Solving MDPs with Skew Symmetric Bilinear Utility Functions

Hugo Gilbert[1,2], **Olivier Spanjaard**[1,2], **Paolo Viappiani**[1,2], **Paul Weng**[3,4]
[1]Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, F-75005, Paris, France
[2]CNRS, UMR 7606, LIP6, F-75005, Paris, France
[3]SYSU-CMU Joint Institute of Engineering, Guangzhou, China
[4]SYSU-CMU Shunde International Joint Research Institute, Shunde, China
{hugo.gilbert,olivier.spanjaard,paolo.viappiani}@lip6.fr, paweng@cmu.edu

Abstract

In this paper we adopt Skew Symmetric Bilinear (SSB) utility functions to compare policies in Markov Decision Processes (MDPs). By considering pairs of alternatives, SSB utility theory generalizes von Neumann and Morgenstern's expected utility (EU) theory to encompass rational decision behaviors that EU cannot accommodate. We provide a game-theoretic analysis of the problem of identifying an SSB-optimal policy in finite horizon MDPs and propose an algorithm based on a double oracle approach for computing an optimal (possibly randomized) policy. Finally, we present and discuss experimental results where SSB-optimal policies are computed for a popular TV contest according to several instantiations of SSB utility functions.

1 Introduction

Decision-theoretic planning deals with planners involving decision-theoretic tools [Blythe, 1999; Boutilier *et al.*, 1999]. As emphasized by Blythe [1999], AI planning and decision theory appear indeed to be complementary, and there has been interest in merging the two approaches for a considerable time. The study of Markov Decision Processes (MDPs) constitute perhaps the biggest part of decision-theoretic planning, because MDPs can be considered as a natural framework both for modeling and solving complex structured decision problems [Puterman, 1994]. In an MDP, scalar rewards, assumed to be additive, are granted along the process, and a policy is evaluated according to the expectation of the sum of rewards. Yet, expectation is far from being the only possible decision criterion. In particular, expectation is not a *risk sensitive* criterion, in the sense that it assumes risk neutrality (e.g., a sure \$500 gain is equivalent to having a probability 1/2 of a \$1000 gain or nothing). One of the topics of decision theory is precisely to provide risk sensitive criteria.

The most popular risk-sensitive criterion in decision theory is the expected utility (EU) model [von Neumann and Morgenstern, 1947]. In this model, an agent is endowed with a utility function u that assigns a numerical value $u(x)$ to each consequence x in the set X of possible outcomes. A probability distribution p is preferred to q, denoted by $p \succ q$, iff $u(p) > u(q)$, where $u(p) = \sum_{x \in X} p(x)u(x)$. According to the shape of the utility function used, the maximization of expected utility will favor risk-averse or risk-seeking behaviors. Solution algorithms for MDPs with expected utility objective functions have been proposed by Liu [2005] and Liu and Koenig [2005; 2006; 2008]. Nevertheless, despite its intuitive appeal, the EU model does not make it possible to account for all rational decision behaviors. For instance, it is unable to explain the paradox of nontransitive dice as designed by statistician Efron and reported by Gardner [1970]. The specific variant we present here is due to Rowett.

Example 1 (Rowett Dice)**.** *Consider a two-player game involving the following set of six-sided dice: die A with sides $(1, 4, 4, 4, 4, 4)$, die B with sides $(3, 3, 3, 3, 3, 6)$ and die C with sides $(2, 2, 2, 5, 5, 5)$. The players, each equipped with a personal set of Rowett dice, simultaneously choose a die to throw; the winner is the player who rolls the highest number. It is easy to realize that die A rolls higher than B most of the time, so die A should be preferred to B. Similarly die B rolls higher than C most of the time, and the same can be said about C against A. In other words, the relation "more likely to win" is not transitive, and in fact it is even cyclic.*

This example can be formalized by characterizing dice A, B, C by probability distributions p_A, p_B, p_C over the set $X = \{1, 2, \ldots, 6\}$ of possible outcomes. The expected utility model is obviously unable to accommodate the above preferences $p_A \succ p_B \succ p_C \succ p_A$ because it is impossible to have $u(p_A) > u(p_B) > u(p_C) > u(p_A)$. Actually, every binary preference relation solely based on a unary functional u over distributions would fail to explain the paradox.

Interestingly, the skew symmetric bilinear (SSB) utility theory [Fishburn, 1984], which extends expected utility theory, enables to accommodate this type of intransitivities by using a *binary* functional φ over distributions, and governing the preference between p and q by the sign of $\varphi(p, q)$, where $p \succ q$ iff $\varphi(p, q) > 0$. It is indeed possible that inequalities $\varphi(p_A, p_B) > 0$, $\varphi(p_B, p_C) > 0$ and $\varphi(p_C, p_A) > 0$ simultaneously hold. As we will see later on, the relation "more likely to win" can be modeled by SSB utility theory.

The possibility of cyclic preferences in SSB utility theory could be seen as a significant barrier to its use in automated decision. However, in a finite set of distributions, there always exists a convex combination of these distributions (probability mixture) that is preferred or indifferent to each mixture in the convex hull [Fishburn, 1984]. For in-

stance, coming back to our dice example, playing die A (resp. B, C) with probability $\frac{3}{13}$ (resp. $\frac{3}{13}$, $\frac{7}{13}$) is a maximal strategy for the relation "more likely to win". Consequently, choice by maximal preference using SSB utility theory is well defined for finite horizon MDPs even if there are preference cycles between deterministic policies. Note that, as emphasized by Fishburn [1984], not only does SSB utility theory enable to explain cases of intransitivity, but it also accommodates reasonable behaviors that involves widely observed violations of the von Neumann-Morgenstern independence axiom [Allais, 1953; Kahneman and Tversky, 1979].

