

Predicting optimal value functions by interpolating reward functions in scalarized multi-objective reinforcement learning

Arpan Kusari¹ and Jonathan P. How²

Abstract—A common approach for defining a reward function for multi-objective reinforcement learning (MORL) problems is the weighted sum of the multiple objectives. The weights are then treated as design parameters dependent on the expertise (and preference) of the person performing the learning, with the typical result that a new solution is required for any change in these settings. This paper investigates the relationship between the reward function and the optimal value function for MORL; specifically addressing the question of how to approximate the optimal value function well beyond the set of weights for which the optimization problem was actually solved, thereby avoiding the need to recompute for any particular choice. We prove that the value function transforms smoothly given a transformation of weights of the reward function (and thus a smooth interpolation in the policy space). A Gaussian process is used to obtain a smooth interpolation over the reward function weights of the optimal value function for three well-known examples: Gridworld, Objectworld and Pendulum. The results show that the interpolation can provide robust values for sample states and actions in both discrete and continuous domain problems. Significant advantages arise from utilizing this interpolation technique in the domain of autonomous vehicles: easy, instant adaptation of user preferences while driving and true randomization of obstacle vehicle behavior preferences during training.

I. INTRODUCTION

Reinforcement learning (RL) is a machine learning technique that provides the basis for decision-making, where a reward provided by the environment leads the agent to behave in a manner so as to maximize the cumulative sum of rewards. The reward function of RL problems often requires optimization of multiple, often conflicting objectives [1]. For example, in the domain of autonomous vehicles, driving preferences have to be balanced between time to goal, comfort and safety [2], which are correlated and its unclear how they influence each other. These conflicting objectives do not yield a single optimal solution, but rather a set of trade-off solutions which balance the objectives [3]. The easiest way to solve the multi-objective problem is to use a linear scalarization function [4] that transforms the given problem into a standard single-objective using a weighted sum of the parameters.

Sutton’s reward hypothesis states “that all of what we mean by goals and purposes can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (reward)”. Thus, the inference

being that any given multi-objective problem can always be transformed into a single objective reward function. The most obvious problem in this case is that the weights used during training are a design parameter and dependent on the preference of the person designing the RL problem. Thus, the trained RL has a set optimal policy (and optimal value function) which is dependent on the weights provided. Having a fixed set of weights can be detrimental to the possibility of adaptation to different user experiences whereby for every instance of change of weights, the process of training (which is tedious and time intensive) needs to be repeated.

A question which arises is: Given a small sparse group of optimal value functions under variable reward functions given by different weights, is it possible to interpolate through the entire space of the reward functions to provide exact estimates of optimal value functions at all possible states and actions?

To the best of our understanding, prior research works focusing on value function interpolation have been used to show convergence of RL algorithms for countable and uncountable spaces. Ref. [5] proposed multilinear interpolation techniques on coarse grid to solve various RL paradigms. Ref. [6] provided convergence of RL algorithms combined with value function interpolation while providing convergence of Q-learning for uncountable spaces. Although it is fairly obvious that changing the reward function would effect the value function directly, we have not found any research work which investigates the relationship and predicts it for weights not previously seen during training.

The majority of MORL approaches consist of single-policy algorithms in order to learn Pareto optimal solutions [7]. Ref. [8] provides a modification of RL to learn all the optimal policies for all linear preference assignments by incorporating the convex hull of the value function. Ref. [9] uses Monte-Carlo Tree Search (MCTS) along with multi-objective indicator by the way of a hypervolume indicator to define action-selection criterion. Ref. [3], which uses multi-objective optimization techniques within a RL framework, creates a multi-policy algorithm that learns a set of Pareto dominating policies in a single run of the algorithm which they call Pareto Q-learning. While our proposed approach is useful for MORL problems, we do not aim to create a different MORL approach in this paper. Rather our research formulation is different than the existing MORL approaches in that we seek to derive value functions at unseen reward weights (in the training phase) from the neighboring interpolations.

