Comput Manag Sci (2014) 11:197–220
DOI 10.1007/s10287-014-0213-y

ORIGINAL PAPER

On distributionally robust multiperiod stochastic optimization

Bita Analui · Georg Ch. Pflug

Received: 12 May 2014 / Accepted: 20 May 2014 / Published online: 4 July 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract This paper considers model uncertainty for multistage stochastic programs. The data and information structure of the baseline model is a tree, on which the decision problem is defined. We consider "ambiguity neighborhoods" around this tree as alternative models which are close to the baseline model. Closeness is defined in terms of a distance for probability trees, called the nested distance. This distance is appropriate for scenario models of multistage stochastic optimization problems as was demonstrated in Pflug and Pichler (SIAM J Optim 22:1–23, 2012). The ambiguity model is formulated as a minimax problem, where the the optimal decision is to be found, which minimizes the maximal objective function within the ambiguity set. We give a setup for studying saddle point properties of the minimax problem. Moreover, we present solution algorithms for finding the minimax decisions at least asymptotically. As an example, we consider a multiperiod stochastic production/inventory control problem with weekly ordering. The stochastic scenario process is given by the random demands for two products. We determine the minimax solution and identify the worst trees within the ambiguity set. It turns out that the probability weights of the worst case trees are concentrated on few very bad scenarios.

B. Analui was partially sponsored by WWTF project: Energy Policies and Risk Management for the 21st Century.

B. Analui · G. Ch. Pflug (✉)
Department of Statistics and Operations Research, University of Vienna, Vienna, Austria
e-mail: georg.pflug@univie.ac.at

B. Analui
IK-Computational Optimization, University of Vienna, Vienna, Austria
e-mail: bita.analui@univie.ac.at

G. Ch. Pflug
International Institute for Applied System Analysis (IIASA), Laxenburg, Austria

Keywords Multistage stochastic optimization · Distributional robustness · Model ambiguity · Nested distance · Inventory management

1 Introduction

The standard assumption in stochastic optimization is that the probability laws of the uncertain parameters are known and only the realizations are unknown at the time of decision making. Experience with applications has shown that the choice of the appropriate probability model is crucial for the quality of the solution. Typically the structure of the parametric model is chosen in a more or less adhoc manner (e.g. by specifying that the data come from some Gaussian sequence) and the parameters of the model are estimated on the basis of past observations. Not only that trusted results of parameter estimates are confidence regions and not point estimates, but also the model class itself can be chosen erroneously. On the basis of the available information a whole set of models could represent the real phenomenon equally well, we call this fact *model ambiguity*. A careful decision maker should then take all these equivalent models into account when looking for the robust decision strategy.

One way to deal with ambiguity is to investigate the *stability* of the optimal solution in stochastic programming: The notion stability refers to continuity properties of the optimal solution with respect to to model parameters, see e.g. (Robinson and Wets 1987; Römisch and Schultz 1991; Rachev and Römisch 2002). However, the solution considered in these stability investigations is always with respect to one single model and the question of how to improve decisions under endogenous model uncertainty is not addressed.

The idea of optimal decisions under ambiguous stochastic models appeared to our knowledge for the first time in a paper by (Scarf 1958). Scarf studied an optimal single product inventory problem under an unknown demand distribution with known mean and variance. The problem was formulated as a linear inventory problem seeking the stockage policy which maximizes the minimum profit considering all demand distributions with given mean and standard deviation.

The utilization of general *minimax* decision rules was pioneered in the mid-1960s by (Žácková a.k.a. Dupačová 1966). This approach was applied to the class of stochastic linear programs with recourse, where results where formulated in terms of two person zero-sum games. The minimax solution was introduced as an optimal pure strategy of the first player in the game and developed further in (Dupačová 1980, 1987). In (Jagannathan 1977) the class of an ambiguity set consisting of all probabilities with given first two moments was studied for linear stochastic problems with simple recourse.

The minimax approach for the ambiguity problem can be regarded as a setup which bridges the gap between the conservatism of robust optimization and the specificity of stochastic optimization. In this setting, the optimal decisions are sought for the worst case probability models by obtaining the best possible decisions under the most adverse circumstances considered. Unfortunately, different names are used in literature for the ambiguity problem. Synonymous names are: *model uncertainty problem, minimax stochastic optimization* and *distributionally robust problem*.

Many parametric and nonparametric proposals for ambiguity sets for two-stage problems have been made and analyzed among which, the probability models are defined by certain properties such as the support and the moment of corresponding probability distributions or neighborhoods with respect to some appropriate distances. A list of popular classes of probability models is introduced in (Dupačová 2001, 2010) and a very fast growing literature dealing with model uncertainty either from theoretical or applied viewpoint can be found in (Chen and Epstein 2002; Calafiore 2007; Shapiro and Kleywegt 2002; Shapiro and Ahmed 2004; Pflug and Wozabal 2007; Thiele 2008; Delage and Ye 2010; Goh and Sim 2010).

In this paper we introduce a concept for distributionally robust decision making for *multistage stochastic optimization* problems. Multistage stochastic optimization is a well established framework for sequential decision making under uncertainty and is successfully applied in various fields such as dynamic portfolio choice, energy production, transportation and telecommunication. In multistage decision models the structure of information plays a crucial role. When time passes, the initially unknown uncertain scenario values can gradually be observed. Stage-by-stage, the amount of information increases and planning decisions have to be made at each time stage based on the available information, i.e., decisions are taken at times $t = 0, \dots, T-1$ with typically increasing levels of information. We denote the random M-dimensional scenario process by $\xi := (\xi_1, \dots, \xi_T)$ and the pertaining multistage decision sequence by $x = (x_0, x_1, \dots, x_{T-1})$. At time instant t, decision x_t must be *non-anticipative,* meaning that it can be based only on information gathered so far, i.e. $x_t = x_t(\xi^t)$ with $\xi^t = (\xi_1, \dots, \xi_t)$. For a broad technical presentation of multistage stochastic programming refer to (Birge and Louveaux 1997; Pflug and Römisch 2007) and (Ruszczynski and Shapiro 2003).

Some literature exists dealing with the parametric ambiguity problem for multistage programs. Goh and Sim (2010) for example, extend the approach proposed by Delage and Ye (2010)—where the authors study distributionally robust stochastic programs when the mean and covariance of the scenario process are themselves subject to uncertainty—to allow for non-anticipativity requirements.

Since for multistage optimization problems, not only the marginal distributions of the scenario process but also the information structure should be taken into account, we argue here that it is quite natural to base the ambiguity set on the nearness of the nested distributions. To this end, we apply the concept of nested distances for the nested distributions. Neglecting the information structure and looking only at the multivariate distributions of the scenario processes lead to counterintuitive examples [cf. (Pflug and Pichler 2012), Example 1 and (Heitsch et al. 2006)]. On the other hand, the nested distance, initially introduced by Pflug and Pichler (2012) is a suitable concept for dealing with the information structure as well.

In contrast to models with uncertain means and variances, our approach is general and nonparametric. This means that any stochastic processes, even non-Markovian ones, may form the baseline model. Notice however that for multistage models no generic confidence sets serving as ambiguity sets can be defined, since conditional distributions cannot be estimated from data without model assumptions.

The paper is organized as follows: In the next section an introduction to risk-neutral multistage stochastic programing and notions of ambiguity and model uncertainty is given. In Sect. 3, the distributionally robust counterpart of a risk-neutral multistage stochastic optimization problem is presented and theoretically discussed. Section 4 is devoted to our proposed solution algorithm. In Sect. 5 we discuss the application of our approach to a classical stochastic multiperiod inventory control problem. Finally Sect. 6 reflects the main results and contains some conclusions.

2 Multistage stochastic optimization

Here, we briefly discuss the risk neutral formulation of multistage linear stochastic optimization problems. Consider the problem

$$\min_x \{\mathbb{E}_{\mathbb{P}}[H(x,\xi)] : \ x \in \mathbb{X}, \ x \lhd \mathfrak{F}; \ (\Omega, \mathfrak{F}, P; \xi) \sim \mathbb{P}\}, \tag{2.1}$$

where H is a real-valued convex cost function, depending on the decision sequence $x = (x_0, \ldots, x_{T-1})$ and the stochastic scenario process $\xi = (\xi_1, \ldots, \ \xi_T)$. The stochastic process ξ describes the economic environment of the decisions (e.g. future prices, demands, external supplies,...) and is defined by its nested distribution. Assume for a moment that this process is defined on a given filtered probability space $(\Omega, \mathfrak{F}, P)$, where $\mathfrak{F} = (\mathcal{F}_1, \ldots, \mathcal{F}_T)$ is a filtration such that ξ_t is measurable w.r.t. $\mathcal{F}_t$, which is denoted by $\xi_t \lhd \mathcal{F}_t$ (and for the whole process $\xi \lhd \mathfrak{F}$) and $\xi(\omega) = (\xi_1(\omega), \ldots, \ \xi_T (\omega))$.

The nested distribution is the family of conditional distributions of ξ_t given $\mathcal{F}_{t-1}$, written as $\xi_t|\mathcal{F}_{t-1}$ collected in the nested structure $((((\xi_T|\mathcal{F}_{T-1}), \xi_{T-1}|\mathcal{F}_{T-2})\ldots), \xi_2|\mathcal{F}_1), \xi_1)$. It turns out that the nested distribution is the right concept to formulate the distribution of the scenario process and the information structure given by the filtration independent of a concrete probability space. That is, two processes which may be defined on different probability spaces, but can—together with the respective filtrations—be mapped to each other by a bijective transformation of the underlying space share the same nested distribution. For a proof and more about the concept of nested distribution see (Pflug 2010) . We denote the nested distribution by $\mathbb{P}$ and notice that it can be concretized to a process ξ defined on a filtered probability space $(\Omega, \mathfrak{F}, P)$ if a concrete model is needed. The notation $(\Omega, \mathfrak{F}, P; \xi) \sim \mathbb{P}$ symbolizes this.

