method for learning representations of human preferences based on raw observations of human behaviour.

Why study IRL?

IRL is not very common in practice. So why focus on IRL?

One reason is that IRL is sufficiently complex that studying IRL will require the development of 'deep' theoretical tools, that may give a greater insight into the properties of reward learning more generally.

Another reason is that there is a duality between IRL and reward optimisation; they both involve reasoning about under what conditions two reward functions lead to the same policy.

Finally, IRL is entirely self-supervised. Thus, if it could be made to work, then it would be much more scaleable than other reward learning methods.

The Main Difficulties For IRL

The IRL problem is typically formalised as the problem of inferring a reward function R from a policy π. To do this, an IRL algorithm needs to have a model of how π relates to R, which is referred to as a *behavioural model*.

However, under most behavioural models, there are typically multiple reward functions which are consistent with each given policy. This means that the reward function is ambiguous, or *partially identifiable*, based on this data source.

To clearly understand the impact of this ambiguity, it is important that the ambiguity can be fully characterised and quantified.

The Main Difficulties For IRL

Another central difficulty in IRL is that the relationship between a person's prefereces and their behaviour is incredibly complex, and practically impossible to model perfectly.

By contrast, most IRL algorithms are based on very simple behavioural models, that typically correspond to some form of noisy optimality.

This means that these behavioural models are *misspecified*, which raises the concern that they might systematically lead to flawed inferences if applied to real-world data.

The Main Difficulties For IRL

Resolving the issue of misspecification in IRL is fundamentally difficult. Of course, we can incorporate findings from behavioural psychology, or use machine learning, to create behavioural models that are more and more accurate. However, it will never be realistically possible to create a behavioural model that is completely free from all forms of misspecification.

For this reason, it is important to understand how sensitive the IRL problem is to misspecification of the behavioural model. Is a mostly accurate behavioural model sufficient to ensure that the inferred reward function likewise is mostly accurate, or can a slight error in the behavioural model lead to a large error in the inferred reward?

Notation and Assumptions

I will use the following notation:

1. An MDP is a tuple $(S, A, \tau, \mu_0, R, \gamma)$.
2. Given S and A, let Π be the set of all policies, and $\mathcal{R}$ the set of all reward functions.
3. The *evaluation function* $J : \Pi \to \mathbb{R}$ gives the expected cumulative discounted reward of each policy.
4. The *ordering of policies* for a reward function R is the ordering induced by J.
5. The optimal Q-function is $Q^\star$.

Moreover, most of the results that I will present assume that S and A are finite, and that all states in S are reachable under τ and μ_0.

Formalising the IRL Problem

We assume that the environment consists of a set of states S, a set of actions A, a transition function $\tau : S \times A \to \Delta(S)$, and an initial state distribution $\mu_0 \in \Delta(S)$.

Moreover, we assume that the preferences of the observed agent are described by a reward function $R : S \times A \times S \to \mathbb{R}$, and that its behaviour is described by a policy $S \to \Delta(A)$.

The task of IRL is then to infer R based on π, S, A, τ, and μ_0.

Behavioural Models

In the current IRL literature, the most common behavioural models are:

1. *Optimality*: We assume that π maximises expected cumulative discounted reward (Ng and Russell, 2000).
2. *Boltzmann Rationality*: We assume that $\mathbb{P}(\pi(s) = a) \propto e^{\beta Q^\star(s,a)}$ (Ramachandran and Amir, 2007).
3. *Maximal Causal Entropy*: We assume that π maximises the causal entropy objective; $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t (R(s_t, a_t, s_{t+1}) + \alpha H(\pi(s_{t+1})))]$, where H is the Shannon entropy function (Ziebart, 2010).

Here γ, β, and α are parameters of the behavioural model.

Partial Identifiability

The Issue of Partial Identifiability

There will often be multiple reward functions that are consistent with a given data source, even in the limit of infinite data.

For example, two rewards may induce exactly the same expert behaviour; in that case, no amount of expert demonstrations can distinguish them.

It is therefore important to characterise this fundamental ambiguity for different reward learning data sources, so that they can be compared, and so that their limitations can be understood.

