SCHRIFTENREIHE DER FAKULTÄT FÜR MATHEMATIK

Risk Analysis of Two Different Stochastic Models Based on a Dice Game

by

Johanna Burtscheidt and Alexander Frank

SM-UDE-819 2018

Received: November 13, 2018

Risk Analysis of Two Different Stochastic Models Based on a Dice Game

Johanna Burtscheidt* Alexander Frank†

November 20, 2018

Abstract

The analysis of stochastic games is an important part of game theory. In such games, separate rounds are frequently uncorrelated and so they can be described as Markov chain or Markov Decision Process. As an example of such stochastic games, we studied a special game of dice with three dice for two (or more) players. Here each participant has to choose one of three options in every round to maximize its score. In doing so they have to make their decision without knowing the realization of some randomness, as among other things the outcome of two dice is unknown. This type of problem can be formulated as a bilevel optimization problem under stochastic uncertainty, in which both the leader and the follower have a lack of information.

We will construct several deterministic bilevel formulations which allow to take risk aversion into account. The objective is to compare two known mathematical methods for a stochastic decision game on different points of views: For a selection of 'good' options, some similar instances of the dice game will be juxtaposed. In addition to that, we compare performance and complexity of our models.

Keywords: Markov Decision Process, Bilevel Optimization Problem, Risk Aversion, Uncertainty, Multi-criteria Optimization
MSC(2010): 60J20, 65C40, 90C15, 91A15, 91A65

1 Introduction

Bilevel programming problems are an example of hierarchical optimization problems, as they arise from an interplay of two decision makers on different levels of a hierarchy. The upper level objective function is dependent on the leader's decision and the solution returned by the lower level modeling the follower's goal. It is assumed, the leader has complete information about the influence of his or her decision on the problem of the follower.

In recent years, however, the inclusion of data uncertainty in problem formulations has become increasingly important as it presents a great difficulty in many real world applications. In stochastic bilevel problems it is assumed that the leader has to make his or her decision without knowing the realization of a certain randomness, while the follower generally decides under full information. Unlike the two-stage stochastic problem, however, a feedback between the lower level and the upper level can be represented. Practically relevant examples include the pricing of electricity swing options ([7]), supply chain planning ([3]), general agency problems ([6]), and have been discussed in the context of transportation ([1]) and economics ([2]).

Another model, searching for decisions and strategies, is a Markov decision process (MDP) ([8]). The concept of these models is to estimate the best option in a current state with a critical look to the future. The underlying mathematical task is independent of previous decisions. This memoryless model is often used to describe stochastic games and is also relevant in robotics and automatic systems. The advantages of MDPs are in constructive modeling and good methods for solving a underlying problem.

*Corresponding author, Faculty of Mathematics, University of Duisburg-Essen, 45117 Essen, Germany; email: johanna.burtscheidt@uni-due.de.
†Faculty of Computer Science, TU Dortmund University, 44221 Dortmund, Germany; email: alexander.frank@cs.tu-dortmund.de.

The present work is on a comparison of two modelings based on a bilevel problem and an MDP for a special class of stochastic games, since both of these models can describe state-discrete stochastic processes. Also the handling of the problem that the first player in the considered dice game of this class has less information than the subsequent player is a hard modeling task and requires strategies from bilevel optimization. In addition, dealing with risk aversion is in our focus and leads to several formulations of our modelings. We will see that we gain a multicriteria optimization task based on different player goals. An additional problem is to separate all "good" solutions as players can choose their own strategy through a given risk function.

How adaptable are the two modelings in relation to some uncertainty? Is it possible to say that one method is better or obviously faster than the other?

Therefor we give a small introduction in bilevel programming and Markov decision processes and concretize the models based on an example presented after a short presentation of the game of dice. Based on the second question we will analyze the running times of the methods depending on generalizations of the game modelings.

2 Dice game

We will analyse a game of dice which is based on the game "Der Wächter bläst vom Turm" ([9, pp.15-16]) (german for "the guard blows from the tower"): You need three dice and a dice cup. Two dice are rolled in the dice cup, they remain hidden. The third cube is placed on top of the cup rim and blown down, so that each player is able to observe it. Every player has to choose exact one of three options:

1. add up the eyes of all three cubes, or
2. multiply the eyes under the cup by the eyes of the third cube, or
3. square the eyes of the blown cube and divide this number by the eyes under the cup.

Then the dice cup is turned over and the player with the lowest score, loses a life point.

The choice of the best option depending on the number of eyes of the third dice $\mathtt{T}$ is very simple, as we can see in Table 1a. $\mathtt{D}_1 + \mathtt{D}_2$ is the sum of the hidden cube's eyes.

2.1 Reformulation of the game of dice

As a result of this determination, we have defined new options

1. square the eyes of the third cube: $\mathtt{T}^2$, or
2. multiply the eyes of the dice under the cup by the eyes of the third cube and add 2: $\mathtt{T} \cdot (\mathtt{D}_1 + \mathtt{D}_2) + 2$, or
3. square the cube's eyes under the cup: $(\mathtt{D}_1 + \mathtt{D}_2)^2$.

In addition, each rolled "6" of the hidden dice is set to "0". The best options for this reformulation can be found in Table 1b. Of course, another definition of the options is also conceivable.

