Mixing Any Cocktail with Limited Ingredients: On the Structure of Payoff Sets in Multi-Objective POMDPs and Its Impact on Randomised Strategies

James C. A. Main
University of Oxford, UK
UMONS – Université de Mons, Belgium

Mickael Randour
F.R.S.-FNRS and UMONS – Université de Mons, Belgium

Abstract

We consider multi-dimensional payoff functions in partially observable Markov decision processes. We study the structure of the set of expected payoff vectors of all strategies (policies) and study what kind are needed to achieve a given expected payoff vector. In general, pure strategies (i.e., not resorting to randomisation) do not suffice for this problem.

We prove that for any payoff function for which the expectation is well-defined under all strategies, it is sufficient to mix (i.e., randomly select a pure strategy at the start of a play and committing to it for the rest of the play) *finitely many* pure strategies to approximate any expected payoff vector up to any precision. Furthermore, for any payoff function for which the expected payoff is finite under all strategies, any expected payoff vector can be obtained exactly by mixing finitely many strategies. These results are already novel in the context of (perfect information) Markov decision processes.

2012 ACM Subject Classification Theory of computation → Probabilistic computation; Theory of computation → Logic and verification

Keywords and phrases partially observable Markov decision process, multiple objectives, randomised strategies, mixed strategies, strategy complexity

Digital Object Identifier 10.4230/LIPIcs.LICS.2026.68

Related Version *Full Version*: `https://arxiv.org/abs/2502.18296` [46]

Funding This work has been supported by EPSRC project EP/Z536179/1 and by the Fonds de la Recherche Scientifique – FNRS (F.R.S.-FNRS) under Grant n° T.0188.23 (PDR ControlleRS).
James C. A. Main: Supported by a Research Fellowship of the F.R.S.-FNRS during the preparation of this work.
Mickael Randour: F.R.S.-FNRS Senior Research Associate and member of the TRAIL institute.

1 Introduction

The model. *Markov decision processes* (MDPs) are a classical framework for decision making in uncertain environments. MDPs are notably used in the fields of formal methods (e.g., [3, 33]) and reinforcement learning (e.g., [58]). An MDP models the interaction of a controllable system with a stochastic environment: at each step of the process, the system selects an action and the state of the process is updated following a probability distribution that depends only on the current state and chosen action. This interaction goes on forever and yields a *play* (i.e., an execution) of the process. The quality of a play is quantified by a *payoff function*, which assigns a numeric value to all plays. The goal is then to compute an *optimal strategy* (policy), i.e., a decision-making procedure for the system that maximises the expected payoff. Such a strategy constitutes a *formal blueprint* for a controller of the modelled system [51].

Imperfect information. In an MDP, the system is perfectly informed of the current state in each step. *Partially observable MDPs* (POMDPs, e.g. [41]) generalise MDPs by relaxing this assumption. In a POMDP, the system perceives an observation at each step from the current state, which may be shared with other states, and strategies must make decisions based *only on observations*. POMDPs are harder to analyse than MDPs in general [21].

Multi-objective (PO)MDPs. More complex specifications, such as simultaneous constraints on the response time and energy consumption of a system, are usually modelled using *multi-dimensional payoff functions*. In this setting, some expected payoff vectors may be incomparable; an analysis of trade-offs between the different dimensions may be necessary.

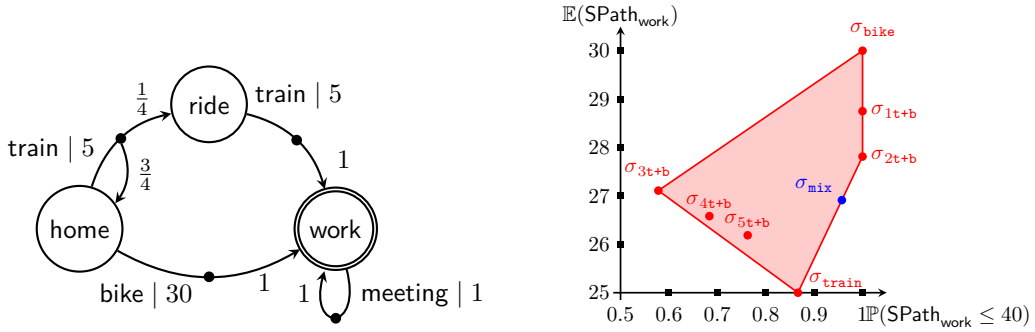
The goal is generally to determine, given a vector, whether it is *achievable*, i.e., whether there exists a strategy whose expected payoff is greater than or equal to the vector in all components (e.g., [32, 23, 53]). A related problem is to compute or approximate the *Pareto curve* of the set of expected payoffs, i.e., the expected payoffs that are *Pareto-optimal* (e.g., [35, 56, 26, 50, 54]). Intuitively, an expected payoff is Pareto-optimal if there is no strategy whose expected payoff is as good on all dimensions and strictly better on one dimension. Alternatively, one can look for strategies with expected payoffs that are optimal for the *lexicographic* order over vectors (e.g., [59, 38, 25, 16]). For instance, in a two-dimensional setting, this equates to finding strategies that maximise the expected payoff on the second dimension among the strategies that maximise the expected payoff on the first dimension.

Strategy complexity. For controller design, simpler strategies, i.e., using limited memory and randomisation, are often deemed preferable, notably for their explainability. In MDPs, for many classical (one-dimensional) specifications, there exist optimal strategies that use no memory and no randomisation. For instance, such strategies suffice to maximise the probability of a parity objective [24] or the expectation of a mean [6] or discounted-sum payoff [57]. For one-dimensional specifications in POMDPs, memory is usually necessary to play optimally while randomisation is not [20].

In contrast to the one-dimensional case, Pareto-optimal strategies in multi-objective MDPs and POMDPs may require both *memory* and *randomisation* (see, e.g., [53, 12]). For this reason, a major problem is understanding what are the simplest relevant strategies, e.g., whether randomisation is necessary, or what is the minimum amount of memory required for the application at hand. A related approach consists in directly searching for simple strategies that satisfy the desired constraints (e.g., [30]).

There is an extensive body of work aimed at understanding memory requirements for strategies in various game-based models. These works provide sufficient conditions on payoffs for which memoryless or finite-memory strategies suffice, or even complete characterisations of such payoffs. For instance, see [44, 7, 10, 18] for games on deterministic graphs, [9, 37] for stochastic games and [36] for MDPs. Most of these works focus on pure strategies (i.e., strategies not resorting to randomisation).

In this work, our focus is on *randomisation requirements* in countable-state multi-objective POMDPs. Instead of identifying bounds on memory, we study whether we can achieve vectors while using strategies with *limited randomisation power*. In general, it is not possible to jointly minimise memory and randomisation requirements: there can be a trade-off between memory and randomisation, even in the one-dimensional case [19, 39, 28, 48].



(a) Transitions are labelled by actions and probabilities. Weights represent the time taken by each action. The target is doubly circled.

(b) A set of expected payoffs. The probability of reaching work before 40 time units is given on the first dimension. The other dimension represents the expected time to reach work.

Figure 1.1 An MDP with two payoffs and its associated set of expected payoffs.

Randomised strategies. A strategy that does not use randomisation is called pure. Formally, a *pure strategy* is a function that assigns actions to histories (i.e., sequences of observations). Randomised strategies can be defined in two ways. On the one hand, a *behavioural strategy* is a function that assigns distributions over actions to histories. In other words, when following a behavioural strategy, the random choices of actions are made at each step of the process. On the other hand, a *mixed strategy* is a distribution over the set of pure strategies. When playing according to a mixed strategy, a pure strategy is randomly chosen at the start of the process and is followed for the entire play.

In POMDPs, behavioural and mixed strategies are equivalent whenever actions are observable; any strategy from one class can be emulated by a strategy of the other. This result is known as *Kuhn's theorem* [43, 2]. We note that distributions over behavioural strategies are no more general than mixed and behavioural strategies [5].

When limiting ourselves to *finite-memory strategies*, the natural analogues of mixed strategies are strictly weaker than the analogues of behavioural strategies [45]. In fact, the emulation of a behavioural strategy that flips a coin at each round requires a mixed strategy that randomises over *uncountably* many pure strategies, infinitely many of which require infinite memory to account for all patterns. It follows that mixed strategies with a finite support, i.e., that mix finitely many pure strategies, are a *strict sub-class* of the set of randomised strategies (regardless of memory). Arguably, this is one of the simplest ways of integrating randomness in strategies. Considering such strategies instead of general (behavioural) strategies amounts to *limiting the randomisation power of strategies*.

A simple example. For the sake of illustration, let us consider the MDP depicted in Fig. 1.1a. This MDP models a situation where a person wants to go to work. They must choose between riding their bicycle or taking the train. However, the train may be delayed with high probability due to an ongoing strike. The goal of the commuter is twofold: maximise the likelihood of reaching work within 40 time units (to reach work on time) and do so as fast as possible on average.

We illustrate the set of expected payoffs for this situation in Fig. 1.1b. We label expected payoffs with their corresponding strategies. On the one hand, σ_{train} and σ_{bike} denote the pure strategies that always choose the action in the subscript. On the other hand, we denote by $\sigma_{\ell\mathsf{t+b}}$ the strategy that attempts to take the train ℓ times before choosing the bicycle.

This simple example highlights the need for randomisation and memory in multi-objective MDPs and POMDPs. When limited to pure strategies, for instance, it is not possible to reach work on time with probability exceeding 90% while guaranteeing an expected commute time lower than 27: any pure strategy whose payoff is not represented will reach work on time with a smaller probability than σ_{train}, and thus will not be satisfactory. However, as suggested by the point labelled by σ_{mix} on the illustration, these constraints can be satisfied by mixing σ_{train} with $\sigma_{\mathsf{2t+b}}$. In fact, here, all expected payoffs can be obtained by finite-support mixed strategies. We explain below how this property generalises. We remark however that, in general, the set of expected payoffs need not be a polytope like in this example, even in a finite MDP (e.g., see Ex. 8).

Our contributions. We study the structure of sets of expected payoffs in countable-state multi-objective POMDPs, focusing on the relationship between what can be obtained with pure and with randomised strategies. Our goal is to provide *general results*, i.e., that apply to a broad class of payoffs. We only consider *universally unambiguously integrable* payoffs (Sect. 3), i.e., payoffs that have a well-defined (possibly infinite) expectation no matter the strategy, a natural requirement when aiming to reason about expectations. We obtain finer results for *universally integrable* payoffs, i.e., payoffs whose expectation is *finite* under all strategies. Our results are already novel for MDPs (with perfect information). We also show that our results for universally integrable payoffs cannot be extended to universally unambiguously integrable payoffs already in *finite MDPs*.

First, we prove that for universally integrable (multi-dimensional) payoffs, for all strategies, there exists a *pure strategy* with a greater or equal expected payoff in the lexicographic sense (Thm. 4).

This first result serves as a building block to one of our main results: any expected payoff vector of a universally integrable (multi-dimensional) payoff is a *convex combination* of expected payoffs of pure strategies (Thm. 7). From a strategic perspective, this means that in multi-objective POMDPs, *finite-support mixed strategies* suffice to exactly obtain any (Pareto-optimal) expected payoff vector. As a corollary, we obtain that any extreme point of a set of expected payoffs of a universally integrable payoff can be obtained by a pure strategy (Cor. 11). These results generalise known properties for classical combinations of objectives for which the set of expected payoffs is a convex polytope (e.g., combinations of ω-regular specifications [32] and universally integrable total-reward payoffs [35]). While none of the previous properties generalise to the whole class of universally unambiguously integrable payoffs (Ex. 6 and 13), we prove that, for such payoffs, convex combinations of expected payoffs of pure strategies can be used to *approximate* any expected payoff (Thm. 14).

In both cases, we can bound the number of strategies that are mixed depending on the number of dimensions d. Depending on the setting, we can match or approximate expected payoffs by mixing no more than $d + 1$ strategies (Thm. 15).

We also provide a sufficient condition to guarantee that the set of expected payoffs is closed. We show that these sets are closed in finite POMDPs for *continuous* universally *square integrable* payoffs, which include *real-valued continuous payoffs* (Thm. 22).

To prove our results, we borrow techniques from several fields. We mainly rely on topology, properties of convex sets (separating and supporting hyperplanes) and the theory of the Lebesgue integral. We assume familiarity with basic properties of the Lebesgue integral. We recall most other required notions in the appendix of the full version of this paper [46].

Applicability. The class of *universally integrable* payoffs is large: it contains most classical payoffs. Indeed, all *bounded* payoffs are de facto in this class. For example, all indicators of objectives are in it. This means that all settings where one considers the *probabilities* of sets of plays (i.e., either an inherently qualitative objective or one arising from fixing a threshold for a quantitative payoff) are in it, for *any* definition of such sets of plays. Classical examples include ω-regular objectives [32], window objectives [11] or percentiles queries [53]. Bounded payoffs also encompass discounted-sum [57] and mean-payoff [15, 26] functions. Heterogeneous combinations, such as, e.g., combinations of energy and mean-payoff [14] also fit under this umbrella.

Payoffs that are completely out of the scope of our results – i.e., not even universally unambiguously integrable – include total payoff [33], average-energy [8] and shortest path [53], all with both positive and negative weights: the expectation is ill-defined when plays of positive and negative infinite payoffs both occur with positive probability. Let us focus on shortest-path objectives. When only non-negative weights are used, a classical restriction [53], the associated payoff is universally unambiguously integrable (but not universally integrable as plays of infinite payoffs may exist). Finally, when the state and action spaces are finite and targets are almost-surely visited (as in the above example), the payoff is continuous and universally square integrable [46, Appendix D].

Our results imply that it is sufficient to *mix a small number of pure strategies* – at most one more strategy than the number of payoffs – to obtain or approximate any expected payoff in multi-objective POMDPs with universally unambiguously integrable payoffs. The complexity of these mixed strategies depends on the considered setting and payoffs. On the one hand, for many classical payoffs that are at least universally unambiguously integrable in finite MDPs, extreme points of the set of expected payoffs can be obtained via pure finite-memory strategies (e.g., ω-regular objectives [32] or mean-payoff [15]). It follows that for these objectives, mixing over few *pure finite-memory* strategies is sufficient to fulfil any achievability requirement. In other words, one of the least expressive models of randomised finite-memory strategies (see the classification of [45]) suffices. On the other hand, in countable MDPs and (finite) POMDPs, infinite memory may be required to play optimally already for classical one-dimensional objectives (e.g., for ω-regular objectives, see [42] for countable MDPs and [17, 4] for finite POMDPs). It follows that complex strategies are needed in the multi-objective setting in general.

Our results also highlight the role of randomisation in strategies for multi-objective POMDPs: it is needed *only* to balance the payoffs on different dimensions in many cases. We note that randomisation can nonetheless be used to reduce strategy complexity in another way that is orthogonal to this work: it can contribute to reducing memory requirements for certain objectives (e.g., [19, 28, 26]).

Related work. We provide a few references, in addition to the main references cited previously. In our proof of Thm. 7, we invoke the hyperplane separation theorem. This theorem also plays a role in approximation schemes of the set of achievable vectors: see, e.g., [35, 50]; a unifying approach is presented in [49].