The aim of this research is to interpolate through the space

*This work was supported by Ford Motor Company

¹Arpan Kusari is with Research and Advanced Engineering, Ford Motor Company, Dearborn, MI 48124, USA akusari@ford.com

²Jonathan P. How is with the MIT Department of Aeronautics and Astronautics, Cambridge, MA 02139, USA jhow@mit.edu

of the value functions as a result of changing the weights of the reward function using a Gaussian Process (GP). The change in weights may be non-uniform, which makes the process highly nonlinear. Thus, it becomes a supervised learning problem where with the increase in the number of objectives, the weight space increases and data points becomes extremely sparse. Finding accurate value function values across problem space would be extremely beneficial for machine learning in general and autonomous vehicles in particular. GPs provide flexible function approximators, capable of learning intricate structure through their covariance kernels [10]. Utilizing the predictive power of GPs to interpolate through the high-dimensional input space should yield accurate value functions at all points of the large state space.

This paper is organized as follows: Section II gives a background of RL and GP, Section III provides the claim along with the mathematical reasoning, Section IV gives the results of the methodology on various standard RL examples, and Section V gives the discussions and conclusions.

II. BACKGROUND

A. Reinforcement learning

In the RL task, at time t , the agent observes a state, $s_t \in \mathcal{S}$, which represents the environmental model of the system. It takes an action, $a_t \in \mathcal{A}$. The agent receives an immediate scalar reward r_t and moves to a new state s_{t+1} . The environment's dynamics are characterized by state transition probabilities $p(s_{t+1}|s_t, a_t)$. This can be formally stated as a Markov Decision Process (MDP) where the next state can be completely defined by the previous state and action (Markov property) and receive a scalar reward for executing the action [11].

The goal of the agent is to maximize the cumulative reward (discounted sum of rewards) or value function:

$$V_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k} \quad (1)$$

where $0 \leq \gamma \leq 1$ is the discount factor and r_t is the reward at time-step t . In terms of a policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$, the value function can be given by Bellman equation as:

$$V_{\pi}(s_t) = \mathbb{E}_{\pi} \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} | \mathcal{S} = s_t \right] \quad (2)$$

$$= \mathbb{E}_{\pi} \left[R_t + \sum_{k=1}^{\infty} \gamma^k R_{t+k} | \mathcal{S} = s_t \right] \quad (3)$$

$$= \sum_a \pi(a_t | s_t) \sum_{s_{t+1}} p(s_{t+1} | s_t, a_t) \times \left[R(s_t, a_t, s_{t+1}) + \gamma \mathbb{E}_{\pi} \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} | \mathcal{S} = s_{t+1} \right] \right] \quad (4)$$

$$= \sum_a \pi(a_t | s_t) \sum_{s_{t+1}} p(s_{t+1} | s_t, a_t) \times \left[R(s_t, a_t, s_{t+1}) + \gamma V_{\pi}(s_{t+1}) \right]. \quad (5)$$

Using Bellman's optimality equation, we can define a policy π which is greater than or equal to any other policy π' , if value function $V_{\pi}(s_t) \geq V_{\pi'}(s_t)$ for all $s_t \in \mathcal{S}$. This policy is known as an optimal policy (π^*) and its value function is known as optimal value function (V^*).

For continuous state space problems, such as arising in control of nonlinear dynamical systems, a common approach to solve the problem is using value function approach [12]. Value-function approach estimates a value function for each action and chooses the "greedy" policy (policy having highest value function) at each time-step. Thus, the value function is updated until it converges to the optimal value function.

B. Gaussian process regression

A stochastic process is a collection of random variables of functions, $\{f(x) : x \in \mathcal{X}\}$, where the variables are collected from a set \mathcal{X} . A GP is a special form of stochastic process, where any finite subset of the random variables has a multivariate Gaussian distribution [13]. In particular, a collection of random variables $\{f(x) : x \in \mathcal{X}\}$ is said to be drawn from a GP with mean function $m(\cdot)$ and covariance function $k(\cdot, \cdot)$, if for any finite set of elements $\{x_1, \dots, x_n\} \in \mathcal{X}$, the associated finite set of random variables $\{f(x_1), \dots, f(x_n)\}$ have distribution,

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix} \right) \quad (6)$$

and the resulting GP is then denoted as

$$f(\cdot) \sim \mathcal{GP}(m(\cdot), k(\cdot, \cdot)).$$

While any real-valued function is suitable for mean function $m(\cdot)$, the kernel function $k(\cdot, \cdot)$ needs to guarantee positive-semidefiniteness.