The considered game is therefore a strategic game in which each player can choose from three pure strategies (options). For the sake of simplicity, we set the number of players to be two, and each player has an own set of dice. The players should not choose their strategy at the same time, but in succession. The second player (follower) can thereby react to the choice of the first player (leader). It creates a sequential game. Henceforth we give functions, variables and other structures an index $_l$ if it depends only on the leader and analogous an index $_f$ for the follower. In choosing their strategy, our players have incomplete

Table 1: Best options

D_1+D_2 \ T	1	2	3	4	5	6
2	1	1|2	2	2|3	3	3
3	1	2	2	2	2	2
	1	2	2	2	2	2
4	1	2	2	2	2	2
	1	2	2	2	2	2
	1	2	2	2	2	2
5	1	2	2	2	2	2
	1	2	2	2	2	2
	1	2	2	2	2	2
	1	2	2	2	2	2
6	1	2	2	2	2	2
	1	2	2	2	2	2
	1	2	2	2	2	2
	1	2	2	2	2	2
	1	2	2	2	2	2
7	1	2	2	2	2	2
	1	2	2	2	2	2
	1	2	2	2	2	2
	1	2	2	2	2	2
	1	2	2	2	2	2
	1	2	2	2	2	2
8	1	2	2	2	2	2
	1	2	2	2	2	2
	1	2	2	2	2	2
	1	2	2	2	2	2
	1	2	2	2	2	2
9	1	2	2	2	2	2
	1	2	2	2	2	2
	1	2	2	2	2	2
	1	2	2	2	2	2
10	1	2	2	2	2	2
	1	2	2	2	2	2
	1	2	2	2	2	2
11	1	2	2	2	2	2
	1	2	2	2	2	2
12	1	2	2	2	2	2

(a) Best options for original game idea

D_1+D_2 \ T	1	2	3	4	5	6
0	2	1	1	1	1	1
1	2	1|2	1	1	1	1
	2	1|2	1	1	1	1
2	2|3	2	1	1	1	1
	2|3	2	1	1	1	1
	2|3	2	1	1	1	1
3	3	3	2	1	1	1
	3	3	2	1	1	1
	3	3	2	1	1	1
	3	3	2	1	1	1
4	3	3	3	2	1	1
	3	3	3	2	1	1
	3	3	3	2	1	1
	3	3	3	2	1	1
	3	3	3	2	1	1
5	3	3	3	3	2	1
	3	3	3	3	2	1
	3	3	3	3	2	1
	3	3	3	3	2	1
	3	3	3	3	2	1
	3	3	3	3	2	1
6	3	3	3	3	3	2
	3	3	3	3	3	2
	3	3	3	3	3	2
	3	3	3	3	3	2
	3	3	3	3	3	2
7	3	3	3	3	3	3
	3	3	3	3	3	3
	3	3	3	3	3	3
	3	3	3	3	3	3
8	3	3	3	3	3	3
	3	3	3	3	3	3
	3	3	3	3	3	3
9	3	3	3	3	3	3
	3	3	3	3	3	3
10	3	3	3	3	3	3

(b) Best options for reformulated game

information because they do not know all eyes of the dice and, in the case of the leader, the strategy of the opponent:

$$\begin{aligned}\text{revelation of } \mathtt{T}_l, \mathtt{T}_f \quad &\rightarrow \quad \underbrace{\text{strategy } \mathtt{A}_l \text{ in dependence of } \mathtt{T}_l, \mathtt{T}_f}_{\mathtt{A}_f \text{ and } \mathtt{D}_{l1}, \mathtt{D}_{l2}, \mathtt{D}_{f1}, \mathtt{D}_{f2} \text{ unknown}} \\ &\rightarrow \quad \underbrace{\text{strategy } \mathtt{A}_f \text{ in dependence of } \mathtt{T}_l, \mathtt{T}_f, \mathtt{A}_l}_{\mathtt{D}_{l1}, \mathtt{D}_{l2}, \mathtt{D}_{f1}, \mathtt{D}_{f2} \text{ unknown}} \\ &\rightarrow \quad \text{revelation of } \mathtt{D}_{l1}, \mathtt{D}_{l2}, \mathtt{D}_{f1}, \mathtt{D}_{f2} \\ &\rightarrow \quad \text{score } \mathtt{V}_l, \mathtt{V}_f.\end{aligned} \tag{1}$$

From now on, the winner of one round gets points. If there is a tie, the leader wins. It is a non-zero-sum game that could be modeled as a Bimatrix game. We get a non-cooperative game, since both player want to score higher than his or her opponent. Every player wants to choose the " best " strategy. Unfortunately, chance and possibly the opponent can influence the game. The art of a player is therefore to take advantage of the opponent's behavior if possible. Otherwise, he or she should at least try to play in such a way that the hostile behavior of the opponent can do as little harm as possible.

But which strategy is the "best" strategy for one of the players depending on the eyes $\mathtt{T}_l$, $\mathtt{T}_f$ of the two third dice? In general and due to the incomplete information, "the" best strategy doesn't exist. Therefore the question arises, how to win:

1. Should be won with the highest possible difference to the opponent, or
2. is it only important that you win?

From these options of evaluation, there are two possible solutions: We are looking for a strategy

1. where an average amount of the difference with which one could win is maximum.
2. that maximizes the number of possible profitable eye combinations.

In the first, qualitative approach we will maximize the expected value of the difference of the scores and receive $\mathtt{V}_l - \mathtt{V}_f$ after the realization of the randomness from the perspective of the leader. In the quantitative approach, we get $\chi(\mathtt{V}_l, \mathtt{V}_f)$ for $\chi : \mathbb{R}_+ \times \mathbb{R}_+ \to \{0,1\}$ with

$$\chi(v_l, v_f) := \begin{cases} 1 & \text{if } v_l - v_f > 0, \\ 1 & \text{if } v_l - v_f = 0 \ (0 \text{ for } \chi(v_f, v_l)), \\ 0 & \text{if } v_l - v_f < 0 \end{cases}$$

after we maximize the probability for the leader to win.

3 Mathematical approaches

For the mathematical modeling of the two approaches, we use, as already stated in the introduction, an MDP and a bilevel optimization problem under stochastic uncertainty, since the players have to choose their decisions in succession. We start with the description of the latter.

