Distributional Multi-Objective Decision Making

Willem Röpke[1] , **Conor F. Hayes**[2] , **Patrick Mannion**[2] , **Enda Howley**[2] , **Ann Nowé**[1] and **Diederik M. Roijers**[1,3]
[1]Vrije Universiteit Brussel, Brussels, Belgium
[2]University of Galway, Galway, Ireland
[3]City of Amsterdam, Amsterdam, The Netherlands
willem.ropke@vub.be, c.hayes13@nuigalway.ie, patrick.mannion@universityofgalway.ie, enda.howley@universityofgalway.ie, ann.nowe@vub.be, diederik.roijers@vub.be

Abstract

For effective decision support in scenarios with conflicting objectives, sets of potentially optimal solutions can be presented to the decision maker. We explore both what policies these sets should contain and how such sets can be computed efficiently. With this in mind, we take a distributional approach and introduce a novel dominance criterion relating return distributions of policies directly. Based on this criterion, we present the distributional undominated set and show that it contains optimal policies otherwise ignored by the Pareto front. In addition, we propose the convex distributional undominated set and prove that it comprises all policies that maximise expected utility for multivariate risk-averse decision makers. We propose a novel algorithm to learn the distributional undominated set and further contribute pruning operators to reduce the set to the convex distributional undominated set. Through experiments, we demonstrate the feasibility and effectiveness of these methods, making this a valuable new approach for decision support in real-world problems.

1 Introduction

Multi-objective sequential decision making is a complex process that involves trade-offs between multiple, often conflicting, objectives. As the preferences over these objectives are typically not known a priori, it is challenging to find a single optimal solution, and instead, a set of solutions that are considered optimal can be presented to the decision maker [Roijers *et al.*, 2013]. To keep decision support tractable, it is necessary to reduce the size of the solution sets as much as possible. Therefore, defining appropriate solution sets that do not retain excess policies while guaranteeing that no concessions are made to optimality, as well as designing corresponding pruning algorithms is essential [Taboada *et al.*, 2007].

A solution set that is often considered appropriate in both multi-objective decision making and multi-objective optimisation is the Pareto front [Roijers *et al.*, 2013]. The Pareto front consists of the policies that lead to Pareto optimal expected payoffs and thus contains all policies which are optimal for decision makers interested in optimising the utility from these expected returns [Hayes *et al.*, 2022a]. However, it is known that the Pareto front does not necessarily contain all optimal policies for problems where the decision maker optimises for their expected utility instead [Hayes *et al.*, 2022c].

To address this limitation, we introduce a novel dominance criterion, called distributional dominance, relating the multivariate return distribution between policies directly. Distributional dominance relies on first-order stochastic dominance, which is known to imply greater expected utility for univariate distributions [Fishburn, 1974; Bawa *et al.*, 1985], and has also been explored for multi-variate distributions [Denuit *et al.*, 2013; Levy, 2016a]. Based on distributional dominance, we propose the *distributional undominated set (DUS)* as a novel solution set and show that it contains all optimal policies for the class of multivariate risk-averse decision makers defined by Richard [1975]. Furthermore, we show that it is a superset of the Pareto front and as a result is a suitable starting set which can be further pruned to smaller subsets for specific scenarios.

While the DUS contains no distributionally dominated policies, it may still contain policies which will never be chosen in the expected utility setting. Therefore, we introduce a second solution set, the *convex distributional undominated set (CDUS)*, which includes only those policies that are undominated by a mixture of policies in the DUS. We find that the CDUS is a subset of the DUS and contains all optimal policies for multivariate risk-averse decision makers. While in general the CDUS and the Pareto front do not coincide, both sets are shown to include the convex hull.

From a computational perspective, we contribute algorithms to prune a set of policies to its DUS or CDUS. As these pruning methods rely on the quality of the input set, we present an extension of the Pareto Q-learning algorithm [Van Moffaert and Nowé, 2014] to learn return distributions and only discard those policies that are not in the DUS. We evaluate our approach on randomly generated MOMDPs of different sizes and compare the sizes of the resulting sets after pruning. As our goal is to use these sets in a decision support scenario, keeping their sizes reasonable and algorithms tractable both in terms of runtime and memory enables decision makers to efficiently select their preferred policy[1].

[1] A full version with supplementary material is available online at https://arxiv.org/abs/2305.05560

2 Background

2.1 Multi-Objective Decision Making

Sequential decision making is often formalised using Markov Decision Processes (MDPs) which provide a mathematical framework for modelling settings in which an agent must choose an action at each time step based on the current state of the system. To address real-world situations where decision makers must consider multiple conflicting objectives, MDPs can be generalised to Multi-Objective Markov Decision Processes (MOMDPs) which allow for vectorial reward functions [Roijers and Whiteson, 2017].

Definition 2.1. A multi-objective Markov decision process is a tuple $M = (\mathcal{S}, \mathcal{A}, T, \gamma, \mathbf{R})$, with $d \geq 1$ objectives, where:

- $\mathcal{S}$ is the state space;
- $\mathcal{A}$ is the set of actions
- $T \colon \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition function;
- $\gamma \in [0, 1]$ is the discount factor;
- $\mathbf{R} \colon \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ is the vectorial reward function.

In a MOMDP, a decision maker takes sequential actions by means of *policy* $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ which maps state-action pairs to a probability. We denote the set of all policies by Π.

