Chi-Optimization
–Novel Approach for Optimization under Uncertainty with Application on Forecast- and Decision Problems–

Uwe Kuss
C-optimal AG Switzerland, ukuss@t-online.de

We propose a novel approach for optimization and decision problems under uncertainty. We first describe it for stochastic optimization under distributional ambiguity with and without data for the random parameter. Distributional ambiguity means that an entire family P of distributions is considered instead of a single one. For our approach, which avoids non-verifiable assumptions and improves upon the model accuracy, P represents the integration of all kind of information at hand being more or less uncertain. This leads to a generalization of the stochastic optimization under distributional ambiguity as well as of the statistical decision approach. When searching for a really suitable solution one has to assume and accept the uncertainty as the given situation in the reality with the consequence that the optimum cannot be achieved, but at best up to an inevitable tolerance or error term. Our approach considers the problem from a completely different point of view, compared to common stochastic optimization approaches under ambiguity, namely from this error or tolerance term. In this way, with the appropriate definition of tolerance, it succeeds in minimizing this term. The result for the Statistical Decision Theory is a convincing optimality property even for finite sample sizes, an important aspect for practical applications. A solution, named c-robust, that follows the same basic principles, is developed in situations where the tolerance becomes too large for the given application but the user can identify a maximum value for this tolerance.

Key words: stochastic optimization under uncertainty; stochastic programming with distributional ambiguity; statistical decision theory; c-optimality, estimation; stochastic optimization; data-driven optimization, statistical operations research
History: April 17, 2021

1. Introduction

At the beginning of his academic career in the seventies, U. Kuss has published about estimation theory as part of mathematical statistics, first of all two, [4] and [5], about Maximum Probability estimators (abbreviated MPe's) of L. Weiss and J. Wolfowitz [11]. Independent of its asymptotic efficiency property, a good estimator like MPe's is an effective condensation of the data, especially if it is a sufficient statistic. An important advantage of MPe's is the possible control of these by a loss function with the result that their limiting risk is minimal.

Because often there exists additional information, which can be used for a better solution of the given problem of statistical estimation theory as described in [11] and [4], U. Kuss presented in his doctoral thesis [7] in 1971 properties of MPe's for finite sample size and proposed a new kind of Bayesian like optimality property, if a set of prior distributions is considered as an information additional to the data. In this thesis, it is shown that the MPe's have not only an asymptotic optimality property under some assumptions, but can also achieve good results for finite sample size, roughly spoken Bayesian like, if they are adapted in certain cases (a paper about small sample size properties of MPe's is also [10] by L. Weiss).

In the following years U. Kuss developed a novel approach in statistical decision theory for finite sample size and with respect to applications, see [6]. In 1979, he meet Jacob Wolfowitz at Tampa, Florida. Professor Wolfowitz offered to support and promote the new approach and method, but unlucky circumstances, and first of all his too early and very tragic death, unfortunately prevented any cooperation.

A view years later, after moving from university to industry (in 1981), U. Kuss applied this new approach and the corresponding methods with real success, at most by adapting it to the given practical situation, but more and more by changing also the concept in some important points. This further development resulted from his experience with the application of the new methods.

In the present paper, we will describe this novel approach with its basic ideas and formulate it, for better understanding, in a more simple manner, i.e., in the main by assuming that the set of possible distributions is finite. Mathematically, this is a strong assumption, but does not pose a real restriction for applications and is commonly encountered in optimization theory as well as in applied mathematical statistics. Moreover the most results can be generalized.

In practice, we typically find information additional to the data, e.g., results of market research, forecasted values by the coworkers of the distribution department, projection of representative orders. The summary or, with other words, integration of all these more or less uncertain information can be presented by a set of prior-distributions resp. as the corresponding set of posterior-distributions, if data are at hand. Why and how to come to this integration and presentation is outside the scope of this paper, but we already mention here, that the method of Stochastic Partial Information will be applied as the main mathematical method (see [3]).

Because we take a set of (prior-)distributions instead of a single one as basis of our approach, we achieve a high flexibility for the modeling of any situation as well as for respecting and integrating all kinds of information, including fuzzy or uncertain. This improves also the application to other sciences.

Then searching for a really suitable solution one has to assume and accept the uncertainty as the given situation in the reality with the consequence that the optimum can be achieved only with an inevitable tolerance or error term. Our novel approach considers the problem of optimization and decision under uncertainty from a completely different point of view, namely from this error or tolerance term. In this way, with the appropriate definition of tolerance, our approach succeeds in minimizing this term. Thereby it realizes a compromise between Bayes and Minimax based on the goal of minimizing the error term. Therefore the solution is more or less like Bayesian or Minimax, dependent on the given situation.

As the basis for the presentation of our new approach, we generalize in Sect. 2 the model of stochastic optimization under uncertainty in order to exhaust all kind of information included the data, the new kind in the main by the definition of Partial Information (Subsect. 2.1). An important addition is in Subsect. 2.2 the appropriate change of our approach in the event that the real average loss deviates significantly from the expected value of the loss, explained by means of a real and simple example.

Then, in the next section, we develop the novel approach called chi-optimal as the solution under uncertainty (Subsect. 3.1). The main result is Theorem 1, which proves the minimization of the tolerance term defined as the inevitable error of optimization, really inevitable because of the uncertainty. Our approach is also able to solve a kernel problem of the statistical decision theory also by minimizing the tolerance, therefore we apply it also to this theory and the estimation problem interpreted as a special case (Subsect. 3.2). The section is completed by three examples, the first gives the solution for the insurance example of Subsect. 2.2 and the following two are practical, thereby the third shows that the chi-optimal solution is different from the robust solution in optimization or Minimax in decision theory as well as from the Bayesian (Subsect. 3.3).

In Sect. 4, we consider the problem if the tolerance (i.e., the inevitable error term) becomes too large for the given application but the user can fix a maximum value for the tolerance. The solution in this situation named c-robust has the property to maximize the size of the admitted PI. As an important basis, we can characterize by topological and analytic methods the size of any PI by only one parameter in an suitable manner. The properties are analogous to that of chi-optimal decisions under the given condition of fixed maximal tolerance (Subsect. 4.2). The next and last subsection completes this section with the proven compatibility of c-robustness with chi-optimization.

Sect. 5, as the last section, rounds off this paper with remarks about asymptotic properties of chi-optimality and the result that the tolerance term of the chi-optimal solution converges to zero under week assumptions. The asymptotic result is based on the corresponding result of c-optimality in the paper [6].

2. Generalized Approach of Optimization under Uncertainty

2.1. Generalized approach of stochastic optimization under uncertainty

In order to formulate the problem to be solved, we start with the Stochastic Optimization Problem (SOP)

$$(\mathcal{M}0) \qquad \min \ \mathbb{E}_P\big[L(a,\xi)\big] \tag{1}$$

$$\text{s.t.} \ \ a \in \mathbb{A} := \{a \in \mathbb{R}^m : \mathrm{G}(a) \le 0\}, \tag{2}$$

where ξ is an l-dimensional random vector with values into $\Xi \subset \mathbb{R}^l$ and its (probability) distribution P. $\mathbb{E}_P$ denotes the expected value with respect to P. Thereby we set, instead of the original loss function l or profit function p, the *standard loss* L, defined for any feasible decision a and value b of the random parameter vector ξ by

$$L(a,b) := \begin{cases} (l(a,b) - l_{inf})/(l_{sup} - l_{inf}) \ , & \text{if a loss function } l(\cdot,\cdot) \text{ is given} \\ (p_{sup} - p(a,b))/(p_{sup} - p_{inf}) \ , & \text{if a profit function } p(\cdot,\cdot) \text{ is given} \end{cases} , \tag{3}$$

where

$$l_{inf} := \inf_{a' \in \mathbb{A}, b' \in \Xi} l(a',b') \ \text{ and } \ l_{sup} := \sup_{a' \in \mathbb{A}, b' \in \Xi} l(a',b')$$

and p_{inf} resp. p_{sup} are defined accordingly.