We model the problem of computing an SSB optimal policy in an MDP as the search for a Nash equilibrium in a zero-sum two-player symmetric game defined from the MDP and the SSB utility function. The set of pure strategies in this game is the set of deterministic policies after transforming the given MDP in an "augmented" MDP [Liu, 2005], and the payoff function is inferred from the SSB utility function. The set of deterministic policies is combinatorial in nature, which makes impractical the generation of the whole payoff matrix of the game. We therefore adopt a double oracle approach [McMahan *et al.*, 2003] that makes it possible to solve the game without generating the whole payoff matrix.

We provide the results of numerical experiments, notably on a famous TV game, *Who Wants to Be a Millionaire?*. Our results illustrate the enhanced possibilities offered by the use of an SSB utility function for controlling the shape of the probability distribution over payoffs.

2 SSB Utility Theory

Any policy in an MDP induces a probability distribution over possible final wealth levels (cumulated reward scores). Comparing policies amounts then to comparing their induced distributions. We assume throughout the paper that the outcome set, denoted $\mathcal{W}$, of policies is the real line, interpreted as wealth, and that the agent's preferences between distributions are described by the SSB model as presented and axiomatized by Fishburn [1984]. In this model, an agent is endowed with a binary functional φ over ordered pairs $(x, y) \in \mathcal{W}^2$ of wealth levels, with $x > y \Leftrightarrow \varphi(x, y) > 0$. The value $\varphi(x, y)$ can be interpreted as the intensity with which the agent prefers x to y. Functional φ is assumed to be skew symmetric, i.e., $\varphi(x, y) = -\varphi(y, x)$ and bilinear w.r.t. the usual mixture operation on distributions. The SSB criterion for comparing probability distributions p and q is then written:

$$\varphi(p, q) = \sum_{x,y \in \mathcal{W}} p(x)q(y)\varphi(x, y)$$

where $p(x)$ (resp. $q(y)$) denotes the probability of wealth level x (resp. y) in distribution p (resp. q). We have $p \succ (resp. \prec) q$ if $\varphi(p, q) > (resp. <) 0$ (strict preference), and $p \sim q$ if $\varphi(p, q) = 0$ (indifference).

The SSB model is very general as it can represent preferences observed in Example 1.

Example 2. *(Example 1 cont'd) Distributions p_A, p_B, p_C are defined in Figure 1. By setting $\varphi(x, y) = 1$ if $x > y$, and $\varphi(x, y) = -1$ if $x < y$, $\varphi(p, q)$ corresponds then to the probability that p beats q minus the probability that q beats p. The obtained SSB utilities in the example are:*

$$\begin{aligned} \varphi(p_A, p_B) &= 25/36 - 11/36 = 14/36, \\ \varphi(p_B, p_C) &= 21/36 - 15/36 = 6/36, \\ \varphi(p_C, p_A) &= 21/36 - 15/36 = 6/36. \end{aligned}$$

Therefore we have $\varphi(p_A, p_B) > 0$, $\varphi(p_B, p_C) > 0$ and $\varphi(p_C, p_A) > 0$, which is consistent with the relation "more likely to win" between dice (i.e., $p_A \succ p_B \succ p_C \succ p_A$).

	1	2	3	4	5	6
p_A	1/6	0	0	5/6	0	0
p_B	0	0	5/6	0	0	1/6
p_C	0	1/2	0	0	1/2	0

Figure 1: Distributions p_A, p_B, p_C.

Moreover, the SSB model can represent different risk attitudes via an adequate choice of functional φ. It accounts for a risk-averse (resp. risk-seeking) behavior (in the weak sense) if the *certainty equivalent* of a distribution p is less (resp. greater) than or equal to its expected value, where the certainty equivalent of a distribution p is the wealth level x such that $\varphi(p, x) = 0$ (which implies $p \sim x$). For existence and unicity of the certainty equivalent, the following conditions on φ should hold for all $x, y, t \in \mathcal{W}$: (1) for $x \neq y$, $\varphi(x, y)/\varphi(y, z)$ is continuous for all $z \in \mathcal{W}$ except $z = y$, (2) $\varphi(y, t)/\varphi(x, t)$ is strictly increasing in t for x, y, t with $x < t < y$. Condition 1 implies that every distribution has at least one certainty equivalent [Fishburn, 1986], while condition 2 implies that the certainty equivalent is unique [Nakamura, 1989]. By defining $\varphi_1(x, y) = \partial\varphi(x, y)/\partial x$, Nakamura showed that, when $\varphi_1(x, x) \neq 0$ exists, φ is weakly risk-averse (risk-seeking) if and only if $\varphi_1(y, y)/\varphi(x, y) \geq (\leq) 1/(x - y)$. We return to this point later in Section 6.