Let $P = \{(x(i), y(i))\}_{i=1}^n$ be a training set of i.i.d. examples from some unknown distribution. In the Gaussian process regression model,

$$y(i) = f(x(i)) + \varepsilon(i), i = \{1, \dots, n\} \quad (7)$$

where the $\varepsilon(i)$ are i.i.d. "noise" variables with independent $\mathcal{N}(0, \sigma^2)$ distributions. We assume a zero-mean Gaussian process prior, $f(\cdot) \sim \mathcal{GP}(0, k(\cdot, \cdot))$ with a covariance function $k(\cdot, \cdot)$. The marginal distribution over any set of input points belonging to \mathcal{X} must have a joint multivariate Gaussian distribution. Therefore, for testing points $Q = \{x^*(i), y^*(i)\}$, the marginal distribution is given as

$$\begin{bmatrix} \vec{f} \\ \vec{f}^* \end{bmatrix} \Big|_{X, X^*} \sim \mathcal{N} \left(0, \begin{bmatrix} K(X, X) & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) \end{bmatrix} \right) \quad (8)$$

where X is the matrix formulation of the training input vector, X^* is the matrix formulation of the test input vector and \vec{f}^* is the compactly written vector formulation of $f(x^*)$. The outputs can therefore be written as:

$$\begin{bmatrix} \vec{y} \\ \vec{y}^* \end{bmatrix} \Big|_{X, X^*} = \begin{bmatrix} \vec{f} \\ \vec{f}^* \end{bmatrix} + \begin{bmatrix} \vec{\varepsilon} \\ \vec{\varepsilon}^* \end{bmatrix}$$

$$\sim \mathcal{N}\left(0, \begin{bmatrix} K(X, X) + \sigma^2 I & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) + \sigma^2 I \end{bmatrix}\right) \quad (9)$$

where $\varepsilon^*(i)$ are i.i.d. ‘‘noise’’ variables with independent $\mathcal{N}(0, \sigma^2)$ distributions. We derive the test outputs from Equation 9 as:

$$\vec{y}^* | \vec{y}, X, X^* \sim \mathcal{N}(\mu^*, \Sigma^*) \quad (10)$$

where

$$\mu^* = K(X^*, X)(K(X, X) + \sigma^2 I)^{-1} \vec{y}$$

and

$$\Sigma^* = K(X^*, X^*) + \sigma^2 I - K(X^*, X)(K(X, X) + \sigma^2 I)^{-1} K(X, X^*)$$

III. METHODOLOGY

A. Value function interpolation

In this section, we focus on providing mathematical justifications for the interpolation of value function based on the weights of the objectives of reward function.

For initial analysis, we wish to prove that given a simple, linear transformation of weights, the value function can be interpolated in an accurate manner. Intuitively, we are trying to derive the intermediate optimal value function giving the optimal policy for some MDP, where the reward is the weighted combination of various different objectives.

Theorem 1: For a reward function R composed of n different objectives, each associated with weight w_i , with the full set given by $\mathbf{w} = \{w_1, w_2, \dots, w_n\}$, such that for a given state $s_t \in S$ and a given action $a_t \in A$, the reward function is

$$R(s_t, a_t) = w_1 r_1(s_t, a_t) + w_2 r_2(s_t, a_t) + \dots + w_n r_n(s_t, a_t). \quad (11)$$

where r_1, r_2, \dots, r_n are normalized reward functions at a given state s_t and action a_t , respectively, the gradient of the state-value function with respect to the weights exists, if all the rewards at the current state and action are finite.

Proof: The optimal value at a state is given by the state-value function

$$V^*(s_t) = \max_{\pi} \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \right], \quad (12)$$

where $R(s_t) = \max_{\pi} R(s_t, a_t)$. Given a particular set of weights, we substitute (11) into (12) to obtain

$$V^*(s_t | \mathbf{w}) = \max_{\pi} \mathbb{E} \left[\sum_{t=0}^{\infty} \sum_{a_t \in A} \gamma^t (w_1 r_1(s_t, a_t) + \dots + w_n r_n(s_t, a_t)) \right]. \quad (13)$$

However, note that for a different set of weights $\mathbf{w}' = w'_1, \dots, w'_n$, the optimal state-value function is

$$V^*(s_t | \mathbf{w}') = \max_{\pi} \mathbb{E} \left[\sum_{t=0}^{\infty} \sum_{a_t \in A} \gamma^t (w'_1 r_1(s_t, a_t) + \dots + w'_n r_n(s_t, a_t)) \right]. \quad (14)$$