We assume, that the suprema and infima in (3) exist and are positive (and, for every a $\in \mathbb{A}$, L(a, .) is measurable and the expected value in (1) exists). This standardization ensures that any solution remains the same for any linear function of the (original) loss or profit function, if the factor is positive. An additional sense of it will be shown and clarified below. *But it is clear that the solutions of the original SOP with original loss or profit function remain the same by this standardization.*

The SOP is called stochastic, because the l-dimensional parameter vector ξ is a random variable, i.e., mapping from Ω into $\mathbb{R}^l$, measurable w.r.t. the σ-algebra $\mathcal{F}$ of a probability space $(\Omega, \mathcal{F}, P)$.

Similar to optimization under distributional ambiguity, we do not assume the knowledge of the underlying probability measure P generating the distribution P of the random parameter vector ξ. We assume only, that P is element of a set $\mathcal{P}$ of probability distributions, maybe the set of all. Thus, we admit also uncertainty and obtain a realistic as well as a more interesting approach. Then we formulate the SOP

$$(\mathcal{M}1) \qquad \min_{a \in \mathbb{A}} \ \mathbb{E}_P\big[L(a,\xi)\big] \qquad P \in \mathcal{P} \tag{4}$$

assuming the existence of the expected value for every P $\in \mathcal{P}$ and a $\in \mathbb{A}$. This kind of generalization of the SOP ($\mathcal{M}0$) not seem to be meaningful, but it is the correct description of the given problem in applications as well as in the reality so a suitable new approach is needed.

In this paper, we want to present this new approach. We assume that the set $\mathcal{P}$ of distributions is finite. Mathematically, this is a strong assumption, but does not pose a real restriction and is commonly encountered in optimization theory as well as in applied mathematical statistics. Moreover the most results can be generalized.

This allows us to write

$$P \in \mathcal{P} = \{P_0, \dots, P_m\}$$

for some finite number $m \in \mathbb{N}$.

Assumption 1. *We assume that $\mathbb{A}$ is compact and that the expectation of L, as a function of a, is continuous for every $P \in \mathcal{P}$.*

The result of any method can only became best if one respect all kind of information, also partial or uncertain information about the set $\mathcal{P}$ of distributions itself, e.g., one knows that a subset of $\mathcal{P}$ is more probable than the complement. The right formulation for any such information is a set of probability distributions on $\mathcal{P}$, usually called *a-priori-distributions.* Because $\mathcal{P}$ is finite, any probability distribution S on it is the element s of the m-dimensional standard simplex SI (therefore often named as probability simplex), formulated as follows:

$$s_i = S(P_i) \qquad \forall i = 0, ..., m, \tag{5}$$

Therefore every such information can be formulated as a subset V of the m-dimensional standard simplex SI.

If data or, mathematically spoken, a random vector $X = (X_1, ..., X_n)$ of n observations for the random l-dimensional parameter vector ξ with value $x = (x_{11}, ..., x_{1l}, ..., x_{n1} ..., x_{nl})$ and n×l - dimensional distribution function F_X is given, we take instead of any distribution s $\in V$ the corresponding conditional distribution $s(x)$, given $X = x$, i.e., for any $i = 0, ..., m$ instead of s_i we set

$$s_i(x) = Prob(P_i | X = x), \tag{6}$$

where Prob is the abbreviation for probability and means the common distribution of S and F_Y.

For the case of discrete random variable ξ the calculation according to the Bayes theorem is simple.

If l = 1 and the distribution $P \in \mathcal{P}$ of ξ has the density $f(.|P)$ with respect to the Lebesgue-measure and the observations $X_1, ..., X_n$ are independent and identically distributed, that reads

$$s_i(x) = \left\{ \frac{g_i(x) s_i}{\sum_{j=0}^{m} g_j(x) s_j} \right.$$

where

$$g_i(x) = \prod_{j=1}^{n} f(x_j | P_i)$$

One applies the conditional distribution, because it includes the additional information by the data, it is also more "concentrated" leading better results for all approaches (see the example in appendix 6.1).

But any of the conditional distributions is also an element of the m-dimensional standard simplex SI, and then every such information can be represented by a subset V of SI also in the situation with data.

Therefore we can treat, in the rest of this paper, both situations simultaneously by proposing that $s = (s_0, ..., s_m) \in V$ is calculated according to equation (6), if data are at hand.

Then we generalize the model ($\mathcal{M}1$) by the following definition and call this generalization ($\mathcal{M}2$).

Definition 1 (Partial information, PI). For any $s \in [0,+\infty[^{m+1}$, we set

$$r(a,s) := \sum_{i=0}^{m} EL(a,P_i)\, s_i \tag{7}$$

with $EL(a,P_i) := \mathbb{E}_{P_i}\big[L(a,\xi)\big]$ for abbreviation. (More formally, r(a,s) is the Bayes-risk for any s as prior resp. posterior-distribution.)

In the context of the SOP resp. $(\mathcal{M}1)$, a (Borel-) measurable subset V of the m-dimensional standard simplex SI is called **partial information (PI)**, if the SOP is as follows:

$$(\mathcal{M}2) \qquad \min_{a\in\mathbb{A}}\ r(a,s) \tag{8}$$

with $s \in V$.

Remark 1. The set of all δ-distributions on $\mathcal{P}$ is a PI. Then, $(\mathcal{M}2)$ is equivalent to $(\mathcal{M}1)$. By this, $(\mathcal{M}1)$ is a special case of $(\mathcal{M}2)$.

Proof. For any i $\in \{0,\dots,m\}$ the corresponding δ-distribution δ_i is defined by

$$\delta_i(P_j) := \begin{cases} 1, \text{ if } j = i \\ 0, \text{ o/w} \end{cases} \tag{9}$$

The set of all δ-distributions is a finite subset of the simplex SI and therefore measurable. □

Now we have formulated a flexible and general model $(\mathcal{M}2)$.

In the literature about stochastic optimization under uncertainty, mainly two research strategies have been proposed to transform the SOP (M1) or (M2) in a mathematically well-defined model: the distributionally robust optimization approaches and the Bayesian approach. Distributionally robust optimization is a minimax approach, where against the worst-case is optimized (see [12]), i.e., $(\mathcal{M}1)$ reads

$$(\mathcal{M}1^{DR}) \qquad \min_{a\in\mathbb{A}}\ \max_{0\le i\le m} EL(a,P_i).$$

In contrast, the Bayesian approach optimizes against the (weighted) average of all distributions (see [1]), i.e.,

$$(\mathcal{M}1^{B}) \qquad \min_{a\in\mathbb{A}}\ \sum_{i=0}^{m} EL(a,P_i)\, w_i.$$

But thereby, the weights or, with other words, the a-prior-distribution must be known. Otherwise, one set usually all weights equal = 1.

2.2. important remark about the expected and the real loss

We want to explain the problem by means of a real and simple example.

Example 1. Someone has to decide whether to buy fire insurance for his own home whose value is \$800,000. The best deal includes an annual payment of \$80, with a maturity of 5 years. The expected value of any damage by fire, i.e., the risk value, can be estimated very well by the insurance company and it is sure, that the \$800 insurance premium is at least two and a half times the risk value. Therefore it is optimal for SOP $(\mathcal{M}1)$ to buy no insurance. But, if a fire resulting in a total loss of his house will occur during this 5 years, the real loss will be \$800,000. Of course, this formally optimal decision is wrong.

What is the reason for this paradoxical result? The expectation of loss is only a good approximation for the real mean of losses, if one considers many decision resp. optimization situations simultaneously as the insurance company does it. Then the sum of losses is minimized by an optimal solution approximately, because optimization of the mean of these results approximately the same, of course. The real average loss is obtained if one calculates the expected value of loss substituting the probabilities by the relative frequencies. In our example as relatively rare case, the real relative frequencies of the random events can be extremely different from its theoretical probabilities and therefore the real loss quite different from the expectation of loss. Clearly, in any similar situation, this paradoxical result can be found.