For a baseline model $(\Omega, \mathfrak{F}, P; \xi) \sim \mathbb{P}$ and an alternative model $(\tilde{\Omega}, \tilde{\mathfrak{F}}, \tilde{P}; \tilde{\xi}) \sim \tilde{\mathbb{P}}$ a concept of distance for the nested distributions has been introduced, which allows to quantify the model error.

Definition 1 (Pflug and Pichler 2012) The multistage (nested) distance of order $r \geq 1$ of two nested-structures $(\Omega, \mathfrak{F}, P; \xi) \sim \mathbb{P}$ and $(\tilde{\Omega}, \tilde{\mathfrak{F}}, \tilde{P}; \tilde{\xi}) \sim \tilde{\mathbb{P}}$ is the optimal value of the optimization problem

$$
\begin{array}{lll}
\min\limits_{\pi} & \left(\int \mathsf{d}(\omega,\tilde{\omega})^r \pi[\mathrm{d}\omega, \mathrm{d}\tilde{\omega}]\right)^{\frac{1}{r}} & \\
\text{subject to} & \pi[A \times \tilde{\Omega} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t](\omega,\tilde{\omega}) = P[A|\mathcal{F}_t](\omega) & (A \in \mathcal{F}_T,\ 1 \le t \le T) \\
 & \pi[\Omega \times B | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t](\omega,\tilde{\omega}) = \tilde{P}[B|\tilde{\mathcal{F}}_t](\tilde{\omega}) & (B \in \tilde{\mathcal{F}}_T,\ 1 \le t \le T).
\end{array}
\tag{2.2}
$$

Here the infimum in (2.2) is taken among all bivariate probability measures π defined on $\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T$ and d is a distance for the realizations of the stochastic scenario processes, typically

$$
\mathsf{d}(\omega,\tilde{\omega}) = \sum_{t=1}^{T} \sum_{m=1}^{M} w_t^m \, |\xi_t^m(\omega) - \tilde{\xi}_t^m(\tilde{\omega})|. \tag{2.3}
$$

where w_t^m are some weights, reflecting discounting in time and reweighting different dimensions of the M-dimensional process ξ. The optimal value of (2.2) is the nested distance of order r and denoted by $\mathrm{dl}_r(\mathbb{P}, \tilde{\mathbb{P}})$.

The nested distance is defined for the nested distributions and is independent of the respective realizations on concrete probability spaces. We repeat that in particular the processes ξ and $\tilde{\xi}$ can be defined on different probability spaces. It has been proved by (Pflug and Pichler 2012) that if the criterion function H is Lipschitz in ξ and convex in x, then the optimal value of the decision problem (2.1) is Lipschitz w.r.t. the nested distance.

While (2.1) describes the general form of a risk neutral multistage stochastic optimization problem, such problems are often formulated in a finite discrete setup, especially for making them tractable by numerical optimization. Finite nested distributions can be represented by *node*- and *arc* valuated *trees*, where the tree structure reflects the filtration, the node valuation represents the values of the stochastic scenario process ξ and the arc valuations encodes the conditional probability distributions. Again we refer to (Pflug 2010) for a thorough treatment of scenario tree (better: equivalence classes of scenario trees) as representations of nested distributions. In the following, we consider scenario trees as finite versions of nested distributions. Trees are characterized by the node sets $\mathcal{N}_t$ per stage t and the predecessor relations $\prec$. If $k \in \mathcal{N}_{t-1}$, $i \in \mathcal{N}_t$ and k is a direct predecessor of i, we write $k = i-$ and $i \in k+$. If m is any predecessor of i we write $m \prec i$.[1] The node set $\mathcal{N}_0$ consists only of the root and the node set $\mathcal{N}_T$ can be identified with the probability space Ω . If $i \in \mathcal{N}_t$ and $k \in \mathcal{N}_{t-1}$ with $k = i-$, then probabilities Q_i sitting on the arcs represent the conditional probabilities of reaching node i from its predecessor node k. The unconditional node probabilities P_i can be calculated by

$$
P_i = Q_i \cdot \prod_{m \prec i} Q_m \tag{2.4}
$$

The probabilities P_i sitting on the leaves $\mathcal{N}_T$ of the tree represent the probability distribution P on $\Omega = \mathcal{N}_T$. The specialization of Definition 1 for the tree situation is

[1] Notation $\mathrm{pred}_s(i)$ denoting the predecessor of i in $\mathcal{N}_s$, with $s < t$ might also be used. If $s = t - 1$ the notation is written as $\mathrm{pred}_{t-1}(i)$ or $i-$.

given by Definition 2. In the following, we only consider the nested distance of order $r = 1$, however all results can be generalized for $r > 1$.

Definition 2 The nested distance of order $r = 1$ between two tree models $\mathbb{P}$ and $\tilde{\mathbb{P}}$ is given by the optimal value of the following large linear program

$$\begin{array}{ll} \mathsf{dl}(\mathbb{P}, \tilde{\mathbb{P}}) = \min\limits_{\pi} & \sum\limits_{i', j' \in \mathcal{N}_T} \mathsf{d}(i', j')\, \pi(i', j') \\ \text{subject to} & \sum\limits_{j \in l+} \pi(i, j|k, l) = Q_i \quad (i \in k+,\ l) \\ & \sum\limits_{i \in k+} \pi(i, j|k, l) = \tilde{Q}_j \quad (k,\ j \in l+)\cdot \\ & \sum\limits_{i,j} \pi(i', j') = 1 \\ & \pi(i', j') \geq 0 \end{array} \tag{2.5}$$

Here $\mathsf{d}(i', j')$ are distances between the leaves $i' \in \mathcal{N}_T$ and $j' \in \tilde{\mathcal{N}}_T$ are given by a distance between the paths leading to i' resp. j' similar to (2.3). $\pi(i', j')$ runs through all joint probability distributions on $\Omega \times \tilde{\Omega} = \mathcal{N}_T \times \tilde{\mathcal{N}}_T$, which we call *transportation plans.*The conditional probabilities in a transportation plan are called *transportation subplans* and are given by

$$\pi(i, j|k, l) = \frac{\sum_{i \prec i',\ j \prec j'} \pi(i', j')}{\sum_{k \prec i',\ l \prec j'} \pi(i', j')}. \tag{2.6}$$

[2]It is easy to see that (2.5) is a linear program in π. In Fig. 1, the nested structure of transportation plan π written as a matrix and induced by two trees of the same height and structure together with the schematic distance matrix d is depicted.

The concept of nested distance provides us with a tool for constructing ambiguity neighborhoods around nested distributions. In the next section the distributionally robust counterpart of model (2.1) is derived and discussed.

3 Multistage distributionally robust stochastic optimization

The *distributionally robust counterpart* of (2.1) is given by

$$\min_{x} \max_{\mathbb{P} \in \mathcal{P}} \{\mathbb{E}_{\mathbb{P}}[H(x, \xi)] :\ x \in \mathbb{X},\ x \lhd \mathfrak{F}\},$$

where $\mathcal{P}$ denotes an ambiguity set of probability models. In the present work we consider balls with radius ϵ around a baseline model $\mathbb{P}$ w.r.t. the nested distance

[2] This quotient necessitates inclusion of constraint $\sum_{i', j'} \pi(i', j') = 1$, otherwise every multiplication of any feasible transportation plan π, would be feasible.

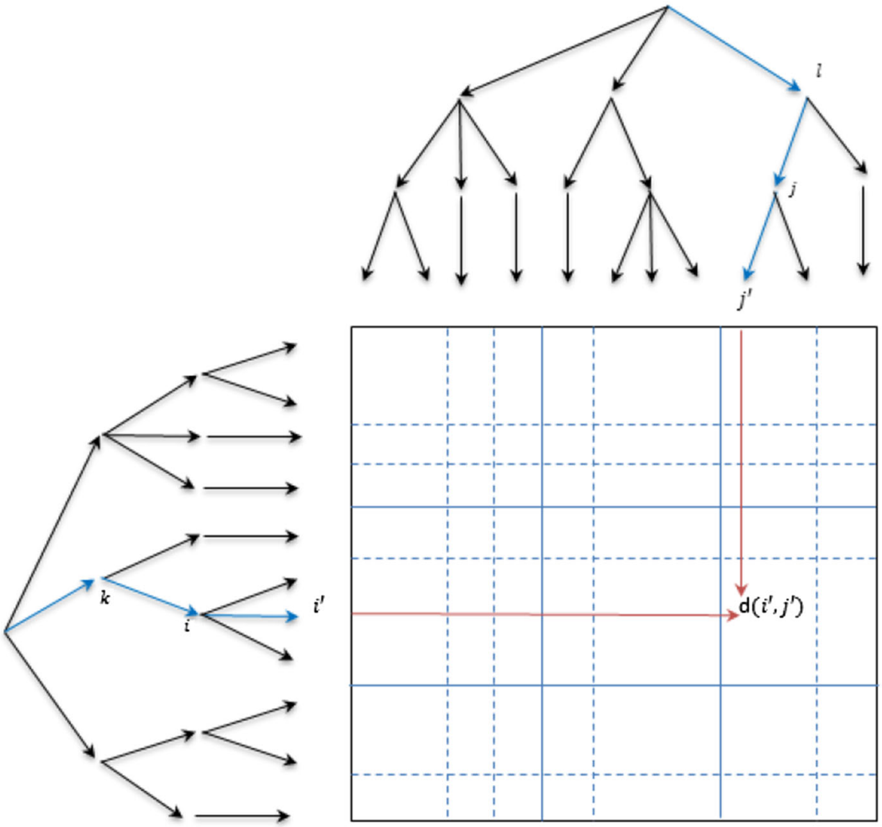

Fig. 1 Visualization of transportation matrix π and distance matrix d for two trees. $i^{'}$, $j^{'}$ are leaf and k, l are generic intermediate nodes

$$\mathcal{P} = \{\tilde{\mathbb{P}} :\ \mathrm{dl}(\mathbb{P}, \tilde{\mathbb{P}}) \le \epsilon\}. \tag{3.1}$$

The distributionally robust counterpart reads now

$$\min_x \{\max_{\tilde{\mathbb{P}}}\ \mathbb{E}_{\tilde{\mathbb{P}}}[H(x,\xi)]\ :\ x \in \mathbb{X},\ x \lhd \tilde{\mathfrak{F}},\ (\tilde{\Omega}, \tilde{\mathfrak{F}}, \tilde{P}; \tilde{\xi}) \sim \tilde{\mathbb{P}},\ \mathrm{dl}(\mathbb{P}, \tilde{\mathbb{P}}) \le \epsilon\}. \tag{3.2}$$

Problem (3.2) is quite difficult to solve. Even in the single-stage case, it requires algorithms for nonconvex optimization such as DC-algorithms, see (Wozabal 2010). For this reason, we will consider a smaller ambiguity set, where we fix the tree structure and only vary the arc probabilities. To this end, introduce the following notation: Let $\mathbb{T}$ denote a tree with given structure and scenario values. The leaf set (the scenarios) of $\mathbb{T}$ is denoted by $\Omega = \mathcal{N}_T$. The probability valuations are given by the scenario probabilities $P = (P_i)_{i \in \mathcal{N}_T}$. The fully valuated tree is denoted by $\mathbb{P}(\mathbb{T}, P)$. Even in cases that the structure and the values of the scenario process are fixed and only the scenario probabilities vary, it would be inconsistent to define simply ambiguity sets as neighborhoods of P, such as

$$\left\{ \tilde{P} : \sum_{i \in \mathcal{N}_T} |P_i - \tilde{P}_i|^r \leq \epsilon^r \right\}, \tag{3.3}$$

since such an ambiguity set does not respect the tree structure.