Subtracting (13) from (14) yields

$$\begin{aligned} \Delta V^*(s_t | \mathbf{w}, \mathbf{w}') &= \\ & \max_{\pi} \mathbb{E} \left[\sum_{t=0}^{\infty} \sum_{a_t \in A} \gamma^t (w'_1 r_1(s_t, a_t) + \dots + w'_n r_n(s_t, a_t)) \right] - \\ & \max_{\pi} \mathbb{E} \left[\sum_{t=0}^{\infty} \sum_{a_t \in A} \gamma^t (w_1 r_1(s_t, a_t) + \dots + w_n r_n(s_t, a_t)) \right] \end{aligned} \quad (15)$$

Using the property $\max(b) - \max(a) \leq \max(b - a)$ yields

$$\begin{aligned} \Delta V^*(s_t | \mathbf{w}, \mathbf{w}') &\leq \max_{\pi} \mathbb{E} \left[\sum_{t=0}^{\infty} \sum_{a_t \in A} \gamma^t ((w'_1 - w_1) r_1(s_t, a_t) + \dots \right. \\ & \quad \left. + (w'_n - w_n) r_n(s_t, a_t)) \right] \end{aligned} \quad (16)$$

Equation (16) can be written in matrix form as

$$\Delta V^*(s_t | \mathbf{w}, \mathbf{w}') \leq \max_{\pi} \mathbb{E} \left[\sum_{t=0}^{\infty} \sum_{a_t \in A} \gamma^t \cdot \Delta \mathbf{w} \cdot \mathbf{r} \right] \quad (17)$$

where $\Delta \mathbf{w} = \mathbf{w}' - \mathbf{w}$ and

$$\mathbf{r} = \begin{bmatrix} r_1(s_t, a_t) \\ \vdots \\ r_n(s_t, a_t) \end{bmatrix}^T$$

Since, $\Delta \mathbf{w}$ is constant for all states and actions, (17) can be rearranged as

$$\frac{\Delta V^*(s_t | \mathbf{w}, \mathbf{w}')}{\Delta w_i} \leq \max_{\pi} \mathbb{E} \left[\sum_{t=0}^{\infty} \sum_{a_t \in A} \gamma^t \cdot \mathbf{1} \cdot \mathbf{r} \right] \quad (18)$$

which gives the approximate gradient of the value function with respect to the i^{th} weight. If all the rewards at the current state and action are finite, then the gradient will exist for that given state of the MDP. Thus, the linear interpolation of weights in reward function leads to smooth interpolation of state-value function. ■

Corollary 1.1: Under linear transformation of weights in reward function, the gradient of the action-value function with respect to the weights exists, if all the rewards at the current state and action are finite.

Proof: For a optimal state-value function $V^*(s)$ that gives the best value at that particular state, the optimal action-value function (optimal value of a state and action combination) is

$$Q^*(s_t, a_t) = R(s_t, a_t) + \gamma \sum_{t=0}^{\infty} p(s_{t+1} | s_t, a_t) V^*(s_{t+1}). \quad (19)$$

Given two different set of weights, the difference in q-value functions can be written as:

$$\Delta Q^*(s_t, a_t) = \Delta \mathbf{w} \cdot \mathbf{r} + \gamma \sum_{t=0}^{\infty} p(s_{t+1} | s_t, a_t) \Delta V^*(s_{t+1}) \quad (20)$$

Replacing ΔV^* from (17) gives

$$\begin{aligned} \Delta Q^*(s_t, a_t) &\leq \Delta \mathbf{w} \cdot \mathbf{r} + \gamma \sum_{t=0}^{\infty} p(s_{t+1} | s_t, a_t) \\ & \quad \max_{\pi} \mathbb{E} \left[\sum_{t=0}^{\infty} \sum_{a_t \in A} \gamma^t \Delta \mathbf{w} \cdot \mathbf{r} \right]. \end{aligned} \quad (21)$$

Therefore the gradient of Q^* with respect to the i^{th} weight is given as:

$$\frac{\Delta Q(s_t, a_t)}{\Delta w_i} \leq \mathbf{1} \cdot \mathbf{r} + \gamma \sum_{t=0}^{\infty} p(s_{t+1} | s_t, a_t) \max_{\pi} \mathbb{E} \left[\sum_{t=0}^{\infty} \sum_{a_t \in A} \gamma^t \cdot \mathbf{1} \cdot \mathbf{r} \right] \quad (22)$$

The shaped reward function is a specific case of the MORL reward function, whereby, the reward function is augmented using an indicator function in which a positive reward is given if the next state is closer to the goal and can be represented as

$$R'(s_t, a_t, s_{t+1}) = R(s_t, a_t, s_{t+1}) + r \cdot \mathbf{I}(d(s_{t+1}, G) < d(s_t, G)) \quad (23)$$

where G is the goal state. Assuming that the goal state is constant across the different weights of the reward function, the added shaped reward remains constant for the given state across weights. Thus, the reward shaping does not pose any problems for interpolating reward functions.