Our approach gives the possibility to reach the right solution also in such a situation. (Here, of course, all sets are finite and therefore ξ discrete.) We take the set Fr of possible relative frequencies for the values $(x_1,...,x_n)$ of ξ instead of the set $\mathcal{P}$ of probabilities and replace in this way the expected loss EL in $(\mathcal{M}1)$ resp. in $(\mathcal{M}2)$ with the mean value of the losses. We denote this mean value of losses substituting the expected standard loss EL by M and we calculate it for any a $\in \mathbb{A}$ and any fr $\in Fr$ as follows:

$$M(a,fr) := \sum_{i=0}^{n} l(a,x_i)\, fr(x_i), \tag{10}$$

where $\mathbb{A} := \{0,1\}$ and "1" indicates the decision for the insurance and "0" the alternative, i.e., a is the number of insurances actually. We remark that we consider in this example the original loss instead of the standard loss for more clearness (see the definition in (3)). As PI one can only choose the set of all δ-distributions (then, $(\mathcal{M}2)$ is equivalent to $(\mathcal{M}1)$).

We can propose that at most one times a fire with total loss of the home is possible, because the investigations and the reconstruction of the house take at least 5 years. Let this value of ξ be denoted by x_1 for simplicity, i.e., the (random) event $\{\xi = x_1\}$ means a fire with total loss of the ho me in these 5 years. Clearly for any element, say fr_1, of Fr holds $fr_1(x_1) = 1$ and then we calculate

$$M(0,fr_1) = 800000 \; and \; M(1,fr_1) = 400.$$

Now, the optimal solution of the SOP is $a = 1$ for almost all decision principles, i.e., the decision for the insurance. For the minimax decision principle is this clear. For a Bayesian like solution is to assume that the standard prior distribution is able defined in this case where relative frequencies instead of probabilities are given. In the next section, more precisely in Subsect. 3.3.1, we will see that this holds also for the χ-optimal decision.

But a method with mathematically exact procedures and algorithms for the problem as described will be subject of any following paper.

3. Chi-Optimality

3.1. Chi-optimization

We want accept the uncertainty as the given situation in reality, because this is the presupposition for a really good solution in practice. But our approach starts, simply spoken, with a completely different view of the problem. We see as the consequent next step the acceptance of an inevitable tolerance or error term in $(\mathcal{M}2)$, which is, of course, the consequence of such a realistic approach. The optimum can be reached only up to a tolerance δ, i.e., for any $a_0 \in A$, we can reach only that

$$r(a_0,s) \le \min_{a\in\mathbb{A}} r(a,s) + \delta \qquad \forall a \in \mathbb{A} \;\wedge\; \forall s \in V \tag{11}$$

holds for a certain $\delta \ge 0$. Thereby, V is the PI given by any application and the functional r defined in (7).

Remark 2. For any $a_0 \in \mathbb{A}$ inequality (11) holds also for the convex and closed hull $\mathbb{V}$ of V. Therefore, we can assume for the remainder of this paper, that any (given) PI is a convex and compact subset $\mathbb{V}$ of the m-dimensional standard simplex.

Proof. $r(a_0, s)$ is a linear and therefore continuous function of s and $\mathbb{V}$ is compact, because every PI is bounded. □

Preparing our new solution we formulate the following definition:

Definition 2 (Tolerance function). For any $a_0 \in \mathbb{A}$, the minimal value of the tolerance term δ (being greater than 0 in almost all situations) is a function of $a_0 \in \mathbb{A}$, defined by

$$T(a_0) = \max_{a \in \mathbb{A}} \max_{s \in \mathbb{V}} \big(r(a_0, s) - r(a, s)\big), \tag{12}$$

T(a_0) is called **tolerance** of a_0 and T **tolerance function**, for the given PI $\mathbb{V}$ (see definition 1). The maximum in (12) exists for all $a_0 \in \mathbb{A}$, because $r(a_0, s)$, as function of s, is linear and therefore continuous and also bounded (because of assumption 1 and the definition of L in (3)) and $\mathbb{V}$ is compact. The maximum over a $\in \mathbb{A}$ exists because of the compactness of $\mathbb{A}$.

For any $a_0 \in \mathbb{A}$, its tolerance value T(a_0) is the inevitable deviation, really inevitable, as the true s resp. the true distribution P is unknown. Therefore the best possible solution is that decision which minimizes the tolerance, i.e., the (best possible) solution of ($\mathcal{M}2$) is a decision a_χ with the property

$$T(a_\chi) \leq T(a) \qquad \forall a \in \mathbb{A}. \tag{13}$$

Definition 3 (chi-optimal solution). We set $\chi := T(a_\chi)$ and call a_χ the **chi-optimal solution** of ($\mathcal{M}2$) for the given PI $\mathbb{V}$.

Remark 3. Let $\min_{a \in \mathbb{A}} r(a, s)$ (as a function of s) be constant, Then the chi-optimal solution fulfills the minimax criterion. If $\mathbb{V}$ is the entire simplex SI, then it fulfills the minimax criterion in the usual 'classical' sense.

Proof. The assertion is a consequence of (12), if one exchange the both maxima.

If $\mathbb{V}$ is the entire simplex SI, the set of δ-distributions is a subset of it and, for this subset, (12) is the minimax criterion in the usual 'classical' sense, if one notice, that the entire simplex is the convex hull of this subset. □

The situation in remark 3 is really extraordinary. Normally there is a fundamental difference between the chi-optimal and the minimax (or robust) approach. But the chi-optimal is also different from the Bayesian. Moreover it produces an optimal compromise between both of these well known approaches. Example 3 below shows that the chi-optimal solution is different from these both.

The following theorem shows that the chi-optimal decision minimizes the risk (defined as the expected value of the loss) for the entire PI $\mathbb{V}$ up to the smallest possible $\delta = \chi$.

Theorem 1. ***T(a_χ) is the smallest tolerance term***, i.e., if any number ϵ with any a(ϵ) $\in \mathbb{A}$ has the property

$$r(a(\epsilon), s) \leq \; r(a, s) \; + \; \epsilon \qquad \forall a \in \mathbb{A} \;\wedge\; \forall s \in V, \tag{14}$$

then $\epsilon \geq T(a_\chi)$ ($= \chi$) holds.

Proof. As a consequence of (14) we obtain

$$\max\big\{r(a(\epsilon), s) - r(a, s) \,|\, a \in \mathbb{A} \;\wedge\; s \in V\big\} = \; r(a(\epsilon), s_a) - r(d_a, s_a) \;\leq\; \epsilon, \tag{15}$$

where (d_a, s_a) is a maximum point of $\big(r(a(\epsilon), s) - r(a, s)\big)$. Therefore, according to the definition of tolerance function,

$$T(a(\epsilon)) \;\leq\; \epsilon.$$

But

$$T(a_\chi) \;\leq\; T(a(\epsilon))$$

holds according to the definition of chi-optimality and (13).

□

3.2. Application to statistical decision and estimation theory

3.2.1. Statistical decision theory

The new approach described in the previous chapters can be formulated also as a generalization of the usual model of statistical decision theory as follows:

- The random vector $X = (X_1, ..., X_n)$ consists of the n observations, their realizations are the data $x = (x_{11}, ..., x_{1l}, ..., x_{n1} ..., x_{nl})$
- The elements of $\mathcal{P}$ are the "stats of nature"
- $\mathbb{A}^*$ is the set of decisions, but more general a compact subset of $\mathbb{R}^m$ (like $\mathbb{A}$) or any finite set.
- $L^*(a, P) := \mathbb{E}_P\big(L(a, \xi)\big)$ is the value of the loss function L^* for any a $\in \mathbb{A}^*$ and P $\in \mathcal{P}$.

The definition of partial information (see in Sect. 2) remains the same, but with r^* instead of r defined as

$$r^*(a, s) = \sum_{i=0}^{m} L^*(a, P_i)\, s_i(x), \tag{16}$$

for any a $\in \mathbb{A}^*$ and any s $\in [0, +\infty[^{m+1}$, the definition of $s_i(x)$ is given in equation (6). The same holds for the definition of the tolerance function $T^* = T^*(a)$ like

$$T^*(a) = \max_{d \in \mathbb{A}^*} \; \max_{s \in \mathbb{V}} \big(r^*(a, s) - r^*(d, s)\big) \tag{17}$$

for any a $\in \mathbb{A}^*$.

Now, analogous to the definition of the chi-optimal solution, we obtain

Definition 4 (chi-optimal decision). If any decision a_χ has the property

$$T^*(a_\chi) \leq T^*(a) \qquad \forall a \in \mathbb{A}^*, \tag{18}$$

then we set $\chi := T^*(a_\chi)$ and call a_χ **chi-optimal decision**.