Specifications with multiple objectives have also been considered in two-player (non-stochastic) games on graphs (e.g., [34, 28]) and two-player stochastic games (e.g., [29, 1]). Closely related to multi-objective specifications are approaches that provide guarantees simultaneously in the worst case and the expected case [13, 27] or expectation optimisation of a payoff among strategies that satisfy some constraint of the expectation of another payoff (e.g., constrained POMDPs [40]) or among strategies that satisfy some probabilistic

constraint [22]. In [40, 22], the payoffs to be optimised payoffs are discounted-sum payoffs; our results imply that in these settings, if there exists an optimal strategy, there exists one that is obtained by mixing finitely many pure strategies.

Regarding randomisation in strategies, [20] studies when randomisation is helpful in strategies or in transitions of games. In particular, the authors show that if there exists an optimal strategy to maximise the probability of an event in a POMDP, then there exists a pure optimal strategy. This property is generalised by our result on lexicographic optimisation in POMDPs.

Outline. Due to space constraints, we only provide an overview of our results. Complete proofs can be found in the full version of this paper [46] Sect. 2 introduces notation and preliminary notions. We introduce payoff functions in Sect. 3 and establish relationships between expected payoffs of general strategies and expected payoffs of pure strategies. We study randomisation requirements for lexicographic optimisation in multi-objective POMDPs in Sect. 4. Sect. 5 presents our results regarding expected payoff sets and achievable sets. We study continuous payoff functions in Sect. 6. We conclude in Sect. 7.

2 Preliminaries

Notation. Let $A' \subseteq A$ and $B' \subseteq B$ be sets and $f\colon A \to B$ be a function. We let $\mathbb{1}_{A'}\colon A \to \{0,1\}$ denote the indicator function of A'. We let $\mathsf{Im}(f)$ denote the image of f. We let $f^{-1}(B') = \{a \in A \mid f(a) \in B'\}$ denote the inverse image of B' by f. For any $b \in B$, we write $f^{-1}(b)$ for $f^{-1}(\{b\})$. We let $|A|$ denote the cardinal of A.

We write $\mathbb{N}$ and $\mathbb{R}$ for the sets of natural and real numbers respectively. We let $\mathbb{N}_{>0} = \mathbb{N} \setminus \{0\}$ and $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$.

Vectors are written in boldface to distinguish them from scalars. Let $d \in \mathbb{N}_{>0}$. We let $\mathbf{0}_d$ and $\mathbf{1}_d \in \mathbb{R}^d$ be the vectors of $\mathbb{R}^d$ where all components are zero and one respectively. We omit the dimension subscript when the dimension is clear from the context. We usually denote linear forms (linear functions with co-domain $\mathbb{R}$) by x^* and y^*. Given $D \subseteq \mathbb{R}^d$, we let $\mathsf{conv}(D)$ denote its convex hull and $\mathsf{extr}(D)$ denote its set of extreme points.

Let $(X, \mathcal{T})$ be a topological space. For all $D \subseteq X$, we let $\mathsf{cl}(D)$ and $\mathsf{int}(D)$ denote the closure and interior of D. The boundary of $D \subseteq X$ is the set $\mathsf{bd}(D) = \mathsf{cl}(D) \setminus \mathsf{int}(D)$. For $D \subseteq \mathbb{R}^d$, we let $\mathsf{ri}(D)$ denote the relative interior of D (i.e., the interior of D relative to the smallest affine set in which D is included).

Probability. Let A be a countable set. We write $\mathcal{D}(A)$ for the set of distributions over A, i.e., the set of functions $\mu\colon A \to [0,1]$ such that $\sum_{a \in A} \mu(a) = 1$. The support of a distribution $\mu \in \mathcal{D}(A)$ is $\mathsf{supp}(\mu) = \{a \in A \mid \mu(a) > 0\}$.

Given a set B and a sigma-algebra $\mathcal{A}$ over B, we denote by $\mathcal{D}(B, \mathcal{A})$ the set of probability distributions over the measurable space $(B, \mathcal{A})$. Let $\mu \in \mathcal{D}(B, \mathcal{A})$ and $f\colon B \to \bar{\mathbb{R}}$ be a measurable function. We say that f is μ-integrable if it is integrable with respect to μ, i.e., if $\int_B |f| \mathrm{d}\mu \in \mathbb{R}$. We extend the Lebesgue integral to non-positive functions in the following way: if f is non-positive, we let $\int_B f \mathrm{d}\mu = -\int_B -f \mathrm{d}\mu$. If f is non-negative, non-positive or μ-integrable, we say that $\int_B f \mathrm{d}\mu$ is the μ-integral of f.

Ordering vectors. We consider the component-wise order and the lexicographic order over $\bar{\mathbb{R}}^d$. Let $\mathbf{q} = (q_j)_{1 \leq j \leq d}$ and $\mathbf{p} = (p_j)_{1 \leq j \leq d} \in \bar{\mathbb{R}}^d$. For the component-wise order, we write $\mathbf{q} \leq \mathbf{p}$ if and only if $q_j \leq p_j$ for all $1 \leq j \leq d$. For the lexicographic ordering over $\bar{\mathbb{R}}^d$, we

write $\mathbf{q} \leq_{\mathsf{lex}} \mathbf{p}$ if and only if $\mathbf{q} = \mathbf{p}$ or $q_j \leq p_j$ where $j = \min\{j' \leq d \mid q_{j'} \neq p_{j'}\}$. We write $\mathbf{q} <_{\mathsf{lex}} \mathbf{p}$ if $\mathbf{q} \leq_{\mathsf{lex}} \mathbf{p}$ and $\mathbf{q} \neq \mathbf{p}$. Given $D \subseteq \bar{\mathbb{R}}^d$, we say that $\mathbf{q} \in D$ is a *Pareto-optimal* element of D if it is maximal for the component-wise order.

Partially observable Markov decision processes. Formally, a *partially observable Markov decision process* (POMDP) is a tuple $\mathfrak{P} = (S, A, \delta, \mathcal{Z}, \mathsf{Obs})$ where S is a countable non-empty set of states, A is a countable non-empty set of actions, $\delta \colon S \times A \to \mathcal{D}(S)$ is a partial probabilistic transition function, $\mathcal{Z}$ is a countable set of observations and $\mathsf{Obs} \colon S \to \mathcal{Z}$ is an observation function. We say that $\mathfrak{P}$ is finite if S and A are finite. For $s \in S$, we let $A(s)$ denote the set of actions $a \in A$ such that $\delta(s, a)$ is defined. If $a \in A(s)$, we say that a is *enabled* in s. We assume that POMDPs have no deadlocks, i.e., there is an action enabled in each state. We also require that any two states that share the same observation have the same enabled actions, i.e., for all s, $t \in S$, $\mathsf{Obs}(s) = \mathsf{Obs}(t)$ implies $A(s) = A(t)$. We fix a POMDP $\mathfrak{P} = (S, A, \delta, \mathcal{Z}, \mathsf{Obs})$ for the rest of the paper.

A *play* of $\mathfrak{P}$ is an infinite sequence $\pi = s_0 a_0 s_1 a_1 \ldots \in (SA)^\omega$ such that $a_\ell \in A(s_\ell)$ and $s_{\ell+1} \in \mathsf{supp}(\delta(s_\ell, a_\ell))$ for all $\ell \in \mathbb{N}$. A *history* is a finite prefix of a play ending in a state. Let $\pi = s_0 a_0 s_1 a_1 \ldots$ be a play. We write $\pi_{\leq r}$ for the history $s_0 a_0 s_1 \ldots a_{r-1} s_r$. We also use this notation for prefixes of histories. We denote the suffix $s_r a_r s_{r+1} a_{r+1} \ldots$ of π by $\pi_{\geq r}$. For any history $h = s_0 a_0 s_1 \ldots s_r$, we let $\mathsf{last}(h) = s_r$. We write $\mathsf{Plays}(\mathfrak{P})$ and $\mathsf{Hist}(\mathfrak{P})$ for the sets of plays and histories of $\mathfrak{P}$. We extend the observation function Obs to histories as follows: for all histories $h = s_0 a_0 s_1 \ldots s_r$, we let $\mathsf{Obs}(h) = \mathsf{Obs}(s_0) a_0 \mathsf{Obs}(s_1) \ldots \mathsf{Obs}(s_r)$.

A *Markov decision process* (MDP) is a POMDP in which the observation function is the identity function. We denote MDPs as tuples $\mathfrak{M} = (S, A, \delta)$, i.e., we drop the observation space and observation function from the notation.

Behavioural strategies. A strategy (or policy) describes how to select actions based on the observation history of the ongoing play. A strategy may resort to randomisation. Formally, a *behavioural strategy* of $\mathfrak{P}$ is a function $\sigma \colon \mathsf{Obs}(\mathsf{Hist}(\mathfrak{P})) \to \mathcal{D}(A)$ such that for all $h \in \mathsf{Hist}(\mathfrak{P})$, $\mathsf{supp}(\sigma(\mathsf{Obs}(h))) \subseteq A(\mathsf{last}(h))$. To lighten notation, we abusively write $\sigma(h)$ for $\sigma(\mathsf{Obs}(h))$ in the sequel.

A strategy is *pure* if it uses no randomisation, i.e., if it maps all observation histories to Dirac distributions. We view pure strategies as functions $\sigma \colon \mathsf{Obs}(\mathsf{Hist}(\mathfrak{P})) \to A$. We let $\Sigma(\mathfrak{P})$ and $\Sigma_{\mathsf{pure}}(\mathfrak{P})$ respectively denote the set of all strategies of $\mathfrak{P}$ and the set of pure strategies of $\mathfrak{P}$.

A play $\pi = s_0 a_0 s_1 \ldots$ is *consistent* with a strategy σ if for all $\ell \in \mathbb{N}$, we have $a_\ell \in \mathsf{supp}(\sigma(\pi_{\leq \ell}))$. An *outcome* of a strategy is any play consistent with the strategy. Consistency of a history with a strategy is defined similarly.

A *cylinder set* is a set of continuations of a history; given $h = s_0 a_0 s_1 \ldots s_r \in \mathsf{Hist}(\mathfrak{P})$, we let $\mathsf{Cyl}(h) = \{\pi \in \mathsf{Plays}(\mathfrak{P}) \mid \pi_{\leq r} = h\}$. We let $\mathcal{A}_{\mathfrak{P}}$ denote the sigma-algebra generated by the cylinder sets. The distribution $\mathbb{P}^\sigma_s \in \mathcal{D}(\mathsf{Plays}(\mathfrak{P}), \mathcal{A}_{\mathfrak{P}})$ induced by a strategy σ from $s \in S$ is defined in the usual way.

Mixed strategies. There exists an alternative definition of randomised strategies: a *mixed strategy* is a distribution over pure strategies. Intuitively, when playing following a mixed strategy, a pure strategy is selected at random at the start of the process and then the play proceeds according to this pure strategy. The distribution $\mathbb{P}^\mu_s$ induced by a mixed strategy μ from $s \in S$ is defined, for all $\Omega \in \mathcal{A}_{\mathfrak{P}}$, by

$$\mathbb{P}^\mu_s(\Omega) = \int_{\tau \in \Sigma_{\mathsf{pure}}(\mathfrak{P})} \mathbb{P}^\tau_s(\Omega) \mathrm{d}\mu(\tau). \tag{2.1}$$

Kuhn's theorem. From a given initial state, any distribution over plays that can be induced by a mixed strategy can also be induced by a behavioural strategy and vice versa. This is a special case of Kuhn's theorem [2]. Due to the equivalent expressiveness of both strategy models, mixed and behavioural strategies can be used interchangeably. For the sake of conciseness, unless otherwise stated, by strategy, we mean a behavioural strategy in the sequel.

3 Payoffs and multi-objective POMDPs

Payoff functions and objectives. A *payoff function* (or payoff for short) is a measurable function $f\colon \mathsf{Plays}(\mathfrak{P}) \to \bar{\mathbb{R}}$; it quantifies the quality of plays. An *objective* is an element of $\mathcal{A}_{\mathfrak{P}}$; it represents a qualitative specification. To each objective Ω, we associate the payoff $\mathbb{1}_\Omega$.

Let f be a payoff, $\sigma \in \Sigma(\mathfrak{P})$ be a strategy and $s \in S$ be an initial state. The $\mathbb{P}_s^\sigma$-integral of f is only formally defined whenever f is non-negative, non-positive or $\mathbb{P}_s^\sigma$-integrable. If f is such a payoff, we let $\mathbb{E}_s^\sigma(f) = \int_{\pi\in\mathsf{Plays}(\mathfrak{P})} f(\pi)\mathrm{d}\mathbb{P}_s^\sigma(\pi)$; $\mathbb{E}_s^\sigma(f)$ is the expected payoff of the strategy σ from s (for f). We also generalise the notion of expected payoff to a broader class of payoff functions as follows. Let $f^+ = \max(f, 0)$ and $f^- = \max(-f, 0)$ denote the non-negative and non-positive parts of f. We say that f has an *unambiguous $\mathbb{P}_s^\sigma$-integral* if $\mathbb{E}_s^\sigma(f^+) \in \mathbb{R}$ or $\mathbb{E}_s^\sigma(f^-) \in \mathbb{R}$. If f has an unambiguous $\mathbb{P}_s^\sigma$-integral, we abuse notation and let $\mathbb{E}_s^\sigma(f) = \mathbb{E}_s^\sigma(f^+) - \mathbb{E}_s^\sigma(f^-)$. We reserve the notation $\mathbb{E}$ for the expectation of payoffs, and use integrals for other probability spaces.

We only consider payoffs for which the expected payoff is (unambiguously) defined for all strategies from all initial states. We say that f is *universally unambiguously integrable* if f has an unambiguous $\mathbb{P}_s^\sigma$-integral for all $\sigma \in \Sigma(\mathfrak{P})$ and $s \in S$. All non-negative and non-positive payoffs are universally unambiguously integrable.

Payoffs that are integrable no matter the strategy and initial state are of particular interest in the sequel. We say that f is *universally integrable* if it is $\mathbb{P}_s^\sigma$-integrable for all $\sigma \in \Sigma(\mathfrak{P})$ and $s \in S$. All bounded functions (e.g., indicators) are universally integrable.

Given a payoff function f, we say that a strategy σ is *optimal* from a state $s \in S$ if for all strategies τ, we have $\mathbb{E}_s^\sigma(f) \geq \mathbb{E}_s^\tau(f)$.

Expectations under mixed strategies. One of our main goals is to show that a simpler class of randomised strategies, i.e., finite-support mixed strategies, suffices when considering POMDPs with multiple payoffs. In our proofs, we rely on the following generalisation of Eq. (2.1).

▶ **Lemma 1.** *Let μ be a mixed strategy, $s \in S$ and f be a universally unambiguously integrable payoff. If $\inf_\tau \mathbb{E}_s^\tau(f) \geq 0$ or $\sup_\tau \mathbb{E}_s^\tau(f) \leq 0$ or f is $\mathbb{P}_s^\mu$-integrable, then the function $\Sigma_{\mathsf{pure}}(\mathfrak{P}) \to \bar{\mathbb{R}}\colon \tau \mapsto \mathbb{E}_s^\tau(f)$ is measurable and*

$$\mathbb{E}_s^\mu(f) = \int_{\tau\in\Sigma_{\mathsf{pure}}(\mathfrak{P})} \mathbb{E}_s^\tau(f)\mathrm{d}\mu(\tau).$$

Proof sketch. Eq. (2.1) implies that the result holds for indicators. We use linearity and the monotone convergence theorem to generalise the result to linear combinations of indicators, then non-negative payoffs and finally universally unambiguously integrable payoffs. ◀

Using Lem. 1, we can prove a characterisation of universally integrable payoffs.