We take a distributional approach [Bellemare *et al.*, 2023; Hayes *et al.*, 2022b] and consider the multivariate return distributions of these policies. The return $\mathbf{Z}^\pi = (Z_1^\pi, \ldots, Z_d^\pi)^\mathrm{T}$ is a random vector where each Z_i^π is the marginal distribution of the i'th objective such that,

$$\mathbb{E}\left[\mathbf{Z}^\pi\right] = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{r}_t \mid \pi, \mu_0\right] = \left(\mathbb{E}\left[Z_1^\pi\right], \ldots, \mathbb{E}\left[Z_d^\pi\right]\right)^\mathrm{T}. \tag{1}$$

For notational simplicity, when considering the expected returns directly we will write this as $\mathbf{V}^\pi = (V_1^\pi, \ldots, V_d^\pi)^\mathrm{T}$.

2.2 Dominance Relations

Multi-objective decision making presents additional complexity compared to traditional decision making, as it is not possible to completely order the return of different policies. Pareto dominance introduces a partial ordering by considering a vector dominant when it is greater or equal for all objectives and strictly greater for at least one objective. We say a policy Pareto dominates a second policy when the expected value of its return distribution is Pareto dominant.

Definition 2.2. Let $\pi, \pi' \in \Pi$. Then π Pareto dominates π', denoted by $\mathbf{V}^\pi \succ_\mathrm{p} \mathbf{V}^{\pi'}$, when $\forall i, V_i^\pi \geq V_i^{\pi'} \wedge \exists i, V_i^\pi > V_i^{\pi'}$.

When the expected return of π is equal to π' or Pareto dominates it, we denote this by $\mathbf{V}^\pi \succeq_\mathrm{p} \mathbf{V}^{\pi'}$.

First-order stochastic dominance (FSD) is a well-known dominance criterion from decision theory and economics, which relates return distributions directly [Levy, 2016b; Denuit *et al.*, 2013]. Let $F_\mathbf{X}(\mathbf{x}) = P(\mathbf{X} \preceq_\mathrm{p} \mathbf{x})$ be the cumulative distribution function (CDF) of a random vector $\mathbf{X}$, denoting the probability that the random vector takes on a value Pareto dominated or equal to $\mathbf{x}$. Informally, we say that $\mathbf{X}$ FSD another distribution $\mathbf{Y}$ when it always has a higher probability of obtaining Pareto dominant returns.

Definition 2.3. A policy π first-order stochastically dominates another policy π', denoted by $\mathbf{Z}^\pi \succeq_\mathrm{FSD} \mathbf{Z}^{\pi'}$, when,

$$\forall \mathbf{v} \in \mathbb{R}^d : F_{\mathbf{Z}^\pi}(\mathbf{v}) \leq F_{\mathbf{Z}^{\pi'}}(\mathbf{v}).$$

2.3 The Utility-Based Approach

We take a utility-based approach to multi-objective decision making [Roijers *et al.*, 2013] and assume that for any decision maker a utility function $u : \mathbb{R}^d \to \mathbb{R}$ exists that represents their preferences over the objectives. We consider the class of strictly monotonically increasing utility functions, denoted by $\mathcal{U}$. Intuitively, such utility functions imply that any decision maker prefers more of each objective, given all else equal.

Definition 2.4. A function $f : \mathbb{R}^d \to \mathbb{R}$ is called strictly monotonically increasing if,

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d : \mathbf{x} \succ_\mathrm{p} \mathbf{y} \implies f(x) > f(y).$$

In the utility-based approach, there is often a need to optimise for an entire class of decision makers or a decision maker for which we do not know the exact utility function. In this case, it is necessary to identify a set of policies that contain an optimal policy for all possible utility functions. A further complication arises from the fact that different optimality criteria exist depending on how the utility is derived [Roijers *et al.*, 2013]. For scenarios where a decision maker's utility is derived from multiple executions of a policy, the scalarised expected returns (SER) criterion can be optimised,

$$V_u^\pi = u\left(\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{r}_t \mid \pi, \mu_0\right]\right). \tag{2}$$

Alternatively, it is possible that the decision maker only executes their policy once and therefore aims to optimise their expected utility. In the utility-based approach, this is known as the expected scalarised returns (ESR) criterion,

$$V_u^\pi = \mathbb{E}\left[u\left(\sum_{t=0}^{\infty} \gamma^t \mathbf{r}_t\right) \mid \pi, \mu_0\right]. \tag{3}$$

It is well-established that, in general, optimal policies under one criterion need not be optimal under the other criterion [Roijers *et al.*, 2013; Vamplew *et al.*, 2022].

2.4 Solution Sets

One of the most common solution sets in the literature is the Pareto front (PF), formally defined in Definition 2.5 [Roijers and Whiteson, 2017]. We stress that this solution set is presented in the context of the SER criterion as it is based on the expected returns of the policies.

Definition 2.5. The Pareto front is the set of all policies that are not Pareto dominated:

$$\mathrm{PF}(\Pi) = \left\{\pi \in \Pi \mid \nexists \pi' \in \Pi, \mathbf{V}^{\pi'} \succ_\mathrm{p} \mathbf{V}^\pi\right\}. \tag{4}$$

A second solution set that is often considered is the convex hull (CH) which contains all policies that are optimal under linear utility functions and is therefore applicable under both SER and ESR [Hayes *et al.*, 2022a]. Additionally, when stochastic policies are allowed, the convex hull can be used to construct all Pareto optimal policies [Vamplew *et al.*, 2009].