3.1 Bilevel problems under stochastic uncertainty

Let $W := \{w \in \mathbb{R}^3_+ : \ |w| = 1\}$ be the feasible set for the upper decision variable x and the lower decision variable y. Since the players have to decide under incomplete information, let $(d_{l1}, d_{l2}, d_{f1}, d_{f2}) =: \omega \in \Omega$ be a random vector defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$. The mapping for determining the score of a player is $c : \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}^3_+$ with

$$c(t, d_1, d_2) := \left(t^2, t(d_1 + d_2) + 2, (d_1 + d_2)^2\right)^\top.$$

From now on we want to look at the outcome of the dice game from the perspective of the leader: He or she has to reckon on his or her opponent's worst solution y, as both players want to score higher than their opponent. We will therefore model the pessimistic approach to the bilevel problem. Because of the two third dice the following bilevel problem under stochastic uncertainty is a parametric optimization problem for all $\mathtt{T}_l, \mathtt{T}_f \in \{1,\dots,6\}$:

$$\big(\mathrm{P}(\mathtt{T}_l,\mathtt{T}_f,\omega)\big) \qquad \max_{x\in W}\left\{c(\mathtt{T}_l,d_{l1},d_{l2})^\top x - \max_{y\in W}\left\{c(\mathtt{T}_f,d_{f1},d_{f2})^\top y : \; y\in\Psi(\mathtt{T}_l,\mathtt{T}_f,x,\omega)\right\}\right\} \ \mathbb{P}\text{-almost surely}$$

with the multifunction $\Psi : \mathbb{R}_+\times\mathbb{R}_+\times\mathbb{R}^3_+\times\Omega \rightrightarrows \mathbb{R}^3_+$,

$$\Psi(t_l,t_f,x,\omega) := \operatorname*{argmax}_{y\in W}\big\{\underbrace{c(t_f,d_{f1},d_{f2})^\top y - c(t_l,d_{l1},d_{l2})^\top x}_{=:\psi(t_l,t_f,x,y,\omega)}\big\}.$$

The interplay of decision and observation during the dice game (1) is mathematically based on a nonanticipativity constraint. This additional condition states the outcome of the random vector ω is not known before the decision of x and y, as leader and follower have to decide before the realization of ω. We assume both decisions may not influence the distribution of ω, i.e. the stochasticity is purely exogenous. In our setting, the follower's decision gives rise to a random variable $\psi(\mathtt{T}_l,\mathtt{T}_f,x,y,\cdot)$ and the lower level problem corresponds a one-stage stochastic program. With the help of a resulting deterministic well defined problem, the set of optimal solutions $\bar\Psi(\mathtt{T}_l,\mathtt{T}_f,x)$ depends only on the parameters $\mathtt{T}_l,\mathtt{T}_f$ and the upper level decision x. Henceforth we assume the leader is conscious of the follower's model. We gain an other random variable $f(\mathtt{T}_l,\mathtt{T}_f,x,\cdot)$ as well as a one-stage stochastic bilevel problem

$$\max_{x\in W}\Big\{\underbrace{c(\mathtt{T}_l,d_{l1},d_{l2})^\top x - \max_{y\in W}\left\{c(\mathtt{T}_f,d_{f1},d_{f2})^\top y : \; y\in\bar\Psi(\mathtt{T}_l,\mathtt{T}_f,x)\right\}}_{=:f(\mathtt{T}_l,\mathtt{T}_f,x,\omega)}\Big\},$$

which arises from $\big(\mathrm{P}(\mathtt{T}_l,\mathtt{T}_f,\omega)\big)$. Deterministic problem formulations for this problem as well as for the lower level problem can be modeled with the help of some special mapping $\mathscr{R} : L^0(\Omega,\mathscr{F},\mathbb{P}) \to \mathbb{R}\cup\{\pm\infty\}$, so that the players can handle the stochastic uncertainty and rank the respective random variables f and ψ:

$$\big(\bar{\mathrm{P}}_{\mathscr{R}_l,\mathscr{R}_f}(t_l,t_f)\big) \qquad \max_{x\in W}\left\{\mathscr{R}_l\big[f_{\mathscr{R}_f}(t_l,t_f,x,\cdot)\big]\right\}$$

and

$$\bar\Psi_{\mathscr{R}_f}(t_l,t_f,x) := \operatorname*{argmax}_{y\in W}\big\{\mathscr{R}_f[\psi(t_l,t_f,x,y,\cdot)]\big\}.$$

Suitable examples are the expectation $\mathbb{E}[\cdot]$, and the excess probability $EP_\eta[\cdot] = \mathbb{P}[\,\cdot > \eta]$ over a fixed target level $\eta\in\mathbb{R}$ ([10]). The latter is a risk measure and allows to take into account risk aversion ([4], [5]). Using these formulations, we can model exactly the two options of evaluation already mentioned in Section 2.1:

1. Risk neutral approach using expected value: Determination of the maximum expected difference between the two scores,
2. Risk averse approach with $\eta := 0$: Determine the maximum probability to win, with the amount of win or defeat play no role.

Let ω be discrete (now it includes the unknown numbers of eyes of the hidden dice) with a finite number realizations $\omega_1,\dots,\omega_{|\Omega|}$. Let $(\Omega,2^\Omega,\mathbb{P})$ be a discrete probability space with sample space $\Omega := \{0,\dots,5\}^4$ and probability measure $\mathbb{P}(\{\omega_i\}) = \frac{1}{|\Omega|}$ for $1\le i\le|\Omega|$. We change set W to $W_3 := \{w\in\mathbb{N}^3_0 : \; |w| = 1\}$ and mapping c so that $c : \{1,\dots,6\}\times\{0,\dots,5\}\times\{0,\dots,5\}\to\mathbb{N}^3_0$. Assume $\Psi(\mathtt{T}_l,\mathtt{T}_l,x,\omega_i)\neq\emptyset$ for any $(\mathtt{T}_l,\mathtt{T}_l,x,\omega_i)\in\{1,\dots,6\}\times\{1,\dots,6\}\times W\times\Omega$.