▶ **Lemma 2.** *Let f be a payoff. Let $s \in S$. The following assertions are equivalent.*
1. *f is $\mathbb{P}_s^\sigma$-integrable for all $\sigma \in \Sigma(\mathfrak{P})$.*
2. *We have $\sup\{\mathbb{E}_s^\sigma(|f|) \mid \sigma \in \Sigma(\mathfrak{P})\} \in \mathbb{R}$.*
3. *We have $\sup\{\mathbb{E}_s^\sigma(|f|) \mid \sigma \in \Sigma_{\mathsf{pure}}(\mathfrak{P})\} \in \mathbb{R}$.*

In particular, f is universally integrable if and only if 2 (resp. 3) holds for all $s \in S$.

Proof sketch. The non-trivial part of the proof is showing that 1 and 3 imply 2. We prove this by contradiction. Assume that 2 does not hold. Then Lem. 1 implies that there are pure strategies τ with arbitrarily large $\mathbb{E}_s^\tau(|f|)$. We can mix such pure strategies to construct a mixed strategy μ such that $\mathbb{E}_s^\mu(|f|) = +\infty$. This shows that 1 and 3 do not hold. ◀

Lem. 1 does not apply to all universally unambiguously integrable payoffs f. The following result, which can be shown using Lem. 2, implies that for all initial states $s \in S$, there exists a constant $\alpha \in \mathbb{R}$ (specified below) such that the expectation of $f + \alpha$ from s is either non-negative for all strategies or non-positive for all strategies, and thus satisfies the assumptions of Lem. 1.

▶ **Lemma 3.** *Let f be a universally unambiguously integrable payoff. For all $s \in S$, $\inf_{\sigma\in\Sigma(\mathfrak{P})} \mathbb{E}_s^\sigma(f) \in \mathbb{R}$ or $\sup_{\sigma\in\Sigma(\mathfrak{P})} \mathbb{E}_s^\sigma(f) \in \mathbb{R}$.*

Let $s \in S$. If $\inf_{\sigma\in\Sigma(\mathfrak{P})} \mathbb{E}_s^\sigma(f) \in \mathbb{R}$, then the expectation of $f - \inf_{\sigma\in\Sigma(\mathfrak{P})} \mathbb{E}_s^\sigma(f)$ from s is non-negative under all strategies. Similarly, if $\sup_{\sigma\in\Sigma(\mathfrak{P})} \mathbb{E}_s^\sigma(f) \in \mathbb{R}$, the payoff $f - \sup_{\sigma\in\Sigma(\mathfrak{P})} \mathbb{E}_s^\sigma(f)$ has a non-positive expectation from s for all strategies. Shifting payoffs by a constant is done without loss of generality when concerned with expectations: any result regarding a payoff $f + \alpha$ can be recovered for f by linearity of the expectation.

Objective and payoff examples. We now present some classical objectives and payoff functions. First, we define reachability objectives, which require that a set of target states be visited. Formally, given a target $T \subseteq S$, we define the *reachability objective* $\mathsf{Reach}(T)$ as the set $\{s_0a_0s_1a_1 \ldots \in \mathsf{Plays}(\mathfrak{P}) \mid \exists\, \ell \in \mathbb{N},\, s_\ell \in T\}$.

We now introduce payoffs that aggregate weights (i.e., costs or rewards) on transitions along a play. We let $w\colon S \times A \to \mathbb{R}$ be a *weight function*. A *total-reward* payoff outputs the sum of weights along a play. Formally, we let $\mathsf{TRew}_w\colon \mathsf{Plays}(\mathfrak{P}) \to \bar{\mathbb{R}}$ be defined, for all plays $\pi = s_0a_0s_1a_1 \ldots$, by $\liminf_{r\to\infty} \sum_{\ell=0}^{r} w(s_\ell, a_\ell)$.

A *shortest-path* payoff can be seen as a quantitative variant of reachability and is defined as the accumulated sum of weights up to the first visit of a set of targets. Formally, given a target T, we let $\mathsf{SPath}_w^T\colon \mathsf{Plays}(\mathfrak{P}) \to \bar{\mathbb{R}}$ be defined by, for all plays $\pi = s_0a_0s_1a_1 \ldots$, $\mathsf{SPath}_w^T(\pi) = +\infty$ if $\pi \notin \mathsf{Reach}(T)$ and, otherwise, $\mathsf{SPath}_w^T(\pi) = \sum_{\ell=0}^{r-1} w(s_\ell, a_\ell)$ where $r = \min\{\ell \in \mathbb{N} \mid s_\ell \in T\}$. Traditionally, the goal is to minimise the shortest-path payoff (i.e., it is a cost function rather than a payoff), hence the infinite payoff whenever the target is not visited.

All total-reward and shortest-path payoffs we consider in upcoming examples are built on a non-negative weight function and are thus universally unambiguously integrable.

We now assume that w is bounded in absolute value (and can take both non-negative and non-positive values). A *discounted-sum payoff* is defined as an accumulated sum of weights multiplied by powers of a discount factor in $[0, 1[$. Formally, given a discount factor $\lambda \in [0, 1[$, we let $\mathsf{DSum}_w^\lambda\colon \mathsf{Plays}(\mathfrak{P}) \to \mathbb{R}$ be the payoff function defined by $\mathsf{DSum}_w^\lambda(\pi) = \sum_{\ell=0}^{\infty} \lambda^\ell w(s_\ell, a_\ell)$ for all plays $\pi = s_0a_0s_1a_1 \ldots \in \mathsf{Plays}(\mathfrak{P})$. The boundedness of w ensures that DSum_w^λ is well-defined, bounded and thus universally integrable.

Multiple payoffs. We present terminology and notation for multi-objective POMDPs, i.e., POMDPs with multiple payoffs. Let $d \in \mathbb{N}_{>0}$. We summarise d payoffs $f_1, \dots, f_d \colon \mathsf{Plays}(\mathfrak{P}) \to \bar{\mathbb{R}}$ as a multi-dimensional payoff $\bar{f} \colon \mathsf{Plays}(\mathfrak{P}) \to \bar{\mathbb{R}}^d$, and write $\bar{f} = (f_j)_{1 \le j \le d}$.

Let $\bar{f} = (f_j)_{1 \le j \le d}$ and let $s \in S$. We say that $\bar{f}$ is universally (resp. unambiguously) integrable whenever f_j is universally (resp. unambiguously) integrable for all $1 \le j \le d$. We now assume that $\bar{f}$ is universally unambiguously integrable (recall that we focus on such payoffs). Given $\Sigma \subseteq \Sigma(\mathfrak{P})$, we let $\mathsf{Pay}_s^{\Sigma}(\bar{f}) = \{\mathbb{E}_s^{\sigma}(\bar{f}) \mid \sigma \in \Sigma\}$ denote the set of expected payoffs of strategies in Σ from s. We let $\mathsf{Pay}_s(\bar{f})$ and $\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f})$ be shorthand for $\mathsf{Pay}_s^{\Sigma(\mathfrak{P})}(\bar{f})$ and $\mathsf{Pay}_s^{\Sigma_{\mathsf{pure}}(\mathfrak{P})}(\bar{f})$ respectively. There can be several Pareto-optimal payoffs in expected payoff sets.

In multi-objective optimisation, the goal is often to ensure a given threshold on each dimension. This is formalised by *achievable vectors*: $\mathbf{q} \in \bar{\mathbb{R}}^d$ is *achievable* (from s) if there exists $\sigma \in \Sigma(\mathfrak{P})$ such that $\mathbf{q} \le \mathbb{E}_s^{\sigma}(\bar{f})$. For any $\Sigma \subseteq \Sigma(\mathfrak{P})$, we let $\mathsf{Ach}_s^{\Sigma}(\mathbf{q}) = \{\mathbf{q} \in \bar{\mathbb{R}}^d \mid \exists\, \sigma \in \Sigma,\ \mathbf{q} \le \mathbb{E}_s^{\sigma}\}$. We define $\mathsf{Ach}_s(\bar{f})$ and $\mathsf{Ach}_s^{\mathsf{pure}}(\bar{f})$ as above.

Convex combinations (without zero coefficients) of elements of $\mathsf{Pay}_s(\bar{f})$ are well-defined (e.g., by Lem. 3) and are themselves in $\mathsf{Pay}_s(\bar{f})$: the expectation of a convex combination of mixed strategies is a convex combination of the expectations of these strategies. In light of this, we (abusively) let $\mathsf{conv}(\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f}))$ denote the set of expected payoffs of mixed strategies with a finite support. We remark that $\mathsf{Pay}_s(\bar{f}) \cap \mathbb{R}^d$ is convex.

We also consider lexicographic optimisation in POMDPs. A strategy σ is *lexicographically optimal* if $\mathbb{E}_s^{\sigma}(\bar{f})$ is the lexicographic maximum of $\mathsf{Pay}_s(\bar{f})$. This notion generalises the notion of optimal strategies for one-dimensional payoffs.

4 Lexicographic POMDPs

In this section, we consider randomisation requirements for lexicographic optimisation of multiple payoff functions in POMDPs. We fix a d-dimensional payoff function $\bar{f} = (f_j)_{1 \le j \le d}$.

The following theorem states that randomisation is not useful for universally integrable payoffs.

▶ **Theorem 4.** *Assume that $\bar{f}$ is universally integrable. Let σ be a strategy and $s \in S$. There exists a pure strategy τ such that $\mathbb{E}_s^{\sigma}(\bar{f}) \le_{\mathsf{lex}} \mathbb{E}_s^{\tau}(\bar{f})$.*

Proof sketch. We first prove that the Lebesgue integral is compatible with the lexicographic order over $\bar{\mathbb{R}}^d$: for any two random vectors $\mathcal{X} = (X_1, \dots, X_d)$ and $\mathcal{Y} = (Y_1, \dots, Y_d)$ of a probability space $(B, \mathcal{A}, \mu)$, if $\mathcal{X} <_{\mathsf{lex}} \mathcal{Y}$ almost-surely, then $\int \mathcal{X} \mathrm{d}\mu <_{\mathsf{lex}} \int \mathcal{Y} \mathrm{d}\mu$. The argument is as follows. Letting

$$j^\star = \min\{1 \le j \le d \mid \mu(X_j \neq Y_j) > 0\},$$

we can show that $\int X_j \mathrm{d}\mu = \int Y_j \mathrm{d}\mu$ for all $1 \le j < j^\star$ and, by exploiting the compatibility of the Lebesgue integral with the order of $\bar{\mathbb{R}}$, that $\int X_{j^\star} \mathrm{d}\mu < \int Y_{j^\star} \mathrm{d}\mu$. This shows that $\int \mathcal{X} \mathrm{d}\mu <_{\mathsf{lex}} \int \mathcal{Y} \mathrm{d}\mu$.

We then assume towards a contradiction that there is no suitable pure strategy τ. By Kuhn's theorem and Lem. 1, we can write $\mathbb{E}_s^{\sigma}(\bar{f})$ as an integral over pure expected payoffs. The contradiction $\mathbb{E}_s^{\sigma}(\bar{f}) <_{\mathsf{lex}} \mathbb{E}_s^{\sigma}(\bar{f})$ follows from the above. ◀

A direct corollary of Thm. 4 is the following.

▶ **Corollary 5.** *Assume that $\bar{f}$ is universally integrable. For all $s \in S$, if there exists a lexicographically optimal strategy from s, then there exists a pure lexicographically optimal strategy.*

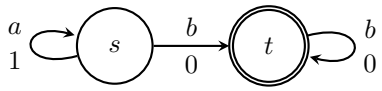


Figure 4.1 An MDP with deterministic transitions. The doubly circled state denotes a target for a reachability objective.

We now show that Thm. 4 does not generalise to all universally unambiguously integrable payoffs already in one-dimensional finite MDPs.

▶ **Example 6.** We consider the MDP $\mathfrak{M}$ depicted in Fig. 4.1. Let w denote the illustrated weight function. Let f be defined by

$$f(\pi) = \begin{cases} \mathsf{TRew}_w(\pi) & \text{if } \pi \in \mathsf{Reach}(\{t\}) \\ 0 & \text{otherwise.} \end{cases}$$

We show that randomisation is necessary to play (lexicographically) optimally from s.

Since transitions of $\mathfrak{M}$ are deterministic, $\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f})$ is the set of payoffs of plays from s. We introduce notation for these plays: for all $r \in \mathbb{N}$, let $\pi_r = (sa)^r s(bt)^\omega$ and let $\pi_\infty = (sa)^\omega$. It follows that $\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f}) = \{\bar{f}(\pi_r) \mid r \in \mathbb{N} \cup \{\infty\}\} = \mathbb{N}$. There also exists a randomised strategy whose expected payoff from s is $+\infty$. For each $r \in \mathbb{N}$, let τ_r be a pure strategy whose outcome from s is π_r. The mixed strategy μ such that, for all $r \in \mathbb{N}$, μ assigns probability $2^{-(r+1)}$ to τ_{2^r}, satisfies $\mathbb{E}_s^\mu(\bar{f}) = \sum_{r=0}^{\infty} \mu(\tau_{2^r}) \cdot \bar{f}(\pi_{2^r}) = +\infty$. This shows that randomisation can be useful when dealing with non-universally-integrable payoffs. ⌟

5 The structure of expected payoff sets

In this section, we study the relationship between the set of expected payoffs of pure strategies and of general strategies through finite-support mixed strategies. We consider universally integrable payoffs in Sect. 5.1. We consider the more general universally unambiguously integrable payoffs in Sect. 5.2. Finally, in Sect. 5.3, we provide bounds on the size of supports of the finite-support mixed strategies appearing in the results of the previous sections.

We fix a d-dimensional payoff function $\bar{f} = (f_j)_{1 \le j \le d}$.

5.1 Universally integrable payoffs

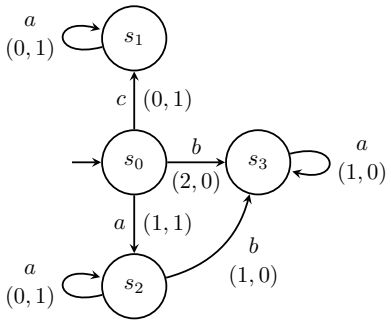
Throughout this section, we assume that $\bar{f}$ is universally integrable. Our main result is the following.

