On Distributionally Robust Multiperiod Stochastic Optimization

Bita Analui*and Georg Ch. Pflug[†]

January 14, 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, 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 find the worst trees within the ambiguity set and determine a solution which is robust w.r.t. model uncertainty. It turns out that the probability weights of the worst case trees are concentrated on few very bad scenarios.

Keywords: Multistage Stochastic Optimization, Distributional Robustness, Model Ambiguity, Nested Distance, Inventory Management

*University of Vienna, Austria. IK-Computational Optimization and WWTF Project, Institute of Statistics and OR. email: bita.analui@univie.ac.at

[†]University of Vienna, Austria, Institute of Statistics and OR and International Institute for Applied System Analysis(IIASA), Laxenburg, Austria, email: georg.pflug@univie.ac.at

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 Gaussian process) and the parameters of the model are estimated on the basis of a sequence 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.

The notion of ambiguity was introduced by (Ellsberg, 1961). In the words of D. Ellsberg, ambiguity is :

> *...a quality depending on the amount, type, reliability and unanimity of information and giving rise to one's "degree of confidence" in an estimate of relative likelihoods. (p. 657).*

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 in an early attempt by (Scarf, 1958). He studies an optimal single product inventory problem under and 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.

More sophisticated approaches assume that true underlying probability model belongs to a given class of models and has motivated the utilization of general *minimax* decision rules;

it was pioneered in the mid-1960s by (Zácková a.k.a. Dupacová, 1966). This approach was applied for the class of stochastic LPs 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. In general, minimax approach is regarded as a body which bridges the gap between the conservatism of robust optimization and the specificity of stochastic optimization where the optimal decisions are sought for the worst case probability models by obtaining the best possible decisions for the most adverse considered circumstances.

There is no unique nomenclature for the ambiguity problem. Synonymous names are: *model uncertainty problem*, *minimax stochsastic optimization* and *distributionally robust problem.*

Many parametric/ 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.

Since we consider multistage decision models, information structure 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 $t = 0, ..., T-1$ with typically different levels of information. We denote the random scenario process by $\xi := (\xi_1, ..., \xi_T)$ and the pertaining multistage decision sequence by $x = (x_0, x_1, ..., x_{T-1})$. The indices of the random process and the decision process differ by one, since at time t, a decision is to be made but the realization of the random process will be observable only at time $t+1$. In addition, decisions at each time step t may depend only on the actual outcomes of the random variables up to time t, they must be *non-anticipative* with respect to the observations after time t, i.e., $x_t = x_t(\xi^t)$ with $\xi^t = (\xi_1, ..., \xi_t)$. For a broad technical presentation of multistage stochastic programming refer to (Birge and Louveaux, 1997), (Pflug and Römisch, 2007) and (Ruszczynski and Shapiro, 2003).

The pecularity of the multistage situation is the fact that model uncertainty has to be defined in terms of the conditional distributions of the scenario process and not just of its multivariate distribution. Some literature exists dealing with distributionally robustness for multistage programs. In (Delage and Ye, 2010) the authors study distributionally robust stochastic programs where the mean and covariance of the primitive uncertainties are themselves subject to uncertainty. In (Goh and Sim, 2010) the approach is extended 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.

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 section 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 section 5 we discuss the application of our approach to a classical stochastic multiperiod inventory control problem. For implementation we considered a problem of reasonable size in order to reflect the technical part of the algorithm in a representable manner. Finally section 6 reflects the main results and conclusions.

2 Multistage Stochastic Optimization

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

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

where H is a real-valued cost function, depending on decision sequence $x = (x_0, ..., x_{T-1})$ and a stochastic scenario process $\xi = (\xi_1, ..., \ \xi_T)$. 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 $\mathfrak{F} = (\mathcal{F}_1, \dots, \mathcal{F}_T)$ 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 $\xi \lhd \mathfrak{F}$). The nested distribution is the collection of conditional distributions of ξ_t given $\mathcal{F}_{t-1}$, written as $\xi_t|\mathcal{F}_{t-1}$, more precisely of the nested structure $(((\xi_T|\mathcal{F}_{T-1}), \xi_{T-1}|\mathcal{F}_{T-2}) \dots)\xi_2|\mathcal{F}_1)\xi_1$. 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, 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 $\mathbb{P} \sim (\Omega, \mathfrak{F}, P, \xi)$ symbolizes this. We assume that the decisions at stage t are non-anticipative, i.e. measurable w.r.t. $\mathcal{F}_t$ and lie in a given nonrandom constraint set $\mathbb{X}_t \subseteq \mathbb{R}^{d_t}$.

For a baseline model $\mathbb{P} \sim (\Omega, \mathfrak{F}, P, \xi)$ and an alternative model $\tilde{\mathbb{P}} \sim (\tilde{\Omega}, \tilde{\mathfrak{F}}, \tilde{P}, \tilde{\xi})$ 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 $r \geq 0$ of two *nested-structures* $\mathbb{P} \sim (\Omega, \mathfrak{F}, P, \xi)$ and $\tilde{\mathbb{P}} \sim (\tilde{\Omega}, \tilde{\mathfrak{F}}, \tilde{P}, \tilde{\xi})$ is the optimal value of the optimization problem

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

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, for instance

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

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. Notice 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 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 valuation 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 and the predecessor relations $\prec$. If $i \in \mathcal{N}_{t-1}$, $j \in \mathcal{N}_t$ and i is a direct predecessor of j, we write $i = j-$ and $j \in i+$. If k is any predecessor of j we write $k \prec j$ [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 $j \in \mathcal{N}_t$ and $i \in \mathcal{N}_{t-1}$ with $i = j-$, then probabilities