The Issue of Partial Identifiability

Identifying a reward function uniquely is often unnecessary, because all plausible reward functions might lead to the same outcome in a given application.

For example, if we want to learn a reward function in order to compute an optimal policy, then it is enough to learn a reward function that has the same optimal policies as the true reward.

It is therefore important to also consider the *ambiguity tolerance* for various applications (such as policy optimisation).

The Issue of Partial Identifiability

Ambiguity and ambiguity tolerance are formally related: both concern *invariances* of objects that can be computed from reward functions to *transformations* of those reward functions.

This will let us reason about ambiguity of different reward learning data sources, and ambiguity tolerance of different applications, within a single unified framework.

Reward Transformations

Many of our results are expressed in terms of *transformations* on the set of all reward functions, $\mathcal{R}$.

Definition

Following Ng et al. (1999), we say that a *potential function* is a function $\Phi : \mathcal{S} \to \mathbb{R}$. Given a discount γ, we say that $R_2 \in \mathcal{R}$ is produced by *potential shaping* of $R_1 \in \mathcal{R}$ if for some potential Φ,

$$R_2(s, a, s') = R_1(s, a, s') + \gamma \cdot \Phi(s') - \Phi(s).$$

Definition

Given a transition function τ, we say that $R_2 \in \mathcal{R}$ is produced by *S'-redistribution* of $R_1 \in \mathcal{R}$ if

$$\mathbb{E}_{S' \sim \tau(s,a)} \left[R_1(s, a, S') \right] = \mathbb{E}_{S' \sim \tau(s,a)} \left[R_2(s, a, S') \right].$$

Select Results

Theorem

Given an MDP and a temperature parameter β, the Boltzmann-rational policy determines R up to S'-redistribution and potential shaping.

Theorem

Given an MDP and a weight α, the maximal causal entropy policy determines R up to S'-redistribution and potential shaping.

In our paper, we fully characterise the ambiguity of 21 different reward objects and include a number of additional results. For more information, see Skalse et al. (2022a).

Misspecification

Misspecification

Our analysis of partial identifiablity assumes that there is no misspecification. That is, if the learning algorithm assumes that the data is computed from a behavioural model f, then we assume that the data in fact is computed from f.

In reality, this will not hold. For example, a human is not Boltzmann-rational. It is therefore crucial to consider what happens if the modelling assumptions are violated.

Framework

Given a partition P of $\mathcal{R}$, and two behavioural models $f, g : \mathcal{R} \to \Pi$, we say that f is P-robust to misspecification with g if:

1. $f(R_1) = g(R_2) \implies R_1 \equiv_P R_2$.
2. $f(R_1) = f(R_2) \implies R_1 \equiv_P R_2$.
3. $\text{Im}(g) \subseteq \text{Im}(f)$.
4. $f \neq g$.

Core Lemma

Say that $\mathrm{Am}(f) \preceq P$ if $f(R_1) = f(R_2) \implies R_1 \equiv_P R_2$.

Lemma

If $\mathrm{Am}(f) \preceq P$*, and* T *is the set of all reward transformations that preserve* P*, then* f *is* P*-robust to misspecification with* g *if and only if* $g = f \circ t$ *for some* $t \in T$ *where* $f \circ t \neq f$*.*

This means that we can derive necessary and sufficient conditions that completely characterise what forms of misspecification a given behavioural model f will tolerate, by first finding the transformations T that characterise P, and then composing f with each element of T!

Other Lemmas

Lemma

If f is not P-robust to misspecification with g, and $\mathrm{Im}(g) \subseteq \mathrm{Im}(f)$, then for any h, $h \circ f$ is not P-robust to misspecification with $h \circ g$.

Lemma

If f is P-robust to misspecification with g then $\mathrm{Am}(g) \preceq P$.

Lemma

If f is P-robust to misspecification with g and $\mathrm{Im}(f) = \mathrm{Im}(g)$ then g is P-robust to misspecification with f.

Lemma

f satisfies that $\mathrm{Am}(f) \preceq P$ but is not P-robust to any misspecification if and only if $\mathrm{Am}(f) = P$.

Which Equivalence Relation Do We Pick?

As for our analysis of partial identifiablity, we can let the partition P correspond to a particular downstream application. However, in this setting, this leads to a combinatorial explosion of cases to consider.