▶ **Theorem 7.** *Assume that $\bar{f}$ is universally integrable. For all $s \in S$, we have $\mathsf{Pay}_s(\bar{f}) = \mathsf{conv}(\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f}))$. In other words, the expected payoff of any strategy is also the expected payoff of a finite-support mixed strategy.*

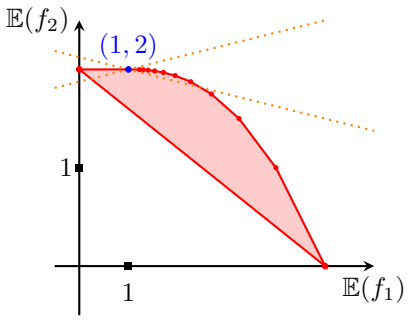
The main tools used in our approach are lexicographic optimisation, and supporting and separating hyperplanes. Fix $s \in S$. To illustrate the main ideas of our proof, we first provide a proof sketch with the assumption that $\mathsf{Pay}_s(\bar{f})$ is a polytope. We comment on the limitations of this argument and how to generalise it below.

Proof sketch (polytope). Let $C^{\mathsf{pure}} = \mathsf{conv}(\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f}))$. The convexity of $\mathsf{Pay}_s(\bar{f})$ implies that $C^{\mathsf{pure}} \subseteq \mathsf{Pay}_s(\bar{f})$. We thus focus on the other inclusion in the following.

All points of a polytope lie in the relative interior of one of its faces. We reason on the relative interior of each face individually. Recall that a face of $\mathsf{Pay}_s(\bar{f})$ is an intersection $\mathsf{Pay}_s(\bar{f}) \cap V$ where V is $\mathbb{R}^d$ or a supporting hyperplane of $\mathsf{Pay}_s(\bar{f})$. We fix such a V. We claim that the face $F = V \cap \mathsf{Pay}_s(\bar{f})$ of $\mathsf{Pay}_s(\bar{f})$ has the same relative interior as $V \cap C^{\mathsf{pure}}$.



(a) An MDP with deterministic transitions. Pairs next to actions represent two-dimensional weights.



(b) The set of expected payoffs for the MDP of Fig. 5.1a for the payoff $f_1 = \mathsf{DSum}^{3/4}_{w_1}$ and $f_2 = \mathsf{DSum}^{1/2}_{w_2}$. This set has infinitely many extreme points close to $(1, 2)$.

Figure 5.1 An MDP with a two-dimensional discounted-sum payoff $\bar{f}$ such that $\mathsf{Pay}_{s_0}(\bar{f})$ is not a polytope.

A convex set of $\mathbb{R}^d$ has the same relative interior as its closure. Therefore, it suffices to show that F and $V \cap C^{\mathsf{pure}}$ have the same closure. We proceed by contradiction: if the closures differ, then it follows from $V \cap C^{\mathsf{pure}} \subseteq F$ that we must have $\mathsf{cl}(F) \not\subseteq \mathsf{cl}(V \cap C^{\mathsf{pure}})$ and therefore there exists $\mathbf{q} \in F \setminus \mathsf{cl}(V \cap C^{\mathsf{pure}})$. Let σ be a strategy such that $\mathbf{q} = \mathbb{E}^{\sigma}_{s}(\bar{f})$. Let x^* be a linear form that is zero if $V = \mathbb{R}^d$ and, otherwise, that witnesses that V is a supporting hyperplane of $\mathsf{Pay}_s(\bar{f})$, i.e., such that $x^*(\mathbf{p}) = \max x^*(\mathsf{Pay}_s(\bar{f}))$ for all $\mathbf{p} \in V$. In particular, $\mathbf{q} \in V$ implies that σ is optimal for $x^* \circ \bar{f}$ from s. Let y^* be a linear form such that $y^*(\mathbf{q}) > y^*(\mathbf{p})$ for all $\mathbf{p} \in \mathsf{cl}(V \cap C^{\mathsf{pure}})$. The existence of such a form is ensured by the hyperplane separation theorem (e.g., [55, Cor. 11.4.2]). We obtain that for all pure strategies τ that are optimal for $x^* \circ \bar{f}$ from s, we have $\mathbb{E}^{\sigma}_{s}(y^* \circ \bar{f}) > \mathbb{E}^{\tau}_{s}(y^* \circ \bar{f})$. It follows that, with respect to the lexicographic order, σ performs strictly better from s than any pure strategy for the payoff $(x^*, y^*) \circ \bar{f}$. This contradicts Thm. 4, and ends the argument. ◀

For non-polytope payoff sets, we replace the faces (i.e., points that maximise a linear form) in the above reasoning with intersections of $\mathsf{Pay}_s(\bar{f})$ with affine sub-spaces such that the elements of this intersection lexicographically maximise a linear function over $\mathsf{Pay}_s(\bar{f})$. Linear forms may no longer suffice to reason on relative interiors as above: given a point $\mathbf{q} \in \mathsf{Pay}_s(\bar{f})$ in the boundary, there may not exist a hyperplane V supporting $\mathsf{Pay}_s(\bar{f})$ at $\mathbf{q}$ such that $\mathbf{q}$ is in the relative interior of $\mathsf{Pay}_s(\bar{f}) \cap V$. We illustrate this below.

▶ **Example 8.** Consider the MDP depicted in Fig. 5.1a and the two discounted payoff function $f_1 = \mathsf{DSum}^{3/4}_{w_1}$ and $f_2 = \mathsf{DSum}^{1/2}_{w_2}$. We illustrate $\mathsf{Pay}_{s_0}(f_1, f_2)$ in Fig. 5.1b. Formally, we can show that $\mathsf{Pay}_{s_0}(f_1, f_2)$ is

$$\mathsf{conv}\left(\{(0,2),(1,2)\} \cup \left\{\left(1 + \frac{3^r}{4^{r-1}}, 2 - \frac{1}{2^{r-1}}\right) \mid r \in \mathbb{N}\right\}\right).$$

We note that $\mathsf{Pay}_{s_0}(f_1, f_2)$ is not a polytope.

The vector $\mathbf{q} = (1, 2)$ is in $\mathsf{bd}(\mathsf{Pay}_{s_0}(f_1, f_2))$ and the only supporting hyperplane at $\mathbf{q}$ is the line of equation $y = 2$. Intuitively, any line passing through $\mathbf{q}$ at an angle intersects the interior of $\mathsf{Pay}_{s_0}(f_1, f_2)$ (see the dotted lines in the figure) and therefore is not a supporting hyperplane.

We observe that $\mathbf{q}$ is the unique element of $\mathsf{Pay}_{s_0}(f_1, f_2)$ that lexicographically maximises the linear function $L\colon (x, y) \mapsto (y, x)$ over $\mathsf{Pay}_{s_0}(f_1, f_2)$. Furthermore, $\mathbf{q} \in \mathsf{ri}(\mathsf{Pay}_{s_0}(f_1, f_2) \cap L^{-1}(L(\mathbf{q}))) = \{\mathbf{q}\}$. We can thus adapt the argument used for polytope sets to $\mathbf{q}$ by using L in place of a linear form. ⌟

The following lemma states that for all $\mathbf{q} \in \mathsf{Pay}_s(\bar{f})$, there exists a linear function that can be used instead of a linear form to generalise the approach used to prove Thm. 7 for polytopes to general payoff sets.

▶ **Lemma 9.** *Let $D \subseteq \mathbb{R}^d$ be convex and let $\mathbf{q} \in D$. There exists $d' \leq d$ and a linear function $L \colon \mathbb{R}^d \to \mathbb{R}^{d'}$ such that $L(\mathbf{q})$ is the lexicographic maximum of $L(D)$ and $\mathbf{q} \in \mathsf{ri}(D \cap L^{-1}(L(\mathbf{q})))$.*

Proof sketch. If $\mathbf{q} \in \mathsf{ri}(D)$, it suffices to let L be the constant zero form. Assume now that $\mathbf{q} \notin \mathsf{ri}(D)$. We repeatedly apply the supporting hyperplane theorem (e.g., [55, Thm. 11.6]) to define a sequence of linear forms $x_1^*, \dots, x_j^*$ such that $\mathbf{q}$ is lexicographically optimal for

$$L_j \colon \mathbf{v} \mapsto (x_1^*(\mathbf{v}), \dots, x_j^*(\mathbf{v})).$$

We stop the induction once $\mathbf{q} \in \mathsf{ri}(D \cap L_j^{-1}(L_j(\mathbf{q})))$. When this holds, we conclude by letting $L = L_j$.

We explain how to guarantee termination of this construction. Assume that L_j has been constructed (for the base case, let L_0 be the constant zero function) and that the termination condition does not hold. We apply the supporting hyperplane theorem to obtain a form x_{j+1}^* defining a hyperplane supporting $D \cap L_j^{-1}(L_j(\mathbf{q}))$ at $\mathbf{q}$ that does not include $D \cap L_j^{-1}(L_j(\mathbf{q}))$. It follows that the dimension of the kernel of L_{j+1} is one less than that of L_j. If the induction proceeds for d iterations, then L_d is bijective and the termination condition follows trivially as $L_d^{-1}(L_d(\mathbf{q})) = \{\mathbf{q}\}$ and singleton sets are equal to their relative interior. ◀

▶ Remark 10. By Lem. 9, for all $\mathbf{q} \in \mathsf{extr}(\mathsf{Pay}_s(\bar{f}))$, there exists a linear function L that reaches its lexicographic maximum over $\mathsf{Pay}_s(\bar{f})$ only at $\mathbf{q}$. This can be used in conjunction with Thm. 4 to prove that all extreme points of $\mathsf{Pay}_s(\bar{f})$ can be attained by pure strategies. We remark however that this is not enough to prove Thm. 7, as payoff sets need not be the convex hull of their extreme points.

We now generalise the proof sketch of Thm. 7 formally.

Proof of Thm. 7. Let $s \in S$ and let $C^{\mathsf{pure}} = \mathsf{conv}(\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f}))$. The inclusion $C^{\mathsf{pure}} \subseteq \mathsf{Pay}_s(\bar{f})$ follows from the convexity of $\mathsf{Pay}_s(\bar{f})$. We focus on the other inclusion.

Let $\mathbf{q} \in \mathsf{Pay}_s(\bar{f})$. Let L be a linear function given by Lem. 9 for $\mathbf{q}$ and $D = \mathsf{Pay}_s(\bar{f})$, i.e., $L(\mathbf{q})$ is the lexicographic maximum of $L(\mathsf{Pay}_s(\bar{f}))$ and $\mathbf{q} \in \mathsf{ri}(\mathsf{Pay}_s(\bar{f}) \cap V)$, where $V = L^{-1}(L(\mathbf{q}))$.

To prove that $\mathbf{q} \in C^{\mathsf{pure}}$, it suffices to show that

$$\mathsf{ri}(\mathsf{Pay}_s(\bar{f}) \cap V) = \mathsf{ri}(C^{\mathsf{pure}} \cap V).$$

As these sets are convex, it suffices to show that

$$\mathsf{cl}(\mathsf{Pay}_s(\bar{f}) \cap V) = \mathsf{cl}(C^{\mathsf{pure}} \cap V),$$

as convex subsets of $\mathbb{R}^d$ have the same relative interior as their closure [55, Thm. 6.3].

The inclusion $C^{\mathsf{pure}} \subseteq \mathsf{Pay}_s(\bar{f})$ implies that $\mathsf{cl}(C^{\mathsf{pure}} \cap V) \subseteq \mathsf{cl}(\mathsf{Pay}_s(\bar{f}) \cap V)$. To prove the other inclusion, it suffices to show that $\mathsf{Pay}_s(\bar{f}) \cap V \subseteq \mathsf{cl}(C^{\mathsf{pure}} \cap V)$. Assume towards a contradiction that there exists $\mathbf{p} \in \mathsf{Pay}_s(\bar{f}) \cap V \setminus \mathsf{cl}(C^{\mathsf{pure}} \cap V)$, and let σ be a strategy such that $\mathbf{p} = \mathbb{E}_s^\sigma(\bar{f})$. We observe that $\mathbf{p} \in V$ implies that σ is lexicographically optimal from s for $L \circ \bar{f}$. We apply the hyperplane separation theorem (e.g., [55, Cor. 11.4.2]) to $\{\mathbf{p}\}$ and $\mathsf{cl}(C^{\mathsf{pure}} \cap V)$ to obtain a linear form x^* such that $x^*(\mathbf{p}) > x^*(\mathbf{p}')$ for all $\mathbf{p}' \in C^{\mathsf{pure}} \cap V$. We conclude that for all pure strategies τ, we have $(L, x^*)(\mathbf{p}) = \mathbb{E}_s^\sigma((L, x^*) \circ \bar{f}) >_{\mathsf{lex}} \mathbb{E}_s^\tau((L, x^*) \circ \bar{f})$, which contradicts Thm. 4 and ends the proof. ◀

We now formulate two corollaries of Theorem 7. The first one relates to extreme points of payoffs sets.

▶ **Corollary 11.** *Assume that $\bar{f}$ is universally integrable. For all $s \in S$, $\mathsf{extr}(\mathsf{Pay}_s(\bar{f})) \subseteq \mathsf{Pay}_s^{\mathsf{pure}}(\bar{f})$, i.e., all extreme points of $\mathsf{Pay}_s(\bar{f})$ are payoffs of pure strategies.*

Proof. By Thm. 7, $\mathsf{Pay}_s(\bar{f}) = \mathsf{conv}(\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f}))$. If some extreme point $\mathbf{q}$ of $\mathsf{conv}(\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f}))$ were not in $\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f})$, then $\mathbf{q}$ would be a convex combination of other elements of $\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f})$, and thus $\mathbf{q}$ would not be an extreme point of $\mathsf{Pay}_s(\bar{f})$. ◀

The second corollary below follows from the convex hull of a compact subset of $\mathbb{R}^d$ being closed.

▶ **Corollary 12.** *Assume that $\bar{f}$ is universally integrable. For all $s \in S$, if $\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f})$ is closed, then $\mathsf{Pay}_s(\bar{f})$ is compact.*

Proof. Let $s \in S$ such that $\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f})$ is closed. By Lem. 2, $\bar{f}$ is universally integrable if and only if $\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f})$ is bounded. It follows that $\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f})$ is compact. Thm. 7 ensures that $\mathsf{Pay}_s(\bar{f}) = \mathsf{conv}(\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f}))$, and thus $\mathsf{Pay}_s(\bar{f})$ is compact (the convex hull of a compact subset of $\mathbb{R}^d$ is compact). ◀

We close this section by observing that Thm. 7 does not generalise to universally unambiguously integrable payoffs.

▶ **Example 13** (Ex. 6 continued)**.** We consider the MDP $\mathfrak{M}$ and payoff function f of Ex. 6. We have shown that $\mathsf{Pay}_s^{\mathsf{pure}}(f) = \mathbb{N}$ and that $+\infty \in \mathsf{Pay}_s(f)$. It follows that $\mathsf{Pay}_s(f)$ is not included in $\mathsf{conv}(\mathsf{Pay}_s^{\mathsf{pure}}(f)) = [0, +\infty[$. ⌟

5.2 Universally unambiguously integrable payoffs

We now relax the assumption that $\bar{f}$ is universally integrable and assume that $\bar{f}$ is universally unambiguously integrable. We establish an approximation variant of Thm. 7 that states that finite-support mixed strategies can be used to approximate the payoff of any arbitrary strategy.