The application to the statistical decision theory is first of all a generalization of the Bayesian approach, but also of the classical. Of course we have also an optimality property corresponding to that for the chi-optimal solution in Theorem 1 formulated in the following

Corollary 1. $T^*(a_\chi$***) is the smallest tolerance term***, i.e., if any non negative real number δ with any $d_1 \in \mathbb{A}$ has the property

$$r^*(d_1, s) \leq \; r^*(a, s) \; + \; \delta \qquad \forall a \in \mathbb{A} \;\wedge\; \forall s \in V, \tag{19}$$

then $\delta \geq T^*(a_\chi)$ $(= \chi)$ holds.

Proof. The proof is also analogous to that of Theorem 1 as follows: As a consequence of (19) we obtain

$$\max\left\{r^*(d_1,s)-r^*(a,s)\mid a\in\mathbb{A}^*\ \wedge\ s\in V\right\}=\ r^*(d_1,s_a)-r^*(d_a,s_a)\ \leq\ \delta, \tag{20}$$

where (d_a,s_a) is the maximum point of $\left(r^*(d_1,s)-r^*(a,s)\right)$. Therefore, according to the definition of tolerance function,

$$T^*(d_1)\ \leq\ \delta.$$

But

$$T^*(a_\chi)\ \leq\ T^*(d_1)$$

holds according to the definition of chi-optimal decision and (18). The last both inequalities together imply the assertion.

□

3.2.2. Statistical estimation theory

As usual, we obtain the (statistical) estimation problem as a special case of decision theory by defining the set $\mathbb{A}^*=\mathcal{P}$. Often one choose then the loss function as a function of the distance between a $\in\mathbb{A}^*$ and P $\in\mathcal{P}$.

Therefore everything in the previous part of this subsection holds analogously here.

3.3. Three examples

3.3.1. Insurance example

We now continue with the insurance example in Subsection 2.2. We denote the tolerance function for M by T_M and define it analogous to that for r (see its definition and (12)) for any $a_0\in\mathbb{A}=\{0,1\}$ as follows:

$$T_M(a_0)=\max_{a\in\{0,1\}}\ \max_{fr\in Fr}\left(M(a_0,fr)-M(a,fr)\right), \tag{21}$$

Now we can calculate the values of T by

$$T_M(0)=800000-400=799600, T_M(1)=400-0=400 \tag{22}$$

In (22), the value zero in the calculation of $T_M(1)$ comes from the decision 0 (no insurance) and the event, that no fire happens, resulting loss zero, because the relative frequency of this event then does not matter. Equation (22) shows that the decision 1, i.e., the decision for the insurance, is χ-optimal being that decision with the minimal tolerance.

3.3.2. Two examples from practice

Now we discuss two practical examples which may illustrate and motivate our new approach additionally.

Example 2. **The Disposition of an Industrial Product**

E1. Main Problem of PPC

In most situations of Production Planning and Control (PPC), the required parameters are not always available, moreover one must use estimated resp. forecasted values instead of data. We consider the PPC of a certain industrial product as an example, in order to develop and declare our new chi-optimal method. Because we want to describe and clarify the core of the problem de facto for the general case, we simplify the underlying situation of real practice a little.

We chose the week as time interval and assume that we are in the 25th calendar week (CW) and that the entire production process needs, e.g., 20 weeks. Now the main task of disposition is to decide the number of pieces to be produced in the next 20 weeks, optimally identical with the total $Or(45)$ of all orders with delivery date 45th CW. But one knows in the present 25th week only a (at most small) part of this total, the rest must be estimated resp. forecasted. Because it is only possible to produce (production) units instead of any exact number of pieces, one has to decide about the number of units (e.g., a unit contains 25 000 pieces).

As the basis we take the total $Or_{-1}(45+corr)$ of of all orders with delivery date (45+corr). CW in the last year, where $corr$ is the correction of a possible shift of the season, if one have any. Typically no correction is needed for the 45th CW. Then we multiply it by the factor Dem estimating the change in demand from the last to this year obtaining

$$T := Or_{-1}(45+corr) * Dem$$

as the forecast of $Or(45)$.

Even having a very good forecast T by this procedure, there exists a deviation of it from the real value $Or(45)$ which is known only 20 weeks later. This deviation $Or(45) - T$, measured in units of course, is a random variable ξ and its distribution is not exactly known, one can, under some assumptions, only approximate it asymptotically. Since T is the best forecast, we choose as decision variable a the correction to T, i.e., $(T+a)$ as the number of units decided to be produced (in the following 20 weeks).

Respecting the long term production plan and because of technological reasons the maximum number of units correcting T is 38, therefore the restriction is

$$|a| \leq 38.$$

and therefore

$$\mathbb{A} = \{-38, -37, \ldots, 37, 38\}.$$

The set of possible distributions of the deviation ξ may be given by the following Table 1.

Difference (quantity)	lower bound of the probability of this difference	upper bound of the probability of this difference
+50	0.30%	0.41%
−50	0.28%	0.40%
+49	0.70%	0.83%
−49	0.48%	0.51%
⋮	⋮	⋮
±0	0.81%	0.98%

Table 1 Distributions of the deviation as typical in practice

E2. Analysis of Deviations

An alternative method is the empirical analysis of the differences observed over a longer time period. For that, let us consider any difference, say +34, i.e., if the real value is 34 units greater than the estimation, then the corresponding probability is to be replaced by the relative frequency of a difference equal to +34. But for the optimization we need an interval for the relative frequency of the future period. Therefore we chose the maximal and the minimal value of all observed relative frequencies.

Thus, respecting the interval of probabilities (see E1), one increases the lower and upper bounds for the relative frequency so that the future relative frequencies lies into this new interval almost surely. Then one obtains the same table as Table 1, but with relative frequencies instead of probabilities. For simplicity, we call each of these also probability.

Here we can see that, in practice, the bounds of the relative frequencies cannot be fixed exactly. But the same holds actually also for these of the probabilities, because these bounds are determined as that of a confidence interval to an appropriate confidence level.

E3. PI and Loss Function for this Example

The PI $\mathbb{V}$ is then calculated by using the data of Table 1.

Let

$$\begin{array}{rl} p_1 & \text{be the probability of a difference of } -50 \\ p_2 & \text{be the probability of a difference of } -49 \\ \vdots & \\ p_{101} & \text{be the probability of a difference of } +50. \end{array}$$

Then, the PI is the collection of all 101-tuples $(p_1, \ldots, p_{101})$, whose coordinates are between a lower and upper bound and whose sum is equal to 1 (100%). This yields a subset $\mathbb{V}$ of the 100-dimensional standard simplex, a PI according to the definition of PI, it is convex and compact (see Remark 2).

Now we determine the loss function. Assuming a loss of 73.5 $ per unit in case of under production and otherwise a loss of 49.0 $ per unit, the loss function is given (for any a $\in \mathbb{A}$) by

$$L(a,\xi) = \begin{cases} 49.0 \cdot (a-\xi), \text{ if } a > \xi \\ 73.5 \cdot (\xi - a), \text{ o/w} \end{cases}$$

and for any a $\in \mathbb{A}$ and s $\in \mathbb{V}$

$$r(a,s) = 49.0 \cdot \sum_{i=1}^{a+50} (a+51-i)s_i + 73.5 \cdot \sum_{i=a+51}^{101} (i-a-51)s_i$$

Now this is a completely formulated example of (M2) and any partial information.

Example 3. **The Disposition of an Industrial Product, simplified and fully calculated**

The data set of example 2 is a realistic one. But we consider a more simple example for better understanding and to make it easier for the calculations. The PI may be given now by Table 2 substituting Table 1, every else remains the same.

Now determine not only the χ-optimal solution resp. decision but also the Bayes and Minimax in order to show that the chi-optimal solution is different from these two.

In the next section we present an additional approach for a special situation which is also of practical importance.