IV. RESULTS

We use three different example tasks with various degrees of complexity to test the validity of our approach. These example tasks have multiple objectives which need to be optimized simultaneously using the RL framework. We change the weights of these objectives and intend to predict the resulting value function using GP regression.

A. Gridworld

The gridworld [14] is a discrete $N \times N$ grid with four actions per state (corresponding to steps in each direction) and each action has a 10% chance of being randomly changed to a different action. If the agent hits a wall, then it stays in the previous state. The goal states correspond to a large terminal reward and there is a negative living reward for each of the other states, which incentivizes the agent to reach the goal as fast as possible. There is a walled state in the (2,2) position. The default terminal rewards are $p = +1$ and $n = -1$ and the default living reward is $l = -0.02$.

The GP regression from Scikit in Python [15] is used to determine the interpolated value function, where the input vector X corresponds to a state vector augmented with the discrete action and the weights of the reward function, and the scalar target y corresponds to the value function. The Matern kernel is utilized for training the GP with default parameters in all the cases. We used other kernels, but we did not find sufficient difference between the kernel choices.

We vary the living reward and terminal rewards for the different experiments. Two kinds of metrics are reported: the mean squared error between the actual value function and predicted value function over all states and all actions and the median value of the standard deviation at the query points. Both interpolated and extrapolated query points are reported, which are presented as representative samples.

TABLE I
PREDICTING VALUE FUNCTIONS FOR LIVING REWARDS

Living reward (l)	Mean squared error	Median sigma
-0.16	1.019e-04	1.732e-03
-0.23	1.529e-05	1.496e-02
-0.37	3.401e-05	1.764e-02
-0.45	9.273e-04	5.111e-02
-0.60	4.999e-02	2.382e-01

1) *Changing the living reward:* We vary the living reward (l) of all states (except the terminal states) to vary the optimal policy (and by virtue the optimal state value function) in a way that the variability is nonlinear. The training is performed by varying the living reward from $l = 0$ to $l = -0.5$ by steps of -0.1 . In the Table I, the results are presented for four different interpolated evaluation living rewards as well as an extrapolated evaluation living reward. The interpolation results are shown to be accurate to the fourth decimal place while the extrapolation is within a feasible error bound. To understand the effect of variability of living reward on the optimal state-value function and the subsequent optimal policy, Figure 1(a) and 1(b) shows the optimal state-value function and policy for extreme living rewards $l = 0$ and $l = -0.5$, respectively. It is clear from the optimal policies that a change in living reward alters the solution to the gridworld problem sufficiently and there is a need to capture the variability in the living reward. Figure 1(c) gives the optimal state-value functions at the neighboring points ($l = -0.2$ and $l = -0.3$) of the example living reward $l = -0.23$, which shows substantial variability in the optimal state-value functions. Figure 1(d) gives the predicted and actual optimal state-value function values at the chosen living reward ($l = -0.23$), which shows that the results are accurate to the third decimal place for all states; thus proving the accuracy of the interpolation for the entire state space.

2) *Changing the negative terminal reward:* The negative terminal reward (n) is varied from $n = -1$ to $n = -5$ with steps of -0.5 . The evaluations are given in the Table II. Again, both interpolation (first four rows) and extrapolation (last row) evaluation cases were considered. Note that with an increase in magnitude of the negative terminal reward, the value function in the other states is not influenced (due to the

TABLE II
PREDICTING VALUE FUNCTIONS FOR NEGATIVE TERMINAL REWARDS

Negative reward (n)	Mean squared error	Median sigma
-1.3	4.098e-03	4.246e-03
-2.2	2.678e-07	8.535e-04
-3.6	3.099e-10	1.282e-04
-4.7	8.290e-08	2.582e-04
-6.0	7.025e-06	2.304e-03