To make our analysis more tractable, we will therefore focus on two specific equivalence relations. We say that $R_1 \equiv_{\text{ORD}} R_2$ if R_1 and R_2 have the same *ordering of policies*, and that $R_1 \equiv_{\text{OPT}} R_2$ if R_1 and R_2 have the same *optimal policies* (in a given environment).

Characterising Equivalent Reward Functions

Theorem

The MDPs $(\mathcal{S}, \mathcal{A}, \tau, \mu_0, R_1, \gamma)$ *and* $(\mathcal{S}, \mathcal{A}, \tau, \mu_0, R_2, \gamma)$ *have the same ordering of policies if and only if* R_1 *and* R_2 *differ by potential shaping (with* γ*), positive linear scaling, and* S'*-redistribution (with* τ*).*

This tells us exactly when two reward functions have the same ordering of policies, i.e. when $R_1 \equiv_{\text{ORD}} R_2$. There is no simple closed-form expression for $\equiv_{\text{OPT}}$, but we can still use it in analysis.

Select Results

Let Π^+ be the set of all policies such that $\pi(a \mid s) > 0$ for all s, a, and let F be the set of all functions $f : \mathcal{R} \to \Pi^+$ that, given R, returns a policy π which satisfies

$$\text{argmax}_{a \in \mathcal{A}} \pi(a \mid s) = \text{argmax}_{a \in \mathcal{A}} Q^\star(s, a),$$

In other words, $F^{\mathcal{M}}$ is the set of functions that generate policies which take each action with positive probability, and that take the optimal actions with the highest probability. This class is quite large, and includes e.g. Boltzmann-rational policies.

Theorem

Let $f \in F$ be surjective onto Π^+. Then f is OPT*-robust to misspecification with g if and only if $g \in F$ and $g \neq f$.*

Select Results

Let $\psi : \mathcal{R} \to \mathbb{R}^+$ be any function from reward functions to positive real numbers, and let $b_\psi : \mathcal{R} \to \Pi^+$ be the function that, given R, returns the Boltzmann-rational policy with temperature $\psi(R)$. Moreover, let $B = \{b_\psi : \psi \in \mathcal{R} \to \mathbb{R}^+\}$ be the set of all such functions b_ψ.

Theorem

If $b_\psi \in B$ then b_ψ is ORD*-robust to misspecification with g if and only if $g \in B$ and $g \neq b_\psi$.*

This means that the Boltzmann-rational model is ORD-robust to misspecification of the temperature parameter β, but not to any other form of misspecification.

Select Results

In the full paper, we also characterise the misspecification robustness of optimal policies and maximal causal entropy policies. We also present several additional results concerning transfer learning and misspecified environment parameters.

Interestingly, we find that a very wide class of behavioural models (including, but not limited to, all the standard behavioural models) lack robustness to misspecification of the discount parameter γ or transition function τ.

For more details, see Skalse and Abate (2023a).

Quantifying the Difference Between Reward Functions

Why Quantify the Difference Between Reward Functions?

When analysing a reward learning algorithm, we want to obtain results regarding *how far* the learnt reward function is from the underlying true reward function. This requires a way to quantify the difference between reward functions.

So far, we have used *equivalence relations*. This is a limitation, because it will not allow us to distinguish between *small* and *large* errors. To do this, we need *metrics*.

Some Naïve Ideas

A simple method might be to measure their L_2-distance. However, two reward functions can have a large L_2-distance, even if they induce the *same* ordering of policies, or a small L_2-distance, even if they induce the *opposite* ordering of policies.

Another option is to evaluate the learnt reward function on a *test set*. However, this can only guarantee that the learnt reward function is accurate on a given data distribution, and when the reward function is *optimised* we necessarily incur a *distributional shift*.

Yet another option is to optimise the learnt reward function, and evaluate the obtained policy according to the true reward function. However, this is very expensive, and it makes it difficult to separate issues with the policy optimisation algorithm from issues with the reward learning algorithm. Moreover, this method cannot be used for theoretical work.