Definition 2.6. The convex hull is the set of all policies that are not Pareto dominated by a convex combination of other policies,

$$\mathrm{CH}(\Pi) = \left\{ \pi \in \Pi \mid \nexists \lambda \in \Delta^{|\Pi|} : \sum_{i=1}^{|\Pi|} \lambda_i \mathbf{V}^{\pi_i} \succ_{\mathrm{p}} \mathbf{V}^{\pi} \right\}. \quad (5)$$

We note that solution sets based on return distributions have also been considered, with for example the ESR set [Hayes *et al.*, 2022c]. In this work, we extend this line of research and provide additional theoretical and computational results.

3 Distributional Decision Making

While most of multi-objective decision making focuses on returning the Pareto front, we demonstrate that this does not cover the full range of optimal policies. Specifically, for decision makers optimising their expected utility, the best policy in the Pareto front may still be significantly worse than a Pareto dominated policy. To overcome this, we propose a novel dominance criterion and subsequently construct a solution set based on this criterion.

3.1 Motivation

To understand why it is necessary to construct these novel solution sets, and in particular why a distributional approach is appropriate, it is helpful to consider a motivating example.

Example 1. Imagine a hospital patient needing to decide on a treatment plan with their doctor. Their objectives are to maximise the efficacy of the treatment, denoted v_1, while also maximising their comfort (i.e. minimise the side-effects), denoted v_2. Unfortunately, these objectives are conflicting. In previous discussions with their doctor, the patient mentioned that they wish to strike a balance between the two. A fitting utility function is the product between the two objectives (Eq. (6)) as it is maximised when values are closer together.

$$u(v_1, v_2) = v_1 \cdot v_2 \quad (6)$$

The doctor then proposes the following two treatment plans.

$$A = \left\{ P(v_1 = 1, v_2 = 0) = \frac{1}{2}, P(v_1 = 0, v_2 = 1) = \frac{1}{2} \right\}$$
$$B = \{ P(v_1 = 0.45, v_2 = 0.45) = 1 \},$$

with $\mathbb{E}[A] = (0.5, 0.5)$ and $\mathbb{E}[B] = (0.45, 0.45)$.

When taking the standard approach and applying Pareto dominance, it is clear that the expected return of A dominates that of B. In contrast, when considering the distributions on the basis of expected utility, A has an expected utility of 0, while B has an expected utility of 0.2025. As the patient will most likely follow the treatment plan only once, they aim to optimise their expected utility and thus prefer distribution B.

As this example shows, it is pertinent to consider exactly what the decision maker aims to optimise for: do they optimise for repeated execution of the same policy, or maximising the expected utility from one execution? In the former case, they may well decide based on the expected value of the distribution. In the latter case, however, taking the full distribution of returns into account is key to effective decision support.

3.2 Distributional Dominance

To address the limitations of Pareto dominance, we introduce the *distributional dominance* criterion. This criterion states that a distribution dominates another when it is first-order stochastic dominant and at least one of the marginal distributions *strictly* first-order stochastic dominates the related marginal distribution of the second distribution.

Definition 3.1. A policy π distributionally dominates another policy π', denoted by $\mathbf{Z}^{\pi} \succ_{\mathrm{d}} \mathbf{Z}^{\pi'}$, when,

$$\mathbf{Z}^{\pi} \succeq_{\mathrm{FSD}} \mathbf{Z}^{\pi'} \wedge \exists i \in [d] : Z_i^{\pi} \succ_{\mathrm{FSD}} Z_i^{\pi'}.$$

One can verify that distributional dominance is equivalent to strict first-order stochastic dominance in the case of random vectors when all variables are independent. In general, however, distributional dominance is a stronger condition than strict first-order stochastic dominance as the condition on the marginal distributions implies strict FSD but is not implied by it. Defining distributional dominance as such enables us to guarantee a strictly greater expected utility for a large class of decision makers and leads to the general solution set discussed in Section 4.

For the class of decision makers with utility functions in $\mathcal{U}$, we show that when a given random vector has *strictly* greater expected utility for all utility functions than a second random vector, this implies distributional dominance.

Theorem 3.1. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be d-dimensional random vectors. Then,*

$$\forall u \in \mathcal{U} : \mathbb{E}\, u(\mathbf{X}) > \mathbb{E}\, u(\mathbf{Y}) \implies \mathbf{X} \succ_d \mathbf{Y}.$$

Proof sketch. We first show an additional lemma stating that the condition implies first-order stochastic dominance. Therefore, the proof reduces to showing the condition on the marginals. It suffices to show that if $\mathbf{X}$ does not distributionally dominate $\mathbf{Y}$, it is always possible to construct a utility function for which $\mathbb{E}\, u(\mathbf{Y})$ is at least as high as $\mathbb{E}\, u(\mathbf{X})$. □

In practice, it is impossible to verify whether the expected utility of a given random vector is always strictly greater than that of a second random vector. On the other hand, we will demonstrate that it is computationally feasible to verify distributional dominance (see Section 4.2). We now show that distributional dominance implies a strictly greater expected utility for a subset of utility functions in $\mathcal{U}$. The condition we impose is referred to as "multivariate risk-aversion", which means that a decision maker in this class will, when confronted with a choice between two lotteries, always avoid the lottery containing the worst possible outcome [Richard, 1975]. Below, we present the theorem and proof for bivariate distributions. We note that for FSD this property has been shown to hold for n-dimensional random vectors as well [Scarsini, 1988].

