Z. Wahrscheinlichkeitstheorie verw. Gebiete
69, 387–420 (1985)

Zeitschrift für
Wahrscheinlichkeitstheorie
und verwandte Gebiete
© Springer-Verlag 1985

Simultaneously Least Favorable Experiments

Part II: Standard Loss Functions and Their Applications

Andreas Buja
Department of Statistics, GN-22, University of Washington, Seattle, Washington 98195, USA

Summary. We continue our investigation into robust finite experiments and decision problems which we started in part I. We give a short derivation of the Huber-Strassen theorem, and also a proof of a new result on the existence of simultaneously least favorable experiments for contaminated classification problems. The main effort goes into the development of tools which allow us to apply the general results of part I. The principal device needed for our derivations is a reparametrization of decision spaces and a corresponding lift of loss functions which we call "standard loss functions" (in analogy to "standard experiments"). Non robust decision theory which deals with linear functionals (expectations) does not need this device, but robust theory based on sublinear functionals (upper expectations) does.

Contents

1. Introduction

In this paper we continue our study of robust finite decision problems, on which we embarked in the previous paper (part I)[1]. Before we continue our technical development, we will try to summarize the concepts and facts of part I, and shed some light on the type of problems we wish to answer in the present paper. At the root of our investigation is the Huber-Strassen theorem [4], of which the two salient points can be described as follows:

[1] Z. Wahrscheinlichkeitstheor. Verw. Gebiete **65**, 367–384 (1984)

a) A robustness aspect is introduced in the testing problem of simple hypothesis versus simple alternative by replacing model distributions P_θ by "inexact" distributions in the form of certain convex sets of distributions. Another way to characterize these "inexact" or "approximate" distributions is by set functions w_θ which are obtained as supremums over these convex sets of probability measures. Huber and Strassen deal more specifically with a kind of set functions which are known as capacities alternating of order 2.

b) The Huber-Strassen theorem states the existence of simultaneously least favorable pairs (Q_1, Q_2) of distributions under the "approximate" model (w_1, w_2). "Simultaneous" here is meant with regard to all testing levels or equivalently all prior weights.

In part I we gave a framework for robust decision theory of general finite experiments, i.e. more than 2 but still finitely many parameter values. We started out with the definition of an "approximate model" (v_θ) in terms of upper expectations. In order to meet the needs for convexity properties, we replaced the set functions w_θ of Huber-Strassen by sublinear functionals v_0, which are just supremums over expectations rather than probabilities.

We then introduced various notions of randomized decision procedures, marred partly by technicalities which have to do with incompatibilities between σ-additivity and topological compactness requirements on sets of decision procedures. LeCam had another way of dealing with this problem, for which we refer to his important paper [7]. Connected with randomizations or transition probabilities is Blackwell's notion of sufficiency, which generalizes the more familiar Halmos-Savage definition. In turn, we defined our concept of "worst-case sufficiency" as a generalization of Blackwell's notion and also LeCam's "approximate sufficiency".

Next we introduced the robust version of Blackwell's standard measures. In finite-parameter experiments, they play the same crucial role as distributions of probability density ratios in testing experiments. The trick is to replace ratios by tuples of densities (q_θ) with regard to the dominating measure $\sum_\theta Q_\theta$, such that one ends up with a mapping of sample space into the unit simplex of $\mathbb{R}^\Theta$. In the case of approximate models (v_θ), there do not exist likelihood ratios or tuples of densities in general, but we can still define an "upper standard functional" on the unit simplex. For details we must refer to part I, Sect. 5. We should also point out that part II can not be understood without a thorough understanding of standard measures and their connection with Bayes risks. Below we will reiterate some statements of part I as far as this subject is concerned.

The seemingly disparate developments of part I lead to a new version of a theorem which is connected with the names of LeCam, Blackwell, Sherman, Stein, and others. In its original form it is a characterization of Blackwell-sufficiency in terms of Bayes risks and standard measures. Basically it asserts the equivalence of the following statements, which express the fact that (Q_θ) is universally less favorable than (P_θ):

1) The experiment (Q_θ) can be obtained as the image of (P_θ) under a randomization.

2) The minimal Bayes risk for (Q_θ) is never better than (i.e. below) the minimal Bayes risk for (P_θ), no matter what the loss function is.

3) The values of the standard measure of (Q_θ) on concave functions on the unit simplex are never below the ones of the standard measure of (P_θ).

LeCam [7] recognized that this theorem can be adapted to more general notions of sufficiency. Our own version in terms of worst-case sufficiency suits the needs of robust decision theory. Especially, it allowed us in part I to derive a structure theorem which characterizes those approximate models and classes of loss functions for which there do exist simultaneously least favorable experiments. The theorem is stated in terms of an additivity condition for the upper standard functional which belongs to the given approximate model.

We now reiterate the definition of standard measures and Bayes risks in more technical terms. Our notation for the latter is:

$$R((Q),\sigma,(W))=\frac{1}{|\Theta|}\sum_\theta Q_\theta\,\sigma(W_\theta)$$

where (Q_θ) is an (exact) experiment, σ is a (randomized) decision procedure, and (W_θ) is a loss function. Obviously we confined ourselves to the uniform prior on the parameter set Θ. The reason is that we can always absorb an arbitrary prior α_θ in the loss function (replacing W_θ by $|\Theta|\cdot\alpha_\theta\cdot W_\theta$), and thus artificially get back to the case of a uniform prior. As mentioned above, let us introduce densities q_θ with regard to the dominating measure $\sum_\theta Q_\theta$. The a posterior expected loss is then:

$$a_t(q(y))=\sum_\theta W_\theta(t)\cdot q_\theta(y) \qquad (*)$$

where $q(y)=(q_\theta(y))$ is the vector of densities. Depending on our needs, we may consider $a_t(z)$ as a linear function on $\mathbb{R}^\Theta$ or an affine function on the unit simplex of $\mathbb{R}^\Theta$. A Bayes procedure consists of deciding such that $t\to a_t(q(y))$ is minimized for given observation y. We therefore define:

$$k(q(y))=\inf_t a_t(q(y)) \qquad (**)$$

where again we may consider $k(z)$ either as a concave function on $\mathbb{R}^\Theta$ or on the unit simplex of $\mathbb{R}^\Theta$. These functions play an all-important role. They summarize the minimal a posteriori expected losses in a model-independent way, based on the loss function alone. With their help, the minimal Bayes risk can now be expressed as follows:

$$\inf_\sigma R((Q),\sigma,(W))=\int k(q(y))\frac{1}{|\Theta|}\sum_\theta Q_\theta(dy).$$

This motivates the definition of a *standard measure* as the distribution of the density vector $q(y)$ under the marginal distribution of the observations, i.e., $\frac{1}{|\Theta|}\sum Q_\theta$. It resides on the unit simplex in $\mathbb{R}^\Theta$. If we denote it by $S^{(Q)}$, we can write the minimal Bayes risk as:

$$\inf_{\sigma} R((Q),\sigma,(W)) = S^{(Q)}(k) = \int k(z)\, S^{(Q)}(dz).$$

The same holds true if we pass to the robust case of approximate models (v_θ) instead of exact ones (Q_θ) (see part I, Sects. 4 and 5). The linear functional $S^{(Q)}$ is then replaced by a sublinear one $s^{(v)}$ (i.e., an upper expectation) which corresponds to the minimal upper Bayes risk of the approximate model (v_θ) for the loss function (W_θ):

$$\inf_{\sigma} R((v),\sigma,(W)) = s^{(v)}(k) = \sup_{(Q)\leqq(v)} S^{(Q)}(k).$$

The main result of part I can now be stated as follows: For an approximate model and a class of loss functions there exists a simultaneously least favorable experiment if and only if the upper standard functional $s^{(v)}$ is additive on the concave functions k which are generated by the loss functions in question.

The intention here in part II is to supply a technique which makes the results of part I applicable. A difficulty stems from the fact that the concave functions k (i.e. basically the minimal a posteriori expected losses) depend in general superadditively rather than linearly on the loss functions $W_\theta(t)$ as can be seen from the formulae (*) and (**) above. This is our main motivation for introducing the concept of a *standard loss function.* It is obtained essentiallly by a reparametrization of the upper tangets a_t of k such that the correspondence becomes linear (Sect. 2).

Another problem in applying the results of part I comes from the difficulty of minimizing upper Bayes risks of approximate models over randomized procedures. For technical reasons, however, minimization over non-randomized decision procedures is feasible, hence one is forced to find out when randomization can be eliminated. It can, – at least for *continuous* standard loss functions, which turns out to be sufficient for our purposes. Continuity is essential since our proof requires a homotopy argument.

With these tools we can proceed to examples. The Huber-Strassen result is an immediate corollary. It is relatively easily obgained since it remains in the binary case ($|\Theta|=2$). A more involved application concerns ε-contamination models in a multiparameter setting ($|\Theta|>2$). Here one is given an exact model (P_θ), but one safeguards against the worst case of the form

$$(Q_\theta) = ((1-\varepsilon)\cdot P_\theta + \varepsilon \cdot H_\theta).$$

Optimal procedures consist of deciding as in the exact case (P_θ) after having cut away those elements of the decision space which would result in too high losses. We call this process *decision-censoring*. There exist loss functions which do not require decision-censoring for a given contamination model. In this case the same decision rule which is optimal under the uncontaminated model (P_θ) will remain optimal even under contamination. This is the content of Sect. 5, in which we will have to go into some nontrivial convex optimization.

As an application, we consider robust classification in Sect. 6. Here it is natural to investigate simultaneous least favorability for families of priors, but this case can be treated the same way as the case of a class of loss functions (see remark above). Curiously enough, for technical reasons it is most natural to consider sets of priors which are contamination neighborhoods in the sense of distributions on the parameter set Θ. We are thus dealing with twofold contamination! Thus let us fix a prior distribution and an amount η of contamination for this prior, and also an amount ε of contamination for a model (P_θ). Then the result is the following:

There exists an experiment in the ε-neighborhood of (P_θ) which is least favorable simultaneously for all elements of the η-neighborhood of the given prior, - if the optimal classification rule for the uncontaminated model (P_θ) does not break down.

Here, breaking down means that throwing away the data and deciding a priori is preferable, which corresponds to decision-censoring the classification loss function. The existence of such an experiment can be proved by mixing the standard loss functions which belong to the η-contaminated priors. By this mixing procedure we obtain a continuous standard loss function which does not have to be decision-censored under ε-contamination of (P_θ) iff none of the contaminated priors cause breakdown. But from the fact that no decision-censoring is required, additivity of the upper Bayes risk on the involved standard loss functions follows. Thus the theory of part I becomes applicable, and we have proved a theorem on the existence of simultaneously least favorable experiments for robust classification.

It is unfortunate that the present theory appears relatively involved, although the ideas underlying our developments are basically simple and can be expressed in intuitive notions such as decision-censoring and breakdown of a classification rule. But from a mathematical point of view, we expand decision theory into the framework of sublinear functionals, abandoning the nice linearity of expectations. The conceptual gain in this endeavor consists in the robustization of decision theory through worst-case considerations based on the notions of approximate models (=parametrized families of sublinear functionals).