As was already said, we restrict ourselves in the following to alternative models, which are defined on the same tree structure of the baseline model, but only vary the probabilities. However we keep the ambiguity set as a ball in the nested distance sense, i.e. we specify (3.1) to

$$\mathcal{B}_\epsilon = \{\tilde{P} : \text{ dl}(\mathbb{P}(\mathbb{T}, P), \mathbb{P}(\mathbb{T}, \tilde{P})) \leq \epsilon\} \tag{3.4}$$

and set

$$\mathcal{P}_\epsilon = \left\{ \mathbb{P}(\mathbb{T}, \tilde{P}) : \ \tilde{P} \in \mathcal{B}_\epsilon \right\}. \tag{3.5}$$

The final formulation of the ambiguity extension problem is now

$$\min_x \max_{\tilde{\mathbb{P}} \in \mathcal{P}_\epsilon} \{\mathbb{E}_{\tilde{\mathbb{P}}}[H(x, \xi)] : \ x \in \mathbb{X}, \ x \lhd \mathfrak{F}\}. \tag{3.6}$$

In the next section, we amplify the ambiguity set to its convex hull in order to apply a minimax theorem and identify a saddle point. In addition, we show that the *worst case* model is also contained in the original ambiguity set $\mathcal{P}_\epsilon$.

3.1 A minimax theorem

The famous minimax theorems (Neumann 1928; Fan 1953; Sion 1958) and all the references therein, assert that the order of min and max can be interchanged in minimax problems, if some geometric properties are fulfilled. In particular, the validity of such theorems is related to convexity/concavity properties of the criterion function and topological and convexity properties of the feasible sets. We are therefore faced with the problem of identifying convexity properties for nested distributions. Notice that it would be incorrect to just form convex combinations of the scenario probabilities, since such a combination is not invariant w.r.t. equivalent permutations of the leaves, i.e. cannot be formulated in terms of the nested distributions. The correct notion of convex combinations however is *compounding*.

Definition 3 If $\mathbb{P}$ and $\tilde{\mathbb{P}}$ are nested distributions, then their compound (with compounding probability λ) is given by

$$\mathcal{C}(\mathbb{P}, \tilde{\mathbb{P}}; \lambda) = \begin{cases} \mathbb{P} & \text{with prob } \lambda \\ \tilde{\mathbb{P}} & \text{with prob } 1 - \lambda \end{cases}.$$

If $\mathbb{P}$ and $\tilde{\mathbb{P}}$ are tree models, then $\mathcal{C}(\mathbb{P}, \tilde{\mathbb{P}}; \lambda)$ is also a tree model, where from a new root subtree $\mathbb{P}$ can be reached with probability λ and subtree $\tilde{\mathbb{P}}$ can be reached with probability $1-\lambda$. Denote by $\mathbb{P}_+ = \mathcal{C}(\mathbb{P}, \mathbb{P}; 1)$ the degenerated compound model, where the baseline model $\mathbb{P}$ is chosen with probability 1. It is equivalent to $\mathbb{P}$, but has an additional root, from which subtree $\mathbb{P}$ can be reached with probability 1.

It turns out, that our ambiguity set $\mathcal{P}_\epsilon$ given by (3.5) is not convex with respect to convex compounding of scenario probability vectors P and $\tilde{P}$. Therefore we consider its closed convex hull $\bar{\mathcal{P}}_\epsilon$. The structure of this convex hull is discussed in the Appendix 7.1. For the extended ambiguity set $\bar{\mathcal{P}}_\epsilon$ we can now prove the following minimax theorem, which follows from general minimax theorems cited above.

Theorem 1 *Let $H(x,\xi)$ be convex in x with a convex and compact decision set $\mathbb{X}$. Then*

$$\min_{x\in\mathbb{X}} \max_{\tilde{\mathbb{P}}\in\bar{\mathcal{P}}_\epsilon} \mathbb{E}_{\tilde{\mathbb{P}}}[H(x,\xi)] = \max_{\tilde{\mathbb{P}}\in\bar{\mathcal{P}}_\epsilon} \min_{x\in\mathbb{X}} \mathbb{E}_{\tilde{\mathbb{P}}}[H(x,\xi)]$$

and there exists a saddle point $(x^*, \tilde{\mathbb{P}}^*)$,i.e.

$$\mathbb{E}_{\tilde{\mathbb{P}}}[H(x^*,\xi)] \le \mathbb{E}_{\tilde{\mathbb{P}}*}[H(x^*,\xi)] \le \mathbb{E}_{\tilde{\mathbb{P}}*}[H(x,\xi)]$$

for all $x\in\mathbb{X}$, $\tilde{\mathbb{P}}\in\bar{\mathcal{P}}_\epsilon$. Moreover, $\tilde{\mathbb{P}}^*$ can be chosen to lie in $\mathcal{P}_\epsilon$ (and not just in $\bar{\mathcal{P}}_\epsilon$).

Proof The proof of this Theorem can be found in the Appendix 7.1. □

In the next section we present a stage-wise approach for constructing the ambiguity neighborhood.

3.2 Ambiguity sets defined by transportation kernels

We have seen, that in its general form, problem (3.6) has a complex structure. In construction of models $\mathbb{P}(\mathbb{T}, \tilde{P})$ only scenario probabilities differ from the baseline model $\mathbb{P}(\mathbb{T}, P)$ as long as the respective nested distance remains small. However, the measurability of decisions x w.r.t. the same $\mathfrak{F}$, i.e. $x \lhd \mathfrak{F}$ ensures the comparability of the decisions of both models (2.1) and (3.6).

In order to describe the nested distance in a recursive form, we introduce the notion of *transportation subkernels*. For arbitrary nodes $k, l \in \mathcal{N}_t$ and $k = i-$, $l = j-$, the subkernel is given by

$$K_t(j|i;k,l) = \frac{\pi(i,j|k,l)}{\sum_j \pi(i,j|k,l)}.$$

It is a probability distribution on the set $l+$, i.e.

$$K_t(j|i;k,l) \ge 0, \quad \sum_{j\in l+} K_t(j|i;k,l) = 1, \ (\forall\ (i,j) \in \mathcal{N}_{t+1}\ k = i-,\ l),$$

The relation between transportation subkernels and transportation subplans is given by:

$$\begin{aligned}\pi(i,j) = {} & K_1(\mathrm{pred}_1(j)|\mathrm{pred}_1(i);1,1)\cdots \\ & \cdot K_{T-2}(\mathrm{pred}_{T-1}(j)|\mathrm{pred}_{T-1}(i);\mathrm{pred}_{T-2}(i),\mathrm{pred}_{T-2}(j)) \\ & \cdot K_{T-1}(j|i;\mathrm{pred}_{T-1}(i),\mathrm{pred}_{T-1}(j))\cdot Q_i\cdot\prod_{m\prec i} Q_m \end{aligned} \tag{3.7}$$

Therefore transportation kernel $K(j|i)$ is the *composition* of subkernels $K_t,\ t = 1\ldots T-1$:

$$\begin{aligned}K(j|i) &= K_1\circ\cdots\circ K_{T-1}(j|i) \\ &= K_1(\mathrm{pred}_1(j)|\mathrm{pred}_1(i);1,1)\cdots \\ &\quad\cdot K_{T-2}(\mathrm{pred}_{T-1}(j)|\mathrm{pred}_{T-1}(i);\mathrm{pred}_{T-2}(i),\mathrm{pred}_{T-2}(j)) \\ &\quad\cdot K_{T-1}(j|i;\mathrm{pred}_{T-1}(i),\mathrm{pred}_{T-1}(j)). \end{aligned} \tag{3.8}$$

For a given baseline probability distribution $P = (P_i)_{i\in\mathcal{N}_T}$ we shall define the new probability distribution $\tilde{P}$ by $\tilde{P}_j = \sum_{i,j\in\mathcal{N}_T} P_i\cdot K(j|i)$, in symbol $\tilde{P} = P\circ K = P\circ(K_1\circ\cdots\circ K_{T-1})$. Then problem (3.2) can now be written in the form

$$\min_{x\in\mathbb{X}}\max_K\{\mathbb{E}_{P\circ K}[H(x,\xi)]\ s.t.\ K = K_1\circ\ldots\circ K_{T-1};\ \sum_{i,j\in\mathcal{N}_T}\mathsf{d}(i,j)\cdot P_i\cdot K(j|i)\le\epsilon\}. \tag{3.9}$$

It is noticeable that the expression $\sum_{i,j\in\mathcal{N}_T}\mathsf{d}(i,j)\cdot P_i\cdot K(j|i)\le\epsilon$ in (3.9) is multilinear in transportation subkernels $K_1,\ldots,K_{T-1}$ and therefore the maximization in (3.9) is a multilinear and hence typically nonconvex optimization. In applications for ambiguity problems the scenario probabilities P (and therefore the conditional probabilities Q) are fixed while $\tilde{P}$ and $\tilde{Q}$, the *worst tree* candidates, are constructed with feasible transportation subkernels such that $\mathrm{dl}(P, P\circ K)\le\epsilon$. In the next section, an exact and an approximative algorithm for solving the problem (3.9) are presented.