Theorem 3.2. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be two-dimensional random vectors. Then* $\forall u \in \mathcal{U}$ *with* $\frac{\partial^2 u(x_1,x_2)}{\partial x_1 \partial x_2} \leq 0$,

$$\mathbf{X} \succ_d \mathbf{Y} \implies \mathbb{E}\, u(\mathbf{X}) > \mathbb{E}\, u(\mathbf{Y}).$$

Proof sketch. The proof utilises the fact that first-order stochastic dominance implies greater or equal expected utility [Hayes *et al.*, 2022c]. We subsequently show that the additional condition on the marginal distributions for distributional dominance implies strictly greater expected utility. □

4 A General Solution Set

We adopt distributional dominance to define the distributional undominated set (DUS). The DUS has two important desiderata: it contains the Pareto front, i.e. the optimal set under SER and contains all optimal policies for multivariate risk-averse decision makers under ESR. The deferred proofs for the theoretical results can be found in the supplementary material.

4.1 Distributional Undominated Set

As the name suggests, the distributional undominated set contains only those policies which are not pairwise distributionally dominated. We define this formally in Definition 4.1.

Definition 4.1. The distributional undominated set is the set of all policies that are not distributionally dominated:

$$\mathrm{DUS}(\Pi) = \left\{\pi \in \Pi \mid \nexists \pi' \in \Pi, \mathbf{Z}^{\pi'} \succ_{\mathrm{d}} \mathbf{Z}^{\pi}\right\}. \qquad (7)$$

From this definition it is clear that all policies which are optimal for multivariate risk-averse decision makers are in the set. To show that the Pareto front is a subset as well, we first introduce Lemma 4.1, stating that distributional dominance implies Pareto dominance.

Lemma 4.1. *For all policies $\pi, \pi' \in \Pi$,*

$$\mathbf{Z}^{\pi} \succ_d \mathbf{Z}^{\pi'} \implies \mathbf{V}^{\pi} \succ_p \mathbf{V}^{\pi'}.$$

Proof sketch. The proof works by utilising a known link between the expected value of a random variable and its cumulative density function. Then, the conditions for distributional dominance imply that the expected value for each marginal distribution is greater or equal and at least one marginal distribution is strictly greater. □

Leveraging Lemma 4.1, it is a straightforward corollary that the Pareto front is a subset of the DUS.

Corollary 4.1.1. *For any family of policies Π, the Pareto front is a subset of the distributional undominated set, i.e.,*

$$PF(\Pi) \subseteq DUS(\Pi).$$

We highlight that our dominance results and solution sets are not restricted to MOMDPs but apply to any stochastic multi-objective decision problem with vector-valued outcomes.

4.2 Computing the DUS

To deal with return distributions computationally, we project distributions to multivariate categorical distributions [Bellemare *et al.*, 2023; Hayes *et al.*, 2022c]. This ensures that finite memory is used, and, importantly, that computations can be performed efficiently. Concretely, to verify first-order stochastic dominance, we need only compare a finite number of points as the CDF is a multivariate step function with steps at $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. Formally, for the categorical distribution $\mathbf{X}$ the cumulative distribution at $\mathbf{x}$ is computed as follows,

$$F_{\mathbf{X}}(\mathbf{x}) = \sum_{\mathbf{v}_i \preceq_{\mathrm{p}} \mathbf{x}} p(\mathbf{v}_i). \qquad (8)$$

Additionally, discrete distributions enable straightforward computation of marginal distributions, thus having all ingredients to check distributional dominance (see Definition 3.1). Then, starting from a given set of policies, the DUS can be computed using a modified version of the Pareto Prune (PPrune) algorithm [Roijers and Whiteson, 2017] that checks for distributional dominance rather than Pareto dominance. We refer to the resulting pruning algorithm as *DPrune*.

5 A Solution Set for ESR

As the DUS is a superset of the Pareto front and further contains optimal policies under ESR, we can intuitively assume that it might grow very large in size, thereby complicating its practical use in decision support systems. When considering SER, it is possible to reduce the set to the Pareto front by utilising existing pruning operators [Roijers and Whiteson, 2017]. We contribute a similar approach for ESR and present both the resulting solution set as well as a pruning algorithm for this purpose.

5.1 Convex Mixture of Distributions

For univariate distributions, it has been shown that a mixture distribution can be constructed that first-order stochastic dominates another distribution if and only if for any decision maker there exists a distribution in the mixture which is preferred over the dominated distribution [Fishburn, 1974; Bawa *et al.*, 1985]. Mixture dominance has also been considered for multivariate distributions [Denuit *et al.*, 2013].

Here, we show that convex distributional dominance implies greater expected utility for multivariate risk-averse decision makers when considering bivariate distributions.

Theorem 5.1. *Let $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\}$ be sets of two-dimensional random vectors. Then,*

$$\exists \lambda \in \Delta^n : \sum_{i=1}^{n} \lambda_i \mathbf{X}_i \succ_d \sum_{i=1}^{n} \lambda_i \mathbf{Y}_i,$$

implies that $\forall u \in \mathcal{U}$ with $\frac{\partial^2 u(x_1,x_2)}{\partial x_1 \partial x_2} \leq 0$,

$$\exists i \in [n] : \mathbb{E}\, u(\mathbf{X}_i) > \mathbb{E}\, u(\mathbf{Y}_i).$$

Proof sketch. The proof follows from Theorem 3.2 and linearity of expectation. □

Observe that in the special case where all random vectors $\mathbf{Y}_i$ are equal, mixture dominance of $\mathbf{Y}$ implies that all decision makers will prefer a random vector $\mathbf{X}_i$ over $\mathbf{Y}$.

5.2 Convex Distributional Undominated Set

We define a final solution set, called the convex distributional undominated set (CDUS), that contains only those policies which are undominated by a mixture of distributions. Theorem 5.1 guarantees that for all decision makers in the class there is an optimal policy contained in the set. We define the CDUS formally below. It follows from this definition that the CDUS is a subset of the DUS.