The SSB model encompasses many decision criteria, e.g.:
- $\varphi(x, y) = x - y$ yields the expectation criterion;
- $\varphi(x, y) = u(x) - u(y)$ yields the EU model;
- $\varphi(x, y) = \delta_\theta(x) - \delta_\theta(y)$, where $\delta_\theta(x) = 1\,(0)$ if $x \geq (<)$ θ, yields the probability threshold criterion [Yu *et al.*, 1998], which states that $p \succ q$ if $\sum_{x \geq \theta} p(x) > \sum_{x \geq \theta} q(x)$;
- $\varphi(x, y) = 1$ (resp. $0, -1$) if $x > y$ (resp. $x = y$, $x < y$) yields the dominance relation of Example 1, which states that $p \succ q$ if $\sum_{x>y} p(x)q(y) > \sum_{y>x} p(x)q(y)$; this is called *probabilistic dominance* (PD) in the sequel.

Consequently, designing and implementing an algorithm for solving MDPs with SSB also gives us a tool to compute optimal policies for all these special cases. Although dedicated algorithms exist for solving MDPs with an EU objective [Liu and Koenig, 2008] or a threshold probability objective [Hou *et al.*, 2014], it is interesting to have a generic algorithm easily adaptable to a large class of decision criteria. The next section presents the augmented MDP framework that will be used to determine SSB optimal policies.

3 Markov Decision Processes with SSB utility

3.1 Background

We study in this paper MDPs with finite state and action spaces, modeling finite horizon problems. As usual, an MDP is formally defined by $\mathcal{M} = (T, \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, c)$ where: T, a positive integer, is the time horizon; $\mathcal{S}$ is a finite collection of states, one of which is designated as the initial state;

$\mathcal{A} = \{\mathcal{A}_s | s \in \mathcal{S}\}$ is a collection of finite sets of possible actions, one set for each state; $\mathcal{P} = \{\mathcal{P}_t | t = 0, \ldots, T-1\}$ is a collection of transition probabilities where $\mathcal{P}_t(s'|s,a)$ is the probability that the state at time step $t+1$ is s' given that the state at time step t was s and that we have performed action a; $\mathcal{R}$ is the set of possible immediate rewards; $c = \{c_t | t = 0, \ldots T-1\}$ is a collection of reward functions where $c_t(s,a,s')$ is the reward obtained if the state at time step $t+1$ is s' given that the state at time step t was s and that we have performed action a. To illustrate our notations on a voluntarily simple sequential decision problem, we (artificially) modify Example 1 from the introduction.

Example 3 (One-agent sequential variant of Rowett dice)**.** *An agent has first to choose whether she wants to throw (action a_1) or not (action a'_1) die A; if this is not the case, she needs to choose between die B (action a_2) or C (action a_3). Whatever die is chosen, we distinguish two cases: success (state s_3) if one of the advantageous faces of the die is rolled (e.g., a face 4 for die A), or failure (state s_2) otherwise. The objective function is the number rolled, to maximize.*

The decision problem can be modeled by the MDP represented in Figure 2 with $T = 2, \mathcal{S} = \{s_1, s'_1, s_2, s_3\}, \mathcal{A} = \{\{a_1, a'_1\}, \{a_2, a_3\}, \cdots, \{a_5\}\}$, $\mathcal{R} = \{0, \cdots, 6\}$ and where c_1, c_2 and c_3 are the chance nodes induced by $\mathcal{P}$. The values $c_t(s,a,s') \,|\, \mathcal{P}_t(s'|s,a)$ (that do not depend on t in this example) are shown along the edges.

We call t-history a succession of state-action pairs of length t, $h_t = (s_0, a_0, s_1, \ldots, s_{t-1}, a_{t-1}, s_t)$.

A decision rule δ_t indicates which action to perform in each state for a given time step t. A decision rule can be *history-dependent* meaning that it takes as argument the entire history generated so far or *Markovian* if it only takes as argument the current state. A decision rule will be *deterministic* if it always prescribes an action per state or *randomized* if it prescribes a probability distribution over actions per state.

A policy π at an horizon T is a sequence of T decision rules $(\delta_0, \ldots, \delta_{T-1})$. Note that our policies are non-stationary in the sense that the decision rules can be different, depending on the time step. A policy can be *history-dependent*, *Markovian*, *deterministic* or *randomized* according to the type of decision rules. We use the notations of Table 1 for the sets of the different types of policies. Importantly, given a set $\Pi = \{\pi_1, \pi_2, \ldots\}$ of policies, we define an enlarged set $\widetilde{\Pi}$ of policies, that denotes the set consisting of mixtures of policies, i.e., $\widetilde{\Pi} = \{\tilde{\pi} = (\pi_1 | \alpha_1, \pi_2 | \alpha_2, \ldots) : \sum_i \alpha_i = 1, \alpha_i \geq 0\}$, where $\tilde{\pi}$ is the *mixed policy*[1] that randomly selects policy π_i with probability α_i.

[1]Not to be confused with the notion of randomized policies.

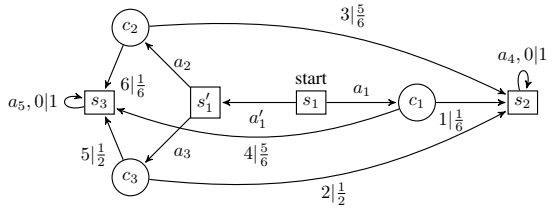

Figure 2: The MDP in Example 3.