2. Definition of Standard Loss Functions

Recall again the meaning of our crucial concave functions k from the previous section. The value of $k(z)$ at $z=(z_\theta)=(q_\theta(y))$ is the minimal a posteriori expected loss for the observation y. In this way, to every loss function is assigned a concave function k, but different loss functions may lead to the same k. This means that they are equivalent from the point of view of Bayes problems (for a fixed prior, here chosen to be uniform on Θ). However, in every equivalence class of loss functions, one can pick a canonical element as follows:

Given a concave function k arising from a bounded loss function, there exists for every point $\zeta \in K$ an affine function a_ζ which is an upper tangent of k at ζ:

$$k(z) \leqq a_\zeta(z) \quad \forall z \in K, \quad \text{and} \quad k(\zeta) = a_\zeta(\zeta) \qquad (*)$$

The affine function a_ζ can be written as a linear combination

$$a_\zeta(z) = \sum_\theta a_{\zeta,\theta} \cdot z_\theta.$$

Now, choose the simplex K as a decision space and define a loss function $\tilde{W}_\theta$ by

$$\tilde{W}_\theta(\zeta) = a_{\zeta,\theta}.$$

By $\tilde{W}(\zeta) = (\tilde{W}_\theta(\zeta))$ we denote the whole vector of losses[2]. The properties (*) can be summarized as follows:

$$\sum_\theta \tilde{W}_\theta(z) \cdot z_\theta = \inf_{\zeta \in K} \sum_\theta \tilde{W}_\theta(\zeta) \cdot z_\theta \tag{**}$$

or in scalar product notation:

$$\langle \tilde{W}(z), z \rangle = \inf_{\zeta \in K} \langle \tilde{W}(\zeta), z \rangle.$$

Loss functions on the decision space $T = K$ satisfying (**) will be called *standard loss functions*. They share some convenient features. E.g. the search for optimal procedures is trivialized in exact models, and a useful additivity property holds, as the next two propositions show:

Proposition 2.1. *For every exact model* (Q_θ) *and every standard loss function* $(\tilde{W}_\theta)$, *the density vector* $q = (q_\theta)$ *with respect to the dominating measure* $\sum_\theta Q_\theta$ *is a Bayes procedure. Especially for every standard model* (S_θ) *the identity map* id: $K \to K$ *is a Bayes procedure.*

Proof. This remark follows from the fact that a Bayes procedure consists in minimizing the posterior expected loss:

$$a_\zeta(q(y)) = \sum_\theta \tilde{W}_\theta(\zeta) \cdot q_\theta(y).$$

For standard loss functions this is achieved by the procedure $y \to \zeta = q(y)$ as is seen from the defining property (**). In Sect. 5 of part I, it was seen that the identity map on K is a density vector for every standard model. □

Proposition 2.2. *Given two standard loss functions* $\tilde{W}_\theta^1(\zeta)$ *and* $\tilde{W}_\theta^2(\zeta)$, *the sum* $\tilde{W}_\theta(\zeta) = \tilde{W}_\theta^1(\zeta) + \tilde{W}_\theta^2(\zeta)$ *is also a standard loss function, and the corresponding concave functions depend additively on them:* $k(z) = k^1(z) + k^2(z)$.

Proof. Immediate consequences of the definition (**). □

Note that for general loss functions on a fixed but arbitrary decision space, the concave functions depend only superadditively on their loss functions: $k(z) \geq k^1(z) + k^2(z)$. The first half of Proposition 2.2 states that the standard loss functions form a convex cone. On a more basic level, the convenience of

[2] Since the upper tangents a_ζ are often not unique, measurability problems could arise. But it is always possible to select a set of tangents a_ζ such that the dependence on ζ becomes measurable

standard loss functions stems also from the fact that they reside on a common decision space K and that there exists a Bayes-equivalent standard loss function for any given loss function.

For a finite decision space $T=\{1,2,\ldots,m\}$, one can provide a direct construction as follows: Let $W_\theta(t)$ be the loss matrix and choose a measurable partition $(A_t)_{t=1\ldots m}$ of K such that

$$A_t \subset \{z \in K \mid \sum_\theta W_\theta(t)\cdot z_\theta = \min_{t'=1\ldots m} \sum_\theta W_\theta(t')\cdot z_\theta\}$$

i.e., for $z\in A_t$ the decision t is optimal since it minimizes the a posteriori expected loss. Then a corresponding standard loss function can be defined by

$$\tilde{W}_\theta(z)=\sum_{t=1}^{m} W_\theta(t)\cdot 1_{A_t}(z)$$

The concave function k arising from a loss matrix on a finite decision set is not smooth as it has edges resulting from the finite minimization. In contrast, if a standard loss function is continuous it must stem from an infinite decision space. This might appear unnatural for finite experiments, but it turns out that continuous standard loss functions are indispensable technical tools as we will see later.

We give a couple of examples which are not arbitrarily chosen. Later on, we will prove theorems on the existence of simultaneously least favorable experiments for both, the first one leading to the Huber-Strassen theorem.

1) Simple Testing. Consider testing experiments, i.e. parameter sets $\Theta=\{1,2\}$. The elements of the unit simplex K are denoted as usual by $z=(z_1,z_2)$, where $z_\theta\geqq 0$ and $z_1+z_2=1$. Let the losses be zero for correct decision, and α_1, respectively $\alpha_2(\geqq 0)$ for erroneously rejecting 1, respectively 2. Using the decision space $T=\{1,2\}$, we are considering the following loss matrix:

$$\begin{pmatrix} W_1^\alpha(1) & W_2^\alpha(1) \\ W_1^\alpha(2) & W_2^\alpha(2) \end{pmatrix} = \begin{pmatrix} 0 & \alpha_2 \\ \alpha_1 & 0 \end{pmatrix}.$$

We obtain then a concave function

$$k^\alpha(z)=\min(\alpha_1 z_1, \alpha_2 z_2), \qquad \alpha=(\alpha_1,\alpha_2).$$

A corresponding standard loss function is:

$$\tilde{W}_1^\alpha(z)=\alpha_1\cdot 1_{[\alpha_1 z_1<\alpha_2 z_2]}(z)$$
$$\tilde{W}_2^\alpha(z)=\alpha_2\cdot 1_{[\alpha_1 z_1\geqq\alpha_2 z_2]}(z).$$

The following generalizes this example:

2) Classification. Let $\Theta=\{1,2,\ldots,m\}$ be an arbitrary finite parameter set. Take as a decision space $T=\{1,2,\ldots,m\}$, and assign losses $\alpha_\theta\geqq 0$ for incorrectly rejecting θ, i.e. for misclassification of θ:

$$\begin{bmatrix} W_1^\alpha(1) & W_2^\alpha(1)\ldots \\ W_1^\alpha(2) & W_2^\alpha(2)\ldots \\ W_1^\alpha(3) & W_2^\alpha(3)\ldots \\ \ldots & \ldots \quad \ldots \end{bmatrix} = \begin{bmatrix} 0 & \alpha_2\ldots \\ \alpha_1 & 0\ldots \\ \alpha_1 & \alpha_2\ldots \\ \ldots & \ldots\ldots \end{bmatrix}.$$

The concave function is:

$$k^\alpha(z) = \min_{t=1\ldots m} \left(\sum_\theta \alpha_\theta z_\theta - \alpha_t z_t\right) = \sum_\theta \alpha_\theta z_\theta - \max_t \alpha_t z_t$$

Again we denote $\alpha = (\alpha_1, \alpha_2, \ldots)$. To find a standard loss function for k^α, choose any measurable partition (A_θ) of K satisfying

$$A_\theta \subset \{z \in K \mid \alpha_\theta z_\theta = \max_{\theta'} \alpha_{\theta'} z_{\theta'}\}.$$

Then we can put:

$$\tilde{W}_\theta^\alpha(z) = \alpha_\theta \cdot 1_{A_\theta^C}(z) = \alpha_{\theta'}(1 - 1_{A_\theta}(z))$$

By constructing a standard loss function for a given decision problem, we actually solve it, as becomes clear from the examples. Only at this stage the convex structure of a decision problem is reduced to its purest form, where we deal just with a posteriori expected losses as upper tangents to the minimal a posteriori expected losses.

3. Continuous Standard Loss Functions

In this section we will elaborate on the important role of continuous standard loss functions. The central topic will be elimination of randomization. For exact models it is a very simple fact that Bayes procedures can be picked nonrandomized. For approximate models, however, matters escalate considerably. The need for nonrandomized procedures is ultimately urgent since minimization of upper Bayes risks over randomized procedures is not feasible, but the possibility of elimination of randomization can be proved only for continuous standard loss functions. Furthermore, the proof involves some topological machinery. What we will show is that for a given randomized procedure, there always exists a nonrandomized procedure which incurs no worse a loss. Unlike risk, loss does not involve any models, hence the result is applicable no matter what the model is (approximate or exact).

Continuous standard loss functions arise in a very natural way by mixing discontinuous ones. Similar to convolution, the mixtures inherit smoothness properties from the mixing measures. The resulting continuous standard loss functions will bring minimal Bayes risks and standard measures to bear, replacing quantities like the Kullback-Leibler information or Hellinger transforms.

At the heart of this section is a kind of multidimensional intermediate value theorem (topological Lemma 3.4). Unlike other theorems of this type it cannot be deduced from the Brouwer fixed point theorem. This is why we have to go through some basic arguments involving topological degrees of continuous mappings. For binary experiments we need actually just the simple intermediate value theorem. The reason why this does not show up in published proofs of the Huber-Strassen theorem is that in this case it is possible to construct a nonrandomized test statistic more or less explicitly [4, Sect. 3]. For general finite approximate models, things are considerably more intricate and we cannot expect the analogue of Huber-Strassen's Sect. 3 to hold. We have to make an effort equivalent to the present section.

Besides the results about elimination of randomization, we will also derive facts (3.9 and 3.10) which explain why convex optimization theory will come into play later on in Sect. 5.

Theorem 3.1. *Let $\tilde{W}_\theta(z)$ be a standard loss function which depends continuously on z. Then there exists for any randomized procedure σ: $Y \to T = K$ a nonrandomized procedure $\tilde{\sigma}$ with a loss which is not worse:*

$$\tilde{\sigma}(\tilde{W}_\theta) \leqq \sigma(\tilde{W}_\theta) \quad \forall\, \theta \in \Theta, \qquad i.e., \quad \tilde{W}_\theta(\tilde{\sigma}(y)) \leqq \int \sigma(y, dz')\, \tilde{W}_\theta(dz')$$

i.e., randomization can be eliminated, no matter what the experiment is.

Proof. In Proposition 3.3 below, we will show that the sets

$$J_y = \{z \in K \mid \tilde{W}_\theta(z) \leqq \int \sigma(y, dz')\, \tilde{W}_\theta(z') \qquad \forall\, \theta \in \Theta\}$$

are nonempty for all $y \in Y$. The remaining problem is to show that for each $y \in Y$ one can pick $z = \tilde{\sigma}(y) \in J_y$ such that $\tilde{\sigma}$: $Y \to K$ is measurable. This can be achieved by a technical device which can be found in Parthasarathy [8, Theorem 4.1] It amounts to replacing K by a compact subset A of the unit interval $[0, 1]$. In A one has natural ways of picking elements of subsets, e.g. the supremum or infimum. We will first transform our problem somewhat in order to defer the technicalities to a lemma which has the flavor of an implicit function theorem. Let

$$g(y, z) = \sum_\theta \left[\int \sigma(y, dz')\, \tilde{W}_\theta(z') - \tilde{W}_\theta(z)\right]^-$$

where $c^- = -c$ for $c \leqq 0$ and $c^- = 0$ for $c > 0$ as usual. We notice that g is measurable in y for every z, continuous in z for every y, and

$$J_y = \{z \in K \mid g(y, z) = 0\}.$$

The proof of Theorem 3.1 is finished by the following:

Lemma 3.2. *Let Y be a measurable space, Z a compact metric space, $g(y, z)$ a real valued function on $Y \times Z$ which is measurable in y for every $z \in Z$ and continuous in z for every $y \in Y$. If finally $\{z \in K \mid g(y, z) = 0\}$ is nonempty for every $y \in Y$, then there exists a measurable mapping f: $Y \to Z$ which solves $g(y, z) = 0$:*

$$g(y, f(y)) = 0 \qquad \forall\, y \in Y.$$

Proof. By [8, Theorem 4.1] there exists a closed (hence compact) subset A of $[0,1]$ and a continuous map h from A onto Z. By means of h, we shift g from Z to A:

$$g^*(y,a)=g(y,h(a)),$$

which is again measurable in y, continuous in a, and

$$J_y^*=\{a\in A\,|\,g^*(y,a)=0\}$$

is nonempty and compact for all $y\in Y$. We put

$$f^*(y)=\inf J_y^*,$$

and notice that $f^*(y)\in J_y^*$ due to compactness, hence

$$g^*(y,f^*(y))=0 \qquad \forall\, y\in Y.$$

Also, f^* is measurable: for this, introduce a countable dense subset A' of A, and consider the following equalities:

$$\begin{aligned}\{y\in Y\,|\,f^*(y)\leqq c\}&=\{y\in Y\,|\,\exists\, a\in A\colon a\leqq c, g^*(y,a)=0\}\\ &=\{y\in Y\,|\inf_{a\leqq c,\,a\in A'}|g^*(y,a)|=0\}\\ &=\bigcap_n \bigcup_{a\leqq c,\,a\in A'}\left\{y\in Y\,\middle|\,|g^*(y,a)|\leqq\frac{1}{n}\right\}.\end{aligned}$$

For the second equality we used continuity of g^* in a. Since g^* is measurable in y and all set operations are countable, f^* has to be measurable.