Difference (quantity)	lower bound of the probability of this difference	upper bound of the probability of this difference
-2	08.21%	09.19%
-1	28.00%	29.12%
± 0	00.00%	51.30%
$+1$	00.00%	10.58%
$+2$	26.15%	35.73%
$+3$	00.55%	01.17%

Table 2 Distributions of the Deviation, a little simplified Example

decision (correction)	tolerance of this decision	criterion
-2	107.49	
-1	046.01	
± 0	025.66	Bayes
$+1$	020.88	Chi-optimal
$+2$	041.76	Minimax
$+3$	090.18	

Table 3 Results of the 3 solution methods

4. C-robustness

4.1. Notations and assumption

From this section on we will describe the method for stochastic programming under uncertainty as well as for decision theory simultaneously by considering a loss function $\mathbb{L}$ for both. For optimization under uncertainty this $\mathbb{L}$ is defined by

$$\mathbb{L}(a,P) = \mathbb{E}_P\big[L(a,\xi)\big] \tag{23}$$

for any a $\in \mathbb{A}$ and P $\in \mathcal{P} = \{P_1,\ldots,P_r\}$. $\mathbb{A}$ is identical with $\mathbb{A}^*$ in the case of decision theory (see the corresponding subsection above). Then r = r(a,s) is defined by

$$r(a,s) = \sum_{i=0}^{m} \mathbb{L}(a,P_i)\, s_i, \tag{24}$$

for any a $\in \mathbb{A}$ and s $\in \mathbb{V}$.

Assumption 2. *We assume from this section on, that the affine dimension of the given PI V being a subset of the m-dimensional standard simplex is m (PI is defined in Sect. 2). That this is no real restriction is proven in Appendix 6.2.*

4.2. C-robust solution

Our work is motivated by problems, where an entire set of distributions $\mathcal{P}$ is given, instead of a single one. In Sect. 3, we treat these problems by computing solutions which minimize the (unavoidable) tolerance for the entire family of distributions $\mathcal{P}$. For the cases where the resulting minimal tolerance c^* is too large (as in some practical situations), we propose a different idea. The idea is then to fix the tolerance c (as user input) and to compute solutions which maximize the resulting family of distributions *on* $\mathcal{P}$ which satisfy this tolerance-criterion. (In practice, one can determine the most appropriate tolerance value c by a simulation procedure.)

Here one can see that the standardization of the loss- or profit-function in (1) has not only advantages for mathematical formulation but also for the choice of a suitable tolerance by the application.

Now we develop the c-robustness as the solution in the case in which the user must or can fix a tolerance level, say c ≥ 0. Let any PI $\mathbb{V}$ be given (being a convex and compact subset of the m-dimensional standard simplex SI, see above).

First, as the mathematical basis, we need a more handy form of the given PI. The following lemma prepare the corresponding theorem.

Lemma 1.

(a) Every convex and compact subset K of R^n having the zero-vector in its interior is homeomorphic to the unit ball in the Taxicab norm

$$U_1 := \{\|z\|_1 \leq 1\} \tag{25}$$

(the Taxicab norm is defined, for any $z \in R^n$, by

$$\|z\|_1 = \sum_{j=1}^{n} |z_j| \ ). \tag{26}$$

(b) One can choose a homeomorphism from K to the unit ball U_1 which is positively homogenous.

Proof. This proof is, of course, in the main similar to the proof of the well-known fact that any convex and compact subset of R^n with nonempty interior is homeomorphic to the (closed) unit ball in R^n (in the Euclidean norm). But we need the unit ball in the Taxicab norm and will have also the property (b), therefore our proof will be different.

We apply, as important element, the Minkowski-functional m_K of K, defined, for any z $\in R^n$, by

$$m_K(z) := inf\{t \mid z/t \in K, t > 0\}. \tag{27}$$

This function is

(1) positively homogenous,

(2) continuous, also at zero,

and has the property

$$K = \{z \mid m_K(z) \leq 1\} \tag{28}$$

(for the proof see [8], chapter 5.12).

Then we choose as homeomorphism h: K → U_1

$$h(z) = \begin{cases} z m_K(z)/\|z\|_1, \text{ if } z \mp 0 \\ \qquad\qquad 0, \text{ o/w} \end{cases} \tag{29}$$

h has the properties (1) and (2) too, the second, because

$$\|h(z)\|_1 = m_K(z)\|z/\|z\|_1\|_1 = m_K(z). \tag{30}$$

The inverse function of h is given by

$$h^{-1}(y) = \begin{cases} y\|y\|_1/m_K(y), \text{ if } y \mp 0 \\ \qquad\qquad 0, \text{ o/w} \end{cases} \tag{31}$$

as one check easily. h is only a homeomorphism, if also the inverse is continuous. It is continuous for every y $\mp$0. Therefore it remains to prove only two claims.

Claim 1: h^{-1} is continuous at zero.

Because zero lies in the interior of K, there exists a positive M < 1 so that

$$U_M := \left\{ \|y\|_1 \leq M \right\} \subset K.$$

But then

$$m_K(y) \geq \|y\|_1 M^{-1} \quad \forall\, y \in U_M$$

and therefore

$$\|h^{-1}(y)\|_1 = \|y\|y\|_1/m_K(y)\|_1 \leq \|y\|_1 M \quad \forall\, y \in U_M \backslash \left\{0\right\}. \tag{32}$$

(32) implies that

$$\lim_{\|y\|_1 \to 0} \|h^{-1}(y)\|_1 = 0 = \|h^{-1}(0)\|_1$$

and this means the continuity of h^{-1} at zero, too.

In order to complete the proof of the lemma it remains to prove the following

Claim 2: h is also onto, i.e.,

$$h(K) = U_1.$$

For any y ∈ h(K) there exists an z ∈ K such that y = h(z). As a consequence of (30) and (28) we have

$$\|y\|_1 = \|h(z)\|_1 = m_K(z) \leq 1$$

and that means y $\in U_1$.

Vice versa let any y $\in U_1$ be given. If we choose z := $h^{-1}(y)$, then

$$m_K(z) = m_K(h^{-1}(y)) = \|y\|_1 \leq 1$$

holds. But

$$m_K(z) \leq 1 \;\Rightarrow\; z \in K \;\Leftrightarrow\; h(z) \in h(K)$$

and therefore

$$y = h(h^{-1}(y)) = h(z) \in h(K)\,.$$

□

Now we can prove the theorem which translate the given PI in a more suitable form.

THEOREM 2. *Let B be the barycenter of the given PI* $\mathbb{V}$*. We denote, for any real number* $\lambda \in [0,1]$,

$$W_Q(\lambda) = \left\{ s \in]0,+\infty[^{m+1} \mid\; s = y/(y_0 + ... + y_m) \textit{ with } y \in]0,+\infty[^{m+1} \wedge -q\lambda \leq Q(y-B) \leq q\lambda \right\} \tag{33}$$

(a) Then there exist a positively homogenous homeomorphism (i.e., a continuous and one to one mapping Q from R^{m+1} *into itself) and a vector* $q \in]0,+\infty[^{m+1}$ *with Taxicab norm norm 1, such that* $\mathbb{V}$ $= W_Q(1)$.

(b) Q is uniquely defined by this property.

Proof. Because of the definition of PI and Remark 2, we have assumed without loss of generality that every PI is convex and compact and having affine dimension m, therefore $\mathbb{V}$ too.

We translate $\mathbb{V}$ by -B to

$$\mathbb{V}_0 := \{y - B \mid y \in \mathbb{V}\}.$$

and denote by $\mathbb{V}_a$ the vector space associated to the minimal affine hull of $\mathbb{V}_0$. According to the definition of affine dimension, there exist vectors $Z_1, ..., Z_m$ which are linear independent and form a basis of $\mathbb{V}_a$. Now the projection Pr of $\mathbb{V}_0$ into the R^m relates (resp. maps) each v $\in \mathbb{V}_0$ to the vector consisting of its coordinates with respect to this basis. Then the interior of Pr($\mathbb{V}_0$) is not empty (see Rockafellar) and contains the zero-vector in its interior, because the zero-vector is the affine image of the barycenter B being element of the relative interior of $\mathbb{V}$. Then, according to Lemma 1, Pr($\mathbb{V}_0$) is homeomorphic to the unit ball

$$U_1 := \{z \in R^m \mid \|z\|_1 \leq 1\}.$$

We denote by H_0 the homeomorphism from Pr($\mathbb{V}_0$) to U_1.