▶ **Theorem 14.** *Assume that $\bar{f}$ is universally unambiguously integrable. Let $s \in S$. We have $\mathsf{cl}(\mathsf{Pay}_s(\bar{f})) = \mathsf{cl}(\mathsf{conv}(\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f})))$. In particular, for all strategies σ, all $\varepsilon > 0$ and all $M \in \mathbb{R}$, there exist finitely many pure strategies $\tau_1, \ldots, \tau_n$ and $\alpha_1, \ldots, \alpha_n \in [0,1]$ with $\alpha_1 + \cdots + \alpha_n = 1$ such that for all $1 \leq j \leq d$:*

- *if $\mathbb{E}_s^\sigma(f_j) = +\infty$, then $\sum_{m=1}^n \alpha_m \mathbb{E}_s^{\tau_m}(f_j) \geq M$,*
- *if $\mathbb{E}_s^\sigma(f_j) = -\infty$, then $\sum_{m=1}^n \alpha_m \mathbb{E}_s^{\tau_m}(f_j) \leq -M$, and,*
- *otherwise, if $\mathbb{E}_s^\sigma(f_j) \in \mathbb{R}$,*

$$\mathbb{E}_s^\sigma(f_j) - \varepsilon \leq \sum_{m=1}^n \alpha_m \mathbb{E}_s^{\tau_m}(f_j) \leq \mathbb{E}_s^\sigma(f_j) + \varepsilon.$$

Proof sketch. Fix $s \in S$, a mixed strategy μ, $\varepsilon > 0$ and $M \in \mathbb{R}$. We reason on the integral formulation of $\mathbb{E}_s^\mu(\bar{f})$ from Lem. 1. Even though Lem. 1 is not applicable to all payoffs, Lem. 3 implies that there exists a vector $\mathbf{v}$ such that $\bar{f} + \mathbf{v}$ satisfies the assumptions of Lem. 1. We can then recover the result for the original payoff using the linearity of the expectation. We thus assume without loss of generality that Lem. 1 applies to all payoffs $f_1, \ldots, f_d$.

For all $1 \leq j \leq d$, we let

$$X_j \colon \Sigma_{\mathsf{pure}}(\mathfrak{P}) \to \bar{\mathbb{R}} \colon \tau \mapsto \mathbb{E}_s^\tau(f_j)$$

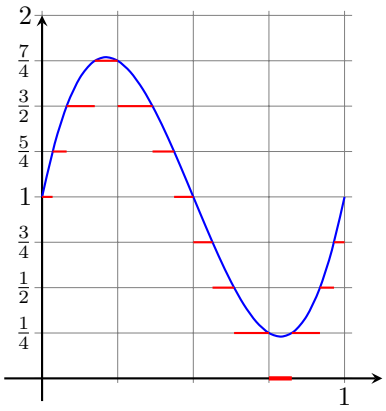


Figure 5.2 Illustration of the approach used to construct the approximation $\mathcal{Y}$ of $\mathcal{X}$ adapted to a function over $[0,1]$. We round the blue function down to the closest multiple of $\frac{1}{4}$ to obtain the red function. This yields a linear combination of indicators that is $\frac{1}{4}$-close in all points to the function in blue.

and $\mathcal{X} = (X_1, \dots, X_d)$. By Lem. 1, we have $\mathbb{E}^{\mu}_{s}(\bar{f}) = \int_{\Sigma_{\mathsf{pure}}(\mathfrak{P})} \mathcal{X} \mathrm{d}\mu$. We sketch the proof when the X_j are (μ-almost-surely) real-valued functions. We comment on the generalisation later.

The broad idea is as follows. First, we approximate $\mathcal{X}$ with a multivariate random variable $\mathcal{Y} = (Y_1, \dots, Y_d)$ over $\Sigma_{\mathsf{pure}}(\mathfrak{P})$. We then approximate the integral of $\mathcal{Y}$ with a convex combination $\sum_{m=0}^{n} \alpha_m \mathcal{Y}(\tau_m) \in \mathsf{conv}(\mathsf{Im}(\mathcal{Y}))$. Finally, we derive the convex combination $\sum_{m=0}^{n} \alpha_m \mathcal{X}(\tau_m)$ from the previous one. The successive approximations above ensure that the last convex combination respects the claims of the theorem.

We now expand this idea. First, we construct $\mathcal{Y}$ such that

$$\mathcal{X} - \frac{\varepsilon}{3}\mathbf{1} \leq \mathcal{Y} \leq \mathcal{X} + \frac{\varepsilon}{3}\mathbf{1}.$$

It follows that, for all $1 \leq j \leq d$, $\int_{\Sigma_{\mathsf{pure}}(\mathfrak{P})} Y_j \mathrm{d}\mu$ is equal to $\mathbb{E}^{\mu}_{s}(f_j)$ whenever $\mathbb{E}^{\mu}_{s}(f_j) \in \{-\infty, +\infty\}$ and otherwise is $\frac{\varepsilon}{3}$-close. Intuitively, for all $1 \leq j \leq d$, we define Y_j by rounding X_j (up or down) to the closest multiple of 2^{-k} for $k \in \mathbb{N}$ such that $2^{-k} \leq \frac{\varepsilon}{3}$. In particular, $\mathcal{Y}$ is an infinite linear combination of indicators (see Fig. 5.2). Its integral is thus (informally) an infinite convex combination of images of $\mathcal{Y}$: there are sequences $(\beta_m)_{m\in\mathbb{N}} \subseteq [0,1]$ and $(\tau_m)_{m\in\mathbb{N}} \subseteq \Sigma_{\mathsf{pure}}(\mathfrak{P})$ such that $\sum_{m=0}^{\infty} \beta_m = 1$ and

$$\int_{\tau \in \Sigma_{\mathsf{pure}}(\mathfrak{P})} \mathcal{Y} \mathrm{d}\mu(\tau) = \sum_{m\in\mathbb{N}} \beta_m \mathcal{Y}(\tau_m).$$

We derive a sequence $(\mathbf{p}^{(n)})_{n\in\mathbb{N}}$ in $\mathsf{conv}(\mathsf{Im}(\mathcal{Y}(\tau_m)))$ that converges to $\int_{\Sigma_{\mathsf{pure}}(\mathfrak{P})} \mathcal{Y} \mathrm{d}\mu$ from this series: we let

$$\mathbf{p}^{(n)} = \sum_{m=0}^{n} \beta_m \mathcal{Y}(\tau_m) + \left(1 - \sum_{m=0}^{n} \beta_m\right) \mathcal{Y}(\tau_0)$$

for all $n \in \mathbb{N}$.

Fix $n \in \mathbb{N}$ large enough such that, for all $1 \leq j \leq d$, component j of $\mathbf{p}^{(n)}$ is $\frac{\varepsilon}{3}$-close to $\int_{\tau \in \Sigma_{\mathsf{pure}}(\mathfrak{P})} Y_j \mathrm{d}\mu(\tau)$ if it is a real number or greater than $M + \varepsilon$ in absolute value otherwise.

The convex combination

$$\mathbf{q} = \sum_{m=0}^{n} \beta_m \mathbb{E}_s^{\tau_m}(\bar{f}) + \left(1 - \sum_{m=0}^{n} \beta_m\right) \mathbb{E}_s^{\tau_0}(\bar{f})$$

is a satisfactory convex combination with respect to the claim of the theorem. Let $1 \leq j \leq d$. If $\mathbb{E}_s^{\mu}(f_j) = +\infty$, we obtain that $q_j \geq M$. Similarly, if $\mathbb{E}_s^{\mu}(f_j) = -\infty$, we obtain that $q_j \leq -M$. Otherwise, we have that q_j is ε-close to $\mathbb{E}_s^{\mu}(f_j)$.

We now briefly discuss the case where some X_j is not μ-almost-surely real-valued. For the sake of illustration, we assume that this only applies to $j = d$ and that $X_d \geq 0$. Therefore, we have $\mathbb{E}_s^{\mu}(f_d) = +\infty$ and there exists some $\tau \in \Sigma_{\mathsf{pure}}(\mathfrak{P})$ such that $\mathbb{E}_s^{\tau}(f_d) = +\infty$ and $\mathbb{E}_s^{\tau}(f_j) \in \mathbb{R}$ for all $1 \leq j \leq d-1$. Let $\tau_1, \ldots, \tau_n$ and $\alpha_1, \ldots, \alpha_n$ given by the theorem for $(f_1, \ldots, f_{d-1})$, $\frac{\varepsilon}{2}$ and $M + \varepsilon$. Let $\mathbf{q} = \sum_{m=1}^{n} \alpha_m \mathbb{E}_s^{\tau_m}(\bar{f})$. By choosing $\eta \in]0,1[$ such that all components of $\mathbb{E}_s^{\tau}(\bar{f})$ other than the last have absolute value no more than $\frac{\varepsilon}{2}$ and the finite components of $(1-\eta)\mathbf{q}$ are $\frac{\varepsilon}{2}$-close to the corresponding components of $\mathbf{q}$, we obtain a suitable convex combination in the form of $\eta \mathbb{E}_s^{\tau}(\bar{f}) + (1-\eta)\mathbf{q}$. ◀

5.3 Bounding the support of mixed strategies

Thm. 7 and 14 state that it suffices to mix finitely many pure strategies to respectively match or approximate the payoff of any strategy. In this section, we provide upper bounds on the number of pure strategies to mix. Our result is analogous to Carathéodory's theorem for convex hulls (e.g., [55, Thm. 17.1]). Carathéodory's theorem implies that for all subsets D in a d-dimensional vector space, any element of $\mathsf{conv}(D)$ can be written as a convex combination of no more than $d+1$ elements of D. Our main result is the following.

▶ **Theorem 15.** *Assume that $\bar{f}$ is universally unambiguously integrable. Let $s \in S$, $\sigma_1, \ldots, \sigma_n$ be pure strategies and $\alpha_1, \ldots, \alpha_n \in [0,1]$ such that $\alpha_1 + \cdots + \alpha_n = 1$. There exist $\beta_1, \ldots, \beta_n \in [0,1]$ with $\beta_1 + \cdots + \beta_n = 1$ and $|\{1 \leq m \leq n \mid \beta_m \neq 0\}| \leq d+1$, and $\gamma_1, \ldots, \gamma_n \in [0,1]$ with $\gamma_1 + \cdots + \gamma_n = 1$ and $|\{1 \leq m \leq n \mid \gamma_m \neq 0\}| \leq d$ such that*

$$\sum_{m=1}^{n} \alpha_m \mathbb{E}_s^{\sigma_m}(\bar{f}) = \sum_{m=1}^{n} \beta_m \mathbb{E}_s^{\sigma_m}(\bar{f}) \leq \sum_{m=1}^{n} \gamma_m \mathbb{E}_s^{\sigma_m}(\bar{f}).$$

Proof sketch. Let

$$\mathbf{q} = (q_j)_{1 \leq j \leq d} = \sum_{m=1}^{n} \alpha_m \mathbb{E}_s^{\sigma_m}(\bar{f}).$$

First, assume that $\mathbf{q} \in \mathbb{R}^d$, i.e., all strategies $\sigma_1, \ldots, \sigma_n$ have a finite expected payoff on all dimensions. The existence of suitable $\beta_1, \ldots, \beta_n \in [0,1]$ is direct by Carathéodory's theorem.

We prove the existence of the coefficients $\gamma_1, \ldots, \gamma_n$ by applying Carathéodory's theorem to a $d-1$ dimensional space. We observe that $\mathbf{q}$ is in the compact polytope $D = \mathsf{conv}(\{\mathbb{E}_s^{\sigma_m}(\bar{f}) \mid 1 \leq m \leq n\})$. In particular, $\mathbf{q}$ is dominated by an element $\mathbf{p}$ lying on a proper face of D (i.e., the intersection of D and a supporting hyperplane of D). For instance, we can consider $\mathbf{q} + \beta \cdot \mathbf{1}$ for $\beta = \max\{\gamma \geq 0 \mid \mathbf{q} + \gamma \mathbf{1} \in D\}$. Because proper faces have dimension no more than $d-1$, Carathéodory's theorem implies that $\mathbf{p}$ is a convex combination of no more than d vectors of the form $\mathbb{E}_s^{\sigma_m}(\bar{f})$, taken among those lying on the considered proper face.

Assume now that $\mathbf{q} \notin \mathbb{R}^d$. In this case, we cannot directly apply Carathéodory's theorem. The idea is to reduce to the previous case. For each $1 \leq j \leq d$, if q_j is infinite, there is $1 \leq m \leq n$ such that $\mathbb{E}_s^{\sigma_m}(f_m) = q_j$. For each infinite component of $\mathbf{q}$, we fix $\alpha_m \mathbb{E}_s^{\sigma_m}(\bar{f})$ in the convex combination for one such m. We then obtain the sought bounds from the above; we reason on the real-valued components of $\mathbf{q}$ in the non-fixed part of the convex combination (after normalising its coefficients to sum to one). ◀

We now highlight two corollaries of Thm. 15. First, we formulate bounds when dealing with universally integrable payoffs.

▶ **Corollary 16.** *Assume that $\bar{f}$ is universally integrable. Let $s \in S$.*

- *For all $\mathbf{q} \in \mathsf{Pay}_s(\bar{f})$, $\mathbf{q}$ is a convex combination of at most $d+1$ elements of $\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f})$.*
- *For all $\mathbf{q} \in \mathsf{Ach}_s(\bar{f})$, there exists a convex combination $\mathbf{p}$ of at most d elements of $\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f})$ such that $\mathbf{q} \leq \mathbf{p}$.*

Proof. By Thm. 7, $\mathsf{Pay}_s(\bar{f}) = \mathsf{conv}(\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f}))$. Furthermore, we recall that $\mathbf{q} \in \mathsf{Ach}_s(\bar{f})$ if and only if there exists a strategy σ such that $\mathbf{q} \leq \mathbb{E}_s^{\sigma}(\bar{f})$. We obtain both claims of the corollary directly by Thm. 15. ◀

Ex. 13 implies that the result for achievable vectors of Cor. 16 does not generalise to universally unambiguously integrable payoffs. Nonetheless, we can show that the result is valid for vectors in $\mathsf{int}(\mathsf{Ach}_s(\bar{f}) \cap \mathbb{R}^d)$ when considering universally unambiguously integrable payoffs. In other words, vectors in $\mathsf{int}(\mathsf{Ach}_s(\bar{f}) \cap \mathbb{R}^d)$ can be achieved by mixing no more than d strategies.

▶ **Corollary 17.** *Assume that $\bar{f}$ is universally unambiguously integrable. Let $s \in S$. For all $\mathbf{q} \in \mathsf{int}(\mathsf{Ach}_s(\bar{f}) \cap \mathbb{R}^d)$, there exists a convex combination $\mathbf{p}$ of at most d elements of $\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f})$ such that $\mathbf{q} \leq \mathbf{p}$.*

Proof sketch. Let $\mathbf{q} \in \mathsf{int}(\mathsf{Ach}_s(\bar{f}) \cap \mathbb{R}^d)$. By definition of the interior, there exists some $\varepsilon > 0$ such that $\mathbf{q} + \varepsilon\mathbf{1} \in \mathsf{Ach}_s(\bar{f})$. Fix a strategy σ witnessing that $\mathbf{q} + \varepsilon\mathbf{1}$ is achievable from s, i.e., such that

$$\mathbb{E}_s^{\sigma}(\bar{f}) \geq \mathbf{q} + \varepsilon\mathbf{1}.$$

It follows from Thm. 14 (applied to $\mathbb{E}_s^{\sigma}(\bar{f})$) that there exists $\mathbf{p}' \in \mathsf{conv}(\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f}))$ such that $\mathbf{p}' \geq \mathbf{q} + \frac{\varepsilon}{2}\mathbf{1}$. By Thm. 15, there exists a convex combination $\mathbf{p}$ of no more than d elements of $\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f})$ such that $\mathbf{p} \geq \mathbf{p}'$, and we conclude by observing that $\mathbf{p} \geq \mathbf{p}' \geq \mathbf{q}$. ◀

6 Continuous payoffs

In this section, we provide a sufficient condition on payoff functions in *finite POMDPs* under which sets of expected payoffs are closed. The main result of this section states that sets of expected payoffs in finite POMDPs are closed for payoffs that are both continuous and universally integrable when squared.