	Markovian	history-dependent
deterministic	Π^t_s	Π_h
randomized	$\Pi^t_{s,r}$	$\Pi_{h,r}$

Table 1: Policy Notations

3.2 Comparing Two Policies

To compute the SSB criterion for comparing two policies, pairs of histories have to be compared (via the φ functional). In order to embed histories in the wealth level space $\mathcal{W}$, we assume that the value of a history is completely defined by the final wealth level accumulated along the history. Let $\mathcal{W}_T = \{w_1, \ldots, w_m\}$ be the set of attainable wealth levels at horizon T. Following Iwamoto [2004], we define the wealth level, $\omega(h_T) \in \mathcal{W}_T$ of a T-history, h_T, as an aggregation of rewards along h_T:

$$\omega(h_0) = \lambda$$
$$\omega(h_t) = \omega(h_{t-1}) \circ c_{t-1}(s_{t-1}, a_{t-1}, s_t)$$

where $\circ$ is a binary operator defined on $\mathcal{W} \times \mathcal{R}$ and $\lambda \in \mathcal{W}$ is a left identity element with respect to $\circ$. Note that no special property is required for $\circ$ as only final wealth levels need to be compared. Therefore, many different ways of evaluating T-histories are compatible with this setting.

The agent's SSB utility function φ defines a preference relation on probability distributions over possible wealth level outcomes $\mathcal{W}_T = \{w_1, \ldots, w_m\}$ and therefore defines a preference relation $\succsim$ on policies :

$$\varphi(\pi, \pi') = \sum_{i=1}^{m} \sum_{j=1}^{m} p^{\pi}_{w_i} p^{\pi'}_{w_j} \varphi(w_i, w_j) \quad (1)$$

$$\pi \succsim \pi' \equiv \varphi(\pi, \pi') \geq 0 \quad (2)$$

where p^{π}_x denotes the probability of x when applying policy π. As $\succsim$ depends on wealth levels, optimal policies will also depend on them. For this reason, we incorporate those values in the state space. Following Liu and Koenig [2008], we transform the given MDP into an augmented MDP whose states are pairs (s, w) where s is a state of the original MDP and $w \in \mathcal{W}$ a wealth level attainable by executing actions in the given MDP.

Example 4. *We illustrate the notion of augmented MDP on our modified Rowett dice problem, represented in Figure 2. In this example wealth levels are combined with the standard summation operator; $w \circ c_t(s,a,s') = w + c_t(s,a,s')$ and $\lambda = 0$. The augmented MDP is represented in Figure 3.*

To avoid any confusion, we denote by $\overline{s} = (s, w)$ a state in the augmented MDP. For instance, $\Pi^t_{\overline{s}}$ (resp. $\Pi^t_{\overline{s},r}$) denotes the set of deterministic (resp. randomized) Markovian

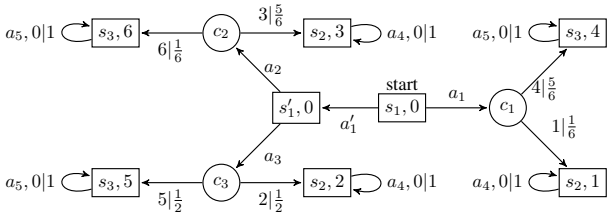

Figure 3: The augmented MDP in Example 4.

policies in the augmented MDP. Note that, in the augmented MDP, Markovian decision rules take into account both the current state and the wealth level accumulated so far.

Let $\overline{\mathcal{S}}_T \subseteq \mathcal{S} \times \mathcal{W}_T$ the set of final states in the augmented MDP. Now, in the augmented MDP, the preference relation over policies only depends on the probability distributions induced by policies over final wealth levels. Those distributions can be computed in the following way. Recall that in any MDP (augmented or not), a policy induces a probability distribution over histories and a fortiori over final states. The probability distribution over final states associated to a policy can be easily computed by a standard dynamic programming procedure. In the augmented MDP, assuming a probability distribution $(p^\pi_{\overline{s}})_{\overline{s}\in\overline{\mathcal{S}}_T}$ has been computed, the associated probability distribution over final wealth levels $(p^\pi_w)_{w\in\mathcal{W}_T}$ can simply be obtained by marginalization: $p^\pi_w = \sum_{\overline{s}=(s,w)\in\overline{\mathcal{S}}_T} p^\pi_{\overline{s}}$.

Collins and McNamara [1998] showed that, provided we are only interested in the probabilities of the final states, it is not a restrictive assumption to focus on mixed policies in $\widetilde{\Pi}^t_{\overline{s}}$. They indeed showed that for any mixed policy $\tilde{\pi}$ in $\widetilde{\Pi}_{h,r}$ there exists a mixed policy $\tilde{\pi}'$ in $\widetilde{\Pi}^t_{\overline{s}}$ such that $p^{\tilde{\pi}}_{\overline{s}} = p^{\tilde{\pi}'}_{\overline{s}}$ for all $\overline{s} \in \overline{\mathcal{S}}_T$.

Besides, from a mixed policy, it is possible to recover an equivalent (w.r.t. SSB) randomized policy. This is useful, in situations where an agent may prefer to apply a randomized policy instead of a mixed strategy. Strauch and Veinott [1966] showed how[2] to compute a randomized policy π in $\Pi^t_{\overline{s},r}$ from any mixed policy $\tilde{\pi}$ in $\widetilde{\Pi}^t_{\overline{s}}$ in such a way that $p^\pi_{\overline{s}} = p^{\tilde{\pi}}_{\overline{s}}$ for all $\overline{s} \in \overline{\mathcal{S}}_T$. Consequently, in the remainder of the paper, we focus on policies in $\widetilde{\Pi}^t_{\overline{s}}$ in the augmented MDP.