The measurable function $f(y)=h(f^*(y))$ solves $g(y,z)=0$. □

Proposition 3.3. *Let $\tilde{W}_\theta(z)$ be a continuous standard loss function and $\mu(dz)$ a probability measure on K. Then there exists $z\in K$ such that*

$$\tilde{W}_\theta(z)\leqq\int\mu(dz')\,\tilde{W}_\theta(z') \qquad \forall\,\theta\in\Theta.$$

Proof. Denote by $\tilde{W}(z)=(\tilde{W}_\theta(z))$ the vector with components $\tilde{W}_\theta(z)$, and by $\tilde{W}(\mu)$ its componentwise μ-integral. Now, the method of proof is to shift everything to the unit sphere $S^{n-1}=\{\zeta\in\mathbb{R}^n\,|\,\|\zeta\|=1\}$, where $n=|\Theta|$. We may assume $\tilde{W}(z)\neq\tilde{W}(\mu)\ \forall\, z\in K$, since otherwise nothing is to prove. Hence we can define a map Δ by

$$\Delta(\zeta)=\frac{\tilde{W}(\mu)-\tilde{W}(z)}{\|\tilde{W}(\mu)-\tilde{W}(z)\|} \quad \text{for } \zeta=\frac{z}{\|z\|} \text{ and } z\in K.$$

Let K' be the image of the unit simplex under normalization:

$$K'=\{\zeta\in S^{n-1}\,|\,\zeta_\theta\geqq 0\ \forall\,\theta\}. \qquad (*)$$

The map $\Delta\colon K'\to S^{n-1}$ is continuous and satisfies $\langle\Delta(\zeta),\zeta\rangle\geqq 0$. This follows from continuity of $\tilde{W}$ and the definition of a standard loss function. Since

$$\tilde{W}_\theta(z)\leqq\tilde{W}_\theta(\mu)\ \forall\,\theta\in\Theta \Leftrightarrow \Delta(\zeta)\in K' \quad \text{for } \zeta=\frac{z}{\|z\|},$$

our problem is to find $\zeta \in K'$ such that $\Delta(\zeta) \in K'$. We may thus summarize the required argument in the following:

Topological Lemma 3.4. *Let K' be the nonnegative face of the unit sphere in $\mathbb{R}^n$. If $\Delta\colon K' \to S^{n-1}$ is a continuous mapping which satisfies*

$$\langle \Delta(\zeta), \zeta \rangle \geqq 0 \qquad \forall\, \zeta \in K',$$

then $\Delta(K') \cap K'$ is nonempty.

The *proof* can be deferred to still another lemma with stronger assumptions and stronger conclusion. If we denote by

$$\begin{aligned} \partial K' &= \{\zeta \in K' \mid \exists\, \theta \in \Theta : \zeta_\theta = 0\} \\ &= \{\zeta \in S^{n-1} \mid \forall\, \theta \in \Theta : \zeta_\theta \geqq 0, \exists\, \theta \in \Theta : \zeta_\theta = 0\} \end{aligned} \tag{**}$$

the boundary of K' as a subspace of S^{n-1}, we recognize that nothing is to prove if $\Delta(\partial K') \cap K'$ is nonempty. Thus the above lemma is proved if we can show the following lemma which is of interest in itself:

Topological Lemma 3.5. *If in addition to the assumptions of the previous lemma $\Delta(\partial K') \cap K'$ is empty, then $K' \subset \Delta(K')$.*

Proof. We have to show that each point in K' is taken on at least once by Δ. A suitable tool for this purpose is the topological degree of continuous mappings. Let Ψ: *closure* $(\Omega) \to \mathbb{R}^p$ be a continuous mapping, where $\Omega \subset \mathbb{R}^p$ is open and bounded. For any point $x \notin \Psi(\partial\Omega)$ the degree $\deg(\Psi, x)$ is well defined, and the two following basic properties hold true:

1) If $\deg(\Psi, x) \neq 0$, then $x \in \Psi(\Omega)$.

2) If $\Psi = \mathrm{id}|\Omega$, then $\deg(\Psi, x) = 1$ for $x \in \Omega$.

3) If $\Phi\colon [0,1] \times \partial\Omega \to \mathbb{R}^p$ is a continuous homotopy, and if Ψ^0 and Ψ^1 are both maps *closure* $(\Omega) \to \mathbb{R}^p$ satisfying $\Psi^0|\partial\Omega = \Phi(0,\cdot)$ and $\Psi^1|\partial\Omega = \Phi(1,\cdot)$, then we have

$$\deg(\Phi(t,\cdot), x) = \text{constant} \quad \forall\, t \in [0,1]$$

for points $x \in \mathbb{R}^p$ satisfying $x \notin \Phi(t, \partial\Omega)$ $\forall\, t \in [0,1]$.

See [1, p. 65–67], and [9, p. 80–81], for (approximately) these facts. In our application we will actually be working with the map $\Delta\colon K' \to S^{n-1}$. We plan to show that its restriction $\Delta|\partial K'$ is homotopic to $\mathrm{id}|\partial K'$ within a punctured sphere. To this situation, 3) can be applied since the punctured sphere is homeomorphic to $\mathbb{R}^{n-1}$. Furthermore, we will construct the homotopy such that it does not touch the interior of K'. By 2), we conclude that all points $\zeta \in \mathrm{interior}(K')$ have degree 1 with regard to the mapping Δ. By 1), any such point is attained by Δ. Since $\Delta(K')$ is compact and hence closed, we even obtain $K' \subset \Delta(K')$.

Now for the details: First we have to show that the map Δ lives in a punctured sphere. Let $e = -1/\sqrt{n} \cdot (1, 1, \ldots, 1)$. The point e cannot be in the image of Δ: we have $\langle e, \zeta \rangle < 0\, \forall\, \zeta \in K'$, but by assumption we also have $\langle \Delta(\zeta), \zeta \rangle \geqq 0$; thus e is not in the image of Δ.

In the main step of the proof, we will construct a sequence of homotopies which deform $\Delta|\partial K'$ into $\mathrm{id}|\partial K'$ without crossing the interior of K' and the point e. We will break it up by first deforming $\Delta|\partial K'$ into a mapping Γ: $\partial K' \to \partial K'$ which can be shown to be homotopic to $\mathrm{id}|\partial K'$. This latter step will be deferred to yet another topological lemma, in whose proof we deform Γ into $\mathrm{id}|\partial K'$ within $\partial K'$, thus neither crossing e nor the interior of K'.

In order to stay within the unit sphere, we will have to normalize all the vectors we use. For ease of notation we let the symbol $\propto$ denote equality of vectors up to normalization. We will always have to make sure that normalization is possible, i.e., the vectors in question do not vanish (a). Then we have to prove that the side constraints are satisfied: neither the interior of K' (b), nor e (c) is crossed.

Let $\Phi_\theta(s,\zeta) \propto \Delta_\theta(\zeta) - s \cdot d(\zeta)$, where $d(\zeta) = \min_\theta \Delta_\theta(\zeta)$. The homotopy parameter is $s \in [0,1]$, whereas $\zeta \in \partial K'$. Notice that $d(\zeta) < 0$ since $\Delta(\zeta)$ is not in K' by assumption. This homotopy deforms $\Phi(s=0,\cdot) = \Delta(\cdot)|\partial K'$ into some mapping $\Gamma(\cdot) = \Phi(s=1,\cdot)$. Now we have to show that the requirements (a)-(c) are satisfied:

(a) Φ is well defined: if $\Delta_\theta(\zeta) - s \cdot d(\zeta) = 0$ for all θ and some $\zeta \in \partial K'$, then all components of $\Delta(\zeta)$ must be identical and negative, hence $\Delta(\zeta) = e$, which is impossible.

(b) Φ does not touch the interior of K': for the component θ at which $\Delta_\theta(\zeta)$ attains the minimum $d(\zeta)$, we have $\Delta_\theta(\zeta) = d(\zeta)$, hence $\Delta_\theta(\zeta) - s \cdot d(\zeta) = (1-s) \cdot d(\zeta) \leqq 0$, which shows that Φ stays outside the interior of K'.

(c) Φ does not touch e: if $\Delta_\theta(\zeta) - s \cdot d(\zeta)$ were constant for all θ at some ζ and s, then $\Delta_\theta(\zeta)$ would have to be constant, too, hence $\Delta(\zeta) = \pm e$ or 0, neither of which is possible.

We have $\Gamma(\partial K') \subset \partial K'$ since

$$\Gamma_\theta(\zeta) = \Phi_\theta(s=1,\zeta) \propto \Delta_\theta(\zeta) - d(\zeta) \geqq 0,$$

and $=0$ for a minimizing component θ. The map Γ satisfies $\langle \Gamma(\zeta), \zeta \rangle > 0$. To prove this, assume the opposite:

$$0 \geqq \langle \Gamma(\zeta), \zeta \rangle \propto \langle \Delta(\zeta), \zeta \rangle - d(\zeta)$$

and hence $0 > d(\zeta) \geqq \langle \Delta(\zeta), \zeta \rangle$, which contradicts the assumptions on Δ.

We would have to show now that Γ can be deformed into $\mathrm{id}|\partial K'$ observing (b) and (c). This will be achieved by the following:

Topological Lemma 3.6. *Let $\partial K'$ be the boundary of the nonnegative face K' on the unit sphere S^{n-1}. Let Γ: $\partial K' \to \partial K'$ be a continuous map satisfying*

$$\langle \Gamma(\zeta), \zeta \rangle > 0 \qquad \forall \zeta \in \partial K',$$

then Γ is homotopic to $\mathrm{id}|\partial K'$ *within $\partial K'$.*

Proof. In a first step, we will deform Γ within $\partial K'$ into a map Λ which satisfies $\Lambda_\theta(\zeta) = 0$ whenever $\zeta_\theta = 0$, i.e., $\Lambda(\zeta)$ comes to lie on the same side of $\partial K'$ as ζ. The second step will straightforwardly deform Λ within $\partial K'$ into $\mathrm{id}|\partial K'$. In both steps we will have to show that the homotopies are welldefined (a) and remain within $\partial K'$ (b).

(1) $\Phi^1_\theta(s,\zeta)\propto\min(1,1-s+\zeta_\theta)\cdot\Gamma_\theta(\zeta)$ for $s\in[0,1]$ and $\zeta\in\partial K'$. This homotopy deforms $\Gamma(\cdot)=\Phi^1(s=0,\cdot)$ into some mapping $\Lambda(\cdot)=\Phi^1(s=1,\cdot)$ which has the above mentioned property: $\Gamma_\theta(\zeta)=0$ for $\zeta_\theta=0$.

(1a) Φ^1 is welldefined: assume that all components $\min(1,1-s+\zeta_\theta)\cdot\Delta^1_\theta(\zeta)$ vanish for some s and ζ; we notice two facts: (i) we have $\Gamma_\theta(\zeta)\geqq 0$ generally, as $\Gamma(\partial K')\subset\partial K'$; (ii) for strictly positive $\Gamma_\theta(\zeta)$ we must have $1-s+\zeta_\theta=0$, and thus $\zeta_\theta=-(1-s)\leqq 0$; (i) and (ii) lead to $\langle\Gamma(\zeta),\zeta\rangle=\sum_\theta\Gamma_\theta(\zeta)\cdot\zeta_\theta\leqq 0$, which contradicts the assumptions on Γ.

(1b) Φ^1 acts within $\partial K'$, since $\Gamma(\zeta)\in\partial K'$ and $\min(1,1-s+\zeta_\theta)\geqq 0$.

(2) $\Phi^2_\theta(\zeta)\propto(1-s)\cdot\Lambda_\theta+s\cdot\zeta_\theta$ for $\zeta\in\partial K'$. We have $\Phi^2(s=0,\cdot)=\Gamma(\cdot)$ and $\Phi^2(s=1,\cdot)=\mathrm{id}\,|\,\partial K'$.

This homotopy is possible only because of the previous step (1), otherwise it would cross the interior of K'.

(2*a*) Φ^2 is welldefined: $(1-s)\cdot\Lambda_\theta(\zeta)+s\cdot\zeta_\theta=0\ \forall\,\theta\in\Theta$ is impossible for $s=1$, since $\zeta\in S^{n-1}$. For $s<1$ we would have $\Lambda_\theta(\zeta)=-\frac{s}{(1-s)}\zeta_\theta\,\forall\,\theta\in\Theta$, which is possible only for $s=\frac{1}{2}$ due to normalization: $\Lambda(\zeta)=-\zeta$. This is in contradiction to $\zeta,\Lambda(\zeta)\in\partial K'$.