In the risk neutral case of the leader, we receive

$$\begin{aligned}
&\left(\bar{\mathrm{P}}_{\mathbb{E},\mathscr{R}_f}(\mathtt{T}_l,\mathtt{T}_f)\right)\\
&\qquad \max_{x\in W_3}\left\{\mathbb{E}\left[f_{\mathscr{R}_f}(\mathtt{T}_l,\mathtt{T}_f,x,\cdot)\right]\right\}\\
&\qquad = \max_{x\in W_3}\left\{\frac{1}{|\Omega|}\sum_{i=1}^{|\Omega|} c(\mathtt{T}_l,d_{li1},d_{li2})^\top x - \max_{y\in W_3}\left\{\frac{1}{|\Omega|}\sum_{i=1}^{|\Omega|} c(\mathtt{T}_f,d_{fi1},d_{fi2})^\top y : \right.\right.\\
&\qquad\qquad\qquad\qquad \left.\left. y\in \operatorname*{argmax}_{\bar{y}\in W_3}\left\{\mathscr{R}_f[\psi(\mathtt{T}_l,\mathtt{T}_f,x,\bar{y},\omega)]\right\}\right\}\right\}.
\end{aligned}$$

For a suitable $M>0$ we get the risk averse case with

$$\begin{aligned}
&\left(\bar{\mathrm{P}}_{EP,\mathscr{R}_f}(\mathtt{T}_l,\mathtt{T}_f)\right)\\
&\qquad \max_{x\in W_3}\left\{\mathbb{P}\left[f_{\mathscr{R}_f}(\mathtt{T}_l,\mathtt{T}_f,x,\cdot)>0\right]\right\}\\
&\qquad = \max_{x\in W_3}\left\{\min_{y\in W_3}\left\{\frac{1}{|\Omega|}\sum_{i=1}^{|\Omega|}\chi\left(c(\mathtt{T}_l,d_{li1},d_{li2})^\top x, c(\mathtt{T}_f,d_{fi1},d_{fi2})^\top y\right) : \right.\right.\\
&\qquad\qquad\qquad \left.\left. y\in \operatorname*{argmax}_{\bar{y}\in W_3}\left\{\mathscr{R}_f[\psi(\mathtt{T}_l,\mathtt{T}_f,x,\bar{y},\omega)]\right\},\ i=1,\dots,|\Omega|\right\}\right\}\\
&\qquad = \max_{x\in W_3}\left\{\min_{y\in W_3}\left\{\max_{\theta\in\{0,1\}^{|\Omega|}}\left\{\frac{1}{|\Omega|}\sum_{i=1}^{|\Omega|}\theta_i : \right.\right.\right.\\
&\qquad\qquad\qquad\qquad c(\mathtt{T}_l,d_{li1},d_{li2})^\top x - c(\mathtt{T}_f,d_{fi1},d_{fi2})^\top y \geq M(\theta_i-1),\\
&\qquad\qquad\qquad\qquad \left.\left.\left. i=1,\dots,|\Omega|\right\}\right\}\right\}\\
&\qquad\qquad\qquad y\in \operatorname*{argmax}_{\bar{y}\in W_3}\left\{\mathscr{R}_f[\psi(\mathtt{T}_l,\mathtt{T}_f,x,\bar{y},\omega)]\right\}
\end{aligned}$$

In Section 5 we will analyse a generalization $\left(\bar{\mathrm{P}}_{\mathscr{R}_f,\mathscr{R}_f}(S)\right)$ of $\left(\bar{\mathrm{P}}_{\mathscr{R}_l,\mathscr{R}_f}(t_l,t_f)\right)$. Let S be the finite set of parameters and A the finite set of options each player can choose from with $|A|=m$. In addition to that we need a discrete probability space $(\tilde{\Omega},2^{\tilde{\Omega}},\tilde{\mathbb{P}})$, the set W_m and a function $c_m: S\times\tilde{\Omega}\to\mathbb{N}_0^m$. A set Z will be generated due to S, A, Ω and c_m.

3.2 Markov decision processes

Another model to handle the underlying problem are Markov decision processes (MDPs). We will start with a generalized formulation of MDPs and then go into detail on how we can transform the stochastic two-player game into a system of Markov models.

A normal MDP can be defined as a 5-tuple $(S,A,\mathbf{P},\mathbf{r},\pi_0)$ with a finite set of states $S=\{1,\dots,n\}$ and a finite set of actions $A=\{1,\dots,m\}$. With the help of the stochastic function $\mathbf{P}: S\times A\times S\to[0,1]$, we get the probability $\mathbf{P}(s,a,s')$ to enter state s' if we choose action a and are in state s. To validate the system and our strategy for the actions, we need a reward function $\mathbf{r}: S\times A\to\mathbb{R}$, where $\mathbf{r}(s,a)$ defines the reward when we take action a in state s. The last coefficient π_0 is the initial distribution of the system.

To illustrate a Markov model, it can be thought of as a graph of round-based stochastic transitions as in Figure 1. The nodes stand for the individual states. Each action has a probability to switch from the current state using action a. Each edge has different rewards depending on the action and the current state. The goal is to find an optimal policy $\pi: S\to A$ that gives us the best action a in a state s. The best option is not always the action with the highest expected reward, as we are interested in the following path. MDPs are memoryless which means that a decision in a state s is not depending on choosen actions in the past. For this reason, we consider only the future possibilities of paths through the state space. There are several ways to solve MDPs in polynomial time, such as using value iteration, policy iteration or linear programming.