4 Solution algorithm—successive programming

To begin with we consider only the general saddle point problem. In a game theroetic interpretation with $f(x,y)$ being the payoff of a zero-sum game, a saddle point (x^*,y^*) is an equlibrium such that neither the decision maker nor the opponent would benefit by deviating from it. Algorithms for calculating saddle points have been of great interest since the seminal work of (Arrow et al. 1958). In the unconstrained case

$$\min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^m} f(x,y), \tag{4.1}$$

under convex-concavity assumptions of f in problem (4.1), for a given x, $f(x,y)$ has an unconstrained maximizer with respect to y and for given y, an unconstrained minimizer with respect to x. Under differentiability, a necessary and sufficient condition for a saddle point $\zeta^* = (x^*,y^*)$ is given by the validity of the simultaneous system of equations: $\mathcal{E}(\zeta) \equiv \begin{bmatrix} \nabla_x \; f(x, y) \\ -\nabla_y \; f(x, y) \end{bmatrix} = 0$. Sometime is even more convenient to solve problem

$$\min_{\zeta} \left\{ \frac{1}{2} \|\mathcal{E}(\zeta)\|_2^2 \right\} \tag{4.2}$$

rather than $\mathcal{E}(\zeta) = 0$, (Rustem and Howe 2002). Authors in (Demynov and Pevnyi 1972) and (Danilin and Panin 1974) proposed a gradient based algorithm for unconstrained problem (4.1) based on direction d_k and step size strategy α_k such that sufficient progress at each iteration is ensured. Besides, in the survey part of (Rustem and Howe 2002), more saddle point computation algorithms are presented and discussed, e.g., quadratic approximation algorithm for constrained problems based on (Qi and Sun 1995), interior point saddle point algorithm for constrained problems as elaborated in (Sasai 1974) and finally a quasi-Newton algorithm for nonlinear systems.

In distributionally robust multiperiod stochastic setting, the algorithm for computing the saddle point, should be tailored in order to fit the complex structure of ambiguity sets and at the same time guarantees the convergence to the equilibrium strategy. Due to the dissimilarity between decision space $\mathbb{X}$ and model space $\mathcal{P}_\epsilon$ in our setting, gradient based algorithms are avoided.[3] For the problem at hand, the criterion function is $F(x, \mathbb{P}) = \mathbb{E}_{\mathbb{P}}[H(x, \xi)]$. We iteratively find a saddle point by stage wise approximating the ambiguity set $\mathcal{P}_\epsilon$ by a finite set. In particular, the following method is proposed and the proof of convergence is given in the Appendix 7.2.

$$\begin{cases} x^{k+1} = \arg\min_{x \in \mathbb{X}} \max_{1 \le l \le k} \; F(x, \; \tilde{\mathbb{P}}^l) \\ \tilde{\mathbb{P}}^{k+1} = \arg\max_{\tilde{\mathbb{P}} \in \mathcal{P}_\epsilon} \; F(x^{k+1}, \tilde{\mathbb{P}}) \end{cases} \tag{4.3}$$

The precise from of the solution procedure is shown in the Algorithm of Table 1. At each iteration a new model $\mathbb{P}(\mathbb{T}, \tilde{P})$, in short: $\tilde{\mathbb{P}}$, which is in ϵ nested distance of baseline model $\mathbb{P}(\mathbb{T}, P)$, in short: $\mathbb{P}$, is included in the model and therefore the size of the problem increases at each iteration.

Proposition 1 *Let $\mathbb{X}$ and $\bar{\mathcal{P}}_\epsilon$ be compact sets and $(x, \mathbb{P}) \mapsto F(x, \mathbb{P}) = \mathbb{E}_{\mathbb{P}}[H(x, \xi)]$ be jointly continuous, then every cluster point of the iteration given by (4.3) is a minimax solution.*

Proof The proof of this Proposition can be found in the Appendix 7.2. □

Step 3 of this algorithm is the crucial part. It involves a nonconvex optimization. For this reason, we approximate step 3 by a stage wise procedure such that feasibility is never violated. This stage wise prodecure is shown in the Algorithm of Table 2.

[3] Notice that even under strict convex-concavity and compactness of $\mathcal{X}$ and $\mathcal{Y}$ the convergence of $\begin{cases} x^{k+1} = \arg\min_{x \in \mathcal{X}} f(x, y^k) \\ y^{k+1} = \arg\max_{y \in \mathcal{Y}} \; f(x^{k+1}, y) \end{cases}$ is not guaranteed.

Table 1 Successive programing algorithm

```
0. Let $k = 0$ and determine the value of $\epsilon$

1. Start with the „base line" model, i.e. $\mathcal{P}_\epsilon^k = \{\mathbb{P}\}$

2. Solve the outer optimization problem:
   $\left\| \begin{array}{l} \min\ u \\ \text{s.t. } \mathbb{E}_{\mathbb{P}}[H(x,\xi)] \le u \text{ for all } \mathbb{P} \in \mathcal{P}_\epsilon^k \\ \quad x \in \mathbb{X}, \\ \quad x \lhd \mathfrak{F} \end{array} \right. \longmapsto (x^k, u^k)$

3. Fix $x^k$ and solve the inner optimization problem:
   $\left\| \begin{array}{l} \max\ \mathbb{E}_{\tilde{\mathbb{P}}}[H(x^k,\xi)] \\ \text{s.t.} \quad \tilde{\mathbb{P}} \in \mathcal{P}_\epsilon^k \end{array} \right. \longmapsto (\tilde{\mathbb{P}}^k) \text{ and } \mathcal{P}_\epsilon^{k+1} = \mathcal{P}_\epsilon^k \cup \{\tilde{\mathbb{P}}^k\}$

4.Stop if there is no improvement in $u^k$, otherwise go to 2.

Note : In practical implementation we might :
- choose a stopping criteria $\theta$ s.t. $u^{k+1} - u^k \ge \theta$, or
- specify in advance the number of iterations $k$ .[4]
```

Table 2 Stage-wise construction of the worst tree

```
As a prerequisite, calculate $n \times n$ distance matrix $\mathsf{d}(i,j) = \sum_{t=0}^{T} \sum_{m=1}^{M} w_t^m \, |\xi_t^m(i) - \xi_t^m(j)|$
and keep it unchanged through the whole procedure.

Suppose we are at $k^{th}$ iteration of Algorithm 1, i.e., the incumbent solution is $x^k$ and is fixed
for now.

Let $K = K_1^{old} \circ \cdots \circ K_{T-1}^{old}$ be the transportation subkernels form the last iteration (At the
beginning start with identity kernels).

For $t = 1..T-1$ do
    Solve $\left\| \begin{array}{ll} \underset{K_t}{\text{maximize}} & (H(x^k,\cdot))^\top \tilde{P}^{k,[t]} \\ \text{s.t.} & \mathbb{d}(\tilde{P}^{k,[t]}, P) \le \frac{\epsilon}{T-1} \\ \text{where} & \tilde{P}^{k,[t]} = P \circ K_1^{new} \circ \cdots \circ K_{t-1}^{new} \circ K_t \circ K_{t+1}^{old} \circ \cdots \circ K_{T-1}^{old} \end{array} \right.$ and call the solution $K_t^{new}$.
```

The proposed algorithm was implemented and the results for a classical multi-period production/inventory control problem are presented and discussed in the next section.

5 Implementation and computational results

5.1 A multiperiod production/inventory control problem

To illustrate the multistage approach towards stochastic modeling and its ambiguity extension and to picture the implications of our proposed saddle point computation algorithm, in this section a simplified multistage stochastic optimization problem—

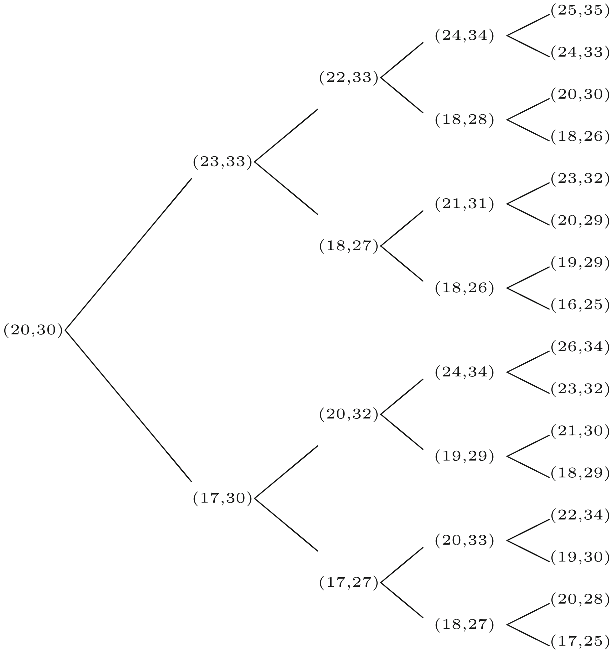

Fig. 2 Demand requirements (product$_1$, product$_2$) and the binary tree structure

a multiperiod production/inventory control problem—is implemented and numerical results are shown.[4]

In this problem the production volume of two products is decided while maximizing the expected net profit derived from selling the products under stochastic demands of the subsequent weeks with fixed selling prices, production, inventory and external supply costs. Deciding on how much of each product types to produce during a particular week forms the decision variables. The production machine is designed to produce both types and there is an overall production capacity. The stochastic demand is characterized in terms of scenarios and a tree terminology is used to describe event probabilities and multistage scenarios. The demand scenarios are represented on a binary tree with not necessarily equal event probabilities. In Fig. 2, the tree structure and demand requirements of both products is depicted. In Fig. 3, however, the demand scenarios of both products and the corresponding scenario probabilities are shown separately to emphasize on two level of demand.