Defining STARC Metrics

Definition

A function $c : \mathcal{R} \to \mathcal{R}$ is a *canonicalisation function* if c is linear, $c(R)$ and R differ by potential shaping and S'-redistribution, and $c(R_1) = c(R_2)$ if and only if R_1 and R_2 only differ by potential shaping and S'-redistribution.

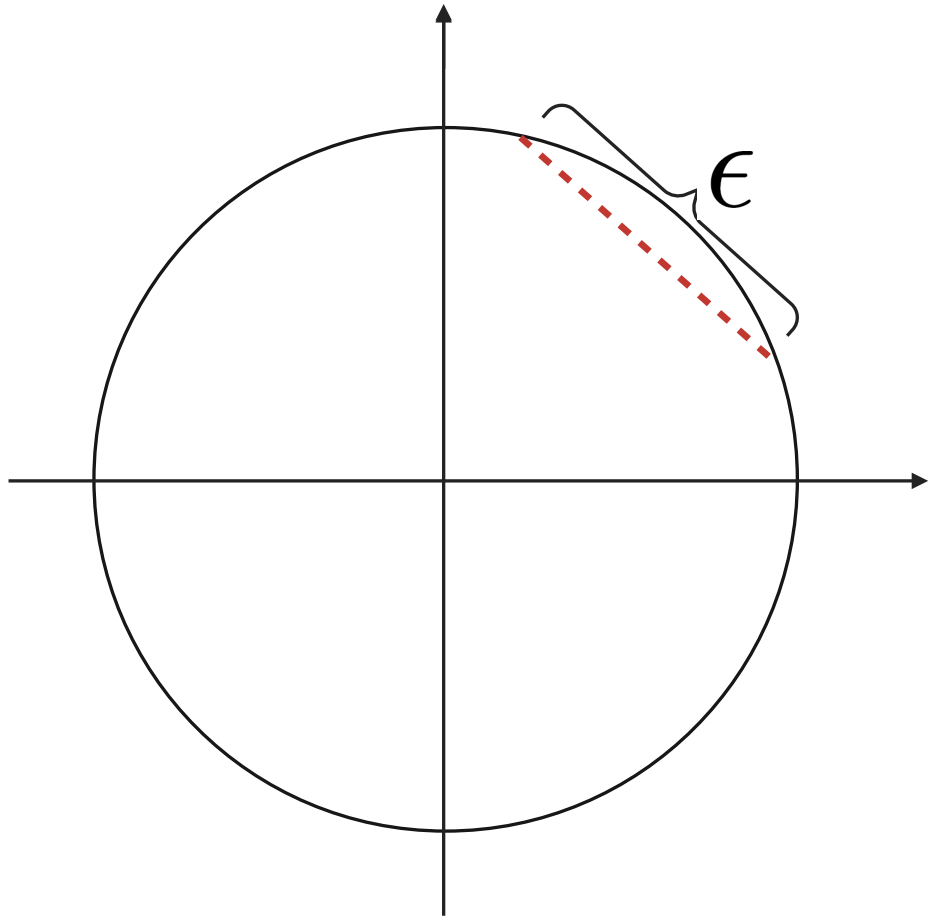
Definition

A metric $m : \mathcal{R} \times \mathcal{R} \to \mathbb{R}$ is *admissible* if there exists a norm p and two (positive) constants u, ℓ such that $\ell \cdot p(x,y) \leq m(x,y) \leq u \cdot p(x,y)$ for all $x, y \in \mathcal{R}$.

Definition

A function $d : \mathcal{R} \times \mathcal{R} \to \mathbb{R}$ is a *STARC metric* (STAndardised Reward Comparison) if there is a canonicalisation function c, a function n that is a norm on $\text{Im}(c)$, and a metric m that is admissible on $\text{Im}(s)$, such that $d(R_1, R_2) = m(s(R_1), s(R_2))$, where $s(R) = c(R)/n(c(R))$ when $n(c(R)) \neq 0$, and $c(R)$ otherwise.

Visualising STARC Metrics I



STARC metrics map $\mathcal{R}$ to a linear subspace, normalise the vectors, and then take a distance with a metric that is similar to a norm.