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

$Q(i,j)$ sitting on the arcs represent the conditional probabilities of reaching node j from node i. The $Q(i,j)$'s are the basis for calculating the unconditional probabilities $P(i)$ for every node. The unconditional 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 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

$$\begin{aligned} \mathrm{dl}(\mathbb{P}, \tilde{\mathbb{P}}) = \min_{\pi} \quad & \sum_{i,j \in \mathcal{N}_T} \mathsf{d}(i,j)\ \pi(i,j) \\ \text{subject to} \quad & \sum_{j \in l+} \pi(i,j|k,l) = Q(k,i) \qquad (k \prec i,\ l) \\ & \sum_{i \in k+} \pi(i,j|m,n) = \tilde{Q}(l,j) \qquad (k,\ l \prec j)\ \cdot \\ & \sum_{i,j} \pi(i,j) = 1 \\ & \pi(i,j) \geq 0 \end{aligned} \tag{2.4}$$

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 given by $\pi(i,j|k,l) = \frac{\pi(i,j)}{\sum_{i' \prec k,\ j' \prec l} \pi(i',j')}$[2], therefore the constraints in (2.4) can be reformulated for the unconditional probabilities P and $\tilde{P}$ as

$$\begin{aligned} P(i) \cdot \textstyle\sum_{i' \prec k,\ j' \prec l}\ \pi(i',j') = P(k) \cdot \sum_{j' \prec l}\ \pi(i,j') \qquad & (k \prec i,\ l) \\ \tilde{P}(j) \cdot \textstyle\sum_{i' \prec k,\ j' \prec l}\ \pi(i',j') = \tilde{P}(l) \cdot \sum_{i' \prec k}\ \pi(i',j) \qquad & (k,\ l \prec j). \end{aligned}$$

[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.

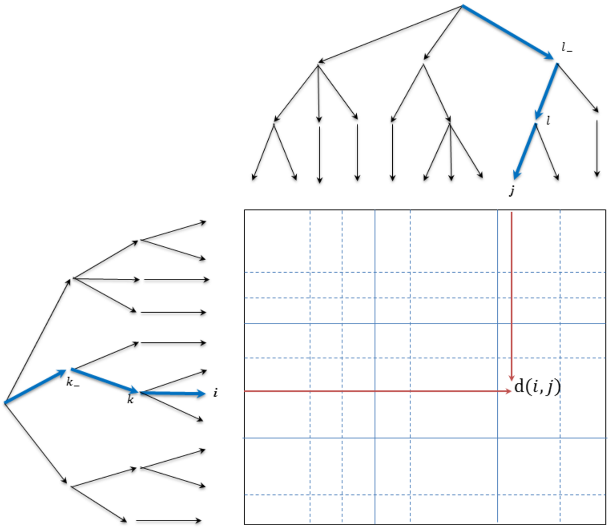

Figure 2.1: Visualization of transportation matrix π and distance matrix d for two trees. i, j are leaf and k, l are generic intermediate nodes.

When P and $\tilde{P}$ are given, reformulated problem (2.4) is indeed a linear program and reads:

$$\begin{aligned}
\mathrm{d\!l}(\mathbb{P}, \tilde{\mathbb{P}}) = \min_{\pi} \quad & \sum_{i,j \in \mathcal{N}_T} \mathsf{d}(i,j)\ \pi(i,j) \\
\text{subject to} \quad & P(i) \cdot \textstyle\sum_{i' \prec k,\ j' \prec l}\ \pi(i',j') = P(k) \cdot \sum_{j' \prec l}\ \pi(i,j') \qquad (k \prec i) \\
& \tilde{P}(j) \cdot \textstyle\sum_{i' \prec k,\ j' \prec l}\ \pi(i',j') = \tilde{P}(l) \cdot \sum_{i' \prec k}\ \pi(i',j) \qquad (l \prec j) \\
& \sum_{i,j} \pi(i,j) = 1 \\
& \pi(i,j) \geq 0
\end{aligned}$$

In Figure 2.1, the nested structure of transportation matrix π (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

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

The distributionally robust counterpart reads now

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

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 valuated by the scenario process. 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

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

The reason is that an ambiguity set of the form (3.3) 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

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

and set

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

The final formulation of the ambiguity extension problem is now

$$\min_x \max_{\tilde{\mathbb{P}} \in \mathcal{P}_\epsilon} \{ \mathbb{E}_{\tilde{\mathbb{P}}}[H(x, \xi)] : \ x \in \mathbb{X}, \ x \lhd \mathfrak{F} \}. \tag{3.5}$$

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.

3.1 A minimax Theorem

The famous minimax theorems (von Neumann, 1928),(Fan, 1953),(Sion, 1958) and all the references therein, assert that the min and the max can be interchanged in (3.5). The validity of such theorems is related to convexity/concavity properties of the criterion function and topological properties of feasible sets. Therefore the question must be answered in what respect nested distributions allow convex combinations. 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 the compound with probability λ is given by

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

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 $\mathbb{P}_+$ 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 $\mathcal{P}_\epsilon$ is not convex (w.r.t. compounding). 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 prove the following minimax theorem, which follows from general minimax theorems cited above.

Theorem 4. *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 a saddle point $(x^*, \tilde{\mathbb{P}}^*)$ exists, i.e.