Now let

$$W_1 := \{y \in]0, +\infty[^{m+1} \mid -q \leq y \leq q\} \tag{34}$$

and thereby the vector q $\in]0, +\infty[^{m+1}$ so, that

$$W(1) := \{s \in]0, +\infty[^{m+1} \mid s = y/(y_0 + ... + y_m)\, with\, y \in]0, +\infty[^{m+1} \wedge -q \leq y \leq q\} \subset \mathbb{V}_0$$

holds with the maximal value of q, in every coordinate.

Because U_1 as well as W_1 is a polygon, there exists a linear mapping, say A_0, from U_1 to W_1.

Since Pr and A_0 are linear mappings,

Q := A_0 o H_0 o Pr

is a homeomorphism from $\mathbb{V}_0$ to W_1, and, with this Q, we obtain

$$\mathbb{V}_0 = \{y \in]0, +\infty[^{m+1} \mid -q \leq Q(y) \leq q\}$$

and therefore

$$\mathbb{V} = \{y \in]0, +\infty[^{m+1} \mid -q \leq Q(y - B) \leq q\}.$$

Now, we apply one both sides of this equality the linear-fractional mapping y $\rightarrow y/(y_0 + ... + y_m)$ from $]0, +\infty[^{m+1}$ into itself. Because $\mathbb{V}$ (as a subset of the simplex SI) is invariant under it, $\mathbb{V} = W_Q(1)$ holds.

According to its definition, Q is also positively homogenous, because H_0 is it (according to Lemma 1) and Pr and A_0 as linear mappings too.

That the assertion (b) is true, if $\mathbb{V}$ is a convex polytop, is a well-known fact. Because $\mathbb{V}$ as a convex and compact subset of the m-dimensional standard simplex can be approximated as closely as desired by inscribed polytopes, this is also true in general. □

In our PPC-example, Q is the identity, in example 2 in R^{101} and in example 3 in R^6. p is the half of upper limit plus the half of the lower and q the upper limit minus p. As it is shown in this example, the main and only really sure information contained in $\mathbb{V}$ is the structure and not the limits. Therefore any λ a little less than 1 is no essential change of the PI $\mathbb{V}$. Now, Theorem 2 shows that this is also true in general. The homeomorphism Q is in the main the Minkowski functional which does not change the geometric form of the PI $\mathbb{V}$.

Therefore one search for a solution that adheres the given tolerance limit c for the largest possible subset $W_Q(\lambda)$ of $\mathbb{V} = W_Q(1)$, i.e., for the maximal value of λ (as $\lambda_1 \leq \lambda_2 \succ W_Q(l_1) \subseteq W_Q(l_2)$). This motivates the following method as the right approach in the situation with given tolerance limit.

First we have to adapt the definition of tolerance function to this situation.

Definition 5 (Tolerance (function) for $W_Q(\lambda)$). For any $a_0 \in \mathbb{A}$ and for any $\lambda \in [0,1]$ we define

$$t(a_0,\lambda) = \max_{a\in\mathbb{A}} \max_{s\in W_Q(\lambda)} \big(r(a_0,s) - r(a,s)\big), \tag{35}$$

t(a_0,λ) is called **tolerance** of a_0 for the given PI $W_Q(\lambda)$.
As in the definition of tolerance function T, the maximum in (35) exists for all $a_0 \in \mathbb{A}$, because $r(a_0,s)$, as function of s, is linear and therefore continuous and also bounded (because of assumption 1 and the definition of L in (3)) and $\mathbb{V}$ is compact. The maximum over a $\in \mathbb{A}$ exists because of the compactness of $\mathbb{A}$ analogously.

The following lemma is the next important step of our method.

Lemma 2. *For any $a_0 \in \mathbb{A}$ its tolerance t(a_0,λ) is an uniformly continuous function of λ.*

Proof. Let any $\epsilon > 0$ be given.

We denote, for any a $\in \mathbb{A}$ and any s $\in W_Q(\lambda)$,

$$R(a,s) := r(a_0,s) - r(a,s).$$

Because R, as function of a and s, is uniformly continuous, because $\mathbb{A} \times W_Q(\lambda)$ is compact, there exists a positive number e such that

$$||(a,s) - (a',s')|| < e \Rightarrow |R(a,s) - R(a',s')| < \epsilon \qquad \forall\ a,a' \in \mathbb{A} \wedge\ s,s' \in W_Q(\lambda). \tag{36}$$

Because the compactness of $\mathbb{A}$, the cover by open spheres of radius e around all of its elements contains a finite sub-cover like

$$\mathbb{A} \subset \bigcup_{j=1}^{J} \{||a - a_j|| < e\}\,. \tag{37}$$

where

$$a_j \in \mathbb{A}\,,\ j = 1,...,J.$$

Then R(a,s), as a function of a $\in \mathbb{A}$, can be approximated uniformly as closely as desired by the values $R(a_j,s_j)$, j = 1, ... , J , because for any a $\in \mathbb{A}$ there exists a i $\in \{1,\ldots,J\}$ such that

$$|R(a,s) - R(a_i,s_i)| < \epsilon \tag{38}$$

holds, independently from s $\in W_Q(\lambda)$ and therefore also independent from $\lambda \in [0,1]$.

Therefore, for the rest of this proof, we can assume that $\mathbb{A}$ is finite , i.e., $\mathbb{A} = \{a_1,\ldots,a_J\}$.

Now, let any $a_k \in \mathbb{A}$ and any λ be given.

Case 1: $\lambda' \leq \lambda$

We consider

$$\Delta(\lambda, a_k) := \max_{s \in W_Q(\lambda)} R(a_k, s). \tag{39}$$

According to Weierstrass' theorem, $R(a_j, .)$ attains the maximum in the right side of (39) at any point $\upsilon(\lambda)$ in $W_Q(\lambda)$. Because $W_Q(\lambda)$ is convex and compact, $\upsilon(\lambda)$ is lying on the edge of $W_Q(\lambda)$, therefore

$$\upsilon(\lambda) = \upsilon'(\lambda)/\|\upsilon'(\lambda)\|_1,$$

with

$$\upsilon'(\lambda) = Q^{-1}(\lambda q^*) + B \tag{40}$$

holds, according to the definition of $W_Q(\lambda)$ in (33) ($\|.\|_1$ is the Taxicab norm defined in (26)). Thereby q^* in (40) means that for a certain subset I of $\{0, \dots, m\}$

$$q_i^* = -q_i \qquad \forall\, i \in I \qquad \wedge \qquad q_j^* = q_j \qquad \forall\, j \notin I.$$

Because Q^{-1} as the inverse of Q is positively homogenous, we can transform the right side of this equation as follows:

$$\upsilon'(\lambda) = \lambda Q^{-1}(q^*) + B. \tag{41}$$

But the latter equation proofs that υ is Lipschitz-continuous. But then there exists a positive δ such that

$$|\lambda' - \lambda| < \delta \Rightarrow |\upsilon(\lambda') - \upsilon(\lambda)| < e. \tag{42}$$

According to (36) we obtain now

$$|\lambda' - \lambda| < \delta \Rightarrow |R(a_k, \upsilon(\lambda')) - R(a_k, \upsilon(\lambda))| < \epsilon$$

and therefore in this case ($\lambda' \le \lambda$)

$$R(a_k, \upsilon(\lambda')) > R(a_k, \upsilon(\lambda)) - \epsilon,$$

i.e., because of (39)

$$\Delta(\lambda', a_k) > \Delta(\lambda, a_k) - \epsilon$$

and therefore

$$\Delta(\lambda, a_k) > \Delta(\lambda', a_k) > \Delta(\lambda, a_k) - \epsilon.$$

But the latter is equivalent to

$$|\Delta(\lambda', a_k) - \Delta(\lambda, a_k)| < \epsilon \tag{43}$$

Case 2: $\lambda' > \lambda$

If we substitute λ by λ' in the proof steps of Case 1 until equation (41) and then interchange λ' and λ, we will get the same result (43) too.

Then $\Delta(., a_k)$ is (uniformly) continuous for all $a_k \in \mathbb{A}$ and therefore also

$$t(a_0, \lambda) = \max_{a_k \in \mathbb{A}} \Delta(\lambda, a_k)$$

as the maximum over a finite number of continuous functions of λ.