Definition 5.1. The CDUS is the set of all policies that are not distributionally dominated by a convex mixture:

$$\mathrm{CDUS}(\Pi) = \left\{\pi \in \Pi \mid \nexists \lambda \in \Delta^{|\Pi|} : \sum_{i=1}^{|\Pi|} \lambda_i \mathbf{Z}^{\pi_i} \succ_{\mathrm{d}} \mathbf{Z}^{\pi}\right\}.$$

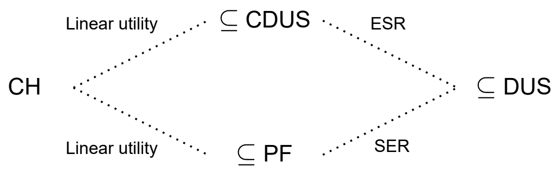

Figure 1: A taxonomy of solution sets in multi-objective decision making.

Given the myriad of solution sets in multi-objective decision making, it is useful to define a complete taxonomy between them. From Corollary 4.1.1, we know that the Pareto front is a subset of the DUS. Additionally, it follows from Definition 5.1 that the CDUS is also a subset of the DUS. Earlier work has shown that the convex hull is a subset of the Pareto front [Roijers and Whiteson, 2017] and we show that this is also true for the CDUS.

Corollary 5.1.1. *For any family of policies* Π*,*

$$\mathit{CH}(\Pi) \subseteq \mathit{CDUS}(\Pi)\,.$$

The final missing piece of the puzzle is the relation between the CDUS and Pareto front. However, here one can find counterexamples which disprove that the CDUS is either a subset or superset of the Pareto front. The landscape of solution sets for multi-objective decision making can then be summarised as shown in Fig. 1.

5.3 Pruning to the CDUS

To prune a set of distributions to its CDUS, we must check for each distribution whether it is dominated by a mixture of the other distributions. Fortunately, this verification is feasible by restating the problem using linear programming. Concretely, we extend an algorithm that checks whether a univariate distribution is convex first-order stochastic dominated to our setting [Bawa *et al.*, 1985]. We show the resulting linear program *CDPrune* in Algorithm 1.

For notational simplicity, we define the size of the set of distributions allowed in the mixture as n. Then the linear program takes in total $n+1$ distributions as input, where the final distribution is the distribution to check. As these distributions are discrete, the CDFs are multivariate step functions that step at a finite number of points. Let D_i be the set of points at which the CDF of distribution i steps. Then $D = \bigcup_{i=1}^{n+1} D_i$ is the union of all such points. We denote $h = |D|$.

The linear program maximises δ, which is the sum of slack variables that make up the difference between the CDFs of the marginal mixture distributions and the marginals of the distribution to check (Eq. (9)). If this procedure leads to a δ greater than zero, this implies that the conditions for distributional dominance are met and the distribution is dominated by the mixture. Note that we may omit an additional constraint on the l slack variables to be greater or equal to zero, as this is implied by the constraint on the s slack variables (Eq. (12)).

When no exact formulation of the joint CDFs is available, we propose an alternative linear program that operates solely on the marginal distributions. In this case, it is necessary to change the first constraint in Eq. (10) to

$$\sum_{i=1}^{n} \lambda_i \prod_{k=1}^{d} F_{\mathcal{Z}_{i,k}}(\mathbf{v}_{j,k}) + s_j = \prod_{k=1}^{d} F_{Z_k}(v_{j,k}), \tag{13}$$

while the second constraint in Eq. (11) is removed altogether. By maximising the sum of s slack variables, the resulting linear program essentially checks for strict first-order stochastic dominance between random vectors with independent variables. One can verify that this implies distributional dominance for independent variables and otherwise may serve as an approximation.

Algorithm 1 CDPrune

Input: A set of return distributions $\mathcal{Z}$ allowed in the mixture and a return distribution $\mathbf{Z}$ to check
Output: Whether the distribution is convex dominated

$$\text{Maximise } \delta = \sum_{i=1}^{n}\sum_{k=1}^{d} l_{i,k} \tag{9}$$

Subject to:

$$\sum_{i=1}^{n} \lambda_i F_{\mathcal{Z}_i}(\mathbf{v}_j) + s_j = F_{\mathbf{Z}}(\mathbf{v}_j) \quad j = 1,\ldots,h \tag{10}$$

$$\sum_{i=1}^{n} \lambda_i F_{\mathcal{Z}_{i,k}}(v_{j,k}) + l_{j,k} = F_{Z_k}(v_{j,k}) \qquad j = 1,\ldots,h \quad k = 1,\ldots,d \tag{11}$$

$$\sum_{i=1}^{n} \lambda_i = 1$$

$$\lambda_i \geq 0 \quad i = 1,\ldots,n$$

$$s_j \geq 0 \quad j = 1,\ldots,h \quad \text{where } s_j \text{ is a slack variable} \tag{12}$$

return TRUE **if** $\delta > 0$ **else** FALSE

6 Computing the Solution Sets

Our final contribution relates theory to practice by designing an algorithm able to learn the DUS in a given MOMDP. We evaluate this algorithm on different sizes of MOMDPs and compare the resulting sizes of the sets when pruned down to the subsets covered in the taxonomy in Fig. 1. All code is available at https://github.com/wilrop/distributional-dominance.