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

Moreover, $\tilde{\mathbb{P}}^* \in \mathcal{P}_\epsilon$ (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 nested neighborhood.

3.2 Ambiguity sets defined by transportation kernels

We have seen, that in its general form, problem (3.5) 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 $\mathfrak{F}$ i.e. $x \lhd \mathfrak{F}$ ensures the comparability of the decisions of both models (2.1) and (3.5).

In order to describe the nested distance in a recursive form, we introduce the notion of transportation subplans. A transportation subplan indexed with pair node (k,l) transports the elements of $k+$, the set of direct successors of k, into $l+$, the set of direct successors of l and must satisfy the following marginal constraints

$$\begin{aligned} &\sum_{l\prec j}\pi(i,j|k,l) = Q(k,i) \\ &\sum_{k\prec i}\pi(i,j|k,l) = \tilde{Q}(l,j). \end{aligned} \tag{3.6}$$

From $P_i = Q(i) \cdot Q(pred_{T-1}(i)) \cdots Q(pred_1(i))$, $\tilde{P}_j = \tilde{Q}(j) \cdot \tilde{Q}(pred_{T-1}(j)) \cdots \tilde{Q}(pred_1(j))$ and satisfying constraints (3.6), all these subplans are *concatenated* to the full transportation plan :

$$\begin{aligned} \pi(i,j) &= \pi(pred_1(i), pred_1(j)|1,1) \cdots \pi(pred_{T-1}(i), pred_{T-1}(j)|pred_{T-2}(i), pred_{T-2}(j)) \cdot \\ &\quad \pi(i,j|pred_{T-1}(i), pred_{T-1}(j)) \qquad (3.7) \end{aligned}$$

Emphasizing here that we are considering *only* the cases where the models are based on the same tree $\mathbb{T}$ implies that only the probabilities vary between these close (in nested distance sense) models. In this case, we would rather use the notion of transportation *subkernels* instead of transportation subplans. For arbitrary nodes $k, l \in \mathcal{N}_t$ and $k \prec i$ the subkernel is the probability distribution in the set $l+(l \prec j)$ such that

$$K_t(j|i;k,l) \geq 0, \ \sum_{l \prec j} K_t(j|i;k,l) = 1, \ (\forall\ (i,j) \in \mathcal{N}_{t+1}\ k \prec i,\ l) \ \ where\ \ K_t(j|i;k,l) = \frac{\pi(i,j|k,l)}{\sum\limits_j \pi(i,j|k,l)},$$

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

$$\begin{aligned} \pi(i,j) &= K_1(pred_1(j)|pred_1(i);1,1) \cdots K_{T-2}(pred_{T-1}(j)|pred_{T-1}(i); pred_{T-2}(i), pred_{T-2}(j)) \cdot \\ &\quad K_{T-1}(j|i; pred_{T-1}(i), pred_{T-1}(j)) \times Q(i) \cdot Q(pred_{T-1}(i)) \times \cdots Q(pred_1(i)) \qquad (3.8) \end{aligned}$$

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

$$\begin{aligned} K(i,j) &= K_1 \circ \cdots \circ K_{T-1}(i,j) \\ &= K_1(pred_1(j)|pred_1(i);1,1) \cdots K_{T-2}(pred_{T-1}(j)|pred_{T-1}(i); pred_{T-2}(i), pred_{T-2}(j)) \\ &\quad \cdot K_{T-1}(j|i; pred_{T-1}(i), pred_{T-1}(j)). \qquad (3.9) \end{aligned}$$

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

$$\min_{x \in \mathbb{X}} \max_K \{\mathbb{E}_{K \circ P}[H(x,\xi)] \ s.t.\ K = K_1 \circ ... \circ K_{T-1}, \ \sum_{i,j \in \mathcal{N}_T} \mathsf{d}(i,j) \cdot K(i,j) \cdot P_i \leq \epsilon\}. \qquad (3.10)$$

It is noticeable that expression $\sum_{i,j\in\mathcal{N}_T} \mathsf{d}(i,j)K(i,j)P_i \leq \epsilon$ in (3.10) is multilinear in transportation subkernels $K_1, ..., K_{T-1}$. In applications for ambiguity problems P and Q are fixed and $\tilde{P}$ and $\tilde{Q}$, regarded as *worst tree* candidates, are constructed with the feasible corresponding transportation supkernels such that $\mathbb{d}(P, K \circ P) \leq \epsilon$. Optimization over these subkernels is done stage-by-stage repeatedly by optimizing (for fixed decisions) the subkernels at each stage. Algorithmically, this procedure is done by successive linear optimization. This Algorithm for multistage stochastic optimization problems has been implemented and the results are being analyzed in the subsequent sections.

4 Solution Algorithm - Successive Convex Programming

To begin with we consider only the general saddle point problem rather than any specific. In a continuous form of $f(x, y)$ where f is convex in $x \in \mathcal{X}$ and concave in $y \in \mathcal{Y}$, a saddle point solution (x^*, y^*) is a tool for decision makers to evaluate the computation of optimal response to the worst strategy. In such equlibria, neither the decision maker, nor the opponent would benefit by deviating from saddle point. Classical methods based on the gradient/subgradient for solving the saddle point problems have been of great interest since the seminal work of (Arrow et al., 1958). In classical setting, several algorithms have been proposed. Consider the unconstrained problem

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

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. A necessary and sufficient condition for a joint optimum is satisfied by $\zeta^* = (x^*, y^*)$ which solves the 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

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

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 (Rustem and Howe, 2002), more saddle point computation algorithms are presented and discussed. 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 Quasai-Newton algorithm for nonlinear systems.