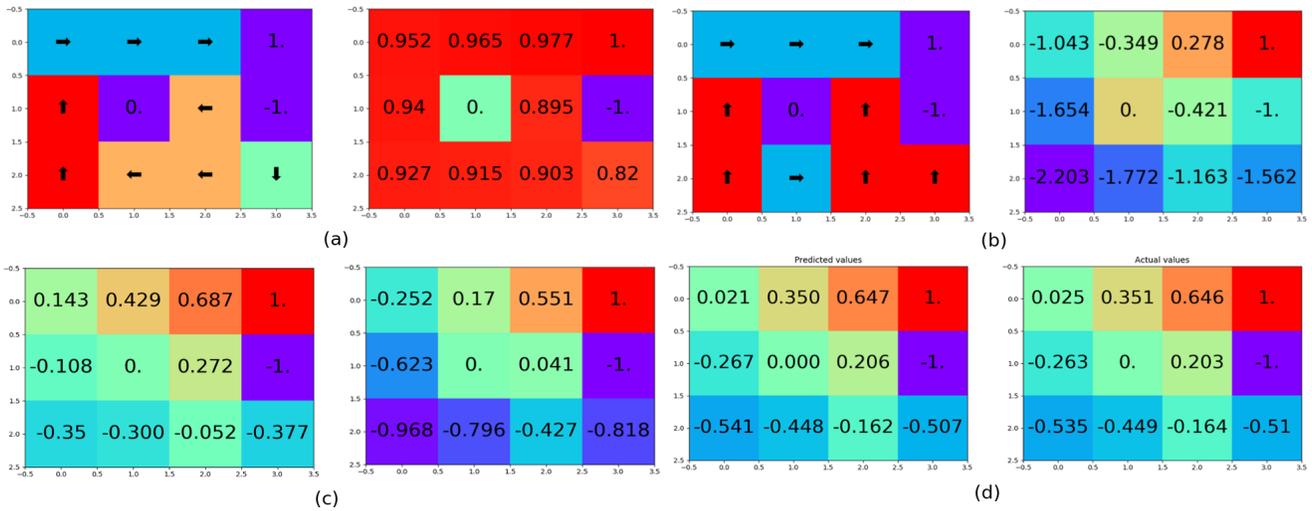


Fig. 1. (a) Optimal policy and optimal value function for living reward (0) and (b) optimal policy and value function for living reward -0.5 . (c) For the interpolation of living reward (-0.23), we show the optimal value functions for two neighboring points with living reward (-0.2) and living reward (-0.3). (d) Predicted and actual optimal function values for living reward (-0.23).

TABLE III

PREDICTING VALUE FUNCTIONS FOR POSITIVE TERMINAL REWARDS

Positive reward (p)	Mean squared error	Median sigma
1.3	3.876e-05	1.193e-03
2.2	6.419e-08	4.502e-04
3.6	2.372e-10	9.059e-04
4.7	6.789e-08	7.603e-04
6.0	1.596e-05	1.053e-03

max operator) and thus the results only reflect the difference in the negative terminal state.

3) *Changing the positive terminal reward*: Table III shows the result when the positive terminal reward (p) is changed from $p = 1$ to $p = 5$ with steps of 0.5 and evaluated at the same random points (positive in this case) as in Table II. The results clearly show that, in both interpolation and extrapolation, the GP is able to recover the value functions.

B. Objectworld

Objectworld [16] is an extension of gridworld that features random objects placed in the grid (Figure 2(a)). The objects are assigned a random outer and inner color (out of C colors) with the state vector being composed of the Euclidean distance to the nearest object with a specific inner or outer color. The true reward is positive in states that are both within 3 cells of outer color 1 and 2 cells of outer color 2, negative within 3 cells of outer color 1, and zero otherwise. Inner colors and all other outer colors are distractors. In the given example, we use two colors, blue and red. Fifteen different objects are placed randomly within the 10×10 grid with randomly chosen inner and outer color. The positive reward is varied from 0.5 to 1 with 0.6, 0.7 and 0.8 points being predicted.

The formulation for GP regression is similar to the ones

TABLE IV

PREDICTING VALUE FUNCTIONS FOR REWARDS IN OBJECTWORLD

Reward	Mean squared error	Median sigma
0.6	7.594e-02	7.673e-03
0.7	4.571e-02	9.608e-03
0.8	2.415e-02	1.526e-03

used in Gridworld. Figure 2(b) shows the actual value function while Figure 2(c) provides the predicted value function. Table IV provides the statistics for the given prediction. The interpolation is not accurate as in gridworld due to the nonlinearity of the reward with respect to the states, but the GP can still recover values close to the actual values, especially in the positive reward region.