5.1.1 Mathematical modeling summary

In Table 3, the symbols defining the parameters, decisions and decision dependent variables of the model are introduced. The full mathematical model in nodal repre-

[4] The numerical example is taken from AIMMS optimization modeling [(Bisschop 2012), Chapter 17.]. However, all computational procedure, solution algorithms and results analysis are implemented in MATLAB R2012a.

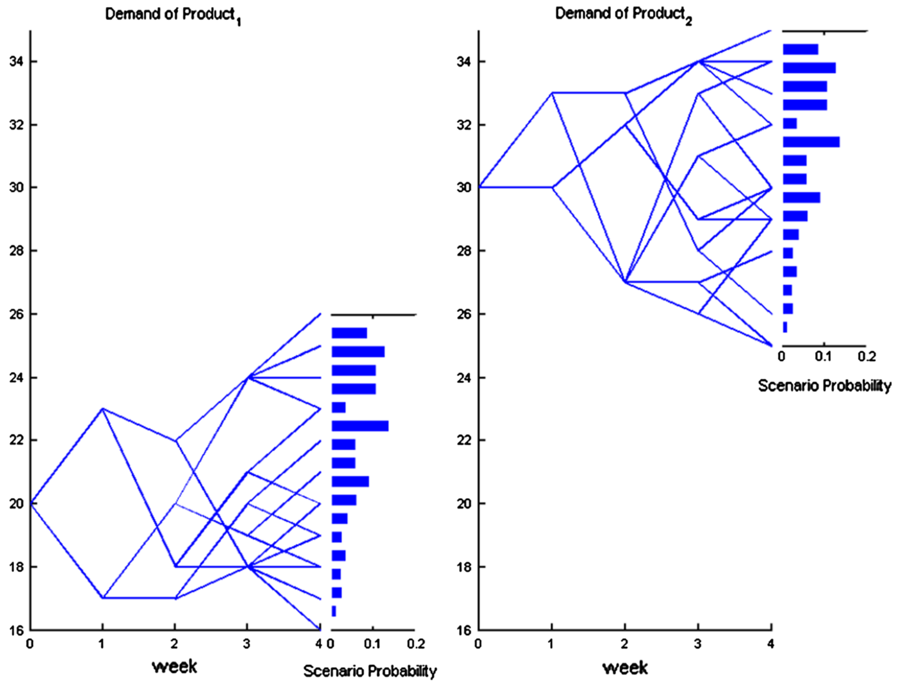

Fig. 3 Demand scenarios of product$_1$ and product$_2$ and the corresponding scenario probabilities

Table 3 Nomenclature

Parameters	
pr^b	Selling price for each product $b = 1, 2$
pc^b	Production cost of each product $b = 1, 2$
ic^b	Inventory cost of each product $b = 1, 2$
ec^b	External supply cost of each product $b = 1, 2$
c	Maximum overall production capacity
$\bar{x}_i$	Maximum inventory capacity
init_b	Initial stock level of product $b = 1, 2$
Random process	
d^b	Demand for product $b = 1, 2$
Decision variables	
x_f^b	Production volume of product b for $b = 1, 2$
Decision dependent variables	
x_i^b	Inventory level of each product $b = 1, 2$
x_e^b	External supply of each product $b = 1, 2$
v	Profit

sentation is formulated too. Note that decisions are only defined for emanating nodes and thus not for leaf (terminating) nodes.

$$
\begin{array}{llll}
\max \sum_n P(n)v(n) & \forall n \in \mathcal{N} & & \\
\text{s.t.} \sum_b x_f^b(n_-) \leq c & \forall n \in \mathcal{N} \backslash \mathcal{N}_0 & (a) & \\
x_i^b(n_-) + x_f^b(n_-) + x_e^b(n) - d^b(n) = x_i^b(n) & \forall b, \forall n \in \mathcal{N} \backslash \mathcal{N}_0 & (b) & \\
\sum_b x_i^b(n) \leq \bar{x}_i & \forall n \in \mathcal{N} & (c) & \\
x_i^b(n_-) + x_e^b(n) \geq d^b(n) & \forall b, \forall n \in \mathcal{N} \backslash \mathcal{N}_0 & (d) & (5.1)\\
\sum_b pr^b d^b(n) - \sum_b [pc^b x_f^b(n_-) + ic^b x_i^b(n) + ec^b x_e^b(n)] = v(n) & \forall n \in \mathcal{N} \backslash \mathcal{N}_0 & (e) & \\
x_i^b,\ x_e^b \in \mathbb{R}_+^{|\mathcal{N}|} & \forall b & & \\
x_f^b \in \mathbb{R}_+^{|\mathcal{N} \backslash \mathcal{N}_T|} & \forall b & & \\
v \in \mathbb{R}^{|\mathcal{N}|} & & &
\end{array}
$$

The objective of this production/ inventory control model is to maximize the total expected net profit ($P(n)$ is the unconditional probability of reaching node $n \in \mathcal{N}$) under the following constraints. Constraint (a) ensures that the total production volume is bounded above with the overall capacity. Whereas, (b) states that the inventory determined at each reachable node by the inventory at the predecessor node plus the production volume at the predecessor node plus the external supply at that not minus the demand pertaining to that node, while (c) illustrates the maximum inventory capacity constraints. Constraint (d) ensures the stochastic demand of both product is met at each node, since because of technicality issues, the product which is produced at the current node, can not be used to satisfy the demand at that node. Finally, constraint (e) is an accounting equation for the net profit position at each node which is derived from the sales' revenue minus the total costs consisting of production, inventory and external supply. The revenues and cost parameters are presented in Table 4. In the next section first the optimal solution of the original multistage problem (5.1) is shown and further the maximin solution of distributionally robust extension of (5.1) is presented and discussed. Distributionally robust extension of this example seeks for equilibrium strategies that ensure the maximum expected net profit under the most adverse demand scenarios.

5.2 Computational results

Optimal solutions of the original problem

Based on the multistage stochastic optimization problem developed in (5.1) and the input data provided, the optimal value of expected net profit is 7,688(€). In Fig. 4,

Table 4 Parameters: revenues, costs and capacities

Product	pr^b(€/U)	pc^b(€/U)	ic^b(€/U)	ec^b(€/U)	init_b	$\bar{x}_i$	c
product$_1$	300	12	5	195	17	52	46
product$_2$	400	10	5	200	35		

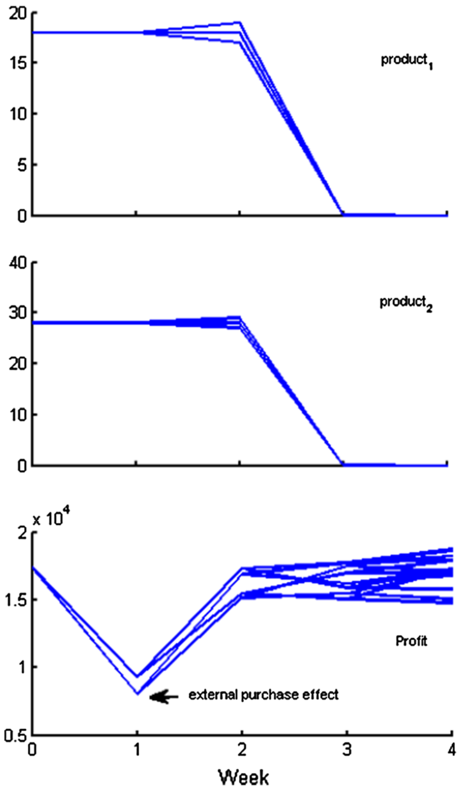

Fig. 4 Optimal solution scenarios

an overview of the optimal scenarios for decision variables x_f^1, x_f^2 and profit v are shown. Solution scenarios for both products follow a rather simple uniform pattern. One direct effect of optimal decisions on profit scenarios is observed in the sudden decrease of net profit levels at stage one, since satisfying the emanating demand at stage two requires a compensatory act by external purchasing for both products.

Worst tree visualizations

The stage wise algorithm for construction of worst trees which was discussed in Sect. 3.2 is implemented. As it is expected by increasing the ambiguity radius the largest probability would be associated to a scenario which for given solutions has the worst outcome. As ϵ increases the worst tree turns out to be less and less complex. The ambiguity sets are constructed for $\epsilon = 1, 6, 11, 16$ the analogy behind the range of varying ϵ empirically is simply ranging between $[\min \mathsf{d}(i, j),\ \max \mathsf{d}(i, j)]$, where $\mathsf{d}(i, j)$, as defined before, is the distance between demand scenarios i and j. Regardless of demand levels for products 1 and 2, in Fig. 5, the scenario probabilities of the respective tree structure for increasing ambiguity radius is depicted. It is observed that at the largest radius, $\epsilon = 16$, remaining scenarios 3 and 4 form the worst tree.