In distributionally robust multiperiod stochastic setting, the algorithm 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 decisions' space $\mathbb{X}$ and models' space $\mathcal{P}_\epsilon$ in our setting, gradient based algorithms are avoided. In addition, direct coordinate wise approach i.e., moving in one step from decisions' space to models' space is subject to oscillation[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 variant is proposed and the proof of convergence is discussed.

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

In the following, the iterative manner of solution procedure is shown in Algorithm 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. It is noticed that *step 3.* itself encompasses the stage wise procedure of the constructing $\mathbb{P}(\mathbb{T}, \tilde{P})s$ discussed in 3.2.

Proposition 5. *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 minimax solution.*

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

[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.

Algorithm 1 **S**uccessive **C**onvex **P**rogramming

```
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:
   $\left\| \begin{array}{ll} \min & u \\ s.t & \mathbb{E}_{\mathbb{P}}[H(x,\xi)] \leq u \quad \textit{for all } \mathbb{P} \in \mathcal{P}_\epsilon^k \\ & x \in \mathbb{X}, \\ & x \lhd \mathfrak{F} \end{array} \right. \longmapsto (x^k, u^k)$

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

4. We 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. $u^{k+1} - u^k \geq \theta$, or
   • specify in advance the number of iterations $k$ (This means that the number of models included
     in set $\mathcal{P}_\epsilon$ to be determined at the beginning).
```

In the following section the proposed algorithm is implemented and computational results for a classical multiperiod production/inventory control problem are presented and discussed.

5 Implementation and Computational Results

5.1 Multiperiod Production/Inventory Control Problem

To picture the implications of our proposed algorithm, in this section a simplified multistage stochastic optimization problem - a multiperiod production/inventory control problem - is implemented and numerical results are shown. This example[4] is used to illustrate the multistage approach to stochastic modeling and its ambiguity extension.

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

[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.

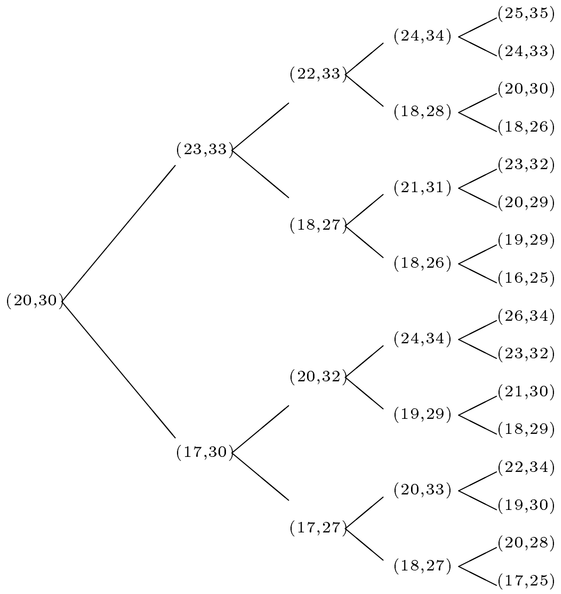

Figure 5.1: Demand requirements (product_1,product_2) and the binary tree structure

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 Figure 5.1, the tree structure and demand requirements of both products is depicted. In Figure 5.2, however, the demand scenarios of both products and the corresponding scenario probabilities are shown separately.

5.1.1 Mathematical Modeling Summary

In Table 1, the symbols defining the parameters, decisions and decision dependent variables of the model are introduced. The full mathematical model in nodal representation is formulated too. Note that decisions are only defined for emanating nodes and thus not for leaf (terminating) nodes.

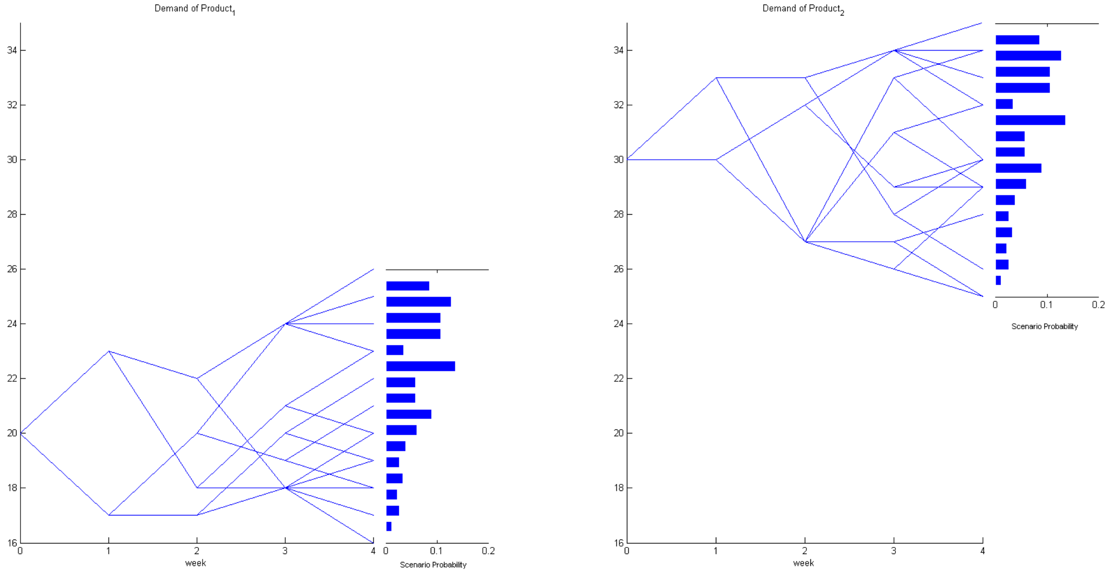

Figure 5.2: Demand scenarios of product$_1$ and product$_2$ and the corresponding scenario probabilities