4 A Game on Policies

Assuming it exists, finding a preferred policy for our criterion is not straightforward. Not only can preference $\succsim$ be intransitive (as illustrated by the Rowett dice problem), but the SSB criterion also does not respect Bellman's principle of optimality. The violation of Bellman's principle of optimality is illustrated by our one-agent sequential variant of the Rowett dice problem represented in Figure 2. Indeed, the PD-optimal policy is the randomized policy $(s_1 \to (a_1|\frac{3}{13}, a'_1|\frac{10}{13}), s'_1 \to (a_2|\frac{3}{10}, a_3|\frac{7}{10}))$ (play dice A, B and C with probabilities $\frac{3}{13}$, $\frac{3}{13}$ and $\frac{7}{13}$). However, the PD-optimal sub-policy when in state s'_1 with $T = 1$ is to play a_2 (die B) with probability 1 as only dice B and C remain and that die B rolls higher than C most of the time. Therefore dynamic programming cannot be used directly and we turn to a game-theoretic analysis of the problem of identifying an SSB-optimal policy from the initial state.

When an MDP and a time step T are fixed, Equations 1 and 2 induce a zero-sum two-player symmetric game where the set of pure strategies can be reduced to $\Pi^t_{\overline{s}}$. Each player $i \in \{1, 2\}$ chooses simultaneously a strategy π_i (pure or mixed). The resulting payoff is then given by $\varphi(\pi_1, \pi_2)$.

In a zero-sum symmetric game of payoff function φ, it is well-known that there exists a symmetric Nash equilibrium (NE). The following holds for an NE (π^*, π^*):

$$\forall \pi, \varphi(\pi^*, \pi) \geq \varphi(\pi^*, \pi^*) = 0.$$

We aim at computing such an NE of the game on policies characterized by payoff function φ (*solving the game*) since strategy π^* will be SSB preferred to any other strategy and will therefore be an SSB optimal policy in the MDP.

Thanks to this game-theoretic view, it is now straightforward to prove that an SSB-optimal policy exists for any finite horizon MDP $(\mathcal{M}, T)$.

Theorem 1. *For any finite horizon MDP $(\mathcal{M}, T)$ and any SSB utility function φ, an SSB-optimal policy exists in $\widetilde{\Pi}^t_{\overline{s}}$.*

Proof. The set $\Pi^t_{\overline{s}}$ is finite ($T, \mathcal{S}$ and $\mathcal{A}$ are finite) and thus the game induced by the SSB criterion restricted to $\Pi^t_{\overline{s}}$ is finite. Therefore, the von Neumann minimax theorem ensures that an optimal strategy exists as a mixed policy in $\widetilde{\Pi}^t_{\overline{s}}$. □

Once realized that an NE of the game on policies of $\Pi^t_{\overline{s}}$ provides us an SSB optimal policy, our aim becomes to solve this game. However, note that the large size of $\Pi^t_{\overline{s}}$ prohibits solving it directly. We address this issue in the next section.

5 Solving the Game

In this section, we first describe the double oracle approach [McMahan *et al.*, 2003] that enables to solve large-size games by avoiding the *ex-ante* enumeration of all pure strategies. Then, we adapt this procedure to our problem by providing a best response procedure (*oracle*) to a given mixed policy.

In our setting, the double oracle approach is implemented by the procedure described in Algorithm 1, which is a simplified version of the initial proposal by McMahan *et al.*.

Algorithm 1: Double Oracle Algorithm

Data: Finite horizon MDP $(\mathcal{M}, T)$, singleton $\Pi' = \{\pi\}$ including an arbitrary policy $\pi \in \Pi^t_{\overline{s}}$
Result: an SSB optimal mixed policy $\tilde{\pi} \in \widetilde{\Pi}^t_{\overline{s}}$

converge = False
while *converge is False* **do**
  Find Nash equilibrium $(\tilde{\pi}, \tilde{\pi}) \in G = (\Pi', \Pi', \varphi|_{\Pi'})$
  Find $\pi = BR(\tilde{\pi}) \in \Pi^t_{\overline{s}}$
  if $\varphi(\pi, \tilde{\pi}) > 0$ **then**
    add π to Π'
  else
    converge =True
return $\tilde{\pi}$

Double oracle approach. The double oracle algorithm finds a Nash equilibrium for a finite two player game where a best response procedure $BR(\cdot)$ exists. Given a mixed strategy $\tilde{\pi}$, $BR(\tilde{\pi})$ returns a pure strategy π (a policy in $\Pi^t_{\overline{s}}$) that