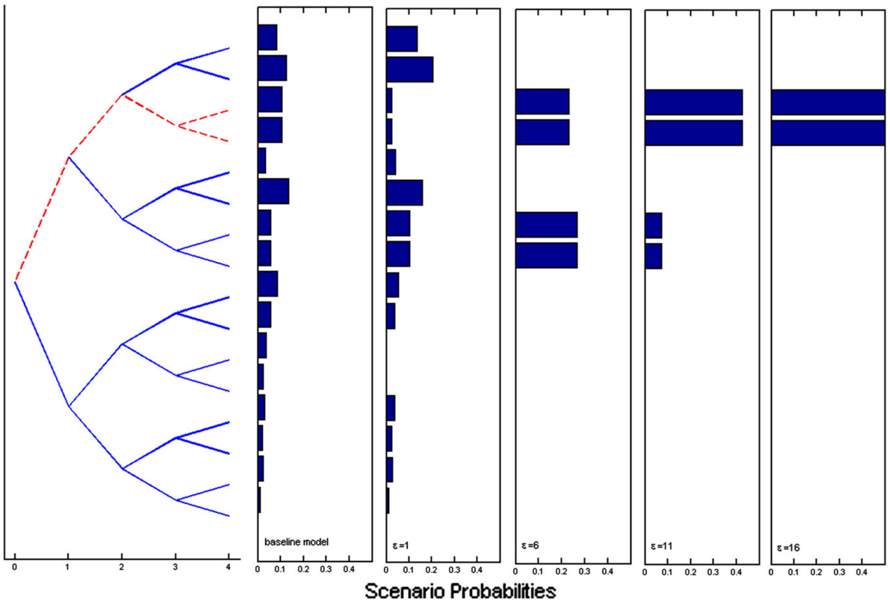

Fig. 5 Tree structure of problem (5.1) and diminishing worst trees for increasing ambiguity radii

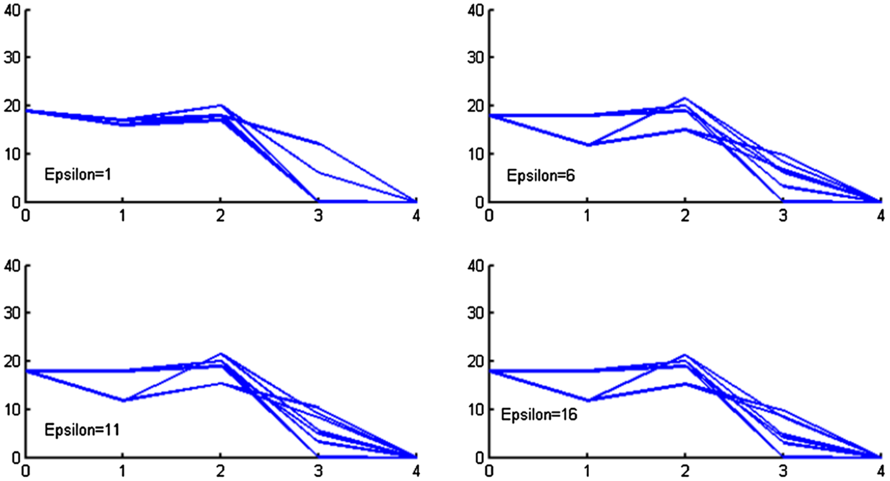

Fig. 6 Sensitivity of decisions under model ambiguity-product$_1$

Maximin solutions of the production scenarios for different ambiguity radii

In Figs. 6 and 7, the maximin solution of production scenarios for product$_1$, product$_2$ and its sensitivity with respect to increasing ambiguity radius is shown. It is noticeable that including rather than only one "baseline" demand model, results in more diverse production scenarios which is observable by comparing, for product$_1$ the graph on

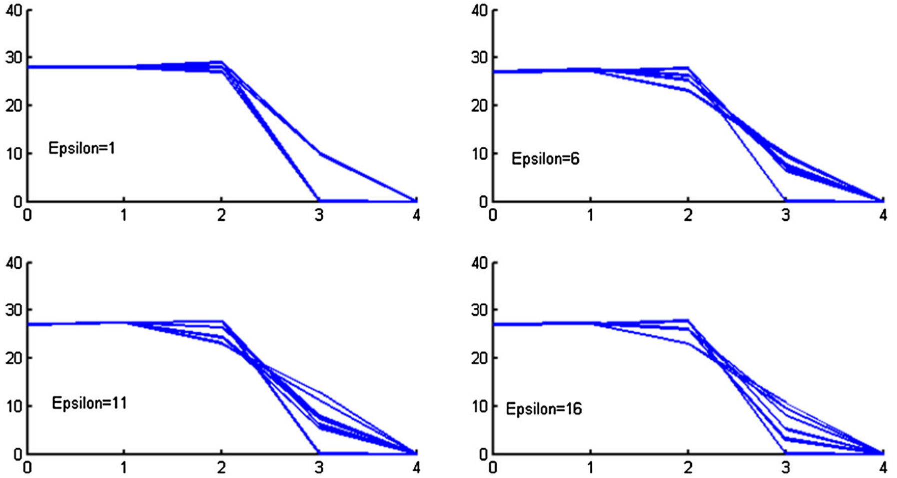

Fig. 7 Sensitivity of decisions under model ambiguity-product$_2$

the top of Fig. 4 with Fig. 6, and for product$_2$ the middle graph in Fig. 4 with Fig. 7 respectively. At first sight these results might seem quite controversial, since worst scenarios are getting a simpler structure as ϵ increases, where the decision scenarios get more complex. This can be explained by the fact that compomise (i.e. distrubutionally robust decisions) tend to be less extreme and more flexible compared to the single model case. As a consequence this phenomenon has also an impact on all decision dependent scenarios, in this example on the external purchase x_e^b and inventory level x_i^b for $b = 1, 2$ and last but not least on the profit scenarios.

The price of ambiguity and the gain of robustness

A fundamental result in the minimax setting of the ambiguity problem is that for $x^*(\mathbb{P})$, the optimal solution of the baseline model and $x^*(\mathcal{P}_\epsilon)$, the minimax solution, we have the following inequalities:

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}[H(x^*(\mathbb{P}), \xi)] \le \mathbb{E}_{\mathbb{P}}[H(x^*(\mathcal{P}_\epsilon), \xi)] &\le \min_x \max_{\tilde{\mathbb{P}} \in \mathcal{P}_\epsilon} \mathbb{E}_{\tilde{\mathbb{P}}}[H(x, \xi)] \\ &\le \max_{\tilde{\mathbb{P}} \in \mathcal{P}_\epsilon} \mathbb{E}_{\tilde{\mathbb{P}}}[H(x^*(\mathbb{P}), \xi)]. \end{aligned} \tag{5.2}$$

Based on these inequalities, one may define the following nonnegative quantities

- The *price of ambiguity*:$\mathbb{E}_{\mathbb{P}}[H(x^*(\mathcal{P}_\epsilon), \xi)] - \mathbb{E}_{\mathbb{P}}[H(x^*(\mathbb{P}), \xi)]$, indicating the loss in optimality, if the baseline model is true and the robust decision is implemented;
- The *gain of robustness*: $\max_{\tilde{\mathbb{P}} \in \mathcal{P}_\epsilon} \mathbb{E}_{\tilde{\mathbb{P}}}[H(x^*(\mathbb{P}), \xi)] - \min_x \max_{\tilde{\mathbb{P}} \in \mathcal{P}_\epsilon} E_{\tilde{\mathbb{P}}}[H(x, \xi)]$, indicating the gain under the worst model, if the robust decision is implemented and not the baseline optimal decision.

Notice that in our example, we maximize profits, i.e. we minimize negative profits, i.e. costs. For this reason, minimax solutions turn into maximin solutions and all

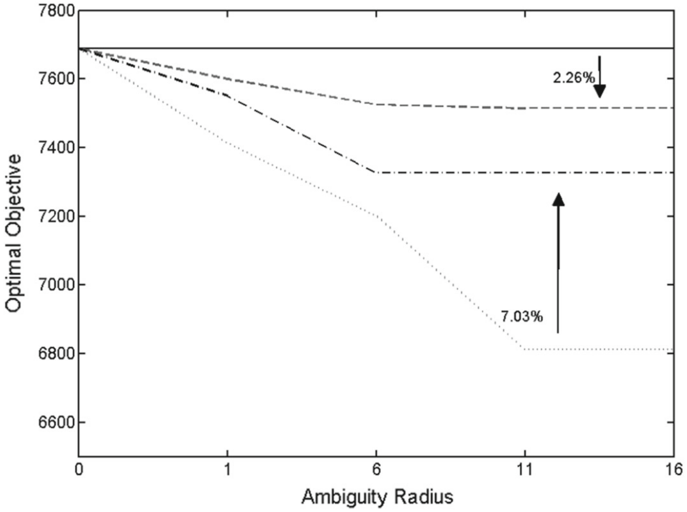

Fig. 8 The price for ambiguity and the gain for robustness as functions of the ambiguity radius

inequalities in (5.2) flip sign. Denoting the profit function to be maximized by $\bar{H}$, we get in the maximin case

- The *price of ambiguity*: $\mathbb{E}_{\mathbb{P}}[\bar{H}(x^*(\mathbb{P}),\xi)] - \mathbb{E}_{\mathbb{P}}[\bar{H}(x^*(\mathcal{P}_\epsilon),\xi)]$;
- The *gain of robustness*: $\max_x \min_{\tilde{\mathbb{P}}\in\mathcal{P}_\epsilon} \mathbb{E}_{\tilde{\mathbb{P}}}[\bar{H}(x,\xi)] - \min_{\tilde{\mathbb{P}}\in\mathcal{P}_\epsilon} \mathbb{E}_{\tilde{\mathbb{P}}}[\bar{H}(x^*(\mathbb{P}),\xi)]$.

In Fig. 8, these values are presented for the production/inventory control problem are presented as functions of the ambiguity radii. The solid line shows the optimal solution $\mathbb{E}_{\mathbb{P}}[\bar{H}(x^*(\mathbb{P}),\xi)]$ of problem (5.1) under the baseline model. The dashed line shows $\epsilon \mapsto \mathbb{E}_{\mathbb{P}}[\bar{H}(x^*(\mathcal{P}_\epsilon),\xi)]$, from which one can see the price of ambiguity, which at the largest ϵ is a 2.26 % decrease in the expected net profit. The dashed-dotted line represents the maximin values $\epsilon \mapsto \max_x \min_{\tilde{\mathbb{P}}\in\mathcal{P}_\epsilon} \mathbb{E}_{\tilde{\mathbb{P}}}[\bar{H}(x,\xi)] = \min_{\tilde{\mathbb{P}}\in\mathcal{P}_\epsilon} \mathbb{E}_{\tilde{\mathbb{P}}}[\bar{H}(x^*(\mathcal{P}_\epsilon,\xi)]$. Finally, the dotted line shows $\epsilon \mapsto \min_{\tilde{\mathbb{P}}\in\mathcal{P}_\epsilon} \mathbb{E}_{\tilde{\mathbb{P}}}[\bar{H}(x^*(\mathbb{P}),\xi)]$ indicating that the gain of robustness is 7.03 %. A closer look shows that there is no more change in the maximin solution when making ϵ larger, i.e., there is an upper bound both for the price of ambiguity and the gain of robustness.