Table 1: 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 $b = 1, 2$
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

$$
\begin{aligned}
\max \sum_n P(n)v(n) & & \forall n \in \mathcal{N} & \quad (5.1)\\
\text{subject to} \sum_b x_f^b(n_-) \le c & & \forall n \in \mathcal{N} \setminus \mathcal{N}_0 & \quad (a)\\
x_i^b(n_-) + x_f^b(n_-) + x_e^b(n) - d^b(n) = x_i^b(n) & & \forall n \in \mathcal{N} \setminus \mathcal{N}_0 & \quad (b)\\
\sum_b x_i^b(n) \le \bar{x}_i & & \forall n \in \mathcal{N} & \quad (c)\\
x_i^b(n_-) + x_e^b(n) \ge d^b(n) & & \forall n \in \mathcal{N} \setminus \mathcal{N}_0 & \quad (d)\\
\sum_b pr^b d^b(n) - \sum_b [pc^b x_f^b(n_-) + ic^b x_i^b(n) + ec^b x_e^b(n)] = v(n) & & \forall n \in \mathcal{N} \setminus \mathcal{N}_0 & \quad (e)\\
x_f^b \ge 0 & & &\\
x_i^b \ge 0 & & &\\
x_e^b \ge 0 & & &
\end{aligned}
$$

The objective of this 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. (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. (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 the cost parameters are presented in Table 2. 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.

Table 2: Parameters : Revenues, Costs and Capacities

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

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 Figure 5.3, 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 section 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 Figure 5.4, 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.

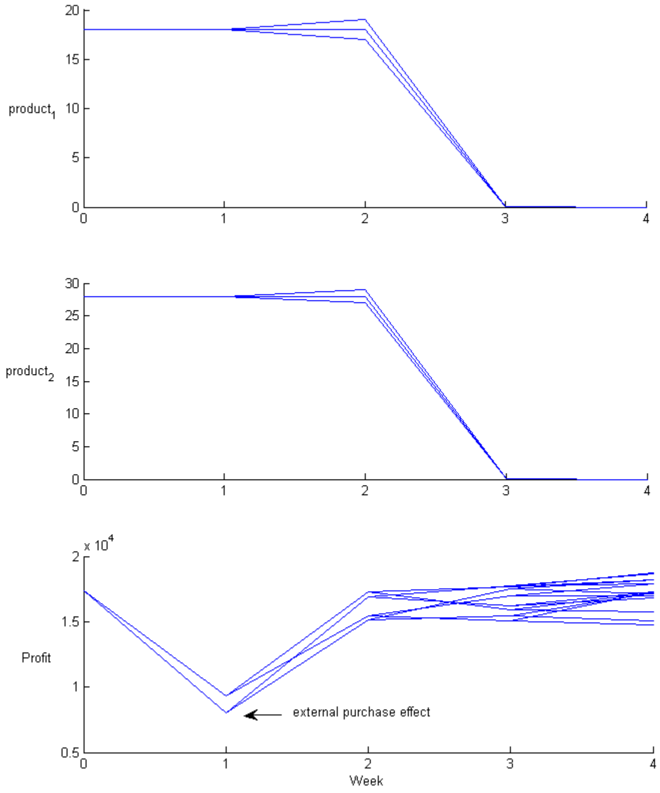

Figure 5.3: Optimal solution scenarios

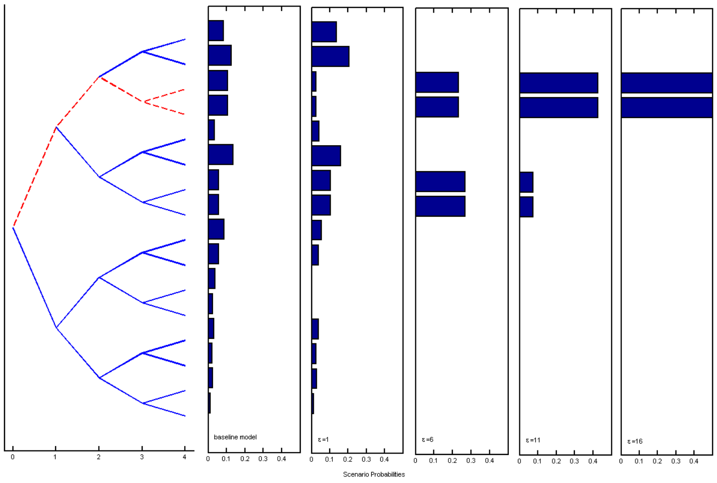

Figure 5.4: Tree structure of problem (5.1) and diminishing worst trees for increasing ambiguity radii