The Occupancy Measure

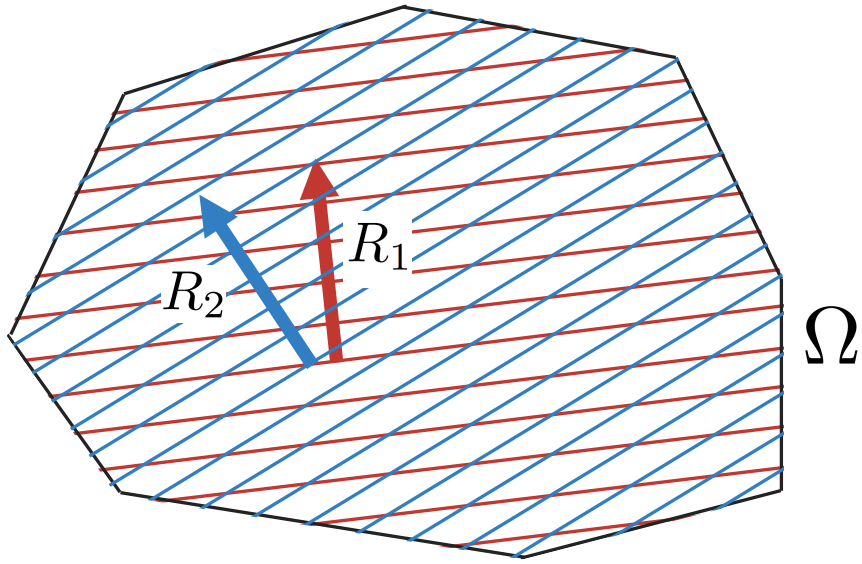
Given an MDP and a policy π, we can define the *occupancy measure* η^π of π as an $|S| \times |A| \times |S|$-dimensional vector where

$$\eta^\pi[s, a, s'] = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s, a_t = a, s_{t+1} = s').$$

We now have that $J(\pi) = \eta^\pi \cdot R$. This means that J can be decomposed into two steps, the first of which is independent of R, and the second of which is linear.

The set Ω of the occupancy measures for all policies forms a convex polytope in an $|S|(|A| - 1)$-dimensional affine subspace.

Visualising STARC Metrics II



The STARC distance between two reward functions R_1, R_2 is bilipschitz equivalent to the angle between R_1 and R_2 after they have been projected onto the space of all occupancy measures Ω.

Examples of a Canonicalisation Functions

Proposition

For any policy π, the function $c : \mathcal{R} \to \mathcal{R}$ given by

$$c(R)(s, a, s') = \mathbb{E}_{S' \sim \tau(s,a)} \left[R(s, a, S') - V^{\pi}(s) + \gamma V^{\pi}(S') \right]$$

is a canonicalisation function.

Proposition

A canonicalisation function $c : \mathcal{R} \to \mathcal{R}$ is minimal for a norm n if for all R we have that $n(c(R)) \leq n(R')$ for all R' such that R and R' differ by potential shaping and S'-redistribution. For any weighted L_2-norm, a minimal canonicalisation function exists and is unique.

Proposition

The function $c : \mathcal{R} \to \mathcal{R}$ which projects reward functions onto the space of occupancy measures Ω is a canonicalisation function.

Basic Properties

Proposition

All STARC metrics are pseudometrics on $\mathcal{R}$.

Proposition

All STARC metrics have the property that $d(R_1, R_2) = 0$ if and only if R_1 and R_2 induce the same ordering of policies.

Soundness

Definition

A pseudometric d on $\mathcal{R}$ is *sound* if there exists a positive constant U, such that for any reward functions R_1 and R_2, if two policies π_1 and π_2 satisfy that $J_2(\pi_2) \geq J_2(\pi_1)$, then

$$J_1(\pi_1) - J_1(\pi_2) \leq U \cdot (\max_{\pi} J_1(\pi) - \min_{\pi} J_1(\pi)) \cdot d(R_1, R_2).$$

Theorem

All STARC metrics are sound.