We introduce the payoffs of interest in Sect. 6.1. In Sect. 6.2, we introduce a topology on the space of randomised strategies to formalise a notion of convergence of strategies. In Sect. 6.3, we formulate our result for continuous universally square integrable payoffs. Sect. 6.4 provides examples illustrating that the results of Sect. 6.3 do not hold for continuous payoffs that are not universally integrable. We leave open whether our results extend to all universally integrable continuous payoffs.

For this section, we fix a d-dimensional payoff $\bar{f} = (f_j)_{1 \leq j \leq d}$.

6.1 Definitions

Continuous payoffs. The set of plays of $\mathfrak{P}$ can be equipped with a natural topology that is metrisable, and that is compact if $\mathfrak{P}$ is finite. This allows us to define continuous and uniformly continuous payoffs. We provide direct definitions of (uniformly) continuous payoffs here. A formal presentation of the topology of $\mathsf{Plays}(\mathfrak{P})$ can be found in the full version of this paper [46].

Intuitively, a payoff is continuous if plays with a long common prefix have a close payoff. Let $f\colon \mathsf{Plays}(\mathfrak{P}) \to \bar{\mathbb{R}}$ be a payoff and let $\pi \in \mathsf{Plays}(\mathfrak{P})$. If $f(\pi) \in \mathbb{R}$, then f is continuous at π if and only if for all $\varepsilon > 0$, there exists $\ell \in \mathbb{N}$ such that for all $\pi' \in \mathsf{Cyl}(\pi_{\leq \ell})$, $|f(\pi) - f(\pi')| < \varepsilon$. If $f(\pi) = +\infty$ (resp. $-\infty$), then f is continuous at π if and only if for all $M \in \mathbb{R}$, there exists $\ell \in \mathbb{N}$ such that for all plays $\pi' \in \mathsf{Cyl}(\pi_{\leq \ell})$, $f(\pi') \geq M$ (resp. $f(\pi') \leq -M$). The payoff f is *continuous* if it is continuous at all plays.

If f is real-valued, then f is *uniformly continuous* if for all $\varepsilon > 0$, there exists $\ell \in \mathbb{N}$ such that for all plays π, $\pi' \in \mathsf{Plays}(\mathfrak{P})$, $\pi_{\leq \ell} = \pi'_{\leq \ell}$ implies that $|f(\pi) - f(\pi')| < \varepsilon$. If $\mathfrak{P}$ is finite, then $\mathsf{Plays}(\mathfrak{P})$ is compact and all continuous real-valued payoffs of $\mathfrak{P}$ are uniformly continuous.

Discounted-sum payoffs are uniformly continuous whenever they are built on a weight function that is bounded in absolute value. Any shortest-path function built on a weight function $w \geq \varepsilon$ for some $\varepsilon > 0$ is continuous.

Square integrable payoffs. We say that a payoff $f\colon \mathsf{Plays}(\mathfrak{P}) \to \bar{\mathbb{R}}$ is *universally square integrable* if its square f^2 is universally integrable. The Cauchy-Schwarz inequality (e.g., [31, Thm. 1.5.2.]) guarantees that any universally square integrable payoff is universally integrable.

In finite POMDPs, real-valued continuous payoffs are universally square integrable (such payoffs are bounded by compactness of $\mathsf{Plays}(\mathfrak{P})$). Universally integrable continuous shortest-path payoffs are also universally square integrable in finite POMDPs.

6.2 A topology on the space of strategies

We introduce a topology on the space of strategies to formalise a notion of convergence for sequences of behavioural strategies. We view $\Sigma(\mathfrak{P})$ as a subset of the product $\mathcal{D}(A)^{\mathsf{Obs}(\mathsf{Hist}(\mathfrak{P}))}$ equipped with the product topology. We endow each copy of $\mathcal{D}(A)$ in this product with the topology given by the metric (induced by the Euclidean norm) $\mathsf{dist}_{\mathsf{pr}}(\mu, \mu') = \sqrt{\sum_{a \in A(s)} |\mu(a) - \mu'(a)|^2}$ for all μ, $\mu' \in \mathcal{D}(A)$.

We can show that $\Sigma(\mathfrak{P})$ is a compact metrisable topological space (it is a closed subset of $\mathcal{D}(A)^{\mathsf{Obs}(\mathsf{Hist}(\mathfrak{P}))}$). Compactness follows from the assumption that $\mathfrak{P}$ is finite: $\mathcal{D}(A)$ is homeomorphic to a closed subset of $[0,1]^{|A|}$. We do not fix a metric over strategies, as it is not necessary in the sequel. Instead, we recall that a sequence of strategies $(\sigma^{(n)})_{n \in \mathbb{N}}$ converges to a strategy σ if and only if, for all $h \in \mathsf{Hist}(\mathfrak{P})$, $(\sigma^{(n)}(h))_{n \in \mathbb{N}}$ converges to $\sigma(h)$.

To prove our result on continuous payoffs, we exploit the fact that close strategies induce distributions that assign similar probabilities to cylinders of histories of bounded length.

▶ **Lemma 18.** *Let σ, τ be strategies, $k \in \mathbb{N}_{>0}$ and $\eta > 0$. Assume that, for all histories h that are at most k states long, $\mathsf{dist}_{\mathsf{pr}}(\sigma(h), \tau(h)) \leq \frac{\eta}{k}$. Then, for all histories h that are at most $k+1$ states long and all $s \in S$, we have $|\mathbb{P}^{\sigma}_{s}(\mathsf{Cyl}(h)) - \mathbb{P}^{\tau}_{s}(\mathsf{Cyl}(h))| \leq \eta$.*

We close this section by observing that $\Sigma_{\mathsf{pure}}(\mathfrak{P})$ is a closed subset of $\Sigma(\mathfrak{P})$. In the following section, we prove that if $\bar{f}$ is universally square integrable and $\Sigma \subseteq \Sigma(\mathfrak{P})$ is closed, then $\mathsf{Pay}^{\Sigma}_{s}(\bar{f})$ is compact. Therefore, by establishing that $\Sigma_{\mathsf{pure}}(\mathfrak{P})$ is closed, we can apply this result to $\Sigma_{\mathsf{pure}}(\mathfrak{P})$.

▶ **Lemma 19.** *The set $\Sigma_{\mathsf{pure}}(\mathfrak{P})$ is a closed subset of $\Sigma(\mathfrak{P})$.*

Proof sketch. Converging sequences of Dirac distributions are ultimately constant, and thus their limit is also a Dirac distribution. This implies that the limit of a converging sequence of pure strategies must be pure. ◀

6.3 Universally square integrable continuous payoffs

We now prove that if $\mathfrak{P}$ is finite and $\bar{f}$ is continuous and universally square integrable, then for all $s \in S$, $\mathsf{Pay}_s(\bar{f})$ and $\mathsf{Ach}_s(\bar{f})$ are closed. Our proof relies on the following property: if $\mathfrak{P}$ is finite and $\bar{f}$ is continuous and universally square integrable, then the function $\sigma \mapsto \mathbb{E}_s^\sigma(\bar{f})$ is continuous for all $s \in S$. We remark that this function is real-valued as any universally square integrable payoff is universally integrable.

We split the proof into two parts: first, we consider the particular case of real-valued continuous payoffs and then generalise to universally square integrable payoffs. We cannot assume that a universally integrable continuous payoff is real-valued without loss of generality: changing the payoff of plays that have an infinite payoff to a real number will violate the continuity property. Therefore, we do not generalise the property for real-valued continuous payoffs through such an assumption.

We first provide a proof sketch for the case of real-valued payoffs.

▶ **Theorem 20.** *Let $s \in S$. Assume that $\mathfrak{P}$ is finite, $\bar{f} \colon \mathsf{Plays}(\mathfrak{P}) \to \mathbb{R}^d$ and $\bar{f}$ is continuous. Then the function $\Sigma(\mathfrak{P}) \to \mathbb{R}^d \colon \sigma \mapsto \mathbb{E}_s^\sigma(\bar{f})$ is continuous.*

Proof sketch. It suffices to prove the claim for one-dimensional payoffs. Let $f \colon \mathsf{Plays}(\mathfrak{P}) \to \mathbb{R}$ be a continuous real-valued payoff. It follows from $\mathfrak{P}$ being finite that $\mathsf{Plays}(\mathfrak{P})$ is compact, and therefore f is uniformly continuous and bounded. Recall that f is uniformly continuous if for all $\varepsilon > 0$, there exists $\ell \in \mathbb{N}$ such that for all plays π, π', $\pi_{\leq \ell} = \pi'_{\leq \ell}$ implies that $|f(\pi) - f(\pi')| < \varepsilon$. In particular, for all $\varepsilon > 0$, f can be ε-approximated by a linear combination of indicator functions of cylinders of histories of a fixed length. This provides a means to ε-approximate $\mathbb{E}_s^\tau(f)$ for any strategy τ as a linear combination of probabilities of cylinders of histories of bounded length. Lem. 18 guarantees that the distributions over plays induced by strategies that are close assign similar probabilities to such cylinders. It follows that if $(\sigma^{(n)})_{n\in\mathbb{N}}$ is a sequence of strategies converging to a strategy σ, through the approximations above, we can obtain that $\mathbb{E}_s^{\sigma^{(n)}}(\bar{f})$ is 3ε-close to $\mathbb{E}_s^\sigma(\bar{f})$ for large values of n, which ends the proof of continuity. ◀

We now generalise Thm. 20 to universally square integrable payoffs. The previous proof relies on the boundedness and uniform continuity of real-valued continuous payoffs, and thus does not carry over to this more general case.

▶ **Theorem 21.** *Let $s \in S$. Assume that $\mathfrak{P}$ is finite and that $\bar{f}$ is continuous and universally square integrable. Then the function $\Sigma(\mathfrak{P}) \to \mathbb{R}^d \colon \sigma \mapsto \mathbb{E}_s^\sigma(\bar{f})$ is continuous.*

Proof sketch. It suffices to prove the theorem for non-negative one-dimensional payoffs: the general case can be recovered by writing payoffs on each dimension as the difference of their non-negative and non-positive parts (which are also continuous).

We let $f \colon \mathsf{Plays}(\mathfrak{P}) \to \bar{\mathbb{R}}$ be a non-negative continuous square integrable payoff, $s \in S$, let $(\sigma^{(n)})_{n\in\mathbb{N}}$ be a sequence of strategies that converges to a strategy σ and $\varepsilon > 0$. Given $M \in \mathbb{R}$, we abbreviate $\{f(\pi) \geq M \mid \pi \in \mathsf{Plays}(\mathfrak{P})\}$ by $\{f \geq M\}$, for all strategies τ, we write $\mathbb{P}_s^\tau(f \geq M)$ instead of $\mathbb{P}_s^\tau(\{f \geq M\})$ and we let $\min(f, M)$ denote the payoff $\pi \mapsto \min\{f(\pi), M\}$.

The goal is to show that $|\mathbb{E}_s^{\sigma^{(n)}}(f) - \mathbb{E}_s^{\sigma}(f))| \leq \varepsilon$ for all large enough values of n. We show that there exists a constant $M \geq 0$ (dependent on ε) such that the real-valued continuous payoff $\min(f, M)$ satisfies $0 \leq \mathbb{E}_s^{\tau}(f) - \mathbb{E}_s^{\tau}(\min(f, M)) \leq \frac{\varepsilon}{3}$ for all strategies $\tau \in \Sigma(\mathfrak{P})$. By the triangular inequality, $|\mathbb{E}_s^{\sigma^{(n)}}(f) - \mathbb{E}_s^{\sigma}(f))|$ is no more than

$$
\begin{aligned}
&|\mathbb{E}_s^{\sigma^{(n)}}(f) - \mathbb{E}_s^{\sigma^{(n)}}(\min(f, M))| \\
&+ |\mathbb{E}_s^{\sigma^{(n)}}(\min(f, M)) - \mathbb{E}_s^{\sigma}(\min(f, M))| \\
&+ |\mathbb{E}_s^{\sigma}(\min(f, M)) - \mathbb{E}_s^{\sigma}(f)|.
\end{aligned}
$$

The first and last terms are no more than $\frac{\varepsilon}{3}$ by choice of M, and the second term is smaller than $\frac{\varepsilon}{3}$ for large values of n by Thm. 20.

Thus, the main hurdle of the proof is establishing the existence of a suitable M. The first step consists in showing that the probability of f being large can be made arbitrarily small, in the sense that for all $\eta > 0$, there exists $M(\eta)$ such that $\mathbb{P}_s^{\tau}(f \geq M(\eta)) \leq \eta$ for all strategies τ. The negation of this property would imply the existence of strategies with an arbitrarily large expected payoff from s, contradicting Lem. 2. It then follows from the Cauchy-Schwarz inequality that, for all strategies τ,

$$\mathbb{E}_s^{\tau}(f - \min(f, M)) \leq \sqrt{\mathbb{E}_s^{\tau}(f^2) \cdot \eta}$$

because $f - \min(f, M) \leq f \cdot \mathbb{1}_{f \geq M}$. Because f^2 is universally integrable, Lem. 2 guarantees that $\sup_{\tau} \mathbb{E}_s^{\tau}(f^2)$ is finite, i.e., we obtain the desired inequality by choosing a small enough η and its associated $M(\eta)$. ◀

We now prove that for universally square integrable payoffs that are continuous, given a closed set of strategies, its set of expected payoffs and achievable set from a state are compact.

▶ **Theorem 22.** *Let $s \in S$ and $\Sigma \subseteq \Sigma(\mathfrak{P})$ be a closed set of strategies. Assume that $\bar{f}$ is a continuous universally square integrable payoff. Then $\mathsf{Pay}_s^{\Sigma}(\bar{f})$ and $\mathsf{Ach}_s^{\Sigma}(\bar{f})$ are compact subsets of $\mathbb{R}^d$ and $\bar{\mathbb{R}}^d$ respectively. In particular, $\mathsf{Pay}_s(\bar{f})$, $\mathsf{Ach}_s(\bar{f})$, $\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f})$ and $\mathsf{Ach}_s^{\mathsf{pure}}(\bar{f})$ are compact.*

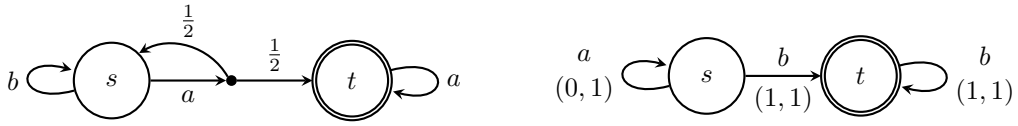
Proof. Since Σ is closed and $\Sigma(\mathfrak{P})$ is compact, we obtain that Σ is compact. It follows from Thm. 21 that $\mathsf{Pay}_s^{\Sigma}(\bar{f})$ is the image of Σ by a continuous function, and thus is compact. The claim regarding $\mathsf{Pay}_s^{\mathsf{pure}}(\bar{f})$ follows from Lem. 19.