C. Pendulum

The pendulum environment [17] is an well-known problem in the control literature in which a pendulum starts from a random orientation and the goal is to keep it upright while applying the minimum amount of force. The state vector is composed of the cosine (and sine) of the angle of the pendulum, and the derivative of the angle. The action is the joint effort as 5 discrete actions linearly spaced within the $[-2, 2]$ range. The reward is

$$R = -(w_1 \cdot \|\theta\|^2 + w_2 \cdot \|\dot{\theta}\|^2 + w_3 \cdot \|a\|^2), \quad (24)$$

where w_1 , w_2 and w_3 are the reward weights for the angle θ , derivative of angle $\dot{\theta}$ and action a respectively. The optimal reward weights given by OpenAI are $[1, 0.1, 0.001]$ respectively. An episode is limited to 1000 timesteps.

A deep Q-network (DQN) was proposed in [18] that combines deep neural networks with RL to solve continuous state discrete action problems. DQN uses a neural network that gives the Q-values for every action and uses a buffer to

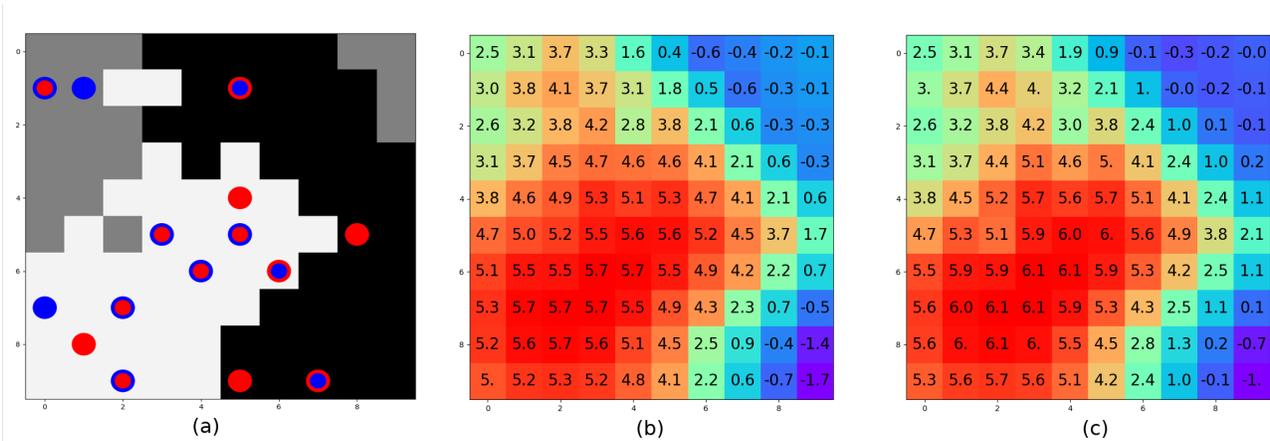


Fig. 2. (a) Objectworld with 15 randomly placed objects in blue and red inner and outer colors chosen randomly; white represents positive reward, black negative reward and grey zero reward (b) Actual value function for positive reward (0.8) (c) Predicted value function

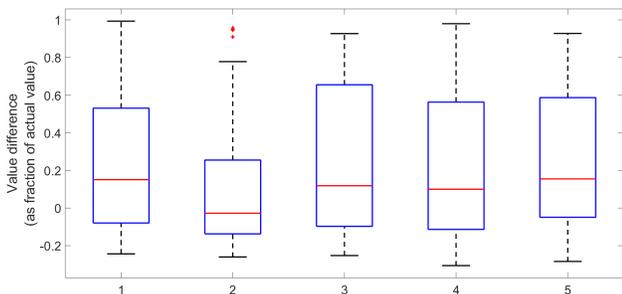


Fig. 3. Boxplots for 5 example episodes showing the difference between the predicted and actual values derived for the weight $w_3 = 0.001$

store old states and actions to sample from to help stabilize training. The pendulum environment is solved using the DQN approach for various $w_3 = \{0.1, 0.01, 0.001, 0.0001\}$ with the evaluation performed at $w_3 = 0.001$. Since this is a continuous state problem, we utilize the trained evaluation model to transition to the next state.