Completeness

Definition

A pseudometric d on $\mathcal{R}$ is *complete* if there exists a positive constant L, such that for any reward functions R_1 and R_2, there exists two policies π_1 and π_2 such that $J_2(\pi_2) \geq J_2(\pi_1)$ and

$$J_1(\pi_1) - J_1(\pi_2) \geq L \cdot (\max_{\pi} J_1(\pi) - \min_{\pi} J_1(\pi)) \cdot d(R_1, R_2),$$

and if $\max_\pi J_1(\pi) - \min_\pi J_1(\pi) = \max_\pi J_2(\pi) - \min_\pi J_2(\pi) = 0$, then we have that $d(R_1, R_2) = 0$.

Theorem

All STARC metrics are complete.

Uniqueness

Proposition

Any two pseudometrics on $\mathcal{R}$ that are both sound and complete are bilipschitz equivalent.

This means that there is a sense in which STARC metrics are unique.

For more details, see Skalse et al. (2024).

A Few Short Results

Reward Hacking

Say that two reward functions R_1 and R_2 are *hackable* if there exists policies π_1, π_2 such that $J_1(\pi_1) > J_1(\pi_2)$ but $J_2(\pi_1) < J_2(\pi_2)$. Otherwise, they are *unhackable*.

Two reward functions R_1, R_2 are *equivalent* if $J_1(\pi_1) \geq J_1(\pi_2)$ iff $J_2(\pi_1) \geq J_2(\pi_2)$ for all π_1, π_2. A reward function R is *trivial* if $J(\pi_1) = J(\pi_2)$ for all π_1, π_2.

Theorem

Two reward functions R_1, R_2 are unhackable if and only if they are equivalent, or at least one of them is trivial.

For more details, see Skalse et al. (2022b).

Risk-Sensitivity

Theorem

Let R_1, R_2 be two reward functions such that

$$G_1(\xi_1) \leq G_1(\xi_2) \iff G_2(\xi_1) \leq G_2(\xi_2)$$

for all $\xi_1, \xi_2 \in (S \times A)^\omega$. Then is an $a > 0$ and a b such that

$$G_2(\xi) = a \cdot G_1(\xi) + b$$

for all $\in (S \times A)^\omega$.

For more details, see Skalse and Abate (2023b).

Multi-Objective Problems

A *MOMDP* is an MDP with more than one reward function. A *MORL objective* O is a function that takes k policy evaluation functions $J_1 \dots J_k$ and returns a total ordering $\prec_O$ over Π.

A MOMDP $M_1 = (S, A, T, \{R_1 \dots R_k\}, \gamma)$ with objective O is equivalent to an MDP $M_2 = (S, A, T, R, \gamma)$ if and only if M_1's policy order is $\prec_{O,M}$. In this case, we say that M_1 with O is *scalarized* by R.

Theorem

If a MOMDP M with objective O is scalarizable, then there exist $w_1 \dots w_k \in \mathbb{R}$ such that M with O is scalarized by the reward $R(s,a) = \sum_{i=1}^{k} w_i \cdot R_i(s,a)$.

This implies that many MORL objectives are not scalarizable! For more details, see Skalse and Abate (2023b).

Thank you!

joar.mvs@gmail.com

References

Ng, A. Y., Harada, D., and Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 278–287, Bled, Slovenia. Morgan Kaufmann Publishers Inc.

Ng, A. Y. and Russell, S. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, volume 1, pages 663–670, Stanford, California, USA. Morgan Kaufmann Publishers Inc.

Ramachandran, D. and Amir, E. (2007). Bayesian inverse reinforcement learning. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 2586–2591, Hyderabad, India. Morgan Kaufmann Publishers Inc.

Skalse, J. and Abate, A. (2023a). Misspecification in inverse reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):15136–15143.

Skalse, J. and Abate, A. (2023b). On the limitations of Markovian rewards to express multi-objective, risk-sensitive, and modal tasks. In Evans, R. J. and Shpitser, I., editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 1974–1984. PMLR.

Skalse, J., Farnik, L., Motwani, S. R., Jenner, E., Gleave, A., and Abate, A. (2024). Starc: A general framework for quantifying differences between reward functions.

Skalse, J., Farrugia-Roberts, M., Russell, S., Abate, A., and Gleave, A. (2022a). Invariance in policy optimisation and partial identifiability in reward learning. *arXiv preprint arXiv:2203.07475*.

Skalse, J., Howe, N., Dima, K., and Krueger, D. (2022b). Defining and characterizing reward hacking. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*.

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