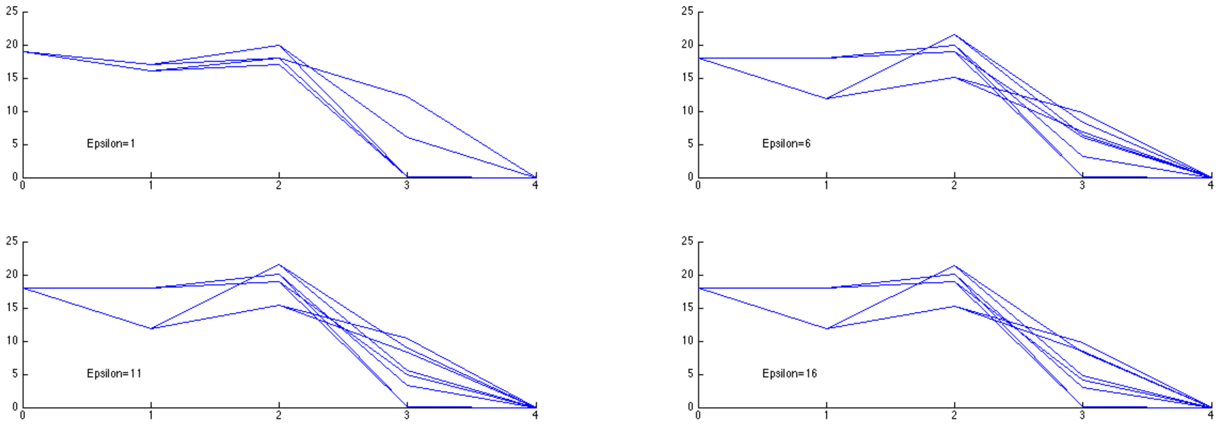

Figure 5.5: Sensitivity of decisions under model ambiguity-product$_1$

Maximin Solutions of the production Scenarios for different ambiguity radii

In Figures 5.5 and 5.6, 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, one direct effect is reflected in more diverse production scenarios which is observable by for product$_1$ by comparing the top graph in Figure 5.3 with Figure 5.5, and for product$_2$ by comparing the middle graph in Figure 5.3 with Figure 5.6. At first this results might seem quite controversial, since worst scenarios are getting simpler and simpler structure as epsilon increases, where the decision scenarios are revealed to be more complex. This might be seen as incorporating more models in the neighborhood of „best guess", those decisions are taken which are good for all included models. This phenomenon also has an impact on all decision dependent scenarios which in this example are the external purchase x_e^b and inventory level x_i^b for $b = 1, 2$ and consequently on profit scenarios.

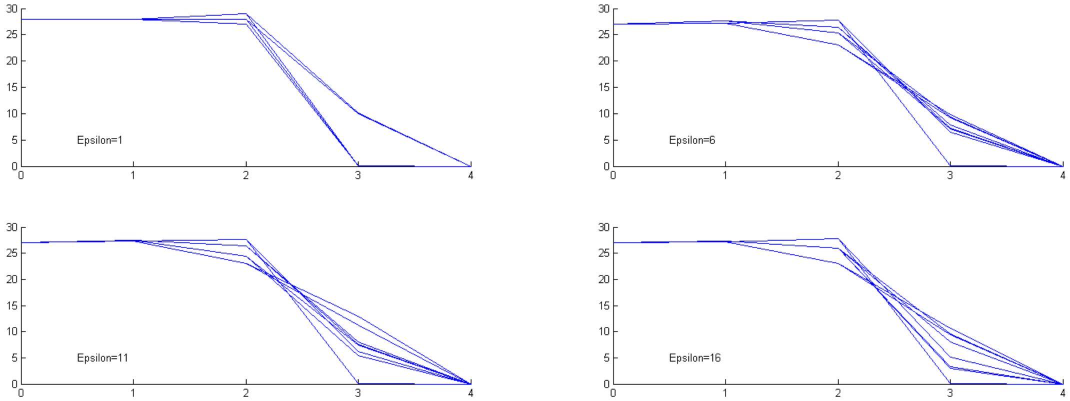

Figure 5.6: Sensitivity of decisions under model ambiguity-product$_2$

Price of Ambiguity

A fundamental result in maximin setting of this example is that at solution x^* we have the following inequality

$$\mathbb{E}_{\tilde{\mathbb{P}}*}[H(x,\xi)] \leq \mathbb{E}_{\tilde{\mathbb{P}}*}[H(x^*,\xi)] \leq \mathbb{E}_{\mathbb{P}}[H(x^*,\xi)] \tag{5.2}$$

This inequality indicates the robust nature of maximin equilibrium in presence of worst cases and , on the right side, ensures the improvement of the optimal solution if worst case does not occur, whereas on the left side, shows the potential deprivations if worst case is realized. In Figure 5.7(left), the numerical results of inequality (5.2) shows the optimal solution to the original problem, distributionally robust problem and worst case problem against the increasing ambiguity radii. It practically shows 1. the price that decision makers pay to obtain the robust solution is 4.71% decrease in expected net profit at the largest neighborhood and 2. how much do they gain by cautiously robustifying themselves against the worst odds. In this specific example, in addition to the increasing ambiguity radius, in Figure 5.7(right), the maximin solution incorporating up to 4 models at each ϵ-neighborhood is pictured. A closer look, shows from a specific radius, there is no more improvement in the maximin solution i.e., there is an upper bound for the *price* to *gain* the robust solution.