[2] The randomized policy π corresponding to mixed policy $\tilde{\pi} = (\pi_1|\alpha_1, \pi_2|\alpha_2, \ldots)$ is obtained by the following:

$$\mathbb{P}(a_t = a|s_t = s, \pi) = \frac{\sum_i \alpha_i \mathbb{P}(s_t = s, a_t = a|\pi_i)}{\sum_{a' \in \mathcal{A}_s} (\sum_k \alpha_k \mathbb{P}(s_t = s, a_t = a'|\pi_k))}$$

maximizes $\varphi(\pi, \tilde{\pi})$. The original double oracle algorithm applies to any zero-sum two-player game. The operation can be described as follows in the symmetric case. The algorithm starts with a small set Π of pure strategies (a singleton in Algorithm 1), and then grows this set in every iteration by applying the best-response oracle to the optimal strategy (given by NE) the players can play in the restricted game $G = (\Pi', \Pi', \varphi|_{\Pi'})$, where Π' is the set of available strategies for both players and $\varphi|_{\Pi'}$ is the restriction of function φ to domain $\Pi' \times \Pi'$. An NE in a zero-sum two-player symmetric game can be computed by linear programming [Raghavan, 1994]. Execution continues until convergence is detected. Convergence is achieved when the best-response oracle does not generate a pure strategy π that is better than the current mixed strategy $\tilde{\pi}$. In other words, convergence is obtained if the payoff $\varphi(\pi, \tilde{\pi})$ given by the best-response oracle is not better than the payoff given by the current NE (0 for a zero-sum two-player symmetric game).

The correctness of best-response-based double oracle algorithms for two-player zero-sum games has been established by McMahan et al [2003]; the intuition for this correctness is as follows. Once the algorithm converges, the current solution must be an equilibrium of the game, because each player's current strategy is a best response to the other player's current strategy. This stems from the fact that the best-response oracle, which searches over all possible strategies, cannot find anything better. Furthermore, the algorithm must converge, because at worst, it will generate all pure strategies. In practice, we expect the restricted-game to stay relatively small as many pure strategies will never enter the restricted strategy set.

Best response oracle. In order to use the double oracle approach, we look for a procedure to find a policy which is a best response to a fixed mixed policy $\tilde{\pi} = (\pi_1|\alpha_1, \ldots, \pi_k|\alpha_k)$. This best response is an optimal policy according to the decision criterion maximizing the value function $v(\pi) = \varphi(\pi, \tilde{\pi})$ and amounts to taking $\tilde{\pi}$ as a reference point.

Let Φ denote the skew-symmetric matrix where $\Phi_{i,j} = \varphi(w_i, w_j)$, and $\mathbf{p}^{\pi}_w$ denote vector $(p^{\pi}_{w_1}, \ldots, p^{\pi}_{w_m})$. With such notations we can express $\varphi(\pi, \tilde{\pi})$ as:

$$\varphi(\pi, \tilde{\pi}) = {}^{\mathrm{t}}\mathbf{p}^{\pi}_w \Phi \mathbf{p}^{\tilde{\pi}}_w.$$

Denoting by e_i the i-th canonical vector, we can interpret $\Phi\mathbf{p}^{\tilde{\pi}}_w$ as a reward function where one would receive $(\Phi\mathbf{p}^{\tilde{\pi}}_w).e_i$ as a reward each time one obtains a final wealth value of w_i, and 0 at each previous time step. More formally, the reward function in the augmented MDP is then defined by:

$$c_t((s, w), a, (s', w')) = 0 \text{ for } t < T - 1$$
$$c_{T-1}((s, w), a, (s', w_i)) = (\Phi\mathbf{p}^{\tilde{\pi}}_w).e_i$$

Using this reward function, maximizing the classic expectation criterion is equivalent to maximizing ${}^{\mathrm{t}}\mathbf{p}^{\pi}_w \Phi \mathbf{p}^{\tilde{\pi}}_w$. A policy maximizing such criterion is classically found by backward induction in the augmented MDP, leading to a deterministic Markovian policy in Π^t_s.

Example 5. *Coming back to the sequential Rowett dice example, we show how to find the best response to the policy $\tilde{\pi}$ that chooses die A with probability* 1. *Assume that one uses the probabilistic dominance criterion, i.e., Φ is the skew-symmetric matrix with ones below the diagonal. In this example, $(w_1, \ldots, w_m) = (1, 2, \ldots, 6)$ and $\mathbf{p}^{\tilde{\pi}}_w = (\frac{1}{6}, 0, 0, \frac{5}{6}, 0, 0)$. Determining $BR(\tilde{\pi})$ amounts then to computing the policy maximizing expectation in the augmented MDP represented in Figure 3, where the reward function is replaced by :*

$$c_0((s, w), a, (s', w')) = 0$$
$$c_1((s, w), a, (s', w_i)) = \Phi\mathbf{p}^{\tilde{\pi}}_w = (-\frac{5}{6}, -\frac{4}{6}, -\frac{4}{6}, \frac{1}{6}, 1, 1).e_i$$

Unsurprisingly, the policy obtained will play die C.

6 Experiments

We provide here experimental results in order to demonstrate the operationality of the method and to provide a deeper insight on the interest of using an SSB utility function.

Who Wants to Be a Millionaire? In this popular television game show, a contestant tries to answer a sequence of 15 multiple-choice questions of increasing difficulty. Questions (four possible answers are given) are played for increasingly large sums, roughly doubling the pot. At each time step, the contestant may decide to walk away with the money currently won. If she answers incorrectly then all winnings are lost besides what has been earned at a "guarantee point" (questions 5 and 10). The player is given the possibility of using 3 lifelines (*50:50*, removing two of the possible choices, *ask the audience* and *call a friend* for suggestions); each can only be used once in the whole game.

We used the two models of the Spanish 2003 version of the game presented by Perea and Puerto [2007].[3] In the first model the probability of answering correctly is a function of the question's number and the lifelines (if any) used; lifelines increase the probability of answering correctly (the model is fitted using real data). This first model is overly simplistic as it does not actually take into account whether the player does

[3]The possible wealth values are 0, 150, 300, 450, 900, 1800, 2100, 2700, 3600, 4500, 9k, 18k, 36k, 72k, 144k, 330k.