(2b) Φ^2 stays within $\partial K'$, since $\zeta,\Lambda(\zeta)\in\partial K'$ and $\Lambda_\theta(\zeta)=0$ whenever $\zeta_\theta=0$. □

Lemma 3.5 can be used to derive the following result, which deals with a type of continuous standard loss functions which arise, e.g., by mixing the standard loss functions of classification problems (Sect. 2, Example 2). These loss functions make intuitive sense, since they put maximal θ-loss $\tilde{W}_\theta(z)$ on those z which are least likely under θ, namely $z_\theta=0$. Recall that $n\cdot z_\theta$ is the θ-density in a standard experiment with regard to a standard measure (part I, Sect. 5). The content of Proposition 3.7 is that the convex surface formed by the loss vectors $\{\tilde{W}(z)\,|\,z\in K\}$ is so wide that any mixture of surface points is still fully below the surface, where "below" is taken in the sense of the usual partial ordering in $\mathbb{R}^n$:

Proposition 3.7. *Let $\tilde{W}_\theta(z)$ be a nontrivial continuous standard loss function in the sense that*

$$\inf_{z'\in K}\tilde{W}_\theta(z')<\sup_{z'\in K}\tilde{W}_\theta(z')\quad\forall\,\theta\in\Theta,$$

and which also satisfies

$$\tilde{W}_\theta(z)=\sup_{z'\in K}\tilde{W}_\theta(z')\qquad\forall\,\theta\in\Theta\qquad\text{for } z_\theta=0.$$

Then for every probability measure μ on K and every direction $\tilde{z}\in K$, there exists $z\in K$ and $\tau\geqq 0$ such that

$$\tilde{W}_\theta(z)=\int\mu(dz')\,\tilde{W}_\theta(dz')-\tau\cdot\tilde{z}\qquad\forall\,\theta\in\Theta.$$

This implies but is stronger than:

$$\tilde{W}_\theta(z)\leqq\int\mu(dz')\,\tilde{W}_\theta(z')\qquad\forall\,\theta\in\Theta.$$

E.g. by choosing $\tilde{z}=e^{(n)}=(0,0,\dots,1)$ *we obtain the existence of* $z\in K$ *such that*

$$\tilde{W}_\theta(z)=\int\mu(dz')\,\tilde{W}_\theta(z') \qquad \forall\,\theta=1,2,\dots,n-1$$
$$\tilde{W}_n(z)\leqq\int\mu(dz')\,\tilde{W}_n(z').$$

Proof. We use the same notation as before. The assertion is obviously equivalent to $K'\subset\Delta(K')$. So we would like to apply Lemma 3.5. Unfortunately this is not readily possible: the assumption $\tilde{W}_\theta(z)=\sup\limits_{z'}\tilde{W}_\theta(z')$ for $z_\theta=0$ translates only into $\Delta_\theta(\zeta)\leqq 0$ for $\zeta_\theta=0$. However, we can replace Δ by slightly perturbed versions Δ^ε which we may obtain by using the measures $\mu^{(\varepsilon)}=(1-\varepsilon)\cdot\mu+\varepsilon\cdot\frac{1}{n}\sum\limits_\theta\delta_{e^{(\theta)}}$. The term $\frac{1}{n}\sum\limits_\theta\delta_{e^{(\theta)}}$ is the mixture of the point masses at the corners $e^{(\theta)}=(0,\dots,1,\dots,0)$ of K. This serves the following purpose: the function $\tilde{W}_\theta(z)$ takes on its minimum at $z=e^{(\theta)}$, and we have by assumption $\inf\tilde{W}_\theta<\sup\tilde{W}_\theta$; thus for $\varepsilon>0$ we get $\int\mu^{(\varepsilon)}(dz')\,\tilde{W}_\theta(z')<\sup\tilde{W}_\theta$, where strict inequality is due to the point mass at $e^{(\theta)}$. Then the mappings $\Delta^{(\varepsilon)}(z)\propto\tilde{W}(\mu^{(\varepsilon)})-\tilde{W}(z)$ are well defined for small enough $\varepsilon>0$, and they satisfy $\Delta_\theta^{(\varepsilon)}(\zeta)<0$ for $\zeta_\theta=0$, hence Lemma 3.5 becomes applicable. We conclude $K'\subset\Delta^{(\varepsilon)}(K')$ for $\varepsilon>0$ ($\leqq 1$). The following elementary lemma allows to conclude $K'\subset\Delta(K')$:

Lemma 3.8. *Assume* U, V *metric spaces,* U *compact. Let* $\Delta^{(n)}$: $U\to V$ *be a sequence of continuous mappings which converge uniformly to* Δ: $U\to V$. *If* $A\subset\Delta^{(n)}(U)$ *for all* n, *then also* $A\subset\Delta(U)$.

Proof. A simple subsequence argument applied to preimages $u^{(n)}$ of $a\in A$ under $\Delta^{(n)}$. □

As a corollary of Proposition 3.3, we will prove a convexity result. It will be crucial later in Sect. 5.

Proposition 3.9. *If* $\tilde{W}$ *is a continuous standard loss function, then an affine function* $a(\cdot)$ *is above* $k(\cdot)$ *on* K *iff* $a(\cdot)$ *is above some* $\tilde{W}$*-tangent, i.e.,* $\exists\,z\in K$: $\langle\tilde{W}(z),\cdot\rangle\leqq a(\cdot)$ *on* K.

Remarks. a) Recall that any affine function $a(\cdot)$ on K is given by a form $a(z)=\sum\limits_\theta a_\theta\cdot z_\theta$, where $a_\theta=a(e^{(\theta)})$. An affine function may as well be identified with a vector $(a_\theta)\in\mathbb{R}^\Theta$, and then we can write $a(z)=\langle a,z\rangle$. Now let:

$$A_k=\{a\in\mathbb{R}^\Theta\,|\,\langle a,\cdot\rangle\geqq k(\cdot)\}$$
$$A_W=\{a\in\mathbb{R}^\Theta\,|\,\exists\,z\in K\colon\ \langle a,\cdot\rangle\geqq\langle\tilde{W}(z),\cdot\rangle\}.$$

The above corollary can be stated as follows: $A_k=A_W$. From the definition of standard loss functions follows that $A_W\subset A_k$ without any conditions on $\tilde{W}$, thus the nontrivial part is the opposite inclusion.

b) A side result is that the set A_W is convex. This is actually the main step in the proof, where Proposition 3.3 is needed.

c) The following geometrical fact can also be deduced from Proposition 3.9: at interior points $z\in K$, there exists only one tangent of k, namely $\tilde{W}(z)$. At boundary points z, the tangent $\tilde{W}(z)$ is the lowest among all tangents.

The *proof* of 3.9 will be in several steps.

1) Both sets A_k, A_W are convex: this is clear for A_k; for convexity of A_W, take $a^{(1)}$, $a^{(2)} \in A_W$; there exist $z^{(1)}, z^{(2)} \in K$ such that $\tilde{W}_\theta(z^{(i)}) \leqq a_\theta\ \forall\, \theta\ (i=1,2)$. By 3.3, there exists z such that

$$\tilde{W}_\theta(z) \leqq \alpha \cdot \tilde{W}_\theta(z^{(1)}) + (1-\alpha) \cdot \tilde{W}_\theta(z^{(2)}),$$

using $\mu = \alpha \cdot \delta_{z^{(1)}} + (1-\alpha) \cdot \delta_{z^{(2)}}$; thus $\tilde{W}_\theta(z) \leqq a_\theta$ for $a = \alpha \cdot a^{(1)} + (1-\alpha) \cdot a^{(2)}$.

2) Both A_k and A_W are closed. To show this in case of A_W, one needs compactness of K and continuity of $\tilde{W}$.

3) Now we show that all linear forms on $\mathbb{R}^\Theta$ attain the same minimum values on both sets A_k, A_W: since both are unbounded upwards, one has $\inf\limits_{a \in A} \langle \eta, a \rangle = -\infty$ for both $A = A_k$ and $A = A_W$, if η has some strictly negative components. Hence we can assume $\eta \in K$. Then $\inf\limits_{a \in A} \langle \eta, a \rangle = k(\eta)$ for both $A = A_k, A_W$.

From 1)-3) we conclude $A_k = A_W$. □

Finally we will establish that continuity of a standard loss function implies continuous differentiability of the corresponding concave function, as one might expect since the standard loss function forms upper tangents:

Proposition 3.10. *Denote the directional derivative of the concave function k at z in the direction $h = z' - z$ by*

$$D_{z;h}\, k(z) = \lim_{\tau \downarrow 0} \frac{1}{\tau} (k(z + \tau h) - k(z)).$$

If the standard loss function $\tilde{W}_\theta$ is continuous, we have

$$D_{z;h}\, k(z) = \langle \tilde{W}(z), h \rangle.$$

The *proof* involves a twofold application of the defining inequality for standard loss functions.

$$\frac{k(z+\tau h) - k(z)}{\tau} = \frac{\langle \tilde{W}(z+\tau h), z+\tau h \rangle - \langle \tilde{W}(z), z \rangle}{\tau}.$$

An application to the term $\langle \tilde{W}(z+\tau h), z+\tau h \rangle$ gives as an upper bound:

$$\leqq \frac{\langle \tilde{W}(z), z+\tau h \rangle - \langle \tilde{W}(z), z \rangle}{\tau} = \langle \tilde{W}(z), h \rangle.$$

Similarly an application to the term $\langle \tilde{W}(z), z \rangle$ gives a lower bound:

$$\geqq \frac{\langle \tilde{W}(z+\tau h), z+\tau h \rangle - \langle \tilde{W}(z+\tau h), z \rangle}{\tau} = \langle \tilde{W}(z+\tau h), h \rangle$$

Since W is continuous, the right hand side converges to $\langle \tilde{W}(z), h \rangle$ as $\tau \downarrow 0$. □

4. The Huber-Strassen Theorem

In our formulation of the Huber-Strassen theorem [4], we deal with certain approximate binary experiments, more specifically pairs of 2-alternating upper expectations (v_1, v_2). The theorem states the existence of universally least favorable pairs (Q_1, Q_2) under the approximate model. The proof can be relatively streamlined due to the theory we developed in part I and the previous section. We will use only arguments which generalize to arbitrary finite experiments. Particularly, we will not have to assume Huber-Strassen's construction of minimax test statistics [4, p. 256 bottom] in order to prove the existence of least favorable pairs (see the introductory remarks to the previous section).

One word should be said about the existence of universally least favorable experiments in case of more than two parameter values ($|\Theta|>2$). A literal analogue to the Huber-Strassen theorem is not going to hold anymore. But partial results of this type can be expected to be true if one suitably shrinks the set of decision problems (loss functions) for which one wishes to find simultaneously least favorable experiments. We will present an instance of such a theorem for some sufficiently small classes of classification loss functions under ε-contaminated experiments later in Sect. 5. Another method to generate such results could feasibly consist in allowing "small" upper expectations only, i.e., upper expectations which are obtained as suprema over small sets of probability measures.

As for the technicalities, we make use of the tools provided by example 1) of Sect. 2:

$$k^\alpha(z)=\min(\alpha_1 z_1, \alpha_2 z_2)$$
$$\tilde{W}_1^\alpha(z)=\alpha_1 \cdot 1_{[\alpha_1 z_1<\alpha_2 z_2]}(z)$$
$$\tilde{W}_2^\alpha(z)=\alpha_2 \cdot 1_{[\alpha_1 z_1 \geqq \alpha_2 z_2]}(z).$$

These are the concave function respectively the standard loss functions of the testing problem with weights proportional to (α_1, α_2), which we can restrict to $(\alpha_1, \alpha_2)=\alpha\in K$, i.e., $\alpha_\theta \geqq 0$ and $\alpha_1+\alpha_2=1$. The convex cone

$$\{a+\sum_i c_i k^{\alpha_i} \mid a \text{ affine function on } K,\ c_i \geqq 0\}$$

is dense in the set of all continuous concave functions on K. Hence, if there exists a simultaneously least favorable pair (Q_1, Q_2) for all k^α then it is universally least favorable, hence "least sufficient". This remark facilitates the proof of the Huber-Strassen theorem which we adapt as follows:

Theorem 4.1. *If (v_1, v_2) is a pair of upper expectations generated by 2-alternating capacities (see part I, Sect. 2), then there exists a "least sufficient" pair (Q_1, Q_2) under (v_1, v_2).*

Proof. We wish to apply Theorem 8.4 of part I. To this end we introduce a finite continuous measure $\lambda(d\alpha)$ on K and consider the mixtures

$$k^\lambda(z)=\int k^\alpha(z)\,\lambda(d\alpha).$$
$$\tilde{W}_\theta^\lambda(z)=\int \tilde{W}_\theta^\alpha(z)\,\lambda(d\alpha).$$

The standard loss function $\tilde{W}_\theta^\lambda$ belongs to k^λ and is continuous because of continuity of λ (i.e. $\lambda\{\alpha\}=0\,\forall\,\alpha$). Hence we can restrict consideration to nonrandomized procedures by Theorem 3.1 of the preceding section. The crucial step is to see that for any measurable map σ: $Y\to K$, we have

$$v_\theta(\tilde{W}_\theta^\lambda\,\sigma)=\int v_\theta(\tilde{W}_\theta^\alpha\,\sigma)\,\lambda(d\alpha).$$

This, however, follows immediately from Proposition 2.8 of part I. So we obtain additivity of the Bayes risk for any nonrandomized procedure:

$$R((v),\sigma,(\tilde{W}^\lambda))=\int R((v),\sigma,(\tilde{W}^\alpha))\,\lambda(d\alpha).$$

Minimizing over σ, we get:

$$s^{(v)}(k^\lambda)\geqq\int s^{(v)}(k^\alpha)\,\lambda(d\alpha).$$

Hence we have equality since "$\leqq$" holds anyway by subadditivity of $s^{(v)}$. By 8.4 of part I, the theorem follows. □

5. The Contamination Model

In this section we will basically solve the robust decision problem for finite experiments under ε-contamination. However, we will be able to achieve this only for continuous standard loss functions. The technical difficulty stems from the need of smooth concave functions k in the convex minimization problem ensuing from the contaminated decision problem. The next section will show how the solutions found for continuous standard loss functions carry over to discontinuous ones via mixing and smoothing of standard loss functions and theorems of simultaneous least favorability. At the end of this section we introduce the notion of *standard singular loss functions robust under ε-contamination.*

In this and the following section we restrict consideration to upper expectations provided by ε-contamination of an exact model (see part I, Sect. 2):

$$v_\theta(f)=(1-\varepsilon)\cdot P_\theta(f)+\varepsilon\cdot\sup f.$$

The corresponding set of probabilities dominated by v_θ is the ε-contamination neighborhood of P_θ:

$$\{(1-\varepsilon)\cdot P_\theta+\varepsilon\cdot H_\theta\,|\,H_\theta \text{ arbitrary probability}\}.$$

The contamination model leads to "censored" procedures. This means that the classical procedures resulting from the exact model (P_θ) are modified if they are based on probability ratios which are too close to zero or infinity. In this way, the upper Bayes risk can be minimized as we will see.