□

Now let any c ≥ 0 be given. The maximal value of c, which makes sense, is χ. Therefore we propose that c $\leq \chi$.

For any $a_0 \in \mathbb{A}$, we choose the following value of λ according to the above formulated and substantiated objective and corresponding procedure as follows:

$$\Lambda(a_0, c) := \max \big\{ \lambda \in [0,1] \,:\, t(a_0, \lambda) = c \big\}. \tag{44}$$

$\Lambda(a_0, c)$ exists, because $t(a_0, \lambda)$ is a (uniformly) continuous function of λ.

Definition 6 (c-robust solution). For some $c \geq 0$, the solution $d(c)$ is called c-robust, if

$$\Lambda\big(d(c), c\big) \geq \Lambda\big(d, c\big) \text{ for all } d \in \mathbb{A}.$$

Then, the c-robustness is well-defined.

With that, $d(c)$ minimizes the mean loss up to c in the largest subset $V\big(\Lambda\big(d(c), c\big)\big)$ of $\mathbb{V}$ with the same structure.

4.3. Properties of C-robustness

First we remark that the c-robust solution is invariant under affine functions of the loss function, if the factor is positive.

But the main property of the c-robust solution (or decision) is the following: We denote $\lambda_c := \Lambda(d(c), c)$. If one substitute V by $W_Q(\lambda_c)$, then all statements about χ-optimality of the previous section apply analogously to c-robustness and the c-optimal solution has the corresponding properties. We formulate as the most important of these properties Theorem 1 again for c-robust solutions resp. decisions as follows:

Corollary 2. ***t(d(c),λ_c) is the smallest tolerance term***, i.e., if any positive number ϵ with any a(ϵ) $\in \mathbb{A}$ has the property

$$r(a(\epsilon), s) \leq \; r(a, s) \; + \; \epsilon \qquad \forall a \in \mathbb{A} \; \wedge \; \forall s \in V, \tag{45}$$

then

$$\epsilon \geq t(d(c), \lambda_c) \qquad (= c)$$

holds.

For simplicity we assume for the following theorem, that Q in equation (33) of Theorem 2 is the identity id in R^r. Under no really restrictive propositions, one can generalize the results for any affine mapping Q, and in the general case, on can approximate any Q by a affine mapping. we abbreviate $W_{id}(\lambda)$ by W(λ).

Theorem 3. *For any $a_0 \in \mathbb{A}$, we abbreviate $\lambda^* := \Lambda(a_0, c)$ and*

$$R_i(a) := EL(a_0, P_i) - EL(a, P_i) \qquad i = 0, ..., m$$

and define w(a,λ^*) $\in W(\lambda^*)$ as follows: For any $i \in \{0, \ldots, m\}$ we set

$$w_i(a, \lambda^*) = z_i / (\sum_{j=0}^{m} z_j), \tag{46}$$

where

$$z_i := \begin{cases} B_i + \lambda^* q_i, \text{ if } R_i(a) \geq 0 \\ B_i - \lambda^* q_i, \text{ o/w} \end{cases} \tag{47}$$

Then the following equation holds

$$t(a_0,\lambda^*) = \max_{a\in\mathbb{A}} \sum_{j=0}^{m} \big(\mathbb{L}(a_0,P_j) - \mathbb{L}(a,P_j)\big) w_j(a,\lambda^*). \tag{48}$$

Proof. Let any a $\in \mathbb{A}$ be given.
Because r(a_0,s) - r(a,s) is a linear (and therefore convex and continuous) function of s and $\in W(\lambda^*)$ is convex and compact, it achieves its maximum at an extreme point $\overline{s}$ of $\in W(\lambda^*)$, according to the Krein-Milman-Theorem, i.e.,

$$\max_{s\in W(\lambda^*)} \big(r(a_0,s) - r(a,s)\big) \;=\; r(a_0,\overline{s}) - r(a,\overline{s}). \tag{49}$$

As a consequence of (33) in Theorem 2, there exists an (m+1)-dimensional non negative vector y satisfying

$$(B - q\lambda^*) \le y \le (B + q\lambda^*)$$

and

$$\overline{s}_i = y_i / (\sum_{j=0}^{m} y_j) \text{ for } \qquad i = 0,\dots,m.$$

By simple calculations one can verify that, for non negative α, x and z and positive b $\geq$ z,

$$\frac{\alpha + x}{b + x} \geq \frac{\alpha}{b} \qquad \wedge \qquad \frac{\alpha - z}{b - z} \leq \frac{\alpha}{b} \tag{50}$$

holds, if $\alpha \leq$ b.
Then, for any j $\in \{0,\ldots,m\}$, we substitute

- α by y_j
- b by $\sum_{i=0}^{m} y_i$
- x by $B_j + \lambda^* * q_j - y_j$
- and
- z by $y_j - B_j + \lambda^* * q_j$

,

obtaining

$$\frac{B_j + \lambda^* * q_j}{\sum_{i=0,i\neq j}^{m} y_i + (B_j + \lambda^* * q_j)} \;\geq\; \frac{y_j}{\sum_{i=0}^{m} y_i} \qquad \wedge \qquad \frac{B_j - \lambda^* * q_j}{\sum_{i=0,i\neq j}^{m} y_i + (B_j - \lambda^* * q_j)} \;\leq\; \frac{y_j}{\sum_{i=0}^{m} y_i}\,. \tag{51}$$

Then, because of (51),

$$\sum_{j=0}^{m} R_j(a)\,\overline{s}_j \;\leq\; \sum_{j=0}^{m} R_j(a) * w_j(a,\lambda^*) \tag{52}$$

holds. Therefore

$$\max_{s\in W(\lambda^*)} \big(r(a_0,s) - r(a,s)\big) \;\leq\; \big(\sum_{j=0}^{m} R_j(a) * w_j(a,\lambda^*). \tag{53}$$

Because in (53) also holds the inverse inequality, we obtain the corresponding equality. But by this the theorem is proven.

□

4.4. Compatibility of c-Robustness with chi-Optimization

Definition 7. We set

$$c^* := \inf \left\{ c \geq 0 \,|\, \mathbb{V} \subset W_Q\big(\Lambda(d(c), c)\big) \right\},$$

i.e., the c^*-robust solution minimizes the mean loss up to c for the entire (given) PI with the smallest c being equal to c^*.
Now we can formulate the theorem proving that the χ-optimal solution is a special c-robust one.

Theorem 4. *χ-optimization and c^*-robustness are equivalent and $\chi = c^*$ (for the same PI $\mathbb{V}$).*

Proof. **Claim 1**: $c^* \leq \chi$.

Let a_χ be χ-optimal. Because $\mathrm{T}(a_\chi) = \chi$. we obtain as a consequence of the definition of χ-optimality and (12)

$$\max_{s \in \mathbb{V}} \big(r(a_\chi, s) - r(a, s) \big) \leq \chi \qquad \forall a \in \mathbb{A}.$$

Therefore

$$\Lambda(a_\chi, \chi) = 1 \geq \Lambda(a, \chi) \qquad \forall a \in \mathbb{A}.$$

Then, according to the definition of c-robustness, a_χ is χ-robust. and, according to the definition of c^*, $c^* \leq \chi$ holds.

Claim 2: $c^* \geq \chi$.

Let a^* be c^*-robust. Then $\mathrm{T}(a^*) = c^*$. Therefore, according to the definition of χ, $c^* \geq \chi$ holds, and this completes the proof.

□

Therefore both methods are compatible.

5. Remark to asymptotic results

In the asymptotic situation, for each $n \in \mathbb{N}$ (the set of natural numbers) we have n observations $X_1, ..., X_n$ for the random l-dimensional parameter vector ξ. We remember on the notations and assumptions formulated in subsection 4.1 which should also remain valid in this section. Additionally, we assume for simplicity that l = 1, what is usually assumed and can also be generalized.