Utilizing a DQN provides no guarantees that the states seen during testing have been visited during training, which can lead to out-of-distribution states. Thus, we have to utilize a robust Student-t likelihood using the GP regression in GPFlow package [19]. The boxplots for the difference in values for 5 sample evaluation episodes are provided in Figure 3. Thus, we use these boxplots to show the value difference as a fraction of the actual value at that state and action. The boxplots show that the GP is able to recover a value close to the actual value (with zero being no difference and greater than 1 meaning that the predicted value is not able to recover the actual value at all) for the majority of the episodes for continuous state domain problems.

V. CONCLUSIONS

This paper shows a direct relationship between the weights of the reward function and the optimal value function for scalarized MORL. This helped us in interpolating through a space of optimal value functions generated using the sparse set of reward functions to estimate the value functions at

sample states. The specific example problems were chosen to understand the value function hypersurface as a function of the reward function. Using GP to interpolate between value functions help us to benefit from prior work in GP regression. Utilizing this relationship would be very beneficial in high-dimensional problems where the instant adaptation of optimal value functions (and thus optimal policies) would save time and cost required for retraining.

The scalarization approach of MORL is restrictive in that it cannot work with objectives where Pareto fronts are non-convex or have discontinuities [20]. MORL is an area of active research that uses algorithms leveraged from the multi-objective optimization literature. However, our paper deals with problems which have a defined convex Pareto front and provides a very simple technique in determining optimal value functions at different weights.

Future work will focus on developing transfer learning of specific behaviors in multi-agent environments with different reward functions based on different weights.

REFERENCES

- [1] D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley, "A survey of multi-objective sequential decision-making," *Journal of Artificial Intelligence Research*, vol. 48, pp. 67–113, 2013.
- [2] G. Prokop, "Modeling human vehicle driving by model predictive online optimization," *Vehicle System Dynamics*, vol. 35, no. 1, pp. 19–53, 2001.
- [3] K. Van Moffaert and A. Nowé, "Multi-objective reinforcement learning using sets of pareto dominating policies," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3483–3512, 2014.
- [4] C.-L. Hwang and A. S. M. Masud, *Multiple objective decision making methods and applications: a state-of-the-art survey*. Springer Science & Business Media, 2012, vol. 164.
- [5] S. Davies, "Multidimensional triangulation and interpolation for reinforcement learning," in *Advances in Neural Information Processing Systems*, 1997, pp. 1005–1011.
- [6] C. Szepesvári, "Convergent reinforcement learning with value function interpolation," Technical Report TR-2001-02, Mindmaker Ltd., Budapest 1121, Konkoly Th. M. u , Tech. Rep., 2001.
- [7] P. Mannion, S. Devlin, K. Mason, J. Duggan, and E. Howley, "Policy invariance under reward transformations for multi-objective reinforcement learning," *Neurocomputing*, vol. 263, pp. 60–73, 2017.
- [8] L. Barrett and S. Narayanan, "Learning all optimal policies with multiple criteria," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 41–47.

- [9] W. Wang and M. Sebag, "Multi-objective monte-carlo tree search," 2012.
- [10] C. Williams and C. E. Rasmussen, "Gaussian processes for regression," in *Advances in neural information processing systems*, 1996, pp. 514–520.
- [11] R. Bellman, "A Markovian decision process," *Journal of Mathematics and Mechanics*, vol. 6, no. 5, pp. 679–684, 1957.
- [12] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in neural information processing systems*, 2000, pp. 1057–1063.
- [13] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer School on Machine Learning*. Springer, 2003, pp. 63–71.
- [14] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [15] 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*, vol. 12, pp. 2825–2830, 2011.
- [16] S. Levine, Z. Popovic, and V. Koltun, "Nonlinear inverse reinforcement learning with gaussian processes," in *Advances in Neural Information Processing Systems*, 2011, pp. 19–27.
- [17] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.
- [18] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.
- [19] A. G. d. G. Matthews, M. van der Wilk, T. Nickson, K. Fujii, A. Boukouvalas, P. Leon-Villagra, Z. Ghahramani, and J. Hensman, "Gpflow: A Gaussian process library using TensorFlow," *Journal of Machine Learning Research*, vol. 18, no. 40, pp. 1–6, apr 2017. [Online]. Available: <http://jmlr.org/papers/v18/16-537.html>
- [20] I. Das and J. E. Dennis, "A closer look at drawbacks of minimizing weighted sums of objectives for pareto set generation in multicriteria optimization problems," *Structural optimization*, vol. 14, no. 1, pp. 63–69, 1997.