We will deal only with *continuous* standard loss functions to avoid randomization (Sect. 3). Given the contamination model (v_θ), the upper Bayes risk of a (nonrandomized) procedure σ: $Y\to K$ is:

$$R((v),\sigma,(\tilde{W}))=\frac{1}{n}\sum_{\theta}[(1-\varepsilon)\cdot P_{\theta}(\tilde{W}_{\theta}\sigma)+\varepsilon\cdot\sup_{y\in Y}\tilde{W}_{\theta}(\sigma(y))]$$
$$=(1-\varepsilon)\cdot R((P),\sigma,(\tilde{W}))+\varepsilon\cdot\frac{1}{n}\sum_{\theta}\sup\tilde{W}_{\theta}\sigma.$$

Obviously an optimal procedure makes a tradeoff between the two terms. It lowers the suprema as long as the decrease is not offset by the increase in $R((P),\sigma,(\tilde{W}))$. Lowering the terms $\sup\tilde{W}_{\theta}\sigma$ means cutting down on the decisions $z=\sigma(y)$ with highest losses $\tilde{W}_{\theta}(z)$. To control this process of decision censoring, we introduce thresholds c_{θ} and define a censored decision space by

$$T_c=\{\zeta\in K\,|\,\tilde{W}_{\theta}(\zeta)\leqq c_{\theta}\,\forall\,\theta\}.$$

Here $c=(c_{\theta})$ denotes the tuple of thresholds. We let it vary within the set

$$\{(c_{\theta})\,|\,T_c\neq\phi\}$$

which coincides with the sets A_k and A_W of the remark after Proposition 3.9. We can construct a concave function k_c which arises from restricting $\tilde{W}$ to T_c:

$$k_c=\inf_{\zeta\in T_c}\sum_{\theta}\tilde{W}_{\theta}(\zeta)\cdot z_{\theta}.$$

It coincides on T_c with the concave function k of $\tilde{W}$, but dominates k outside:

$$k_c\geqq k,\qquad k_c|T_c=k|T_c.$$

With these definitions, we can reformulate the minimization of Bayes risks as follows:

Proposition 5.1. *The minimal upper Bayes risk for the contamination model can be found by a minimization over the thresholds* $c=(c_{\theta})$:

$$s^{(v)}(k)=\inf_{c}\left[(1-\varepsilon)\cdot S^{(P)}(k_c)+\varepsilon\cdot\frac{1}{n}\sum_{\theta}c_{\theta}\right].$$

There exists a minimizing tuple c.

Proof. 1) That the minimum is attained follows from continuity of $c\to S^{(P)}(k_c)$ and lower compactness of $\{(c_{\theta})\,|\,T_c\neq\phi\}$. Continuity in turn follows from an application of the dominated convergence theorem and continuity of $c\to k_c(z)$ for every z.

2) The trivial part of the asserted equality is:

$$s^{(v)}(k)\leqq s^{(v)}(k_c)\leqq(1-\varepsilon)\cdot S^{(P)}(k_c)+\varepsilon\cdot\frac{1}{n}\sum_{\theta}c_{\theta}.$$

3) For the reverse inequality, choose a nearly optimal procedure σ:

$$s^{(v)}(k)+\eta\geqq R((v),\sigma,(\tilde{W}))$$

for given $\eta>0$. By Theorem 3.1, we know that we can assume that σ is non-randomized. Put $c_\theta=\sup \tilde{W}_\theta(\sigma(\cdot))$. Then we obtain:

$$\begin{aligned} R((v),\sigma,(\tilde{W}))&=R((v),\sigma,(\tilde{W}\,|\,T_c)) \\ &=(1-\varepsilon)\cdot R((P),\sigma,(\tilde{W}\,|\,T_c))+\varepsilon\cdot\frac{1}{n}\sum_\theta c_\theta \\ &\geqq(1-\varepsilon)\cdot S^{(P)}(k_c)+\varepsilon\cdot\frac{1}{n}\sum_\theta c_\theta. \end{aligned}$$

Since $\eta>0$ was arbitrary, the remaining inequality follows. It is crucial here to see how the need for a non-randomized σ comes in: if σ were randomized, the construction of the thresholds

$$c_\theta=\sup \tilde{W}_\theta\,\sigma=\sup_y \int \tilde{W}_\theta(z)\,\sigma(y,dz)$$

would not allow to pull through the first of the above equalities/inequalities (namely $R((v),\sigma,(\tilde{W}))=R((v),\sigma,(\tilde{W}\,|\,T_c))$), since the support of the measures $\sigma(\cdot,dz)$ would not necessarily have to be within the set T_c. □

We will pursue now the problem of optimal censoring, i.e., we want to find characterizations of optimal thresholds $c=(c_\theta)$. For this, we have to study the dependence of

$$(1-\varepsilon)\cdot S^{(P)}(k_c)+\varepsilon\cdot\frac{1}{n}\sum_\theta c_\theta$$

on c, the hard part being obviously k_c. It will turn out that this expression is convex in c. This fact will allow us to use directional derivatives in order to characterize global minima. The crucial point will be the calculation of directional derivatives for $c\to k_c(z)$, which can be done by relating them to Lagrange multipliers. These can be found from the dual optimization problem of the constrained minimization which defines $k_c(z)$. Our references for convex optimization are [5] and [6]. The next few steps will essentially consist of adaptations of some classical optimization results. Proposition 3.9 will lead us in the first step. It states that the set of upper tangents provided by the standard loss function $\tilde{W}(z)$ is complete if $\tilde{W}(z)$ is continuous.

As in the remark a) after Proposition 3.9, denote by A_k the cone of affine functions above k on K:

$$A_k=\{a\in\mathbb{R}^\Theta\,|\,\langle a,z\rangle\geqq k(z)\ \forall\, z\in K\}.$$

Then we can write by 3.9:

$$k_c(z)=\inf\{\langle a,c\rangle\,|\,a\in A_k,\ a\leqq c\} \qquad (*)$$

where "$a\leqq c$" is the componentwise partial ordering for vectors: $a_\theta\leqq c_\theta\ \forall\,\theta$.

Thus we obtain $k_c(z)$ by minimization of the affine (and hence convex) function $a\to\langle a,z\rangle$ over the convex set A_k under the constraints $a\leqq c$. In optimization theory, the map $c\to k_c(z)$ is called *primal function* [6, p. 216] or perturbation function [5, p. 37] of the given minimum problem.

Lemma 5.2. *Both* $k_c(z)$ *and* $(1-\varepsilon)\cdot S^{(P)}(k_c)+\varepsilon\cdot\frac{1}{n}\sum_\theta c_\theta$ *are convex in c.*

The *proof* of convexity of $k_c(z)$ is not hard and can be obtained from [5, p. 37] or [6, p. 216]. By mixing $k_c(z)$ over z with regard to the measure $S^{(P)}(dz)$, and by adding the affine function $\varepsilon\cdot\frac{1}{n}\sum_\theta c_\theta$, the lemma follows. □

The above lemma grants the existence of directional derivatives [5, p. 16–19] with regard to c for directions $h=c'-c$, both $c,c'\in A_k$:

Definition.

$$D_{c;h}k_c(z)=\lim_{\tau\downarrow 0}\frac{1}{\tau}\cdot(k_{c+\tau h}(z)-k_c(z)).$$

The difference quotient on the right hand side stays uniformly bounded due to boundedness of $\tilde{W}(z)$ as can be seen from the proof of Proposition 3.10. Thus, when integrating $k_c(z)$ with regard to $S^{(P)}$, one can interchange differentiation and integration:

$$D_{c;h}\left((1-\varepsilon)\cdot S^{(P)}(k_c)+\varepsilon\cdot\frac{1}{n}\sum_\theta c_\theta\right)=(1-\varepsilon)\cdot S^{(P)}(D_{c;h}k_c)+\varepsilon\cdot\frac{1}{n}\sum_\theta h_\theta.$$

Directional derivatives of convex functions depend subadditively on the directions: $D_{c;h_1+h_2}\leqq D_{c;h_1}+D_{c;h_2}$. Additivity is equivalent to differentiability. For convex functions, global minima can be characterized by non-negative directional derivatives. Thus we obtain:

Proposition 5.3. *The thresholds* $c=(c_\theta)$ *are optimal iff*

$$0\leqq S^{(P)}(D_{c;h}k_c)+\frac{\varepsilon}{1-\varepsilon}\cdot\frac{1}{n}\sum_\theta h_\theta$$

for all directions $h=c'-c$, *and* $c,c'\in A_k$.

We see that everything depends on whether we are able to calculate $D_{c;h}(k_c)$. To pursue this further, we will relate directional derivatives to so called subgradients and subderivatives [5, p. 20]. A *subgradient* of $c\to k_c(z)$ at $c\in\text{interior}(A_k)$ is a vector $\Lambda\in\mathbb{R}^\Theta$ which satisfies

$$k_{c'}(z)-k_c(z)\geqq\langle\Lambda,c'-c\rangle\qquad\forall c'\in A_k.$$

(Analogously, we can say that the standard loss function vector $\tilde{W}(z)$ is a supergradient of $z\to k(z)$, due to its defining property.) The set of all subgradients Λ at c will be denoted $\partial_c k_c(z)$. In [5], this is called the *subderivative* or subdifferential.

Lemma 5.4. *Directional derivatives of* $c\to k_c(z)$ *at* $c\in$*interior*(A_k) *can be obtained by*

$$D_{c;h}k_c(z)=\sup\{\langle\Lambda,h\rangle\,|\,\Lambda\in\partial_c k_c(z)\}.$$

The supremum is attained.

A proof is in [5, p. 27]. Implicitly we use here the fact that convex functions on finite dimensional spaces are always continuous on the interior of their domain, see [5, p. 28] or [6, p. 194].

Subgradients would not be of any use if they were not easier to obtain than directional derivatives. The theory of constrained optimization provides us with a link between subgradients and Lagrange multipliers, which we introduce now:

When applied to the minimization $(*)$ of $a \to \langle a, z\rangle$ over the convex set A_k under the constraints $a \leqq c$, the fundamental Lagrange multiplier theorem [6, p. 217] grants the following:

Lemma 5.5. *For $c \in interior(A_k)$, there does exist a Lagrange multiplier or Kuhn-Tucker vector $\Lambda \geqq 0$ (i.e. $\Lambda_0 \geqq 0 \ \forall\, \theta$), satisfying*

$$k_c(z) = \inf_{a \in A_k} (\langle a, z\rangle + \langle a - c, \Lambda\rangle).$$

The *proof* follows from the cited reference. (Notice that the constraint operator G in [6, p. 216] specializes to $G(a) = a - c$, and hence the assumption of Theorem 1 [6, p. 217] concerning the existence of $a \in A_k$ satisfying $G(a) < 0$ translates into the requirement of c being in the interior of A_k.)