As formulated in [6], in the asymptotic theory or application the existence of a sequence of uniformly consistent estimators is assumed directly or indirectly. In this paper is also explained what that means. Here, we need, because of the finiteness of $\mathcal{P}$, only the consistency, i.e., the existence of any estimator-sequence $(T_n)_{n \in \mathbb{N}}$ and any null sequence (e_n) with

$$\lim_{n \to \infty} P_i\big(\|T_n(X_1, ..., X_n) - F_i\|_s > e_n \big) = 0 \qquad \forall \qquad i = 0, ..., m. \tag{54}$$

Thereby F_i denotes the cumulative distribution function of P_i, $\|...\|_s$ the supremum norm and the estimator T_n is a measurable mapping $T_n = T_n(X_1, ..., X_n)$ from the sample space into the normed space of bounded continuous functions.

In the so called classical case of independent and identically distributed real-valued observations, these assumptions are fulfilled, if every F_i has a continuous density function. Because we can choose for any n the empirical distribution function as T_n and any null sequence $e_n = \mathrm{O}(\ n^{-\frac{1}{2}})$ (i.e., the sequence (e_n) divided by the sequence $(n^{-\frac{1}{2}})$ diverges to infinity), e.g., $e_n = \log(\log(\mathrm{n}))\ n^{-\frac{1}{2}}$. Then (54) is true according to the Lemma 2 in [2]. In this paper is explained all, too.

Under these weak assumptions the tolerance of the χ-optimal decision sequence converges to zero, precisely holds the following

Theorem 5. *Let $(a_{\chi,n})_{n\in\mathbb{N}}$ be the sequence of χ-optimal decision and P_r the right distribution of ξ. Then*

$$P_r - \lim_{n\to\infty} T\big(a_{\chi,n}(X_1,...,X_n)\big) = 0 \tag{55}$$

and for discrete distribution of ξ

$$\lim_{n\to\infty} T\big(a_{\chi,n}(X_1,...,X_n)\big) = 0 \qquad \text{almost surely} \tag{56}$$

holds.

Proof. First we remark, that, for any natural number n $a_{\chi,n}$ is a measurable function of $X_1,...,X_n$ with values into $\mathbb{A}$ usually called decision function. This is true under the given assumptions, in the main, as $a_{\chi,n}(X_1,...,X_n)$ is minimal point of a uniformly continuous function and because of the compactness of $\mathbb{A}$ (for more details and a proof see [9]).

The convergence in (54) is also uniformly, because $\mathcal{P}$ is finite. Then all assumptions of Lemma 5 and Corollary 9 of [6] are fulfilled and therefore there exists a sequence $(d_{0,n})$ of decision functions with the property

$$P_r - \lim T\big(d_{0,n}(X_1,...,X_n)\big) = 0, \tag{57}$$

i.e., the convergence of the tolerance to zero in probability, because the convergence in Corollary 9 of this paper is uniform in (d_n), as the proofs of Lemma 5 and of the Theorems 6 and 7 show.

But

$$T\big(a_{\chi,n}(X_1(\omega),...,X_n(\omega))\big) \;\le\; T\big(d_{0,n}(X_1(\omega),...,X_n(\omega))\big)$$

holds according to the definition of χ-optimality and therefore (55).

Finally, (56) is a well-known consequence of (55), because the convergence in probability is also almost sure for discrete random variable.

□

This result is not surprising, as asymptotic considerations make a little more sense for an infinite set $\mathcal{P}$, normally parametrized, in order to reach estimators or decision functions with fast convergence to the asymptotic optimum. More precisely the speed of convergence is the speed of convergence to zero of

$$\max_{a\in\mathbb{A}} \max_{s\in\mathbb{V}} \big(r(a_{0,n},s) - r(a_n,s)\big)$$

(r(a,s) see (7)), but this term is exactly the value of the tolerance function which is minimized by the chi-optimal sequence. Therefore the latter results the maximal speed of convergence to zero.

6. Appendices

6.1. Example showing the Effect of observations

The family of distributions may be normal distribution with unknown mean μ and known standard deviation σ^2; *i.e.*,

$$\xi \sim \mathcal{N}(\mu,\sigma^2).$$

We assume that the mean μ is contained in the interval $[-a,a]$ with positive a, and that L is given by

$$L(\mathbf{a},\xi) := c(\mathbf{a}-\xi)^2 \tag{58}$$

For some known RV $\xi \sim P = \mathcal{N}(\mu,\sigma^2)$, the expected value of (58) is evaluated as

$$\mathbb{E}_P\big[L(\mathbf{a},\xi)\big] = (c_1+c_2)\frac{\sigma}{\sqrt{2\pi}}\exp^{-\frac{(a-\mu)^2}{2\sigma^2}} - c_2 a + (c_1+c_2)(a-\mu)F^{\mathrm{SN}}\Big(\frac{a-\mu}{\sigma}\Big) + c_2\mu$$

with standard normal CDF $F^{\mathrm{SN}}(a)$.

6.2. Proof of the Remark in Assumption 2

Proof of this Remark. Let l < m be the affine dimension of $\mathbb{V}$. According to the definition of affine dimension, there exists a orthogonal basis of the affine hull of it consisting of l vectors lying in it. By dividing every vector of this basis by his Taxicab norm (definition of Taxicab norm see (26)), we obtain a new basis $Z_1,...,Z_l \in \mathbb{V}$ of vectors with Taxicab norm 1.

Let any s be given:

Then we can represent s with the basis as follows

$$s = \sum_{k=1}^{l} \alpha(s)_k \, Z_k,$$

i.e.,

$$s_i = \sum_{k=1}^{l} \alpha(s)_k \, Z_{k,i} \, \forall \, i \in \{0, \dots, m\}$$

($Z_{k,i}$ denotes the i-th coordinate of the basis vector Z_k). We can choose the vectors of the basis so that the coordinates $\alpha(s)_k$ of every s $\in \mathbb{V}$ are non negative.

Therefore, if we substitute $\mathcal{P}$ by $\{P_1^*, \dots, P_l^*\}$, with

$$P_k^* := \sum_{i=0}^{m} \, Z_{k,i} \, \forall \, k \in \{1, \dots, l\} \, ,$$

and $\mathbb{V}$ by

$$\mathbb{V}^* := \big\{ (\alpha(s)_1, \dots, \alpha(s)_l) \mid s \in \mathbb{V} \big\},$$

then $\mathbb{V}^*$ is a convex and compact subset of the l-dimensional standard simplex and a partial information with affine dimension l.

□

References

[1] Bernado, J.M., A.F.M. Smith. 1994. *Bayesian Theory*. Wiley.

[2] Dvoretzky, A., J. Kiefer, J. Wolfowitz. 1964. Asymptotic minimax of the sample distribution function and the classical multinomial estimator. *The Annals of Mathematical Statistics* **27** 125–134.

[3] Kofler, E., G. Menges, R. Fahrion, S. Huschens, U. Kuss. 1980. Stochastische partielle information (spi). *Statistische Hefte (Statistical Papers)* **21** 160–167.

[4] Kuss, U. 1972. Contributions to maximum probability estimators. *Z. Wahrscheinlichkeitstheorie verw. Geb.* **24** 123–133.

[5] Kuss, U. 1975. Maximum probability estimators in the case of exponential distribution. *Metrika* **22** 129–146.

[6] Kuss, U. 1985. C-optimal decisions with optimality properties for finite sample size. *Contributions to Econometrics and Statistics Today*. Springer, 162–176.

[7] Kuss, Uwe. 1971. Maximum probability-methode und optimale angenaeherte bayes-loesung bei festem stichprobenumfang in der schaetz- und testtheorie. Ph.D. thesis, Ruprecht-Karls-Universität Heidelberg.

[8] Luenberger, David G. 1997. *Optimization by vector space methods*. John Wiley & Sons.

[9] Reiss, R.-D. 1973. On the measurability and consistency of maximum likelihood estimates for unimodal densities. *The Annals of Statistics* **1** 888–901.

[10] Weiss, Lionel. 1983. Small-sample properties of maximum probability estimators. *Stochastic Processes and their Applications* **14**(3) 267–277.

[11] Weiss, Lionel, J Wolfowitz. 1969. Maximum probability estimators with a general loss function. *Probability and Information Theory*. Springer, 232–256.

[12] Žáčková, J. 1966. On minimax solutions of stochastic linear programming. *Časopis pro pěstování matematiky* **91** 423–430.