By definition, $\mathsf{Ach}_s^{\Sigma}(\bar{f})$ is the downward closure of $\mathsf{Pay}_s^{\Sigma}(\bar{f})$, i.e., the set of vectors that are smaller than or equal to some vector in $\mathsf{Pay}_s^{\Sigma}(\bar{f})$ for $\leq$. The downward closure of a compact subset of $\bar{\mathbb{R}}^d$ is compact. By combining this with the above, we obtain that $\mathsf{Ach}_s^{\Sigma}(\bar{f})$ is compact. ◀

Thm. 22 provides a sufficient condition on payoffs that guarantees that we can dominate any expected payoff by a Pareto-optimal expected payoff. This property does not hold for all universally integrable payoffs in general, e.g., in the one-dimensional setting, optimal strategies need not exist.

6.4 Non-universally integrable continuous payoffs

We now present counterexamples to Thm. 21 and 22 when the assumption of universal square integrability is relaxed. The MDPs used in these counterexamples are depicted in Fig. 6.1.



(a) An MDP with one randomised transition. Weights are omitted and are all 1.

(b) An MDP with two-dimensional weights.

Figure 6.1 MDPs for counter-examples to Thm. 21 (left) and 22 (right) without the universally square integrable assumption.

For the first example, we provide an MDP with a continuous shortest-path payoff witnessing that Thm. 21 does not generalise to shortest-path payoffs that are not universally (square) integrable.

▶ **Example 23.** We consider the MDP $\mathfrak{M}$ depicted in Figure 6.1a, the constant weight function $w = 1$ and the shortest-path function $\mathsf{SPath}_w^{\{t\}}$. We provide a sequence of pure strategies $(\sigma^{(n)})_{n\in\mathbb{N}}$ converging to a pure strategy σ such that $\lim_{n\to\infty} \mathbb{E}_s^{\sigma^{(n)}}(\mathsf{SPath}_w^{\{t\}}) \neq \mathbb{E}_s^{\sigma}(\mathsf{SPath}_w^{\{t\}})$. For all $n \in \mathbb{N}$, we define $\sigma^{(n)}$ as the pure strategy such that, for all $h \in \mathsf{Hist}(\mathfrak{M})$, $\sigma^{(n)}(h) = b$ if $\mathsf{last}(h) = s$ and there are at least n actions in h, and otherwise, $\sigma^{(n)}(h) = a$. Intuitively, when starting in s, $\sigma^{(n)}$ uses action a for the first n rounds of the play and then uses b for all subsequent rounds. The pure (memoryless) strategy σ assigning action a to all histories is easily checked to be $\lim_{n\to\infty} \sigma^{(n)}$.

Let $n \in \mathbb{N}$. Let $\pi_n = (sa)^n(sb)^\omega$. The definition of $\sigma^{(n)}$ implies that

$$\mathbb{P}_s^{\sigma^{(n)}}(\{\pi_n\}) = \mathbb{P}_s^{\sigma^{(n)}}(\mathsf{Cyl}((sa)^n s)) = \frac{1}{2^n}.$$

It follows from $\mathsf{SPath}_w^{\{t\}}(\pi_n) = +\infty$ that $\mathbb{E}_s^{\sigma^{(n)}}(\mathsf{SPath}_w^{\{t\}}) = +\infty$. We conclude that $\lim_{n\to\infty} \mathbb{E}_s^{\sigma^{(n)}}(\mathsf{SPath}_w^{\{t\}}) = +\infty$.

We now show that $\mathbb{E}_s^{\sigma}(\mathsf{SPath}_w^{\{t\}}) = 2 \neq +\infty$. It follows from $\mathbb{P}_s^{\sigma}(\mathsf{Reach}(\{t\})) = 1$ that $\mathbb{E}_s^{\sigma}(\mathsf{SPath}_w^{\{t\}})$ is the unique solution of the equation $x = 1 + \frac{1}{2}x$ (see, e.g., [3]), i.e., $\mathbb{E}_s^{\sigma}(\mathsf{SPath}_w^{\{t\}}) = 2$. We have shown that $\lim_{n\to\infty} \mathbb{E}_s^{\sigma^{(n)}}(\mathsf{SPath}_w^{\{t\}}) \neq \mathbb{E}_s^{\sigma}(\mathsf{SPath}_w^{\{t\}})$. ⌟

We now present a counter-example to Thm. 22 when the considered payoffs are not universally integrable: we provide a continuous payoff such that its set of expected payoffs and achievable sets from a given state are not closed.

▶ **Example 24.** We consider the MDP $\mathfrak{M}$ and the weight function $w = (w_1, w_2)$ depicted in Fig. 6.1b. Let $\bar{f} = (f_1, f_2)$ be the two-dimensional payoff such that $f_1 = \mathsf{DSum}_{w_1}^{1/2}$ and $f_2 = \mathsf{SPath}_{w_2}^{\{t\}}$. We show that $(2, +\infty) \in \mathsf{cl}(\mathsf{Pay}_s(\bar{f})) \setminus \mathsf{Pay}_s(\bar{f})$ to prove that $\mathsf{Pay}_s(\bar{f})$ is not closed.

First, we show that $(2, +\infty) \in \mathsf{cl}(\mathsf{Pay}_s(\bar{f}))$. The vectors $(0, +\infty) = \bar{f}((sa)^\omega)$ and $(2, 1) = \bar{f}(s(bt)^\omega)$ are payoffs of pure strategies as there is no randomisation in the transitions of $\mathfrak{M}$. By convexity of $\mathsf{Pay}_s(\bar{f})$, we obtain that for all $\eta \in]0, 1[$, $(2\eta, +\infty) \in \mathsf{Pay}_s(\bar{f})$. It follows that $(2, +\infty) \in \mathsf{cl}(\mathsf{Pay}_s(\bar{f}))$.

Second, we prove that $(2, +\infty) \notin \mathsf{Pay}_s(\bar{f})$. The only play π from s such that $f_1(\pi) = 2$ is the play $s(bt)^\omega$. All other plays π' starting in s are such that $f_1(\pi') < 2$. Indeed, for all $\ell \in \mathbb{N}$, we have $f_1((sa)^\ell s(bt)^\omega) = \sum_{\ell' \geq \ell} \frac{1}{2^{\ell'}} = \frac{1}{2^{\ell-1}} < 2$. It follows that, to obtain a payoff of 2 on the first dimension from s, we require a strategy whose only outcome is $s(bt)^\omega$. This implies that $(2, +\infty) \notin \mathsf{Pay}_s(\bar{f})$. We have shown that $\mathsf{Pay}_s(\bar{f})$ is not closed. ⌟

7 Conclusion

We have investigated randomisation requirements and the structure of expected payoff sets in multi-objective POMDPs. First, we have shown that for universally integrable payoffs, pure strategies suffice for lexicographic optimisation in countable POMDPs. This generalises a result of [20] showing that pure strategies are sufficient to optimise the probability of objectives in countable-state POMDPs. Second, we have shown that for universally integrable payoffs and universally unambiguously integrable payoffs respectively, finite-support mixed strategies can be used to exactly obtain or approximate the expected payoff of any strategy. This generalises known results on finite MDPs for specific classes of payoffs to general combinations of payoffs (see Sect. 1), and shows that these properties hold up to countable POMDPs. This provides new insights regarding strategy complexity in multi-objective countable POMDPs on which there is limited work. Existing works on strategy complexity in countable (fully observable) MDPs have mainly focused on memory requirements for one-dimensional objectives (e.g., [42, 47]), for which randomisation is not necessary but it may be necessary to use infinite memory.

We now comment on the techniques and results used throughout the paper and their use in related literature. First, we observe that we make use of both mixed and behavioural strategies in our proofs. This highlights the usefulness of Kuhn's theorem to reason on strategic requirements. On the one hand, we have used mixed strategies to prove Thm. 4 on lexicographic POMDPs, which in turn is used to prove Thm. 7, and to prove Thm. 14 on payoff sets of universally unambiguously integrable payoffs. The idea of writing expected payoffs as an integral over the payoffs of pure strategies also appears in [20] to prove that pure strategies suffice to optimise the probability of objectives in POMDPs – this is generalised to lexicographic optimisation of universally integrable payoffs by Thm. 4. On the other hand, we have reasoned on behavioural strategies to prove Thm. 21 and 22 on continuous payoffs. Thm. 20 and its proof approach can be seen as a generalisation of the result and techniques for discounted-sum payoffs of [23]: we generalise the idea of exploiting the discounting to approximate payoffs by linear combinations of cylinder indicators through uniform continuity.

Second, we note that the use of the hyperplane separation theorem, invoked in our proof of Thm. 7, is classical in the analysis of multi-objective (PO)MDPs. Typically, separating hyperplanes are used in reductions from multi-dimensional payoffs to one-dimensional payoffs. For instance, an argument relying on such a reduction is used in [32] to establish memory requirements for multi-objective MDPs with reachability objectives. The approximation schemes of achievable sets mentioned in the introduction [35, 50, 49] also use such reductions to a single dimension. The main difference between our proof approach for Thm. 7 and these existing approaches is that we use the hyperplane separation theorem in a reduction to lexicographic optimisation instead of one-dimensional optimisation.

Finally, we comment on possible extensions of this work. The results of Sect. 4 and 5 provide insight on the randomisation requirements of multi-objective POMDPs. Another facet of strategy complexity is the memory required to play (almost) Pareto-optimally (see, e.g., [52] and references therein). One could study the structure of payoff sets when restricted to finite-memory strategies, which are strategies usually deemed to be implementable [30]. Additional restrictions on payoffs or the setting are likely necessary to obtain finite-memory analogues of the main results of this paper. For instance, a finite-memory variation of Thm. 4 would not hold directly, as there exist non-classical one-dimensional objectives in finite MDPs for which optimal randomised finite-memory exist, but no pure finite-memory strategy is even close to optimal. One such objective would be to require a non-ultimately periodic play

in a MDP with one state and two self-loops labelled by different actions; playing uniformly at random is sufficient to ensure that this objective is almost-surely satisfied but no pure finite-memory strategy can accomplish this. We note that a non-ultimately periodic play in this example can be constructed without randomisation by using infinite memory, so this example does not contradict the results on pure strategies given in this paper.

Sect. 6 provides a sufficient condition ensuring that all achievable vectors are dominated by a Pareto-optimal payoff. A related problem would be to characterise the payoff functions for which this holds.

References

1 Pranav Ashok, Krishnendu Chatterjee, Jan Kretínský, Maximilian Weininger, and Tobias Winkler. Approximating values of generalized-reachability stochastic games. In Holger Hermanns, Lijun Zhang, Naoki Kobayashi, and Dale Miller, editors, *Proceedings of the 35th Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2020, Saarbrücken, Germany, July 8–11, 2020*, pages 102–115. ACM, 2020. doi:10.1145/3373718.3394761.

2 Robert J . Aumann. Mixed and behavior strategies in infinite extensive games. In Melvin Dresher, Lloyd S. Shapley, and Albert William Tucker, editors, *Advances in Game Theory. (AM-52), Volume 52*, pages 627–650. Princeton University Press, 1964. doi:10.1515/9781400882014-029.

3 Christel Baier and Joost-Pieter Katoen. *Principles of model checking*. MIT Press, 2008.

4 Marius Belly, Nathanaël Fijalkow, Hugo Gimbert, Florian Horn, Guillermo A. Pérez, and Pierre Vandenhove. Revelations: A decidable class of POMDPs with omega-regular objectives. In Toby Walsh, Julie Shah, and Zico Kolter, editors, *Proceedings of the 39th AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, February 25 – March 4, 2025*, pages 26454–26462. AAAI Press, 2025. doi:10.1609/AAAI.V39I25.34845.

5 Nathalie Bertrand, Blaise Genest, and Hugo Gimbert. Qualitative determinacy and decidability of stochastic games with signals. *Journal of the ACM*, 64(5):33:1–33:48, 2017. doi:10.1145/3107926.

6 Kark-Josef Bierth. An expected average reward criterion. *Stochastic processes and their applications*, 26:123–140, 1987.

7 Patricia Bouyer, Stéphane Le Roux, Youssouf Oualhadj, Mickael Randour, and Pierre Vandenhove. Games where you can play optimally with arena-independent finite memory. *Logical Methods in Computer Science*, 18(1), 2022. doi:10.46298/lmcs-18(1:11)2022.

8 Patricia Bouyer, Nicolas Markey, Mickael Randour, Kim G. Larsen, and Simon Laursen. Average-energy games. *Acta Informatica*, 55(2):91–127, 2018. doi:10.1007/s00236-016-0274-1.

9 Patricia Bouyer, Youssouf Oualhadj, Mickael Randour, and Pierre Vandenhove. Arena-independent finite-memory determinacy in stochastic games. *Log. Methods Comput. Sci.*, 19(4), 2023. doi:10.46298/LMCS-19(4:18)2023.

10 Patricia Bouyer, Mickael Randour, and Pierre Vandenhove. Characterizing omega-regularity through finite-memory determinacy of games on infinite graphs. *TheoretiCS*, 2, 2023. doi:10.46298/THEORETICS.23.1.

11 Thomas Brihaye, Florent Delgrange, Youssouf Oualhadj, and Mickael Randour. Life is random, time is not: Markov decision processes with window objectives. *Logical Methods in Computer Science*, 16(4), 2020. URL: https://lmcs.episciences.org/6975.

12 Thomas Brihaye, Aline Goeminne, James C. A. Main, and Mickael Randour. Reachability games and friends: A journey through the lens of memory and complexity (invited talk). In Patricia Bouyer and Srikanth Srinivasan, editors, *Proceedings of the 43rd IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science, FSTTCS 2023, IIIT Hyderabad, Telangana, India, December 18–20, 2023*, volume 284 of *LIPIcs*, pages 1:1–1:26. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023. doi:10.4230/LIPIcs.FSTTCS.2023.1.

13 Véronique Bruyère, Emmanuel Filiot, Mickael Randour, and Jean-François Raskin. Meet your expectations with guarantees: Beyond worst-case synthesis in quantitative games. *Information and Computation*, 254:259–295, 2017. doi:10.1016/j.ic.2016.10.011.

14 Véronique Bruyère, Quentin Hautem, Mickael Randour, and Jean-François Raskin. Energy mean-payoff games. In Wan J. Fokkink and Rob van Glabbeek, editors, *Proceedings of the 30th International Conference on Concurrency Theory, CONCUR 2019, Amsterdam, the Netherlands, August 27–30, 2019*, volume 140 of *LIPIcs*, pages 21:1–21:17. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019. doi:10.4230/LIPIcs.CONCUR.2019.21.