Our interest in Lagrange multipliers stems from the fact that they are basically the subgradients of the *primal function* (modulo sign):

Lemma 5.6. *If $c \in interior(A_k)$, the set $-\partial_c k_c(z)$ coincides exactly with the set of Lagrange multipliers for the minimization problem $(*)$. All elements Λ of $-\partial_c k_c(z)$ are non-negative.*

An explicit *proof* under slightly different assumptions is in [5, p. 38], or more implicitly in [6, p. 218 and p. 222 bottom].

We proceed with a characterization of Lagrange multipliers in terms of the dual of the optimization problem $(*)$. The *dual function* $\varphi_{c,z}(\Lambda)$ [6, *p.* 223f.] is defined as:

$$\varphi_{c,z}(\Lambda) = \inf_{a \in A_k} (\langle a, z\rangle + \langle a - c, \Lambda\rangle).$$

The following lemma will allow us to explicitly determine the Lagrange multipliers:

Lemma 5.7. *If $c \in interior(A_k)$, the dual function $\Lambda \to \varphi_{c,z}(\Lambda)$ attains its maximum. The set of maxima coincides with the set of Lagrange multipliers for the minimization problem $(*)$.*

The *proof* is contained in [6, p. 224f.].

The preceding lemmas are the basic machinery we need. We continue with the calculation of the dual function. It can be written as:

$$\varphi_{c,z}(\Lambda) = \inf_{a \in A_k} \langle a, z + \Lambda\rangle - \langle c, \Lambda\rangle.$$

The infimum over $a \in A_k$ is taken on at $a = \tilde{W}\left(\frac{z + \Lambda}{1 + \sum \Lambda_\theta}\right)$ as one can see from the definition of standard loss functions. We realize at this point that we should extend both the concave function $k(z)$ and the standard loss function $\tilde{W}(z)$ from K to $\mathbb{R}_+ \cdot K$. This is sensibly done by extending the latter by constancy along rays $\mathbb{R}_+ \cdot z$, and $k(z)$ via positive homogeneity. We make this precise in the following:

Definition and Lemma 5.8. *If the standard loss function $\tilde{W}(z)$ is extended from the simplex K to the cone $\mathbb{R}_+ \cdot K$ via*

$$\tilde{W}(\zeta)=\tilde{W}\left(\frac{\zeta}{\sum \zeta_\theta}\right) \qquad \forall \zeta \geqq 0 \ (\neq 0),$$

then the corresponding concave function $k(z)$ can be extended via

$$k(\zeta)=\langle \tilde{W}(\zeta), \zeta \rangle$$

to a positively homogeneous and concave function on the $z \geqq 0$ in $\mathbb{R}^{\Theta}$. We preserve thus the crucial property

$$k(\zeta)=\inf_{\zeta} \langle \tilde{W}(\zeta'), \zeta \rangle$$

where ζ' may vary over K or all $z \geqq 0$ in $\mathbb{R}^{\Theta}$.

The *proof* is obvious, and we gave a formal statement of the above for its importance only. In terms of the extensions of k and $\tilde{W}$, we obtain:

Lemma 5.9. *The dual function of the minimization problem* $(*)$ *is*

$$\Lambda \to \varphi_{c,z}(\Lambda)=k(z+\Lambda)-\langle c, \Lambda \rangle.$$

Fortunately, there do exist theorems for optimization over convex cones. To this end, however, we need the dual function and hence $k(z)$ differentiable, i.e., directional derivatives must behave additively in the directions. This assumption is met in our case if the standard loss function is continuous, see Proposition 3.10. The conditions of the following proposition are sufficient even without the differentiability assumption, but they are not necessary then.

Proposition 5.10. *Necessary and sufficient conditions for $\Lambda \geqq 0$ to maximize the dual function are:*

$$\tilde{W}(z+\Lambda) \leqq c, \qquad \langle \tilde{W}(z+\Lambda), \Lambda \rangle = \langle c, \Lambda \rangle.$$

The *proof* involves an application of a very useful lemma in [6, p. 227]. Applied to our context with the appropriate interchanges of concave/convex and $\geqq/\leqq$, it gives the following necessary and sufficient conditions for Λ to take on a maximum of the dual function $\varphi_{c,z}(\Lambda)$:

$$D_{\Lambda;\Lambda'}\, \varphi_{c,z}(\Lambda) \leqq 0 \qquad \forall \Lambda' \geqq 0$$
$$D_{\Lambda;\Lambda}\, \varphi_{c,z}(\Lambda) = 0.$$

By the previous lemma, this translates into:

$$\langle \tilde{W}(z+\Lambda'), \Lambda' \rangle \leqq \langle c, \Lambda' \rangle \qquad \forall \Lambda' \geqq 0$$
$$\langle \tilde{W}(z+\Lambda), \Lambda \rangle = \langle c, \Lambda \rangle.$$

The first condition boils down to constraint satisfaction (use for Λ the canonical basis vectors $e^{(\theta)}$): $\tilde{W}_\theta(z+\Lambda) \leqq c_\theta \; \forall \, \theta$. □

Our search for ways of calculating directional derivatives $D_{c;h} k_c(z)$ of the primal function $c \to k_c(z)$ can now be summarized as follows:

Proposition 5.11. *For thresholds $c=(c_\theta) \in interior(A_k)$ we have:*

$$D_{c;h} k_c(z) = \max\{\langle h, -\Lambda\rangle \mid \tilde{W}(z+\Lambda) \leqq c, \; \langle \tilde{W}(z+\Lambda), \Lambda\rangle = \langle c, \Lambda\rangle, \; \Lambda \geqq 0\}.$$

If one can say so, Proposition 5.3 and 5.11 solve the problem of characterizing procedures which are minimax under ε-contamination in the sense that they minimize the worst-case Bayes risk $R((v), \sigma, (\tilde{W}))$. The propositions describe the optimal decision-censoring thresholds c_θ. As outlined at the beginning of this section, one has to cut out (censor) those decisions which result in standard losses above the optimal thresholds, and with the remaining decisions one should decide according to Bayes procedures.

We would like to continue with the question of what standard loss functions do not require censoring for a given model (P_θ) and contamination ε. Such standard loss functions may justifiably be called *robust* for this situation:

Definition. *A standard loss function $\tilde{W}$ is called robust for the model (P_θ) under ε-contamination if the decision-censoring thresholds $c_\theta = \sup_z \tilde{W}_\theta(z)$ are optimal, in other words, no decision-censoring is need.*

For thresholds $c_\theta \geqq \sup \tilde{W}_\theta$ no censoring happens. This is expressed by $k = k_c$. Thus, when characterizing robust standard loss functions by directional derivatives with regard to thresholds, we do not have to consider directions other than negative ones:

Proposition 5.12. *Let $c_\theta = \sup \tilde{W}_\theta$ be the trivial thresholds. The standard loss function $\tilde{W}$ is robust for (P_θ) under ε-contamination iff:*

$$S^{(P)}(D_{c;-h} k_c) \geqq \frac{\varepsilon}{1-\varepsilon} \cdot \frac{1}{n} \qquad \forall \, h \in K \;\; (n = |\Theta|).$$

If $c \to k_c(z)$ is differentiable from below $\forall \, z$, i.e., if

$$D_{c;-h} k_c(z) = \sum_\theta h_\theta \cdot D_{c;-e^{(\theta)}} k_c(z) \qquad \forall \, h \in K, \; z \in K,$$

then robustness can be characterized by the lower partial derivatives:

$$S^{(P)}(D_{c;-e^{(\theta)}} k_c) \geqq \frac{\varepsilon}{1-\varepsilon} \cdot \frac{1}{n} \qquad \forall \, \theta.$$

This last proposition motivates a further investigation into when differentiability from below at the trivial supremum thresholds is met, and how the lower partial derivatives are calculated. We would like to rely on Proposition 5.11, but we have to make sure that the assumption $c \in interior(A_k)$ is satisfied. The following lemma will take care of that:

Lemma 5.13. *Assume the standard loss function $\tilde{W}$ has no constant components: $\inf \tilde{W}_\theta(\cdot) < \sup \tilde{W}_\theta(\cdot)\ \forall\, \theta$. Then there exists $z \in K$ such that $\tilde{W}_\theta(z) < \sup \tilde{W}_\theta(\cdot)\ \forall\, \theta$. In other words, the vector c of trivial thresholds $c_\theta = \sup \tilde{W}_\theta(\cdot)$ is interior in A_k. (For the remainder of this section, we will always assume that this condition is satisfied.)*

Proof. For each θ there exists $z^{(\theta)} \in K$ such that $\tilde{W}_\theta(z^{(\theta)}) < \sup \tilde{W}_\theta(\cdot)$. Then consider the measure $\mu = \frac{1}{|\Theta|} \cdot \sum_\theta \delta_{z^{(\theta)}}$. We notice that $\int \tilde{W}_\theta(z)\, \mu(dz) < \sup \tilde{W}_\theta(\cdot)\ \forall\, \theta$. By Proposition 3.3, we obtain the existence of $z \in K$ for which $\tilde{W}_\theta(z) \leqq \int \tilde{W}_\theta(z')\, \mu(dz')$. This finishes the proof. □

If the trivial sup-thresholds $c_\theta = \sup \tilde{W}_\theta(\cdot)$ are interior in A_k, we can characterize the lower directional derivatives by Proposition 5.11:

$$D_{c;\,-h}\, k_c(z) = \max\{\langle h, \Lambda\rangle \mid \langle \tilde{W}(z+\Lambda), \Lambda\rangle = \langle c, \Lambda\rangle,\ \Lambda \geqq 0\}.$$

This expression is additive in h iff there exists a largest vector Λ which attains the maximum on the right hand side simultaneously for all $h \in K$. Thus:

Proposition 5.14. *The primal function $c \to k_c(z)$ is lower differentiable in c at the trivial thresholds $c_\theta = \sup \tilde{W}_\theta(\cdot)$ iff there exists a largest vector in the set*

$$\{\Lambda \in \mathbb{R}^\Theta \mid \Lambda \geqq 0,\ \langle \tilde{W}(z+\Lambda), \Lambda\rangle = \langle c, \Lambda\rangle\}.$$

We wish to calculate the lower partial derivatives at the trivial thresholds. Simplifications in Proposition 5.11 occur in this case. When calculating $D_{c;\,-e^{(\theta)}}\, k_c(z)$ at $c_\theta = \tilde{W}_\theta(\cdot)$, it does not matter whether we raise $c_{\theta'}$ to even larger values than $\sup \tilde{W}_{\theta'}(\cdot)$ for $\theta' \neq \theta$. The point of this is to force $\Lambda_{\theta'}$ for $\theta' \neq \theta$ to zero through the condition $\langle \tilde{W}(z+\Lambda), \Lambda\rangle = \langle c, \Lambda\rangle$ in 5.11. Thus this condition simplifies to

$$\tilde{W}_\theta(z + \lambda \cdot e^{(\theta)}) = \sup \tilde{W}_\theta(\cdot).$$

Since $\Lambda_{\theta'} = 0\ \forall\, \theta' \neq \theta$, we may write $\Lambda = \lambda \cdot e^{(\theta)}$. To summarize:

Proposition 5.15. *At the trivial thresholds $c_\theta = \sup \tilde{W}_\theta(\cdot)$, we have*

$$D_{c;\,-e^{(\theta)}}\, k_c(z) = \max\{\lambda \geqq 0 \mid \tilde{W}_\theta(z + \lambda \cdot e^{(\theta)}) = \sup \tilde{W}_\theta(\cdot)\}.$$

As an application of lower partial derivatives, we will derive a theorem which shows that a certain degree of robustness in loss functions for finite experiments is a rather likely case. Our question will be: given an exact model (P_θ) and a continuous standard loss function $\tilde{W}$, does there exist a contamination amount $\varepsilon > 0$ for which $\tilde{W}$ is robust? If we abbreviate $I_{(P),k} = \inf_{h \in K} S^{(P)}(D_{c;\,-h}\, k_c)$ for the moment, then Proposition 5.12 gives us $\varepsilon = I_{(P),k} \Big/ \left(I_{(P),k} + \frac{1}{n}\right)$ as the largest amount of contamination for which we have robustness. The quantity $I_{(P),k}$ is zero iff $\tilde{W}$ is non-robust for any $\varepsilon > 0$. When comparing with the lower partial derivatives, we have in a first cut only the following obvious inequality:

$$\inf_{z\in K} S^{(P)}(D_{c;\,-h}\, k_c(z)) \leqq \inf_{\theta\in\Theta} S^{(P)}(D_{c;\,-e^{(\theta)}}\, k_c(z)).$$

We have equality if $c\to k_c(z)$ is lower differentiable at the trivial thresholds. However, it is possible to obtain lower bounds for the left hand side in terms of the right hand side:

Lemma 5.16. $\inf\limits_{h\in K} S^{(P)}(D_{c;\,-h}\, k_c(z)) \geqq \frac{1}{n} \cdot \min\limits_{\theta\in\Theta} S^{(P)}(D_{c;\,-e^{(\theta)}}\, k_c(z))$.