6.1 Distributional Multi-Objective Q-Learning

Pareto Q-learning (PQL) is a classical algorithm used in multi-objective reinforcement learning to learn the Pareto front [Van Moffaert and Nowé, 2014]. We find that the general framework of PQL lends itself nicely to learning the DUS. Our algorithm, DIstributional Multi-Objective Q-learning (DIMOQ) is shown in Algorithm 2.

Algorithm 2 DIMOQ

```
Input: The state space $\mathcal{S}$, actions space $\mathcal{A}$ and discount factor $\gamma$
Output: The DUS
Initialise all $Q(s, a)$ as empty sets
Initialise all $\mathbf{R}(s, a, s')$ as Dirac delta distributions
Estimate $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ from random walks
for each episode do
    Initialise state $s$
    repeat
        Take an action $a \sim \pi(a|s)$
        Observe the next state $s' \in \mathcal{S}$ and reward $\mathbf{r} \in \mathbb{R}^d$
        $ND(s, a, s') \leftarrow \textsc{DPrune}\left(\bigcup_{a' \in \mathcal{A}} Q(s', a')\right)$
        Update the reward distribution $\mathbf{R}(s, a, s')$ with $\mathbf{r}$
        $s \leftarrow s'$
    until $s$ is terminal
end for
return $\textsc{DPrune}\left(\bigcup_{a \in \mathcal{A}} Q(0, a)\right)$
```

The algorithm first initialises the Q-sets containing undominated distributions to empty sets and reward distributions to Dirac delta distributions at zero. During training, the agent follows an ϵ-greedy policy and learns the immediate reward distributions $\mathbf{R}(s, a, s')$ separate from the expected future reward distributions $ND(s, a, s')$. Learning the immediate reward distribution is done by recording the empirical distribution, while learning the future reward distribution is done using a modified version of the Q-update rule employed for PQL (see Eq. (14)). Note, however, that for DIMOQ the pruning operator for the distributions in the next state is *DPrune* rather than *PPrune*.

Dealing With Stochasticity

The Q-learning update in PQL is described for deterministic environments. As we deal with fundamentally stochastic environments, we propose an alternative formulation in Eq. (14).

$$Q(s, a) \leftarrow \bigoplus_{s'} T(s'|s, a) \left[\mathbf{R}(s, a, s') + \gamma ND(s, a, s')\right] \tag{14}$$

First, the term $[\mathbf{R}(s, a, s') + \gamma ND(s, a, s')]$ constructs a set of expected return distributions when the state-action pair leads to s'. Next, the $\bigoplus_{s'} T(s'|s, a)$ constructs mixture policies over all next states s' where each distribution is weighted according to its transition probability $T(s'|s, a)$.

In a learning setting, the transition probabilities are not assumed to be given. As such, we perform a number of random walks before training to estimate these probabilities. During learning, we do not update the transition function anymore, to avoid creating unnecessary distributions which will never be observed again due to drift in the probabilities.

Action Selection

The second adaptation necessary to learn the DUS rather than the Pareto front is the action scoring and selection mechanism. Even for PQL, this is complicated as it is not obvious what metric to use to determine the quality of a set of Q-values. Several set evaluation mechanisms have been proposed for this, such as for example the hypervolume metric [Guerreiro *et al.*, 2021] or using a Chebyshev scalarisation function [Van Moffaert *et al.*, 2013]. We note that these approaches can be extended to DIMOQ as well by computing the expected value of the distribution first and then continuing with one of the aforementioned scoring metrics.

Name	States	Actions	Next states	Timesteps	Set limit
Small	5	2	$[1, 2]$	3	10
Medium	10	3	$[1, 2]$	5	15
Large	15	4	$[1, 2]$	7	20

Table 1: Configuration of the generated MOMDPs. Timesteps refer to the maximum time horizon after which the episode is terminated.

In addition to the classical scoring methods, we propose using a linear utility function as a baseline and scoring a set of distributions by its mean expected utility. As linear scalarisation can be done efficiently, this results in a performant scoring method. An additional advantage of this approach is that when more information about the shape of the utility function is known, the linear utility baseline can be substituted with a better approximation.

Limiting Set Size

Due to stochasticity in the environment and because the Q-update rule in Eq. (14) performs all possible combinations, the Q-sets in the algorithm are quick to explode in size. To constrain the size of the sets, we propose two mechanisms.

First, we limit the precision of the distributions that are learned. This approach was demonstrated to be successful in multi-objective dynamic programming as well [Mandow *et al.*, 2022]. Second, we set a fixed limit on the set size. Whenever this limit is crossed, we perform agglomerative clustering where the number of clusters equals the maximum set size. As input for the clustering, we compute the pairwise distances between all distributions. In experiments, we compute the Jensen-Shannon distance between the flattened distributions. Alternatively, one could use the cost of optimal transport between pairs of distributions.

6.2 Empirical Results

We evaluate DIMOQ (Algorithm 2) and CDPrune (Algorithm 1) on randomly generated MOMDPs of different sizes shown in Table 1. For each size category, we repeat the experiment with seeds one through five and perform $50,000$ random walks to estimate T followed by $2,000$ training episodes. All experiments considered two objectives, used a discount factor of 1 and limited the precision of distributions to three decimals. Finally, the experiments were run on a single core of an Intel Xeon Gold 6148 processor, with a maximum RAM requirement of 2GB.

We observe that the runtimes for DIMOQ shown in Table 2 are heavily influenced by the size of the MOMDP. Additionally, there is a large variance in runtime across different seeds. We find that these differences cannot solely be attributed to having a more complex transition function, but are most likely due to the interplay between the transition function and the reward function. Specifically, if transitions result

Name	Mean	SD	Min	Max
Small	00:01:21	00:00:25	00:00:58	00:02:01
Medium	01:49:11	00:47:07	00:17:41	02:31:18
Large	17:01:25	06:02:35	09:46:06	27:55:55

Table 2: Runtime for DIMOQ on randomly generated MOMDPs.