The problem with MDPs is that we only optimize for a single goal and receive only one policy. If we know the leader's strategy, it is easy to build an MDP that will provide us the optimal solution of the follower. But if you want to construct an MDP for the leader, there is a huge problem because you don't know the

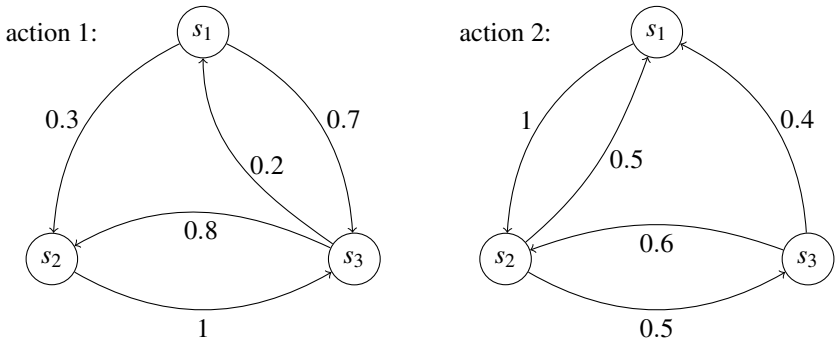

Figure 1: Example of an MDP with 3 states and 2 actions without rewards

reaction of the follower. In that case we get an partially observable Markov decision process (POMDP) which can also be formulated as a set of MDPs. So we have to think about every counter-strategy of the follower and take the best solution over all possibilities. Fortunately, our problem is in a special class of problems which are nonanticipative. This means that the follower's action in state s depends only on the leader's action in s and not on decisions of other leaders in other states.

Now we will look in detail at a model of a stochastic two-player game. We assume that we know a state set (set of parameters) S in which players have to choose their actions (options) $a \in A$. In addition, the set of ending states Z and the associated rewards are known.
In the considered game of dice there is a state $z_{4,12} \in Z$ in which the first player has the score $i := \mathtt{V}_l = 4$ and the second player has the score $k := \mathtt{V}_f = 12$ and we want to compare the two different reward functions $\mathbf{r}_{\mathbb{E}}, \mathbf{r}_{\mathbb{P}}$: For the leader, the rewards are

1. $\mathbf{r}_{\mathbb{E}}(z_{i,k}) := i - k$ and
2. $\mathbf{r}_{\mathbb{P}}(z_{i,k}) = \chi(i,k)$.

In addition, the probabilities $\mathbb{P}(z|s,a_l,a_f)$ are given.
The class of nonanticipative round-based stochastic games for two players can be solved with MDPs. Before we can compute the policies for leader and follower, we need to generate counter-strategies. So first we fix the leader for all actions $a \in A$ and build m MDPs. A fixed MDP consists of $\mathscr{M}_j = (S \cup Z, A, \mathbf{P}_j, r, \pi_0)$ with $\mathbf{P}_j(s,a,z) = \mathbb{P}(z|s,j,a)$. By solving all fixed MDPs we get the counter-strategies $\pi_j : S \to A$. Based on that we are able to build the leader's MDP $\mathscr{M}_l = (S \cup Z, A, \mathbf{P}_l, r, \pi_0)$ by using the counter-strategies. A transition matrix is calculated by $\mathbf{P}_l(s,a,z) = \mathbb{P}(z|s,a,\pi_a(s))$. The solution of this modeling gives us the leader's policy π_l. It can be used to pick out optimal counter-strategies and get the policy of the follower π_f.
This concept can be extended to more general nonanticipative round-based, stochastic games for two players, we will use in Section 5. Let be given $(S, Z, A_l, A_f, P, r_l, r_f, \pi_0)$, where

- S is the finite set of states of decision,
- Z is the finite set of (goal) states,
- A_l, A_f are the finite sets of options for leader and follower,
- $P(s,a_l,a_f,z) = \mathbb{P}(z|s,a_l,a_f)$,
- $r_l(z)$, $r_f(z) \in \mathbb{R}$ are the rewards in the goal states, and
- π_0 is the initial distribution over S.

Algorithm 1 Leader-Follower-MDPs

GIVEN: $\big(S, Z, A_l, A_f, P, r_l, r_f, \pi_0\big)$
for $j \in A_l$ **do**
    build $\mathscr{M}_j = (S \cup Z, A_f, \mathbf{P}_j, r_f, \pi_0)$ with $\mathbf{P}_j(s,a,z) = P(s,j,a,z)$
    solve $\mathscr{M}_j$ to get $\pi_j : S \to A_f$
build $\mathscr{M}_l = (S \cup Z, A_l, \mathbf{P}_l, r_l, \pi_0)$ with $\mathbf{P}_l(s,a,z) = P(s,a,\pi_a(s),z)$
solve $\mathscr{M}_l$ to get $\pi_l : S \to A_l$
link $\pi_f : S \to A_f$ by using $\pi_f(s) = \pi_{\pi_l(s)}(s)$
OUTPUT π_l, π_f

Every MDP can be build up and solved in polynomial time. The solution structure of combined MDPs, which is given by Leader-Follower-MDPs, also gives us a solution in polynomial time. Various types of main functions and goals only modify the rewards of the system of MDPs. States and transitions are unaffected.

Algorithm 1 gives out the most egoistic solution for the two-player game. This means, that both players gives a maximum weighting on the optimization of their own goal. Of course, further possibilities for weighting the non-cooperative game are possible. Sometimes it is more profitable to decrease the expected gains of your opponent depending on an own handicap.

If we want to get all Pareto-optimal decisions, we have to compare the different solutions after the *for*-loop in algorithm 1. There exist several methods to separate a Pareto frontier from a finite set of vectors, like lexicographical sorting and an elimination process.

4 Results of game of dice

The deterministic bilevel formulations and the MDPs for the mentioned class of games were implemented in MATLAB, with the free MATLAB toolbox YALMIP additionally used for the bilevel modelings. Our computations were performed on a machine with a 2.6 GHz 4-Core processor and 8 GB main memory running Ubuntu Linux or on a machine with a 3.0 GHz 20-Core processor and 126 GB main memory running Debian Linux.
In order to model the game situation of the reformulated dice game, we used the parameters (states of decision) $(\mathtt{T}_l, \mathtt{T}_f) \in S = \{1,\dots,6\} \times \{1,\dots,6\}$ and the $m = 3$ options for both players as described in Section 2.1. The options were evaluated using the four possible combinations of the two approaches in Section 2.1, i.e. the objective functions $\mathbb{E}$ and EP or the reward functions $\mathbf{r}_{\mathbb{E}}$ and $\mathbf{r}_{\mathbb{P}}$.