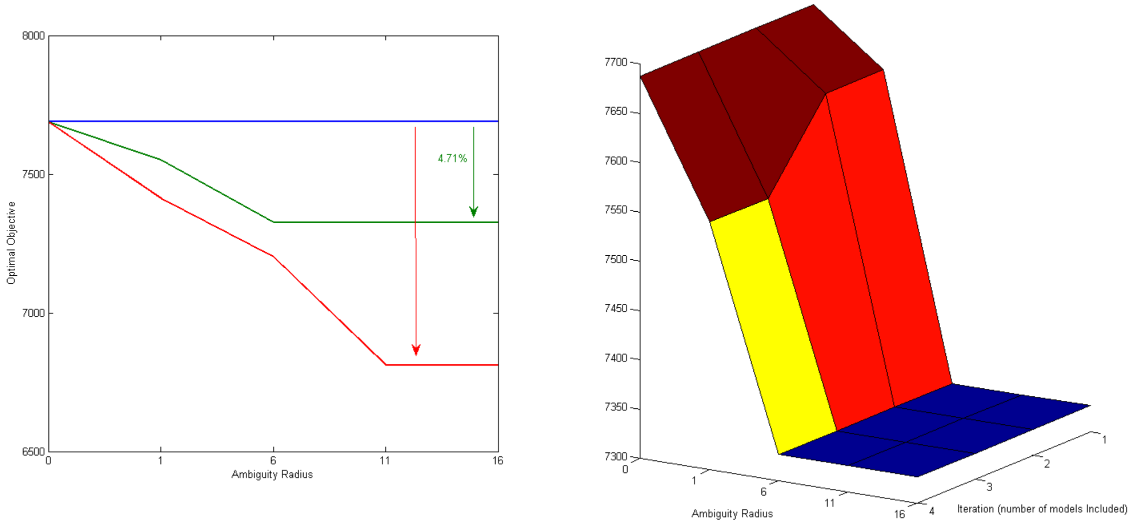

Figure 5.7: Left: The price to ***pay*** for ***gaining*** robust solution Right: Ambiguity surface

6 Concluding Remarks and further work

In this paper an algorithm to robustify a multistage stochastic optimization problem with ambiguity about the underlying probability model 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 in the underlying model, however the algorithm has the possibility for this extension subject to more technical complication. It can be expected that by high performance and parallel computing methods ambiguity problems for quite large trees can be solved.

The nested distance is a very new concept that appropriately incorporates the filtration structure in the multistage stochastic optimization models. In our approach, we considered minimax w.r.t worst case (a bunch of them) which is getting a simpler and simpler structure as epsilon increases (maybe at the largest radius even sits on a single scenario) . The decisions, however, shown to be more and more complex. The reason for this phenomena might be the inclusion of more models and decisions should be taken which are optimal for all models, hence it turns out that we have more bushy decision scenarios when the ambiguity radius gets larger and larger. Moreover, it is seen that there is a threshold for epsilon range, at which no improvements appear in decisions and objective function.

References

K. J. Arrow, L. Hurwicz, and H. Uzawa. *Studies in Linear and Non-Linear Programming.* Stanford University Press, 1958.

J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming.* Springer, 1997.

J. Bisschop. *AIMMS Optimization Modelling.* Paragon Decision Technology, 2012.

G. Calafiore. Ambiguous risk measures and optimal robust portfolios. *SIAM Journal of Control Optimization*, 18(3):853–877, 2007.

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

Y. M. Danilin and V. M. Panin. Methods for searching saddle points. *Kibernetika*, 3:119–124, 1974.

E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problem. *Operations Research*, 58:596–612, 2010.

V. F. Demynov and A. B. Pevnyi. Numerical methods for finding saddle points. *USSR Computational Mathematics and Mathematical Physics*, 12:1099–1127, 1972.

J. Dupačová. On minimax decision rule in stochastic linear programing. *Studies in Mathematical programming*, pages 47–60, 1980.

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

J. Dupačová. Stochastic programming: Minimax approach. *Encyclopedia of Optimization (Floudas, Ch. A. and Pardalos, P.M. eds)*, V:327–330, 2001.

J. Dupačová. Uncertainties in minimax stochastic programs. *Optimization*, 1:191–220, 2010.

D. Ellsberg. Risk, ambiguity and savage axioms. *The Quarterly Journal of Economics*, 75(4): 643–669, 1961.

K. Fan. Minimax theorems. *Proc. Nat. Acad. Sci.*, 39:42–47, 1953.

J. Goh and M. Sim. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58:902–917, 2010.

H. Heitsch, W. Römisch, and C. Strugarek. Stability of multistage stochastic programs. *SIAM Journal of Optimization*, 17:511–525, 2006.

R. Jagannathan. Minimax procedure for a class of linear programs under uncertainty. *Operations Research*, 25:173–177, 1977.

G. Ch. Pflug. Version-independence and nested distributions in multistage stochastic optimization. *SIAM Journal on Optimization*, 20(3):1406–1420, 2010.

G. Ch. Pflug and A Pichler. A distance for multistage stochastic optimization models. *SIAM Journal on Optimization*, 22(1):1–23, 2012.

G. Ch. Pflug and W. Römisch. *Modelling, Measuring and Managing Risk*. Singapore: World Scientific., 1 edition, 2007.

G. Ch. Pflug and D. Wozabal. Ambiguity in portfolio selection. *Quantitative Finance*, 7(4): 435–442, 2007.

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

S. Rachev and W. Römisch. Quantitative stability in stochastic programming: the method of probability metrics. *Mathematics of Operations Research*, 27:798–818, 2002.

W. Römisch and R. Schultz. Stability analysis for stochastic programs. *Annals of Operations Research*, 30:241–266, 1991.