Name	DUS	CDUS	PF	CH
Small	13.0 ± 10.73	$95.71\% \pm 8.57$	$39.88\% \pm 16.45$	$36.07\% \pm 20.12$
Medium	372.2 ± 211.88	$61.27\% \pm 12.16$	$6.28\% \pm 6.70$	$2.87\% \pm 3.48$
Large	639.0 ± 221.71	$53.00\% \pm 5.68$	$3.43\% \pm 2.01$	$1.33\% \pm 0.82$

Table 3: The relative sizes of the pruned subsets.

in a large number of undominated returns each iteration needs to perform a large number of combinations. It is clear however that scaling becomes an issue for DIMOQ when going to larger action and state spaces. As such, we plan to investigate the use of function approximation to further extend DIMOQ to larger MOMDPs. Additionally, we note that MOMDPs modelled after real-world scenarios will likely contain more structure and are thus interesting to study for future work.

In Table 3 we show the average size of the DUS, as well as what percentage of the DUS belongs to the CDUS, Pareto front and convex hull on average. We observe a similar pattern, namely that larger MOMDPs lead to larger solution sets. Interestingly though, larger MOMDPs also allow for a greater percentage of policies to be pruned for the smaller solution sets, which is beneficial for their use in decision support.

We highlight that although the CDUS is often substantially smaller than the DUS, the Pareto front and convex hull are much smaller than either. Intuitively, this is because when both objectives are to be maximised, Pareto optimal policies can only occur on the upper right hand region of the objective space, while policies in the DUS and CDUS may still exist in the Pareto dominated part of the space. However, recall from Example 1 that these policies may still be optimal under ESR. We visualise this in Fig. 2 where the expected values for the final distributions from one representative experiment are plotted.

Finally, we remark that while the CDUS cannot be guaranteed to be a superset of the Pareto front in general, in all experiments this was in fact the case. This is also apparent from the results in Fig. 2. An interesting direction for future work is to specify the exact conditions under which this relation is guaranteed to hold.

7 Related Work

Stochastic dominance has long been employed in areas of finance and economics [Levy, 2016b] and has more recently also found use in solving decision making problems through reinforcement learning (RL). In single-objective settings, Epshteyn and DeJong [2006] employ stochastic dominance to learn optimal policies in MDPs with incomplete specifications. Martin et al. [2020] define a risk-aware distributional algorithm that utilises stochastic dominance at decision time to determine the best action. Techniques from stochastic dominance have also been used to analyse the theoretical properties of distributional RL [Rowland *et al.*, 2018].

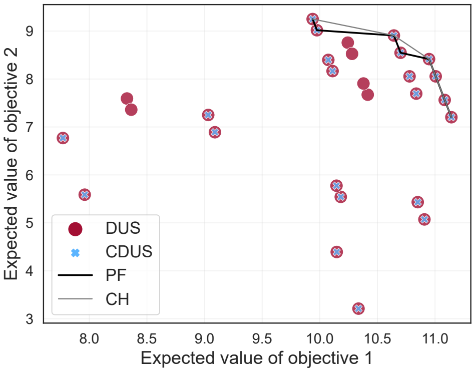

Figure 2: The resulting solution sets for a sample experiment. Policies in the dominated part of the objective space may still be optimal for certain decision makers and can thus not be excluded a priori.

The distributional approach in general has become an active area of research for both single-objective and multi-objective settings. For a thorough overview of techniques in single-objective settings, we refer to a recent textbook on the matter [Bellemare *et al.*, 2023]. In multi-objective settings, Hayes et al. [2021] and Reymond et al. [2023] define single-policy multi-objective RL algorithms that can learn policies for nonlinear utility functions under the ESR criterion. Furthermore, Hayes et al. [2022b] outline a multi-policy multi-objective distributional value iteration algorithm that computes a set of policies for the ESR criterion, known as the ESR set. The ESR set is the first solution set for use in multi-objective sequential decision making under the ESR criterion and leverages strict first-order stochastic dominance to determine whether a policy is included in the set. This set was shown to contain all optimal policies for multivariate risk-averse decision makers, but implicitly assumes all variables in the random vector to be independent [Hayes *et al.*, 2022c].

8 Conclusion

We investigate multi-objective decision making and find that existing solution sets frequently fall short in specific use cases. To resolve this, we first propose the distributional undominated set. We show that this set contains both the Pareto front as well as all optimal policies for multivariate risk-averse decision makers optimising their expected utility. We subsequently present the convex distributional undominated set, which aims to target the expected utility setting in particular. From this, we determine a taxonomy of existing solution sets in multi-objective decision making.

To facilitate the application of these concepts, we present computational approaches for learning the distributional undominated set and pruning operators to reduce the set to the convex distributional undominated set. Through experiments, we demonstrate the feasibility and effectiveness of these methods. As such, this work offers a promising approach to decision support in real-world problems.

Acknowledgments

WR is supported by the Research Foundation – Flanders (FWO), grant numbers 1197622N. CH is funded by the University of Galway Hardiman Scholarship. This research was supported by funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program.

References

[Bawa *et al.*, 1985] Vijay S. Bawa, James N. Bodurtha, M. R. Rao, and Hira L. Suri. On determination of stochastic dominance optimal sets. *The Journal of Finance*, 40(2):417–431, 1985.