The upper two tables of Table 2 contain the results for the case where e.g. both players choose the qualitative approach. Table 2a shows the best options for the leader depending on $\mathtt{T}_l, \mathtt{T}_f$ and Table 2b the corresponding values for the follower. For the cases where players choose different approaches, there exist parameter values $s \in S$ where two options are given, such as at $(5,3)$ of Table 2g. For this purpose, the corresponding values of EP and $\mathbf{r}_{\mathbb{P}}$ are given in Table 2i below. When specifying several options for one parameter value, the first value is for the more egoistic solution of the MDPs, the second for the pessimistic solution of the bilevel formulation. However, in Table 2c and 2d with $\mathtt{T}_l = 6$, these are the first values.
There are two reasons why we get several best options:

a) Objective function or reward function reaches the same values for several options: The corresponding probability to win for the options in $(6,5)$ of Table 2d and in $(6,5)$ of Table 2l is 27.78%. Interestingly, the chosen options of the follower produce different maximum expected differences for the leader in the first case (Table 2e).

b) Different combinations of solutions of leader and follower are elements of the corresponding Pareto frontier: In Figure 2 we graph the possible combinations of solutions for leader and follower, as they can be found in $(6,4)$ of Table 2d. The dotted surface is the dominated set of solutions. Combination of options $(2/3)$, which means that the leader will choose option $\mathtt{A}_l = 2$ and the follower reacts with option $\mathtt{A}_f = 3$, is dominated by the combination $(1/3)$. The leader gets a lower maximum expected

Table 2: Best options and values of the functions depending on the best options

T_l \ T_f	1	2	3	4	5	6
1	3	3	3	3	3	3
2	3	3	3	3	3	3
3	3	3	3	3	3	3
4	3	3	3	3	3	3
5	3	3	3	3	3	3
6	1	1	1	1	1	1

(a) Leader: qualitative approach

T_l \ T_f	1	2	3	4	5	6
1	3	3	3	3	3	1
2	3	3	3	3	3	1
3	3	3	3	3	3	1
4	3	3	3	3	3	1
5	3	3	3	3	3	1
6	3	3	3	3	3	1

(b) Follower: qualitative approach

T_l \ T_f	1	2	3	4	5	6
1	3	3	3	3	3	3
2	3	3	3	3	3	3
3	3	3	3	3	3	3
4	3	3	3	3	3	3
5	1|3	1|3	1|3	1|3	2|3	3
6	1	1	1	1|3	1	1

(c) Leader: qualitative approach

T_l \ T_f	1	2	3	4	5	6
1	3	3	3	2	2	1
2	3	3	3	2	2	1
3	3	3	3	2	2	1
4	3	3	3	2	2	1
5	3	3	3	3|2	3|2	1
6	3	3	3	3|2	3|2	2

(d) Follower: quantitative approach

T_l \ T_f	1 - 3	4	5
5	-5.83 | 0	-5.83 | 8.83	-3.83 | 3.83
6		5.17 | 8.83	5.17 | 9

(e) Expected values for differences in Table 2c

T_l \ T_f	1 - 3	4	5
5	41.67 | 44.37	41.67 | 44.75	47.92 | 52.08
6		27.78 | 44.75	27.78 | 27.78

(f) Probabilities (in %) for differences in Table 2d

T_l \ T_f	1	2	3	4	5	6
1	3	3	3	3	3	3
2	3	3	3	3	3	3
3	3	3	3	3	3	3
4	3	3	3	3	3	3
5	3|1	3|1	3|1	3|1	3|1	3
6	1	1	1	1	1	1

(g) Leader: quantitative approach

T_l \ T_f	1	2	3	4	5	6
1	3	3	3	3	3	1
2	3	3	3	3	3	1
3	3	3	3	3	3	1
4	3	3	3	3	3	1
5	3	3	3	3	3	1
6	3	3	3	3	3	1

(h) Follower: qualitative approach

T_l \ T_f	1 - 5
5	55.63 | 58.33

(i) Probabilities (in %) for differences in Table 2g

T_l \ T_f	1 - 5
5	0 | 5.83

(j) Expected values for differences in Table 2h

T_l \ T_f	1	2	3	4	5	6
1	3	3	3	3	3	3
2	3	3	3	3	3	3
3	3	3	3	3	3	3
4	3	3	3	3	3	3
5	1	1	1	1	2	3
6	1	1	1	1	1	1

(k) Leader: quantitative approach

T_l \ T_f	1	2	3	4	5	6
1	3	3	3	2	2	1
2	3	3	3	2	2	1
3	3	3	3	2	2	1
4	3	3	3	2	2	1
5	3	3	3	3	3	1
6	3	3	3	3	2|3	2

(l) Follower: quantitative approach

T_l \ T_f	5
6	72.22

(m) Probability (in %) for difference in Table 2k

T_l \ T_f	5
6	27.78

(n) Probability (in %) for difference in Table 2l

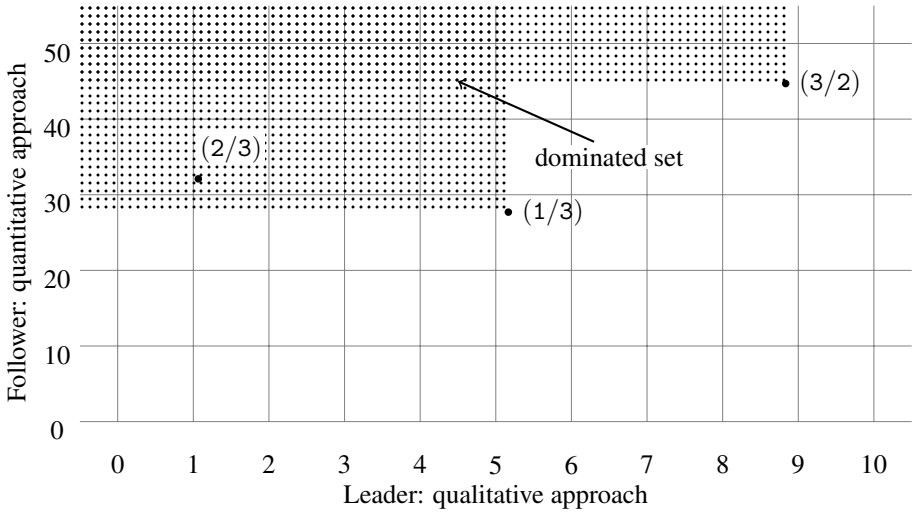

Figure 2: Pareto-optimal options for leader/follower and dominated options

difference and also increases the probability to win for his or her opponent. Non-dominated pairs of actions are called Pareto-optimal and are solutions to our problem. An exact solution for a player depends on his or her chosen approach.