	Exp	PD	PD	PD	RA	Th
p	1	0.07	0.39	0.54	1	1
t=1	NL	NL	NL	NL	NL	NL
t=2	NL	NL	NL	NL	NL	NL
t=3	NL	NL	NL	NL	NL	NL
t=4	NL	50:50	50:50	50:50	50:50	NL
t=5	NL	P	P	P & A	P&A	A
t=6	NL	NL	NL	NL	NL	50:50
t=7	NL	A	NL	NL	NL	P
t=8	NL	S	A	S	NL	S
t=9	50:50		S		NL	
t=10	P				NL	
t=11	NL				NL	
t=12	A				S	
t=13	S					

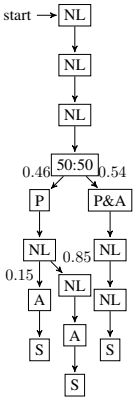

Figure 4: First model of the game. (Left) Optimal mixed policies according to the following criteria : Exp=Expectation, PD=Probabilistic Dominance, RA=Risk-averse, Th=Threshold. (Right) Optimal randomized policy according to the PD criterion (NL=No Lifelines, P=Call a friend, A=Audience, S=Stop)

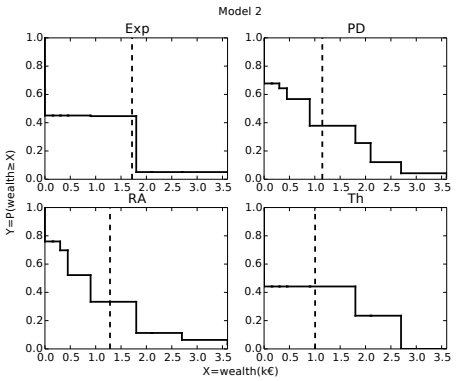

Figure 5: Second model of the game. Decumulative probability distribution of wealth according to the different SSB utility functions. The dashed vertical line indicates the expected wealth.

in fact know the answer or not. The second model represents the hesitation of the contestant by distinguishing four epistemic cases, corresponding to the number of answers (among the four given) that are believed possible correct answers for the current question. At each step of the game, when a new question is asked, a categorical distribution dictates the probability of each of the epistemic cases. As the game proceeds, questions are more difficult, and the distribution is then skewed towards hesitating between larger sets of answers.

We computed the optimal policies for the two models according to several instantiations of the SSB utility function: the expectation (Exp), probabilistic dominance (PD), threshold probability (Th) criteria (threshold set to 2700) and a risk averse SSB utility function (RA) defined by $\varphi_{RA}(x, y) = (x - y)/(x + y)^{\frac{2}{3}}$, which is indeed risk averse since Nakamura's condition holds (Section 2).

Given the simplicity of the first model of the game, policies can be compactly displayed; the optimal policies are reported on the left of Figure 4. The Exp, RA and Th criteria are associated with deterministic policies, while the PD-optimal policy can be seen either as a mixed policy or as a randomized policy (see right of Figure 4). According to intuition, the preferred policies for PD and RA make use of lifelines much earlier in the game, in order to secure a significant gain; the preferred policy for Th uses the lifelines to reach 2700 with high probability and stops playing thereafter. Finally, with Exp, one keeps aside the lifelines for later (more difficult) questions, even at a cost of a premature end of the game.

When using the second model, the state space is much larger and policies are too complex to be represented compactly. As we are interested in comparing their overall performance, we plot in Figure 5 the decumulative distribution of wealth (as the pot sky-rockets if the contestant reaches the very last questions, but this happens for all policies with low probabilities, we plot only the 9 first wealth levels for emphasizing the differences between the wealth distributions).

Unsurprisingly, the expectation-optimal policy yields the highest wealth expectancy (2387 and 1717 in the two models of the game). While the optimal policy according to PD achieves a lower expected value, it scores better very often: it achieves a wealth level at least as good as 75% (resp. 70%) of the time and strictly better 44% (resp. 48%) of the time, when considering the first (resp. second) model. The policy obtained with the RA criterion is safer than the expectation-optimal policy. Regarding the threshold-optimal policy, it obtains with higher probability (23%) the threshold objective as can be seen in Figure 5 (second model).

Regarding the computational aspect, the initial MDP in this problem consists of 9 (resp. 136) states for the first model (resp. second model) and the corresponding augmented MDP has 33 (resp. 496) states. The computation time (resp. number of iterations of the double oracle algorithm) for finding each optimal policy for the second model of the game was respectively of 7.6s (resp. 1) for the Exp criterion, 9.0s (resp. 8) for PD criterion, 7.8s (resp 3) for the RA criterion and 7.6s (resp 1) for the Th criterion.[4]

Other domains. To have a deeper insight into the operationality of our method, we have carried out some preliminary experiments on other domains, notably on a simple grid world (GW) domain (with randomly generated instances) and a cancer clinical trials (CCT) domain (with a model proposed by Cheng *et al.* [2011]). In both domains, we used the probabilistic dominance criterion. In the GW domain, the initial state space includes from 100 to 400 states, and the augmented state space from thousands to tens of thousands states. The computation times vary from a few seconds to half an hour. Regarding the CCT domain, the initial state space includes 5078 states, and the augmented state space 5129 states. The method takes about two minutes to compute an optimal policy. Generally speaking, provided the structure of the MDP prevents the augmentation of the state space to be too costly (as can bee seen in the two previous examples, there is an important variability in the increase of the number of states after augmentation of the MDP), we believe the method is rather scalable. A more thorough study of its scalability is underway.