[Bellemare *et al.*, 2023] Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2023. http://www.distributional-rl.org.

[Denuit *et al.*, 2013] Michel Denuit, Louis Eeckhoudt, Ilia Tsetlin, and Robert L. Winkler. Multivariate Concave and Convex Stochastic Dominance. In Francesca Biagini, Andreas Richter, and Harris Schlesinger, editors, *Risk Measures and Attitudes*, pages 11–32. Springer London, London, 2013.

[Epshteyn and DeJong, 2006] Arkady Epshteyn and Gerald DeJong. Qualitative reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 305–312, New York, NY, USA, 2006. Association for Computing Machinery.

[Fishburn, 1974] Peter C Fishburn. Convex stochastic dominance with continuous distribution functions. *Journal of Economic Theory*, 7(2):143–158, February 1974.

[Guerreiro *et al.*, 2021] Andreia P. Guerreiro, Carlos M. Fonseca, and Luís Paquete. The hypervolume indicator: Computational problems and algorithms. *ACM Comput. Surv.*, 54(6), jul 2021.

[Hayes *et al.*, 2021] Conor F Hayes, Mathieu Reymond, Diederik M Roijers, Enda Howley, and Patrick Mannion. Distributional monte carlo tree search for risk-aware and multi-objective reinforcement learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1530–1532, 2021.

[Hayes *et al.*, 2022a] Conor F. Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai A. Irissappane, Patrick Mannion, Ann Nowé, Gabriel Ramos, Marcello Restelli, Peter Vamplew, and Diederik M. Roijers. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1):26, April 2022.

[Hayes *et al.*, 2022b] Conor F Hayes, Diederik M Roijers, Enda Howley, and Patrick Mannion. Decision-theoretic planning for the expected scalarised returns. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 1621–1623, 2022.

[Hayes *et al.*, 2022c] Conor F. Hayes, Timothy Verstraeten, Diederik M. Roijers, Enda Howley, and Patrick Mannion. Expected scalarised returns dominance: A new solution concept for multi-objective decision making. *Neural Computing and Applications*, July 2022.

[Levy, 2016a] Haim Levy. Bivariate FSD (BFSD). In *Stochastic Dominance: Investment Decision Making under Uncertainty*, pages 441–465. Springer International Publishing, Cham, 2016.

[Levy, 2016b] Haim Levy. Stochastic dominance decision rules. In *Stochastic Dominance: Investment Decision Making under Uncertainty*, pages 41–124. Springer International Publishing, Cham, 2016.

[Mandow *et al.*, 2022] L. Mandow, J. L. Perez-de-la-Cruz, and N. Pozas. Multi-objective dynamic programming with limited precision. *Journal of Global Optimization*, 82(3):595–614, March 2022.

[Martin *et al.*, 2020] John Martin, Michal Lyskawinski, Xiaohu Li, and Brendan Englot. Stochastically dominant distributional reinforcement learning. In *International Conference on Machine Learning*, pages 6745–6754. PMLR, 2020.

[Reymond *et al.*, 2023] Mathieu Reymond, Conor F. Hayes, Denis Steckelmacher, Diederik M. Roijers, and Ann Nowé. Actor-critic multi-objective reinforcement learning for non-linear utility functions. *Autonomous Agents and Multi-Agent Systems*, 37(2):23, April 2023.

[Richard, 1975] Scott F. Richard. Multivariate Risk Aversion, Utility Independence and Separable Utility Functions. *Management Science*, 22(1):12–21, September 1975.

[Roijers and Whiteson, 2017] Diederik M. Roijers and Shimon Whiteson. Multi-objective decision making. In *Synthesis Lectures on Artificial Intelligence and Machine Learning*, volume 34, pages 129–129. Morgan and Claypool, 2017.

[Roijers *et al.*, 2013] Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

[Rowland *et al.*, 2018] Mark Rowland, Marc Bellemare, Will Dabney, Remi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 29–37. PMLR, April 2018.

[Scarsini, 1988] Marco Scarsini. Dominance Conditions for Multivariate Utility Functions. *Management Science*, 34(4):454–460, 1988.

[Taboada *et al.*, 2007] Heidi A. Taboada, Fatema Baheranwala, David W. Coit, and Naruemon Wattanapongsakorn. Practical solutions for multi-objective optimization: An application to system reliability design problems. *Selected*

Papers Presented at the Fourth International Conference on Quality and Reliability, 92(3):314–322, March 2007.

[Vamplew *et al.*, 2009] Peter Vamplew, Richard Dazeley, Ewan Barker, and Andrei Kelarev. Constructing Stochastic Mixture Policies for Episodic Multiobjective Reinforcement Learning Tasks. In Ann Nicholson and Xiaodong Li, editors, *AI 2009: Advances in Artificial Intelligence*, pages 340–349, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[Vamplew *et al.*, 2022] Peter Vamplew, Cameron Foale, and Richard Dazeley. The impact of environmental stochasticity on value-based multiobjective reinforcement learning. *Neural Computing and Applications*, 34(3):1783–1799, February 2022.

[Van Moffaert and Nowé, 2014] Kristof Van Moffaert and Ann Nowé. Multi-objective reinforcement learning using sets of pareto dominating policies. *Journal of Machine Learning Research*, 15(107):3663–3692, 2014.

[Van Moffaert *et al.*, 2013] Kristof Van Moffaert, Madalina M. Drugan, and Ann Nowé. Scalarized multi-objective reinforcement learning: Novel design techniques. In *2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pages 191–199, 2013.