Let us return to a question from Section 2.1: Which strategy is the "best" strategy for a player depending on the parameters? For the valuation approaches we considered, we get unique best options for some parameter values (Table 3). For the other parameter values, as well as in general, the "best" options depend

Table 3: Summary of unique best options

T_l \ T_f	1	2	3	4	5	6
1	3	3	3	3	3	3
2	3	3	3	3	3	3
3	3	3	3	3	3	3
4	3	3	3	3	3	3
5						3
6	1	1	1		1	1

T_l \ T_f	1	2	3	4	5	6
1	3	3	3			1
2	3	3	3			1
3	3	3	3			1
4	3	3	3			1
5	3	3	3			1
6	3	3	3			

on the valuation approaches of both players, as one may assume based on the considerations at the end of Section 2.1.

Now we address the question of runtimes in Section 1. The elapsed times of the implementations of the deterministic bilevel formulations and of the MDPs are as follows:

- (qualitative approach/qualitative approach): 9,63 s 7,52 s
- (qualitative approach/quantitative approach): 11,22 s 7,71 s
- (quantitative approach/qualitative approach): 9,84 s 7,25 s
- (quantitative approach/quantitative approach): 11,07 s 7,72 s

Within these times, the underlying bilevel problems or MDPs were solved and the Pareto frontiers computed. It turns out that the runtimes to the MDPs are almost constant, independent of the chosen valuation approach. The timing differences of the bilevel formulations can be attributed to the more complex calculations for the excess probability from the point of view of the follower. Further investigations on the runtimes of the two modelings will be introduced in the next subsection.

5 Results of general formulations

Now we take a closer look at the runtimes for a more general class of problems. For our generalized modelings as described at the end of Section 3.1 and 3.2, we will compare the running times of computing individual solutions and the Pareto-optimal set of option combinations. For this, randomized test instances with a variable number of decision and target states and options were be generated. In addition, a given number of equally sized instances was solved and we took the mean of the running times.

Considering that, we used a tuple (S,Z,A_l,A_f,P,r_l,r_f) with a set of parameters or decision states S with $n = |S|$, a set of goal states $Z = \{z_{\mathtt{v}_l,\mathtt{v}_f} \in \mathbb{Z} : z_{\mathtt{v}_l,\mathtt{v}_f} = \mathtt{v}_l - \mathtt{v}_f\}$, sets of options of the two players A_l and A_f ($m = |A_l| = |A_f|$), a mapping for determining the score $\mathtt{v}$ of a player $c_m : S \times \tilde{\Omega} \to \mathbb{N}_0^m$, a probability function $P : S \times A_l \times A_f \times Z \to [0,1]$ for the transition to $z \in Z$ and reward functions r_l, r_f in the goal states. We assume that every decision state has the same entree probability so we renounce of π_0.
For our test instances, we modified $n \in \{30,40,50,60\}$, $|Z| \in \{1000,1500,\ldots,3000\}$ and $m \in \{3,\ldots,7\}$. The transitions are sparse functions, means that only 10-30 percentage of the goal states can be arrived from one decision state. The distribution is randomly generated and the rewards are normally distributed.

In Figure 3 we fixed $n = 60$ and compared the running times for a different number of options and goal states. The MDP method computes the whole Pareto-set, whereas the bilevel formulation computes only the pessimistic solution. We get almost constant solutions for the bilevel modeling, which means that the number of options and goal states is nearly negligible for the performance. A comparison of different results depending on $n \in \{30,40,50\}$ shows that only the number of decision states influences the running time. The plots for the MDP method are just about linear functions. Here, higher quantities of options and goal states increase the running times. Interestingly enough, the performance is hardly influenced by the number of decision states. So we get the counter part to the bilevel (BP) formulation.
In Figure 4 we compared single, egoistic solutions of the MDP method with the bilevel solutions including the Pareto frontier. We can see that the runtimes for a single solution or for the entire Pareto frontier are nearly the same for the MDP method. Thus, it is possible to compute the Pareto frontier in an effective manner, so that the running time is almost constant.

6 Conclusion

We compared modelings for bilevel problems under uncertainty and for MDPs based on a game of dice from a class of nonanticipative round-based stochastic two-player games. Both methods compute partly different results, as the modelings have different weightings for reaching their goals. The modeling of the MDPs calculates the most egoistic solution and the deterministic bilevel formulation gives out the pessimistic solution of the leader. However, all solutions are Pareto-optimal.
Interestingly the runtimes of the models are as different as they can be. The implementation of the MDPs is nearly unaffected by the number of decision states and the bilevel programming is nearly constant for an increasing number of options and goal states. Depending on the dimension of the underlying problem, it makes sense to choose one or the other modeling.

Acknowledgements Thanks to the *Deutsche Forschungsgesllschaft* for their support within the Research Training Group 1855 *Discrete Optimization of Technical Systems under Uncertainty*.

References

[1] S. M. Alizadeh, P. Marcotte and G. Savard, Two-stage Stochastic Bilevel Programming over a Transportation Network, Transportation Research Part B: Methodological, 58, pp. 58-105 (2013).