6 Concluding remarks and further work

In this paper a concept and an algorithm to robustify a multistage stochastic optimization problem under model ambiguity is presented. This robust counterpart of the original problem is constructed by the worst case approach with respect to the probability models which are in an ϵ nested neighborhood of a baseline model. We considered only fixed scenario values and assumed the changes only in the probability values. However the approach has the possibility for extensions subject to more technical complication. It is expected that by high performance and parallel computing methods, ambiguity problems for quite large trees can be solved.

The nested distance is a new concept that appropriately incorporates the filtration structure in the multistage stochastic optimization models. In our approach, we considered minimax w.r.t a neighborhood of a baseline model. It was observed that the worst case model gets a simpler and simpler structure as epsilon increases. The decisions, however, are shown to be more and more complex. The reason for this phenomena might be in the fact that by including more models, the decisions should be taken which more flexibility and decisions lying on bounds whould be avoided. We were able to quantify the cost (and reward) of robustness. Moreover, it is seen that there is a typically threshold for the ambiguity radius, from which onwards no improvements appear in minimax decisions and the objective function.

It is desirable to relate the ambiguity radius to the quality of the statistical information about the model. However, the situation is much more complex than in the single- or two-stage case. Since conditional distributions are typically estimated on the basis of a parametric model, the ambiguity radius is only indirectly related to the confidence sets of the parameters. This relation is the topic of further research.

7 Appendix

7.1 The proof of Theorem 1.

We fix a finite tree $\mathbb{T}$ with a given structure and with the values of the scenario process sitting on its nodes. By determining the scenario probabilities $P = (P_i)_{i \in \mathcal{N}_T}$ the corresponding nested distribution $\mathbb{P}(\mathbb{T}, P)$ is formed. The alternative models are $\mathbb{P}(\mathbb{T}, \tilde{P})$ with a variant $\tilde{P}$ of the scenario probabilities. The notion of compound can be generalized to infinitely many elements: Let $\mathfrak{P}$ be the family of all probability measures on $\mathcal{N}_T$, which is—since $\mathcal{N}_T$ is a finite set—a simplex. Let Λ be a probability measure from $\mathfrak{P}$. The compound $\mathcal{C}(\mathbb{P}(\mathbb{T}, \tilde{P}), \Lambda)$ is defined as

$$\mathcal{C}(\mathbb{P}(\mathbb{T}, \cdot), \Lambda) = \mathbb{P}(\mathbb{T}, \tilde{P}) \quad \text{where } \tilde{P} \text{ is distributed according to } \Lambda,$$

meaning that the compound is obtained by first sampling a distribution $\tilde{P}$ according to Λ and then taking the model $\mathbb{P}(\mathbb{T}, \tilde{P})$. Refer to Fig. 9. in which $\mathcal{C}(\mathbb{P}(\mathbb{T}, \cdot), \Lambda)$ is illustrated for probability measure Λ with finite support . If Λ sits on $\tilde{P}^{(1)}, \tilde{P}^{(2)}, .., \tilde{P}^{(k)}$ with probabilities λ_l for $1 \leq l \leq k$, then compound model has k nodes at stage 1 and to the lth node of stage 1 the subtree $\mathbb{P}(\mathbb{T}, \tilde{P}^{(l)})$ is associated, i.e.

$$\mathcal{C}(\mathbb{P}(\mathbb{T}, \cdot), \Lambda) = \sum_{l=1}^{k} \lambda_l \mathbb{P}(\mathbb{T}, \tilde{P}^{(l)})$$

where the convex combination $\sum_{l=1}^{k} \lambda_l \mathbb{P}(\mathbb{T}, P^{(l)})$ is in the sense of compounding. Notice that the tree of $\mathcal{C}(\mathbb{P}(\mathbb{T}, \tilde{P}_\lambda), \Lambda)$ is of height $T + 1$. Thus original tree $\mathbb{P}(\mathbb{T}, P)$ to be comparable with $\mathcal{C}(\mathbb{P}(\mathbb{T}, \tilde{P}_\lambda), \Lambda)$, we assume that a further root (with probability *one*) is appended to the tree of $\mathbb{P}(\mathbb{T}, P)$ and denote this extended tree by $\mathbb{P}_+(\mathbb{T}, P)$. In the following, we write $\mathbb{P}(\mathbb{T}, \Lambda)$ for $\mathcal{C}(\mathbb{P}(\mathbb{T}, \cdot), \Lambda)$.

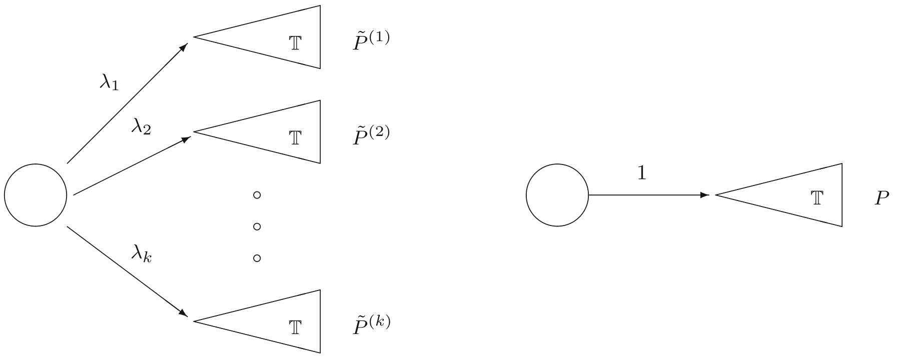

Fig. 9 The compound convex structure of trees $\mathbb{P}(\mathbb{T}, \tilde{P}^{(l)})$ and augmented tree $\mathbb{P}_{+}(\mathbb{T}, P)$

The convex hull of the set

$$\mathcal{P}_\epsilon = \left\{\mathbb{P}(\mathbb{T}, \tilde{P}) : \ \tilde{P} \in \mathcal{B}_\epsilon\right\}$$

with

$$\mathcal{B}_\epsilon = \{\tilde{P} : \ \mathrm{dl}(\mathbb{P}(\mathbb{T}, P), \mathbb{P}(\mathbb{T}, \tilde{P})) \le \epsilon\}$$

is the set

$$\bar{\mathcal{P}}_\epsilon = \{\mathcal{C}(\mathbb{P}(\mathbb{T}, \cdot), \Lambda) : \ \Lambda \text{ is a probability measure on } \mathcal{B}_\epsilon\}. \tag{7.1}$$

The convexified problem (3.2) is rewritten to

$$\min_{x \in \mathbb{X}} \max_{\tilde{\mathbb{P}} \in \bar{\mathcal{P}}_\epsilon} \{\mathbb{E}_{\tilde{\mathbb{P}}}[H(x, \xi)] \text{ s.t. } x \lhd \mathfrak{F}, \ (\Omega, \mathfrak{F}, \tilde{P}; \xi) \sim \tilde{\mathbb{P}}\}. \tag{7.2}$$

Notice that in the formulation (7.2) the decision variables x must coincide in all randomly sampled subproblems, cf. Fig. 9. By safeguarding ourselves against any random selection of elements of $\mathcal{B}_\epsilon$, we automatically safeguard ourselves against the worst case in $\mathcal{B}_\epsilon$. The next step is to calculate the nested distance between two elements of $\bar{\mathcal{P}}_\epsilon$. For two leaves i resp. j of the tree $\mathbb{T}$ the distance is defined as the distance of the corresponding paths leading to i resp. j, i.e.,

$$\mathsf{d}(i, j) = \sum_{t=1}^{T} \sum_{m=1}^{M} w_t^m |\xi_t^m(i) - \xi_t^m(j)|$$

Assume that for all $i \neq j$, there exist constants $c, C > 0$ such that $c \le \mathsf{d}(i, j) \le C$. Let

$$\left\| P - \tilde{P} \right\| = \sum_{i \in \mathcal{N}_T} \left| P_i - \tilde{P}_i \right| = 2 - 2 \sum_{i \in \mathcal{N}_T} \min(P_i, \tilde{P}_i).$$

It follows that

$$\frac{c}{2} \cdot \left\| P - \tilde{P} \right\| \le \mathrm{dl}(\mathbb{P}(\mathbb{T}, P), \mathbb{P}(\mathbb{T}, \tilde{P})) \le \frac{C}{2} \cdot \left\| P - \tilde{P} \right\|. \tag{7.3}$$

In order to show (7.3) notice that an optimal transportation plan can transport a mass of $\min(P_i, \tilde{P}_i)$ from i to i with distance 0. Thus only the masses $1 - \sum_{i \in \mathcal{N}_T} \min(P_i, \tilde{P}_i)$ have to be transported, over distances which lie between c and C, whence the assertion follows. Notice well that the use of the distance $\|P - \tilde{P}\|$ is only to demonstrate compactness. While the topologies generated by the two metrics $\|P - \tilde{P}\|$ and $\mathrm{dl}(\mathbb{P}(\mathbb{T}, P), \mathbb{P}(\mathbb{T}, \tilde{P}))$ are the same [due to relation (7.3)], balls are quite different in the two metrics and only the latter metric is appropriate for nested distributions. Next we see that $\bar{\mathcal{P}}_\epsilon$ is compact, since it is the continuous image of the set of all probability measures on $\mathcal{B}_\epsilon$, which is a compact set, since $\mathcal{B}_\epsilon$ itself is compact. Thus all conditions for the validity of the basic von Neumann Minimax Theorem are fulfilled and a saddle point $(x^*, \mathbb{P}(\mathbb{T}, \Lambda^*))$ must exist. Now we prove the equation

$$\mathrm{dl}(\mathbb{P}(\mathbb{T}, \Lambda), \mathbb{P}_+(\mathbb{T}, P)) = \int \mathrm{dl}(\mathbb{P}(\mathbb{T}, \tilde{P}), \mathbb{P}(\mathbb{T}, P)) \; \Lambda(\mathrm{d}\tilde{P}). \tag{7.4}$$