15 Tomáš Brázdil, Václav Brožek, Krishnendu Chatterjee, Vojtěch Forejt, and Antonín Kučera. Markov decision processes with multiple long-run average objectives. *Logical Methods in Computer Science*, Volume 10, Issue 1, February 2014. doi:10.2168/LMCS-10(1:13)2014.

16 Damien Busatto-Gaston, Debraj Chakraborty, Anirban Majumdar, Sayan Mukherjee, Guillermo A. Pérez, and Jean-François Raskin. Bi-objective lexicographic optimization in Markov decision processes with related objectives. In Étienne André and Jun Sun, editors, *Proceedings (Part I) of the 21st Interval Symposium on Automated Technology for Verification and Analysis, ATVA 2023, Singapore, October 24–27, 2023*, volume 14215 of *Lecture Notes in Computer Science*, pages 203–223. Springer, 2023. doi:10.1007/978-3-031-45329-8_10.

17 Arnaud Carayol, Axel Haddad, and Olivier Serre. Randomization in automata on infinite trees. *ACM Transactions on Computational Logic*, 15(3):24:1–24:33, 2014. doi:10.1145/2629336.

18 Antonio Casares and Pierre Ohlmann. Characterising memory in infinite games. In Kousha Etessami, Uriel Feige, and Gabriele Puppis, editors, *Proceedings of the 50th International Colloquium on Automata, Languages, and Programming, ICALP 2023, July 10–14, 2023, Paderborn, Germany*, volume 261 of *LIPIcs*, pages 122:1–122:18. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023. doi:10.4230/LIPIcs.ICALP.2023.122.

19 Krishnendu Chatterjee, Luca de Alfaro, and Thomas A. Henzinger. Trading memory for randomness. In *Proceedings of the 1st International Conference on Quantitative Evaluation of Systems, QEST 2004, Enschede, The Netherlands, 27–30 September 2004*, pages 206–217. IEEE Computer Society, 2004. doi:10.1109/QEST.2004.1348035.

20 Krishnendu Chatterjee, Laurent Doyen, Hugo Gimbert, and Thomas A. Henzinger. Randomness for free. *Information and Computation*, 245:3–16, 2015. doi:10.1016/j.ic.2015.06.003.

21 Krishnendu Chatterjee, Laurent Doyen, and Thomas A. Henzinger. Qualitative analysis of partially-observable Markov decision processes. In Petr Hlinený and Antonín Kucera, editors, *Proceedings of the 35th International Symposium on Mathematical Foundations of Computer Science, MFCS 2010, Brno, Czech Republic, August 23–27, 2010*, volume 6281 of *Lecture Notes in Computer Science*, pages 258–269. Springer, 2010. doi:10.1007/978-3-642-15155-2_24.

22 Krishnendu Chatterjee, Adrián Elgyütt, Petr Novotný, and Owen Rouillé. Expectation optimization with probabilistic guarantees in POMDPs with discounted-sum objectives. In Jérôme Lang, editor, *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, July 13–19, 2018*, pages 4692–4699. ijcai.org, 2018. doi:10.24963/IJCAI.2018/652.

23 Krishnendu Chatterjee, Vojtech Forejt, and Dominik Wojtczak. Multi-objective discounted reward verification in graphs and MDPs. In Kenneth L. McMillan, Aart Middeldorp, and Andrei Voronkov, editors, *Proceedings of the 19th International Conference on Logic for Programming, Artificial Intelligence, and Reasoning, LPAR-19, Stellenbosch, South Africa, December 14–19, 2013*, volume 8312 of *Lecture Notes in Computer Science*, pages 228–242. Springer, 2013. doi:10.1007/978-3-642-45221-5_17.

24 Krishnendu Chatterjee, Marcin Jurdzinski, and Thomas A. Henzinger. Quantitative stochastic parity games. In J. Ian Munro, editor, *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2004, New Orleans, Louisiana, USA, January 11–14, 2004*, pages 121–130. SIAM, 2004. URL: http://dl.acm.org/citation.cfm?id=982792.982808.

25 Krishnendu Chatterjee, Joost-Pieter Katoen, Stefanie Mohr, Maximilian Weininger, and Tobias Winkler. Stochastic games with lexicographic objectives. *Formal methods in system design*, pages 1–41, 2023. doi:10.1007/s10703-023-00411-4.

26 Krishnendu Chatterjee, Zuzana Kretínská, and Jan Kretínský. Unifying two views on multiple mean-payoff objectives in Markov decision processes. *Logical Methods in Computer Science*, 13(2), 2017. doi:10.23638/LMCS-13(2:15)2017.

27 Krishnendu Chatterjee, Petr Novotný, Guillermo A. Pérez, Jean-François Raskin, and Dorde Zikelic. Optimizing expectation with guarantees in POMDPs. In Satinder Singh and Shaul Markovitch, editors, *Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, California, USA, February 4–9, 2017*, pages 3725–3732. AAAI Press, 2017. doi:10.1609/AAAI.V31I1.11046.

28 Krishnendu Chatterjee, Mickael Randour, and Jean-François Raskin. Strategy synthesis for multi-dimensional quantitative objectives. *Acta Informatica*, 51(3-4):129–163, 2014. doi:10.1007/s00236-013-0182-6.

29 Taolue Chen, Vojtech Forejt, Marta Z. Kwiatkowska, Aistis Simaitis, and Clemens Wiltsche. On stochastic games with multiple objectives. In Krishnendu Chatterjee and Jirí Sgall, editors, *Proceedings of the 38th International Symposium on Mathematical Foundations of Computer Science, MFCS 2013, Klosterneuburg, Austria, August 26–30, 2013*, volume 8087 of *Lecture Notes in Computer Science*, pages 266–277. Springer, 2013. doi:10.1007/978-3-642-40313-2_25.

30 Florent Delgrange, Joost-Pieter Katoen, Tim Quatmann, and Mickael Randour. Simple strategies in multi-objective MDPs. In Armin Biere and David Parker, editors, *Proceedings (Part I) of the 26th International Conference on Tools and Algorithms for the Construction and Analysis of Systems, TACAS 2020, Held as Part of ETAPS 2020, Dublin, Ireland, April 25–30, 2020*, volume 12078 of *Lecture Notes in Computer Science*, pages 346–364. Springer, 2020. doi:10.1007/978-3-030-45190-5_19.

31 Rick Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5th edition, 2019. doi:10.1017/9781108591034.

32 Kousha Etessami, Marta Z. Kwiatkowska, Moshe Y. Vardi, and Mihalis Yannakakis. Multi-objective model checking of Markov decision processes. *Logical Methods in Computer Science*, 4(4), 2008. doi:10.2168/LMCS-4(4:8)2008.

33 Nathanaël Fijalkow, C. Aiswarya, Guy Avni, Nathalie Bertrand, Patricia Bouyer, Romain Brenguier, Arnaud Carayol, Antonio Casares, John Fearnley, Hugo Gimbert, Thomas A. Henzinger, Paul Gastin, Florian Horn, Rasmus Ibsen-Jensen, Nicolas Markey, Benjamin Monmege, Petr Novotný, Pierre Ohlmann, Mickael Randour, Ocan Sankur, Sylvain Schmitz, Olivier Serre, Mateusz Skomra, Nathalie Sznajder, and Pierre Vandenhove. *Games on Graphs: From Logic and Automata to Algorithms*, volume abs/2305.10546. Cambridge University Press, 2025. In press, Cambridge University Press. doi:10.48550/arXiv.2305.10546.

34 Nathanaël Fijalkow and Florian Horn. Les jeux d'accessibilité généralisée. *Technique et Science Informatiques*, 32(9-10):931–949, 2013. doi:10.3166/tsi.32.931-949.

35 Vojtech Forejt, Marta Z. Kwiatkowska, and David Parker. Pareto curves for probabilistic model checking. In Supratik Chakraborty and Madhavan Mukund, editors, *Proceedings of the 10th International Symposium on Automated Technology for Verification and Analysis, ATVA 2014, Thiruvananthapuram, India, October 3–6, 2012*, volume 7561 of *Lecture Notes in Computer Science*, pages 317–332. Springer, 2012. doi:10.1007/978-3-642-33386-6_25.

36 Hugo Gimbert. Pure stationary optimal strategies in Markov decision processes. In Wolfgang Thomas and Pascal Weil, editors, *Proceedings of the 24th Annual Symposium on Theoretical Aspects of Computer Science, STACS 2007, Aachen, Germany, February 22–24, 2007*, volume 4393, pages 200–211. Springer, 2007. doi:10.1007/978-3-540-70918-3_18.

37 Hugo Gimbert and Edon Kelmendi. Submixing and shift-invariant stochastic games. *International Journal of Game Theory*, 52(4):1179–1214, 2023. doi:10.1007/S00182-023-00860-5.

38 Ernst Moritz Hahn, Mateo Perez, Sven Schewe, Fabio Somenzi, Ashutosh Trivedi, and Dominik Wojtczak. Model-free reinforcement learning for lexicographic omega-regular objectives. In Marieke Huisman, Corina S. Pasareanu, and Naijun Zhan, editors, *Proceedings of the 24th International Symposium on Formal Methods, FM 2021, Virtual Event, November 20–26, 2021*, volume 13047 of *Lecture Notes in Computer Science*, pages 142–159. Springer, 2021. doi:10.1007/978-3-030-90870-6_8.

39 Florian Horn. Random fruits on the Zielonka tree. In Susanne Albers and Jean-Yves Marion, editors, *Proceedings of the 26th International Symposium on Theoretical Aspects of Computer Science, STACS 2009, Freiburg, Germany, February 26–28, 2009*, volume 3 of *LIPIcs*, pages 541–552. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany, 2009. doi:10.4230/LIPIcs.STACS.2009.1848.

40 Joshua D. Isom, Sean P. Meyn, and Richard D. Braatz. Piecewise linear dynamic programming for constrained POMDPs. In Dieter Fox and Carla P. Gomes, editors, *Proceedings of the 23rd AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13–17, 2008*, pages 291–296. AAAI Press, 2008. URL: http://www.aaai.org/Library/AAAI/2008/aaai08-046.php.

41 Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998. doi:10.1016/S0004-3702(98)00023-X.

42 Stefan Kiefer, Richard Mayr, Mahsa Shirmohammadi, and Patrick Totzke. Strategy complexity of parity objectives in countable MDPs. In Igor Konnov and Laura Kovács, editors, *Proceedings of the 31st International Conference on Concurrency Theory, CONCUR 2020, Vienna, Austria, September 1–4, 2020*, LIPIcs, pages 39:1–39:17. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020. doi:10.4230/LIPIcs.CONCUR.2020.39.

43 Harold W Kuhn. Extensive games and the problem of information. *Contributions to the Theory of Games*, 2(28):193–216, 1953.

44 Stéphane Le Roux, Arno Pauly, and Mickael Randour. Extending finite-memory determinacy by Boolean combination of winning conditions. In Sumit Ganguly and Paritosh K. Pandya, editors, *Proceedings of the 38th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science, FSTTCS 2018, Ahmedabad, India, December 11–13, 2018*, volume 122 of *LIPIcs*, pages 38:1–38:20. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018. doi:10.4230/LIPIcs.FSTTCS.2018.38.

45 James C. A. Main and Mickael Randour. Different strokes in randomised strategies: Revisiting Kuhn's theorem under finite-memory assumptions. *Information and Computation*, 301:105229, 2024. doi:10.1016/J.IC.2024.105229.

46 James C. A. Main and Mickael Randour. Mixing any cocktail with limited ingredients: On the structure of payoff sets in multi-objective POMDPs and its impact on randomised strategies. *CoRR*, abs/2502.18296, 2025. doi:10.48550/arXiv.2502.18296.

47 Richard Mayr and Eric Munday. Strategy complexity of point payoff, mean payoff and total payoff objectives in countable mdps. *Logical Methods in Computer Science*, 19(1), 2023. doi:10.46298/LMCS-19(1:16)2023.

48 Benjamin Monmege, Julie Parreaux, and Pierre-Alain Reynier. Reaching your goal optimally by playing at random with no memory. In Igor Konnov and Laura Kovács, editors, *Proceedings of the 31st International Conference on Concurrency Theory, CONCUR 2020, Vienna, Austria, September 1–4, 2020*, volume 171 of *LIPIcs*, pages 26:1–26:21. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020. doi:10.4230/LIPIcs.CONCUR.2020.26.

49 Tim Quatmann. *Verification of multi-objective Markov models*. PhD thesis, RWTH Aachen University, Germany, 2023. URL: https://publications.rwth-aachen.de/record/971553.

50 Tim Quatmann and Joost-Pieter Katoen. Multi-objective optimization of long-run average and total rewards. In Jan Friso Groote and Kim Guldstrand Larsen, editors, *Proceedings (Part I) of the 27th International Conference on Tools and Algorithms for the Construction and Analysis of Systems, TACAS 2021, Held as Part of ETAPS 2021, Luxemburg City, Luxemburg, March 27–April 1, 2021*, volume 12651 of *Lecture Notes in Computer Science*, pages 230–249. Springer, 2021. doi:10.1007/978-3-030-72016-2_13.

51 Mickael Randour. Automated synthesis of reliable and efficient systems through game theory: A case study. In *Proc. of ECCS 2012*, Springer Proceedings in Complexity XVII, pages 731–738. Springer, 2013. doi:10.1007/978-3-319-00395-5_90.

52 Mickael Randour. Simplicity lies in the eye of the beholder: A strategic perspective on controllers in reactive synthesis. In Pierre Ganty and Alessio Mansutti, editors, *Reachability Problems - 19th International Conference, RP 2025, Madrid, Spain, October 1-3, 2025, Proceedings*, volume 16230 of *Lecture Notes in Computer Science*, pages 31–48. Springer, 2025. doi:10.1007/978-3-032-09524-4_3.

53 Mickael Randour, Jean-François Raskin, and Ocan Sankur. Percentile queries in multi-dimensional Markov decision processes. *Formal methods in system design*, 50(2-3):207–248, 2017. doi:10.1007/s10703-016-0262-7.

54 Mathieu Reymond, Eugenio Bargiacchi, and Ann Nowé. Pareto conditioned networks. In Piotr Faliszewski, Viviana Mascardi, Catherine Pelachaud, and Matthew E. Taylor, editors, *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, May 9–13, 2022*, pages 1110–1118. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2022. doi:10.5555/3535850.3535974.

55 R. Tyrrell Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1970.

56 Diederik Marijn Roijers, Shimon Whiteson, and Frans A. Oliehoek. Point-based planning for multi-objective POMDPs. In Qiang Yang and Michael J. Wooldridge, editors, *Proceedings of the 24th International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25–31, 2015*, pages 1666–1672. AAAI Press, 2015. URL: http://ijcai.org/Abstract/15/238.

57 Lloyd S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10):1095–1100, 1953. doi:10.1073/pnas.39.10.1095.

58 Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

59 Kyle Hollins Wray and Shlomo Zilberstein. Multi-objective POMDPs with lexicographic reward preferences. In Qiang Yang and Michael J. Wooldridge, editors, *Proceedings of the 24th International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25–31, 2015*, pages 1719–1725. AAAI Press, 2015. URL: http://ijcai.org/Abstract/15/245.