[2] J. M. Arroyo, M. Carrion and A. J. Conejo, A Bilevel Stochastic Programming Approach for Retailer Futures Market Trading, Power Systems, IEEE Transactions on, 24(3), pp. 1446-1456 (2009).

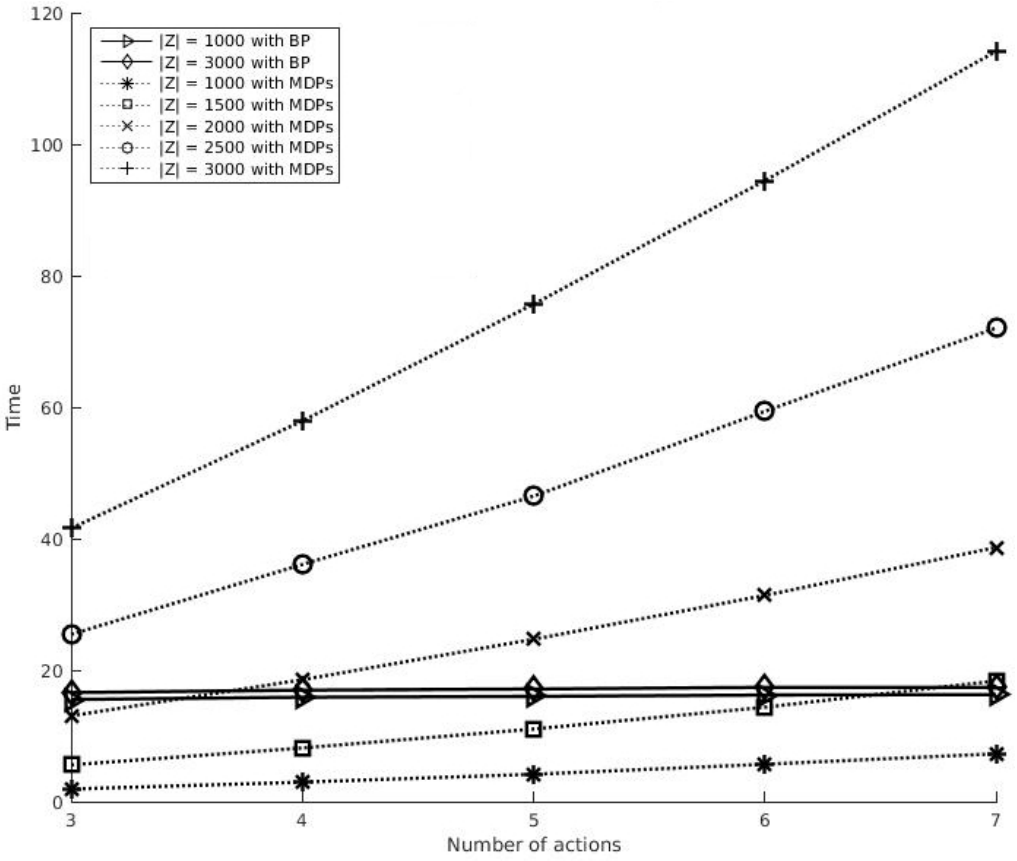

Figure 3: Running times for generalized modelings with $n = 60$, $|Z| = 1000$-3000 and $m = 3$-7, MDP method for Pareto frontier, BP formulation for single solution

[3] M.-B. Aryanezhad, E. Roghanian and S. J. Sadjadi, A Probabilistic Bi-level Linear Multi-objective Programming Problem to Supply Chain Planning, Applied Mathematics and Computation, 188(1), pp. 786-800 (2007).

[4] D. Dentcheva, A. P. Ruszczyński and A. Shapiro, Lectures on Stochastic Programming: Modeling and Theory, MPS SIAM Series on Optimization, 9, SIAM, Philadelphia, 2nd Edition (2014).

[5] H. Föllmer and A. Schied, Convex Measures of Risk and Trading Constraints, Finance and Stochastics, 6(4), pp. 429-447 (2002).

[6] A. A. Gaivoronski and A. Werner, Stochastic Programming Perspective on the Agency Problems under Uncertainty, In: Managing Safety of Heterogeneous Systems, pp. 137-167. Springer, Berlin Heidelberg (2012).

[7] P. Gross and G. C. Pflug, Behavioral Pricing of Energy Swing Options by Stochastic Bilevel Optimization, Energy Systems, pp. 1-26 (2016).

[8] M. L. Puterman, Markov Decision Processes. Wiley, (2005)

[9] Reutlinger Generalanzeiger, Der Reutlinger Mutscheltag: Geschichte des Reutlinger Mutscheltages und Beschreibung der gebräuchlichsten Spiele, Verlag des Reutlinger Generalanzeigers, Reutlingen (n.d.)

[10] R. Schultz and S. Tiedemann, Risk Aversion via Excess Probabilities in Stochastic Programs with Mixed-Integer Recourse, SIAM Journal on Optimization, 14, pp. 115-138 (2003).

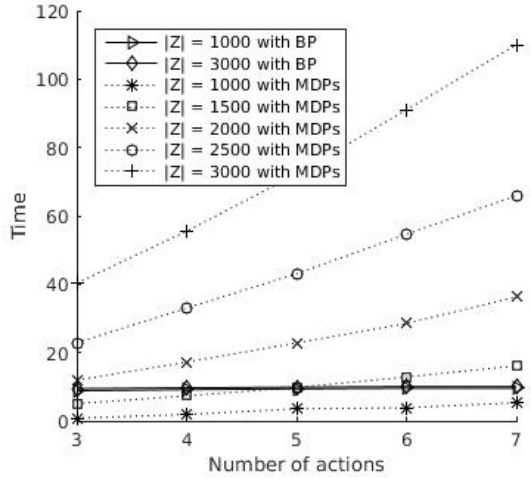

Figure 4: Running times for generalized modelings with $n = 30$, $|Z| = 1000$-3000 and $m = 3$-7, MDP method for single solution, BP formulation for Pareto frontier