In order to see this, assume first that Λ is finite, say $\mathbb{P}(\mathbb{T}, \Lambda) = \sum_{l=1}^{k} \lambda_l \mathbb{P}(\mathbb{T}, \tilde{P}^{(l)})$. Then:

$$\begin{aligned} \mathrm{dl}(\mathbb{P}(\mathbb{T}, \Lambda), \mathbb{P}_+(\mathbb{T}, P)) &= \mathrm{dl}\left(\sum_{l=1}^{k} \lambda_l \mathbb{P}(\mathbb{T}, \tilde{P}^{(l)}), \mathbb{P}_+(\mathbb{T}, P) \right) \\ &= \sum_{l=1}^{k} \lambda_l [\mathrm{dl}(\mathbb{P}(\mathbb{T}, \tilde{P}^{(l)}), \mathbb{P}(\mathbb{T}, P))] \quad . \end{aligned}$$

If Λ is not finite, it can be approximated by finite measures and therefore the relation (7.4) holds in general. Finally, we show that the worse case model $\tilde{\mathbb{P}}^*$ happens at a single tree and not a mixture of trees: Let x^* be the minimax decision, i.e.

$$\mathbb{E}_{\tilde{\mathbb{P}}}[H(x^*, \xi)] \le \mathbb{E}_{\tilde{\mathbb{P}}^*}[H(x^*, \xi)] \le \mathbb{E}_{\tilde{\mathbb{P}}^*}[H(x, \xi)].$$

Let the saddle point model be $\tilde{\mathbb{P}}^* = \mathbb{P}(\mathbb{T}, \Lambda^*)$. The support of Λ^* is closed (hence compact) and the continuous function $\tilde{P} \mapsto \mathbb{E}_{\mathbb{P}(\mathbb{T}, \tilde{P})}[H(x^*, \xi)]$ takes its maximum at some distribution $\tilde{P}^*$. Since $\mathrm{dl}(\mathbb{P}(\mathbb{T}, \tilde{P}^*), \mathbb{P}(\mathbb{T}, P)) \le \epsilon$ by construction, $\mathbb{P}(\mathbb{T}, \tilde{P}^*) \in \mathcal{P}_\epsilon$ and therefore $\mathbb{E}_{\mathbb{P}(\mathbb{T}, \tilde{P}^*)}[H(x^*, \xi)] \le \mathbb{E}_{\tilde{\mathbb{P}}^*}[H(x^*, \xi)]$. On the other hand,

$$\mathbb{E}_{\tilde{\mathbb{P}}^*}[H(x^*, \xi)] = \int \mathbb{E}_{\mathbb{P}(\mathbb{T}, \tilde{P})}[H(x^*, \xi)] \, \mathrm{d}\Lambda(\tilde{P}) \le \mathbb{E}_{\mathbb{P}(\mathbb{T}, \tilde{P}^*)}[H(x^*, \xi)].$$

Consequently, $\mathbb{E}_{\mathbb{P}(\mathbb{T},\tilde{P}^*)}[H(x^*,\xi)] = \mathbb{E}_{\tilde{\mathbb{P}}^*}[H(x^*,\xi)]$, which shows that the saddle point model can be chosen from $\mathcal{P}_\epsilon$. This concludes the proof.

7.2 The proof of Proposition 1.

Here we prove the convergence of iterative procedure

$$\begin{cases} x^{k+1} \in \arg\min_{x\in\mathbb{X}} \max_{1\leq l\leq k} \ F(x,\ \tilde{\mathbb{P}}^l) \\ \tilde{\mathbb{P}}^{k+1} \in \arg\max_{\tilde{\mathbb{P}}\in\mathcal{P}_\epsilon} F(x^{k+1},\tilde{\mathbb{P}}) \end{cases}.$$

Denote by $F^k = \max_{1\leq l\leq k} F(x^{k+1},\ \mathbb{P}^l)$, then $F^{k+1} = \max_{1\leq l\leq k+1} F(x^{k+2},\ \tilde{\mathbb{P}}^l)$ and by monotonicity $F^{k+1} \geq F^k$. Since the function F is bounded, F^k converges to $F^*{:=}\sup F^k$. Moreover, by compactness, the sequence x^k has one or several cluster points. Let x^* such a cluster point. We show that $F^* = \max_{\tilde{\mathbb{P}}\in\mathcal{P}_\epsilon} F(x^*,\ \tilde{\mathbb{P}})$. Since always $F^* \leq \max_{\tilde{\mathbb{P}}\in\mathcal{P}_\epsilon} F(x^*,\ \tilde{\mathbb{P}})$, suppose that $F^* < \max_{\tilde{\mathbb{P}}\in\mathcal{P}_\epsilon} F(x^*,\ \tilde{\mathbb{P}})$. Then there must exist a $\tilde{\mathbb{P}}^+$ such that $F(x^*,\ \tilde{\mathbb{P}}^+) > F^*$. By continuity this inequality must then hold in a neighborhood of x^* and therefor there must exist a x^k for which the same inequality holds. However, this contradicts the construction of the iteration. Finally, we show that $x^* \in \arg\min_{x\in\mathbb{X}} \max_{\tilde{\mathbb{P}}\in\mathcal{P}_\epsilon} F(x,\ \tilde{\mathbb{P}})$. If not, there must exist a x^+ such that $\max_{\tilde{\mathbb{P}}\in\mathcal{P}_\epsilon} F(x^+,\ \tilde{\mathbb{P}}) < \max_{\tilde{\mathbb{P}}\in\mathcal{P}_\epsilon} F(x^*,\ \tilde{\mathbb{P}})$. Hence, by construction $\max_{1\leq l\leq k} \ F(x^+,\ \tilde{\mathbb{P}}^l) \geq \max_{1\leq l\leq k} \ F(x^{k+1},\ \tilde{\mathbb{P}}^l) = F^k$ and letting k tend to infinity, one sees that $\max_{\tilde{\mathbb{P}}\in\mathcal{P}_\epsilon} F(x^+,\ \tilde{\mathbb{P}}) \geq F^* = \max_{\tilde{\mathbb{P}}\in\mathcal{P}_\epsilon} F(x^*,\ \tilde{\mathbb{P}})$ and this is a contradiction which shows that x^* is the cluster point and thus every cluster point is a solution of the minimax problem.

References

Arrow KJ, Hurwicz L, Uzawa H (1958) Studies in linear and non-linear programming. Stanford University Press, CA

Birge JR, Louveaux F (1997) Introduction to stochastic programming. Springer, Berlin

Bisschop J (2012) AIMMS optimization modelling. Paragon Decision Technology, USA

Calafiore G (2007) Ambiguous risk measures and optimal robust portfolios. SIAM J Control Optim 18(3):853–877

Chen Z, Epstein L (2002) Ambiguity, risk and asset returns in continuous time. Econometrics 70(4):1403–1443

Danilin YM, Panin VM (1974) Methods for searching saddle points. Kibernetika 3:119–124

Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problem. Oper Res 58:596–612

Demynov VF, Pevnyi AB (1972) Numerical methods for finding saddle points. USSR Comput Math Math Phys 12:1099–1127

Dupačová J (1980) On minimax decision rule in stochastic linear programing. Stud Math Program pp 47–60

Dupačová J (2001) Stochastic programming: Minimax approach. In: Floudas Ch. A, Pardalos PM (eds) Encyclopedia of Optimization, vol V, pp 327–330

Dupačová J (1987) The minimax approach to stochastic programming and an illustrative application. Stochastics 20:73–88

Dupačová J (2010) Uncertainties in minimax stochastic programs. Optimization 1:191–220

Fan K (1953) Minimax theorems. Proc Nat Acad Sci 39:42–47

Goh J, Sim M (2010) Distributionally robust optimization and its tractable approximations. Oper Res 58:902–917

Heitsch H, Römisch W, Strugarek C (2006) Stability of multistage stochastic programs. SIAM J Optim 17:511–525

Jagannathan R (1977) Minimax procedure for a class of linear programs under uncertainty. Oper Res 25:173–177

Pflug GCh, Römisch W (2007) Modelling, measuring and managing risk, 1st edn. World Scientific, Singapore

Pflug GCh, Wozabal D (2007) Ambiguity in portfolio selection. Quant Financ 7(4):435–442

Pflug GCh (2010) Version-independence and nested distributions in multistage stochastic optimization. SIAM J Optim 20(3):1406–1420

Pflug GCh, Pichler A (2012) A distance for multistage stochastic optimization models. SIAM J Optim 22(1):1–23

Qi L, Sun W (1995) An iterative method for the minimax problem. Minimax and Applications (Kluwer), art. Du and Pardalos

Rachev S, Römisch W (2002) Quantitative stability in stochastic programming: the method of probability metrics. Math Oper Res 27:798–818

Robinson W, Wets R (1987) Stability in two stage stochastic programming. SIAM J Control Optim 25:1409–1416

Römisch W, Schultz R (1991) Stability analysis for stochastic programs. Ann Oper Res 30:241–266

Rustem B, Howe M (2002) Algorithms for worst-case design and applications to risk management. princeton University Press, Princeton

Ruszczynski A, Shapiro A (2003) Stochastic Programming, 1st edn. ser. Handbooks in Operations Research and Management science, Amsterdam

Sasai H (1974) An interior penalty method for minimax for problems with constraints. SIAM J Control Optim 12:643–649

Scarf H (1958) Studies in the mathematical theory of an inventory problems. Stanford Univeristy Press, CA

Shapiro A, Kleywegt A (2002) Minimax analysis of stochastic problems. Optim Methods Softw 17:523–542

Shapiro A, Ahmed Sh (2004) On a class of minimax stochastic programs. SIAM J Optim 14:1237–1249

Sion M (1958) On general minimax theorems. Pac J Math 8:171–176

Thiele A (2008) Robust stochastic programming with uncertain probabilities. IMA J Manag Math 19:289–321

von Neumann J (1928) Zur Theorie der Gesellschaftsspiele. Math Ann 100:295–320

Wozabal D (2010) A framework for optimization under ambiguity. Ann Oper Res (online First)

Žácková a.k.a. Dupačová J (1966) On minimax solutions of stochastic linear programming problems. Casopis pro Pestovani Mathematiky 91:423–430