W. Robinson and R. Wets. Stability in two stage stochastic programming. *SIAM Journal of Control Optimization*, 25:1409–1416, 1987.

B. Rustem and M. Howe. *Algorithms for Worst-Case Design and Applications to Risk Management.* Princeton University Press, 2002.

A. Ruszczynski and A. Shapiro. *Stochastic Programming, 1st ed.* ser. Handbooks in Operations Research and Management science Amsterdam, 2003.

H. Sasai. An interior penalty method for minimax for problems with constraints. *SIAM Journal of Control Optimization*, 12:643–649, 1974.

H. Scarf. *Studies in the mathematical Theory of an Inventory Problems.* Stanford Univeristy Press, 1958.

A. Shapiro and Sh. Ahmed. On a class of minimax stochastic programs. *SIAM Journal of Optimization*, 14:1237–1249, 2004.

A. Shapiro and A. Kleywegt. Minimax analysis of stochastic problems. *Optimization Methods and Software*, 17:523–542, 2002.

M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8:171–176, 1958.

A. Thiele. Robust stochastic programming with uncertain probabilities. *IMA Journal of Management Mathematics*, 19:289–321, 2008.

J. von Neumann. Zur theorie der gesellschaftsspiele. *Math. Ann.*, 100:295–320, 1928.

D. Wozabal. A framework for optimization under ambiguity. *Annals of Operations Research (online First)*, 2010.

J. Záacková a.k.a. Dupacová. On minimax solutions of stochastic linear programming problems. *Casopis pro Pestovani Mathematiky*, 91:423–430, 1966.

7 Appendix

7.1 The proof of Theorem 4.

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 Figure 7.1. 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 l-th 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 $\mathcal{C}(\mathbb{P}(\mathbb{T}, \cdot), \Lambda)$.

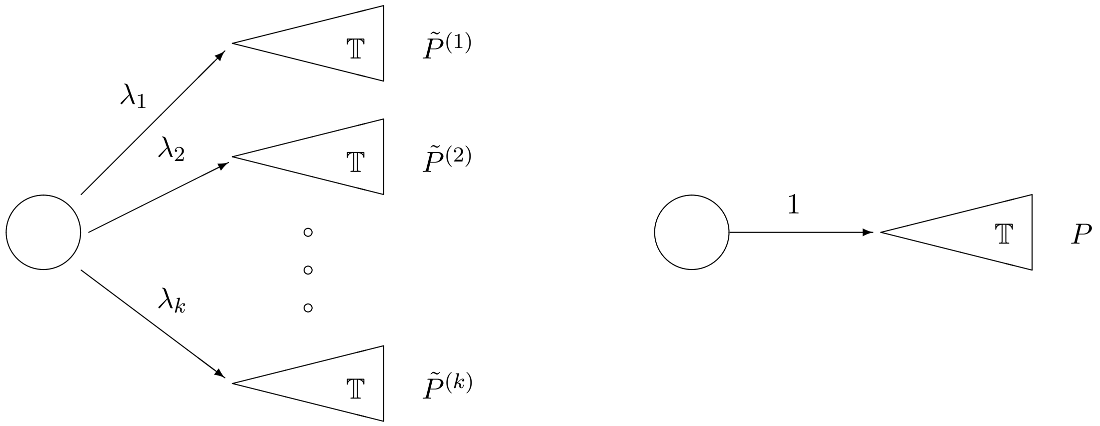

Figure 7.1: The compound convex structure of trees $\mathbb{P}(\mathbb{T}, \tilde{P}^{(l)})$ and augmented tree $\mathbb{P}_+(\mathbb{T}, P)$

The convex hull of the set

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

with

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

is the set

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

The convexified problem (3.2) is rewritten to

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

Notice that in the formulation (7.2) the decision variables x must coincide in all randomly sampled subproblems, cf. Figure 7.1. 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.,

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

Assume that for all $i \neq j$,there exist constants c,$C > 0$ such that $c \leq d(i,j) \leq C$. 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

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

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 $1 - \sum\limits_{i \in \mathcal{N}_T} \min(P_i, \tilde{P}_i)$ have to be transported, over distances which lie between c and C, whence the assertion follows. Notice well that the use of the distance $\left\| P - \tilde{P} \right\|$ is only to demonstrate compactness. While the topologies generated by the two metrics $\left\| P - \tilde{P} \right\|$ 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 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 $\mathbb{P}(\mathbb{T},\Lambda) = \sum\limits_{l=1}^{k} \lambda_l \mathbb{P}(\mathbb{T},\tilde{P}^{(l)})$. Then:

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

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 $\mathbb{d}(\mathbb{P}(\mathbb{T},\tilde{P}^*),\mathbb{P}(\mathbb{T},P)) \le \epsilon$ 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 5.

Here we prove the convergence of iterative procedure

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

Denote by $F^k = \max_{1\le l\le k} F(x^{k+1},\, \mathbb{P}^l)$, then $F^{k+1} = \max_{1\le l\le k+1} F(x^{k+2},\, \mathbb{P}^l)$ and by monotonicity $F^{k+1} \ge F^k$. 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 $F^* \le \max_{\tilde{\mathbb{P}}\in\mathcal{P}_\epsilon} F(x^*,\, \tilde{\mathbb{P}})$, 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 $\max_{1\le l\le k}\ F(x^+,\, \tilde{\mathbb{P}}^l) \ge \max_{1\le l\le k}\ F(x^{k+1},\, \tilde{\mathbb{P}}^l) = F^k$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.