7 Conclusion

Skew Symmetric Bilinear utility functions (SSB) is a useful general decision model that encompasses many decision criteria (e.g., EU, threshold probability, probabilistic dominance...). We showed that there exists an optimal (potentially randomized) Markovian policy in an augmented MDPs where preferences are described with an SSB utility function and we proposed an iterative solution method based on a game-theoretic view of the optimization of an SSB utility function.

Our current work can be extended in several natural ways. First, a theoretical study in order to guarantee an upper bound on the number of iterations of the double oracle algorithm is needed. Besides, it would be useful to extend this work to factored MDP in order to tackle large size problems. Moreover, it would be interesting to investigate the use of SSB utility functions in reinforcement learning settings.

[4] All times are wall-clock times on a 2,4 GHz Intel Core i5 machine with 8G main memory. Our implementation is in Python, with an external call to GUROBI version 5.6.3 in order to solve the linear programs required to find the Nash equilibria.

References

[Allais, 1953] M. Allais. Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'ecole americaine. *Econometrica*, 21(4):pp. 503–546, 1953.

[Blythe, 1999] J. Blythe. Decision-theoretic planning. *AI Magazine*, 20(2):37–54, 1999.

[Boutilier *et al.*, 1999] C. Boutilier, T. Dean, and S. Hanks. Decision theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, 11:1–94, 1999.

[Cheng *et al.*, 2011] W. Cheng, J. Fürnkranz, E. Hüllermeier, and S.-H. Park. Preference-based policy iteration: Leveraging preference learning for reinforcement learning. In *Proc. of European Conference on Machine Learning and Knowledge Discovery in Databases ECML PKDD Part I*, pages 312–327, 2011.

[Collins and McNamara, 1998] E.J. Collins and J.M. McNamara. Finite-horizon dynamic optimisation when the terminal reward is a concave functional of the distribution of the final state. *Adv. in Appl. Probab.*, 30(1):122–136, 1998.

[Fishburn, 1984] P.C. Fishburn. SSB utility theory: an economic perspective. *Mathematical Social Sciences*, 8(1):63 – 94, 1984.

[Fishburn, 1986] P.C. Fishburn. Implicit mean value and certainty equivalence. *Econometrica*, 54(5):1197–1205, 1986.

[Gardner, 1970] M. Gardner. Mathematical games: The paradox of nontransitive dice and the elusive principle of indifference. *Sci. Amer.*, 223:110–114, Dec. 1970.

[Hou *et al.*, 2014] P. Hou, W. Yeoh, and P. Varakantham. Revisiting risk-sensitive MDPs: New algorithms and results. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*, pages 136–144, 2014.

[Iwamoto, 2004] S. Iwamoto. Stochastic optimization of forward recursive functions. *Journal of Mathematical Analysis and Applications*, 292(1):73 – 83, 2004.

[Kahneman and Tversky, 1979] D. Kahneman and A. Tversky. Prospect theory: An analysis of decisions under risk. *Econometrica*, pages 263–291, 1979.

[Liu and Koenig, 2005] Y. Liu and S. Koenig. Existence and finiteness conditions for risk-sensitive planning: Results and conjectures. In *In UAI*, pages 354–363, 2005.

[Liu and Koenig, 2006] Y. Liu and S. Koenig. Functional value iteration for decision-theoretic planning with general utility functions. In *AAAI*, pages 1186–1193. AAAI Press, 2006.

[Liu and Koenig, 2008] Y. Liu and S. Koenig. An exact algorithm for solving MDPs under risk-sensitive planning objectives with one-switch utility functions. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '08, pages 453–460, 2008.

[Liu, 2005] Y. Liu. Risk-sensitive planning with one-switch utility functions: Value iteration. In *In AAAI*, pages 993–999, 2005.

[McMahan *et al.*, 2003] H.B. McMahan, G.J. Gordon, and A. Blum. Planning in the presence of cost functions controlled by an adversary. In *ICML*, pages 536–543, 2003.

[Nakamura, 1989] Y. Nakamura. Risk attitudes for nonlinear measurable utility. *Annals of Operations Research*, 19:pp. 311–333, 1989.

[Perea and Puerto, 2007] F. Perea and J. Puerto. Dynamic programming analysis of the TV game who wants to be a millionaire? *European Journal of Operational Research*, 183(2):805 – 811, 2007.

[Puterman, 1994] M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.

[Raghavan, 1994] T.E.S Raghavan. Zero-sum two-person games. *Handbook of game theory*, 2:735–768, 1994.

[Strauch and Veinott, 1966] R. Strauch and Jr. A.F. Veinott. A property of sequential control processes. Technical report, Rand McNally, Chicago, 1966.

[von Neumann and Morgenstern, 1947] J. von Neumann and O. Morgenstern. *Theory of games and economic behaviour*. Princeton University Press, 1947.

[Yu *et al.*, 1998] S.X. Yu, Y. Lin, and P. Yan. Optimization models for the first arrival target distribution function in discrete time. *Journal of Mathematical Analysis and Applications*, 225(1):193–223, 1998.