Proof. For $h=\sum\limits_{\theta} h_\theta \cdot e^{(\theta)} \in K$ and $\tau>0$, we have $k_{c-\tau\cdot h} \geqq k_{c-\tau h_\theta e^{(\theta)}}\ \forall\,\theta$, since the constraints satisfy trivially $c-\tau h \leqq c-\tau h_\theta e^{(\theta)}$. Differentiation with regard to τ at $+0$ and integration with regard to $S^{(P)}(dz)$ yields:

$$S^{(P)}(D_{c;\,-h}\, k_c) \geqq h_\theta \cdot S^{(P)}(D_{c;\,-e^{(\theta)}}\, k_c) \qquad \forall\,\theta.$$

Replacing $S^{(P)}(D_{c;\,-e^{(\theta)}}\, k_c)$ by $\min\limits_{\theta} S^{(P)}(D_{c;\,-e^{(\theta)}}\, k_c)$ on the right hand side and then taking $\max\limits_{\theta} h_\theta$ we get:

$$S^{(P)}(D_{c;\,-h}\, k_c) \geqq (\max_{\theta} h_\theta) \cdot \min_{\theta} S^{(P)}(D_{c;\,-e^{(\theta)}}\, k_c).$$

Finally we notice that $\max\limits_{\theta} h_\theta \geqq \frac{1}{n}$ for $h\in K$. This finishes the proof. □

This rough estimate is enough to obtain the following criterion:

Proposition 5.17. *There exists a contamination amount* $\varepsilon>0$ *for which the continuous standard loss function* $\tilde{W}$ *is robust under the model* (P_θ) *iff*

$$S^{(P)}(\{z\in K \mid \exists\,\lambda>0\colon \tilde{W}_\theta(z+\lambda\cdot e^{(\theta)}) = \sup \tilde{W}_\theta(\cdot)\}) \ > \ 0 \qquad \forall\,\theta.$$

Proof. By Lemma 5.16 and the remarks beforehand, there exists $\varepsilon>0$ for which we have robustness iff $\inf\limits_{\theta} S^{(P)}(D_{c;\,-e^{(\theta)}}\, k_c(z)) > 0\ \forall\,\theta$. By 5.15, we get:

$$D_{c;\,-e^{(\theta)}}\, k_c(z) > 0 \qquad \text{iff} \ \ \exists\,\lambda>0\colon \tilde{W}_\theta(z+\lambda e^{(\theta)} = \sup \tilde{W}_\theta(\cdot).$$

A standard argument from measure theory finishes the proof. □

Here are some remarks on the relevance of Proposition 5.17. There are two items which have an impact on the condition in 5.17: 1) the loss function $(\tilde{W}_\theta)$, and 2) the experiment (P_θ).

ad 1): The standard loss function $(\tilde{W}_\theta)$ takes on its supremum someplace on the boundary face $\{z \mid z_\theta = 0\}$ of the simplex K (or the positive quadrant of $\mathbb{R}^\Theta$ if $\tilde{W}$ is extended according to 5.8). This is seen by restricting $\tilde{W}_\theta(z)$ to segments $[e^{(\theta)}, z]$ on which the maximum is taken on at z, due to concavity of k[3]. Intuitively, this means that we encounter the greatest θ-loss for those "standard observations" z which give least evidence to the parameter θ: $z_\theta = 0$[4]. For those

[3] Recall that $\tilde{W}_\theta(z)$ is the value of the upper tangent $\zeta \to \langle \tilde{W}_\theta(z), \zeta\rangle$ of k at z, evaluated at $e^{(\theta)}$

[4] Recall again that $n\cdot z_\theta$ is the density of the standard experiment $(S_\theta^{(P)})$ with regard to the standard measure $S^{(P)}$. See part I, Sect. 5

discontinuous standard loss functions which arise from loss matrices on finite decision sets $T=\{1, 2, 3, \ldots, n\}$, it is clear that the sets

$$\{z \in K \mid \exists\, \lambda > 0 \colon \tilde{W}_\theta(z+\lambda \cdot e^{(\theta)}) = \sup \tilde{W}_\theta(\cdot)\}$$

extend into the interior of K. If we plan to use continuous standard loss functions as approximations to discontinuous ones in some sense, we may actually consider this as a frequent case, and it appears not unlikely that the condition in Proposition 5.17 is satisfied in many cases.

ad 2): It seems intuitively clear that one has the more robustness the more informative an experiment (P_θ) is. Some ways of formalizing information are in terms of Blackwell-sufficiency (part I, Sect. 4), simultaneous comparison of Bayes risks (part I, Sect. 6), and standard measures (part I, Sect. 5). The theorem of Blackwell-Sherman-Stein grants the equivalence of all these partial orderings of experiments (part I, Sect. 7). An intuitive interpretation of high information in an experiment (P_θ) via its standard measure $S^{(P)}(dz)$ is that it should take on small values on concave functions $k(z)$, i.e., its mass should concentrate along the boundaries of K as much as possible. This is exactly what the condition of Proposition 5.17 asks for.

Last in this section, we will prove a theorem on the existence of simultaneously least favorable experiments, suitable for application to contaminated classification problems. The proof uses Theorem 8.4 of part I, and relies on mixing arbitrary (e.g., discontinuous) standard loss functions such that the mixture is continuous, as was the case in Sect. 4 for the Huber-Strassen theorem. As pointed out at the beginning of Sect. 3, mixing is likely to produce continuous standard loss functions, since they may inherit the smoothness properties of the mixing measure.

Theorem 5.18. *Let $\tilde{W}_\theta^\alpha(z)$ be a parametrized family of standard loss functions, not necessarily continuous in z. Make the following assumptions:*

1) *Let $\tilde{W}_\theta^\alpha(z)$ be measurable in α for each z, and the corresponding concave functions $k^\alpha(z)$ depend continuously on α.*

2) *Every $\tilde{W}_\theta^\alpha(\cdot)$ takes on its supremum over z at some common point $z^{(\theta)} \in K$:*

$$\tilde{W}_\theta^\alpha(z^{(\theta)}) = \sup_z \tilde{W}_\theta^\alpha(z) \qquad \forall\, \theta \in \Theta,\ \alpha.$$

3) *Let $\lambda(d\alpha)$ be a probability measure on the parameters α. Assume that the mixed standard loss function*

$$\tilde{W}_\theta^\lambda(z) = \int \tilde{W}_\theta^\alpha(z)\, \lambda(d\alpha)$$

is continuous and robust for (P_θ) under ε-contamination.

Then there exists a least favorable experiment $(Q_\theta) = ((1-\varepsilon) \cdot P_\theta + \varepsilon \cdot H_\theta)$ simultaneously for all $\tilde{W}^\alpha$.

Proof. Robustness means that the trivial supremum thresholds are optimal, hence, by 5.1:

$$s^{(v)}(k^\lambda) = (1-\varepsilon) \cdot S^{(P)}(k^\lambda) + \varepsilon \cdot \frac{1}{n} \sum_\theta \sup \tilde{W}_\theta^\lambda(\cdot)$$

where k^λ is the λ-mixture of k^α. The first term on the right hand side is additive:

$$S^{(P)}(k^\lambda)=\int S^{(P)}(k^\alpha)(d\alpha).$$

To evaluate the second term, we note that

$$\sum_\theta \sup \tilde{W}_\theta^\lambda(\cdot)=\sum_\theta \sup \tilde{W}_\theta^\alpha(\cdot) \qquad \forall\, \alpha$$

since suprema are taken on at identical z's across all α's. This leads to the following:

$$s^{(v)}(k^\lambda)=\int \left[(1-\varepsilon)\cdot S^{(P)}(k^\alpha)+\varepsilon\cdot\frac{1}{n}\sum_\theta \sup \tilde{W}_\theta^\alpha(\cdot)\right]\lambda(d\alpha).$$

If we denote by $p=(p_\theta(\cdot))$ the vector of densities of the probability measures P_θ under the dominating measure $\sum_\theta P_\theta$ (see part I, Sect. 5), then we obtain for the integrand:

$$(1-\varepsilon)\cdot S^{(P)}(k^\alpha)+\varepsilon\cdot\frac{1}{n}\sum_\theta \sup \tilde{W}_\theta^\alpha=R((P),p,(\tilde{W}^\alpha))+\varepsilon\cdot\frac{1}{n}\sum_\theta \sup \tilde{W}_\theta^\alpha$$
$$\geqq R((v),p,(\tilde{W}^\alpha))\geqq s^{(v)}(k^\alpha).$$

Thus we get the inequality:

$$s^{(v)}(k^\lambda)\geqq \int s^{(v)}(k^\alpha)\,\lambda(d\alpha).$$

The reverse inequality holds true anyway due to subadditivity of $s^{(v)}$, hence the assumptions of Theorem 8.4 of part I are satisfied, and the assertion of 5.18 follows. □

6. Classification Under Contamination

As an application of the previous section, we wish to present a result on simultaneously least favorable experiments for certain sets of classification loss functions. We believe that it is the first theorem which goes beyond the Huber-Strassen theorem and binary experiments, although in another way it resembles more Huber's earlier results [2, 3], due to our use of ε-contamination rather than capacities. How much can be generalized remains an open question. As mentioned in Sect. 4, for general finite experiments it is unlikely that there exist least favorable experiments simultaneously for *all* loss functions (i.e., all decision problems). We will have to be somewhat more modest and consider smaller sets of loss functions, or choose other, less satisfactory upper expectations than the ones given by, e.g., ε-contamination. We will stay with ε-contaminated models and introduce some peculiar subsets of loss functions. Recall from Sect. 2 that a classification loss function is given by a set of parameters α_θ which can be interpreted either 1) as losses for erroneously rejecting θ given the uniform prior which puts weight $\frac{1}{n}$ on each parameter value θ, or 2) as prior weights on the parameter values θ given the uniform losses $\frac{1}{n}$ for erroneously rejecting θ. Both interpretations lead to the same standard loss functions and Bayes risks. We prefer to think of α_θ as prior weight, because the sets of classification loss functions we wish to consider are given by contamination neighborhoods of (α_θ) as probability distribution on Θ. This means that we will fix a prior $\alpha^*=(\alpha_\theta^*)$ and contaminate it as follows:

$$\alpha_\theta=(1-\eta)\cdot\alpha_\theta^*+\eta\cdot\zeta_\theta$$

where $\zeta\in K$ is an arbitrary contaminating prior on Θ. We will thus be dealing with two types of contamination: ε-contamination on the side of the model, and η-contamination on the side of the priors. We will show the existence of least favorable experiments in the ε-contamination neighborhood of (P_θ) simultaneously for all η-contaminated priors, at least if ε and η are not too large. The precise condition is that the classical procedure (based on maximum weighted densities under (P_θ)) should not break down under both contaminations, where breakdown means that throwing away the data and deciding a priori is preferable under the worst possible contamination.

Here are the relevant notations and facts which we partly introduced in Sect. 2 already. Let the following be the set of "standard observations" (i.e., elements of the simplex K) for which one may decide for θ:

$$B_\theta=\{(\alpha,z)\in K\times K\,|\,\alpha_\theta z_\theta=\sup_{\theta_1}\alpha_{\theta_1}z_{\theta_1}\}.$$

Unfortunately, these sets do not quite form a partition of K into disjoint sets. As in Sect. 2, we introduce therefor measurable sets $A_\theta\subset B_\theta$ which form a partition, such as $A_\theta=B_\theta-\bigcup_{\theta_1<\theta}B_{\theta_1}$. The cross sections will be denoted by:

$$A_\theta(z)=\{\alpha\in K\,|\,(\alpha,z)\in A_\theta\},\qquad B_\theta(z)=\{\alpha\in K\,|\,(\alpha,z)\in B_\theta\}.$$

Due to symmetry in α and z, we do not have to distinguish these sets from α-cross sections: $A_\theta(\alpha)=\{z\in K\,|\,(\alpha,z)\in A_\theta\}$, and same for $B_\theta(\alpha)$. For later we notice that *closure* $(A_\theta(\alpha))=B_\theta(\alpha)$.

According to Sect. 2, an instance of a standard loss function of a classification problem with prior weights α_θ is given by:

$$\tilde{W}_\theta^\alpha(z)=\alpha_\theta\cdot 1_{A_\theta^c}(\alpha,z)=\alpha_\theta\cdot 1_{A_\theta^c(\alpha)}(z).$$

In what follows, we wish to work towards meeting the assumptions of Theorem 5.18. Thus we introduce mixtures of the above standard loss functions with regard to measures $\lambda(d\alpha)$ on K:

$$\tilde{W}_\theta^\lambda(z)=\int\tilde{W}_\theta^\alpha(z)\,\lambda(d\alpha)=\int\alpha_\theta\cdot 1_{A_\theta^c(z)}(\alpha)\,\lambda(d\alpha).$$

Without loss of generality we may restrict ourselves to probability measures λ.

Our next concern is to pick λ such that $\tilde{W}_\theta^\lambda$ is continuous. A sufficient condition for this to hold is that the boundary $\partial A_\theta(z)$ is a null set with regard to λ for all $z \in K$. This is seen by the dominated convergence theorem since:

$$1_{A_\theta^C(z^{(n)})}(\alpha) \to 1_{A_\theta^C(z)}(\alpha) \quad \lambda\text{-a.s. if } z^{(n)} \to z \quad \text{and} \quad \alpha \notin \partial A_\theta(z),$$

which implies almost sure convergence with regard to λ since $\partial A_\theta(z)$ is a null set. This latter condition is met, e.g., if λ is absolutely continuous with regard to the uniform measure on the unit simplex K.

In a next step we specify the mixing measure λ further. This is where the contamination neighborhoods of priors come in. As before, let $\alpha^* \in K$ be a fixed prior on Θ, and denote by $(1-\eta) \cdot \alpha^* + \eta \cdot K$ the set of all η-contaminated priors, where we always assume $\eta > 0$. We may now pick $\lambda(d\alpha)$ as the uniform distribution on $(1-\eta) \cdot \alpha^* + \eta \cdot K$, but all we really need is that λ have $(1-\eta) \cdot \alpha^* + \eta \cdot K$ as its support and the resulting standard loss function be continuous (which is the case here since this λ is absolutely continuous with regard to the uniform distribution on K).

Consider next the concave functions obtained by λ-mixing:

$$k^\lambda(z) = \int k^\alpha(z) \, \lambda(d\alpha) = \langle W^\lambda(z), z \rangle,$$

and also the primal function $c \to k_c^\lambda(z)$ which goes with it. For the mixing measures λ specified above, we will be able to show that the primal function $c \to k_c^\lambda(z)$ is lower differentiable at the trivial supremum thresholds $c_\theta = \sup \tilde{W}_\theta^\lambda(\cdot)$, and we can also calculate the lower derivatives explicitly. This and Proposition 5.12 will give us a simple criterion for robustness of the mixture standard loss function $\tilde{W}_\theta^\lambda$. With Theorem 5.18 we will finally arrive at a theorem of the anticipated sort. First, however, we have to face some technicalities for the sake of proving lower differentiability via Proposition 5.14. This will be done by chopping up the arguments into a series of lemmas. To make use of 5.14, we have to determine what are the sets:

$$\{\Lambda \in \mathbb{R}^\Theta \mid \Lambda \geqq 0, \langle \tilde{W}^\lambda(z+\Lambda), \Lambda \rangle = \langle c, \Lambda \rangle\}.$$

This will be answered by the following:

Lemma 6.1. *Assume as always in this section that the mixing measure $\lambda(d\alpha)$ has the set $(1-\eta)\alpha^* + \eta \cdot K$ as its support, and that $\tilde{W}_\theta^\lambda$ is continuous. Denote by*

$$\alpha^{(\theta)} = (1-\eta) \cdot \alpha^* + \eta \cdot e^{(\theta)}$$

the extremal points of the set $(1-\eta) \cdot \alpha^* + \eta \cdot K$, and let always $c_\theta = \sup \tilde{W}_\theta^\lambda(\cdot)$ *be the trivial thresholds. For $\Lambda \geqq 0$ the following statements are equivalent:*

1) $\langle \tilde{W}^\lambda(z+\Lambda), \Lambda \rangle = \langle c, \Lambda \rangle$.

2) $\forall \, \theta \in \Theta$: $\quad \Lambda_\theta = 0 \quad$ *or* $\; \alpha_\theta^{(\theta)} \cdot (z_\theta + \Lambda_\theta) \leqq \max\limits_{\theta_1 \neq \theta} a_{\theta_1}^{(\theta)} \cdot z_{\theta_1}$.

3) $\forall \, \theta \in \Theta$: $\quad \Lambda_\theta \leqq \dfrac{1}{\alpha_\theta^{(\theta)}} \cdot \max\limits_{\theta_1} \alpha_{\theta_1}^{(\theta)} \cdot (\mathrm{z}_{\theta_1} + (1-\delta_{\theta,\theta_1}) \cdot \Lambda_{\theta_1}) - \mathrm{z}_\theta$.

Remark 6.2. *The extremal priors of the η-contamination neighborhood satisfy the following inequalities:*

1) $\alpha_\theta^{(\theta)}>0\ \forall\,\theta, \quad \alpha_\theta^{(\theta_1)}<\alpha_\theta^{(\theta)}\ \forall\,\theta_1\neq\theta,$
$\alpha_\theta^{(\theta_1)}=\alpha_\theta^{(\theta_2)} \quad \forall\,\theta_1\neq\theta,\ \theta_2\neq\theta.$

2) $\alpha_{\theta_1}^{(\theta)}\cdot\alpha_{\theta_2}^{(\theta_1)}\leqq\alpha_{\theta_1}^{(\theta_1)}\cdot\alpha_{\theta_2}^{(\theta)} \quad \forall\,\theta,\theta_1,\theta_2.$

Proof of Remark 6.2. 1) These inequalities follow immediately from the definitions:

$$\alpha_\theta^{(\theta)}=(1-\eta)\cdot\alpha_\theta^*+\eta, \quad \text{and} \quad \alpha_{\theta_1}^{(\theta)}=(1-\eta)\cdot\alpha_{\theta_1}^* \quad \text{for } \theta_1\neq\theta,$$

and the assumption $\eta>0$. 2) If $\theta_2=\theta_1$ we have equality. If $\theta_2\neq\theta_1$ we get $\alpha_{\theta_2}^{(\theta_1)}\leqq\alpha_{\theta_2}^{(\theta)}$ and $\alpha_{\theta_1}^{(\theta)}\leqq\alpha_{\theta_1}^{(\theta_1)}$, from which 6.2 follows. □

Proof of Lemma 6.1. 1)⇔2): The first condition translates into:

$$\forall\,\theta\colon \Lambda_\theta=0 \quad \text{or} \quad \tilde{W}_\theta^\lambda(z+\Lambda)=c_\theta.$$

To the second equality we may apply the following equivalence:

$$\tilde{W}_\theta^\lambda(z)=c_\theta \quad \text{iff} \quad \alpha_\theta^{(\theta)}z_\theta\leqq\max_{\theta_1\neq\theta}\alpha_{\theta_1}^{(\theta)}z_{\theta_1},$$

which gives us condition 2). The proof of the latter equivalence is obtained through a sequence of reformulations:

$$\begin{aligned}
\tilde{W}_\theta^\lambda(z)=c_\theta \quad &\text{iff } (1-\eta)\cdot\alpha^*+\eta\cdot K \subset \operatorname{closure}(A_\theta^C(z))\\
&\text{iff } (1-\eta)\cdot\alpha^*+\eta\cdot K \subset \bigcup_{\theta_1\neq\theta} B_{\theta_1}(z)\\
&\text{iff } \forall\,\alpha\in(1-\eta)\cdot\alpha^*+\eta\cdot K \quad \exists\,\theta_1\neq\theta\colon \ \alpha_{\theta_1}z_{\theta_1}=\max_{\theta_2}\alpha_{\theta_2}z_{\theta_2}\\
&\text{iff } \forall\,\alpha\in(1-\eta)\cdot\alpha^*+\eta\cdot K\colon \ \alpha_\theta z_\theta\leqq\max_{\theta_1\neq\theta}a_{\theta_1}z_{\theta_1}\\
&\text{iff } \alpha_\theta^{(\theta)}z_\theta\leqq\max_{\theta_1\neq\theta}\alpha_{\theta_1}^{(\theta)}z_{\theta_1}.
\end{aligned}$$

The last equivalence is due to the following inequalities:

$$\forall\,\alpha\in(1-\eta)\cdot\alpha^*+\eta\cdot K\colon \quad \alpha_\theta^{(\theta)}\geqq\alpha_\theta \quad \text{and} \quad \alpha_{\theta_1}^{(\theta)}\leqq\alpha_{\theta_1} \quad \text{for } \theta_1\neq\theta,$$

which immediately follow from the definitions.

2)⇔3): This is straightforward by rewriting the inequality in 2) with Λ_θ alone on the left hand side. This is always possible since $\alpha_\theta^{(\theta)}>0$. The maximum in 3) is extended over all θ_1 and includes especially $\theta_1=\theta$, which results in a zero and covers the case $\Lambda_\theta=0$. □

Lemma 6.3. *The set of* $\Lambda\geqq0$ *which satisfy one of the equivalent conditions in* 6.1 *has a largest element which is given by:*

$$\Lambda_\theta=\frac{1}{\alpha_\theta^{(\theta)}}\cdot\max_{\theta_1}\alpha_{\theta_1}^{(\theta)}z_{\theta_1}-z_\theta.$$

Proof. We introduce an auxiliary mapping $f(\Lambda)$ from $\{\Lambda\in\mathbb{R}^\Theta\mid\Lambda\geqq0\}$ into itself as follows:

$$f_\theta(\Lambda)=\frac{1}{\alpha_\theta^{(\theta)}}\cdot\max_{\theta_1}\alpha_{\theta_1}^{(\theta)}\cdot(z_{\theta_1}+(1-\delta_{\theta,\theta_1})\cdot\Lambda_{\theta_1})-z_\theta.$$

Lemma 6.3 asserts that the set $\{\Lambda \geqq 0 \mid \Lambda \leqq f(\Lambda)\}$ has a greatest element, namely $\Lambda = f(0)$. It is immediate that $f(0)$ is in this set, since $0 \leqq f(0)$ and f is monotone increasing, and hence $f(0) \leqq f(f(0))$. There remains to show that $f(0)$ is the largest element:

$$\Lambda \leqq f(\Lambda) \Rightarrow \Lambda \leqq f(0).$$

Assume the left hand side, or equivalently (by 6.1):

$$\forall\, \theta \in \Theta\colon \Lambda_\theta = 0 \quad \text{or} \quad \exists\, \theta_1 \neq \theta\colon \alpha_\theta^{(\theta)} \cdot (z_\theta + \Lambda_\theta) \leqq \alpha_{\theta_1}^{(\theta)} z_{\theta_1}.$$

If for given $\Lambda \geqq 0$ we introduce the set $\Theta^{(\Lambda)} = \{\theta \in \Theta \mid \Lambda_\theta > 0\}$, the above assumption is equivalent to the existence of a map $g\colon \Theta^{(\Lambda)} \rightarrow \Theta$ such that for $\theta_1 = g(\theta)$ we have:

$$\theta_1 \neq \theta \quad \text{and} \quad \alpha_\theta^{(\theta)} \cdot (z_\theta + \Lambda_\theta) \leqq \alpha_{\theta_1}^{(\theta)} \cdot (z_{\theta_1} + \Lambda_{\theta_1}).$$

The conclusion $\Lambda \leqq f(0)$ we are aiming at can be restated as:

$$\forall\, \theta \in \Theta\colon \quad \alpha_\theta^{(\theta)} \cdot (z_\theta + \Lambda_\theta) \leqq \max_{\theta_1} \alpha_{\theta_1}^{(\theta)} z_{\theta_1}. \qquad (*)$$

Our idea of proof is to consider iterates g^m of g as far as they exist, i.e., as $\theta, g(\theta), g^2(\theta), \ldots, g^{m-1}(\theta) \in \Theta^{(\Lambda)}$. We will show below that

a) such a sequence of iterates cannot end up in a loop, i.e., $g^k(\theta) \neq g^m(\theta)$ for $k \neq m$, and

b) for $\theta_1 = g^m(\theta)$ we have $\alpha_\theta^{(\theta)} \cdot (z_\theta + \Lambda_\theta) \leqq \alpha_{\theta_1}^{(\theta)} \cdot (z_{\theta_1} + \Lambda_{\theta_1})$.

From a) and finiteness of the set $\Theta^{(\Lambda)}$ follows that for every $\theta \in \Theta^{(\Lambda)}$ there does exist a maximal sequence of iterates $\theta, g(\theta), g^2(\theta), \ldots, g^m(\theta)$ such that $\theta_1 = g^m(\theta)$ is no longer in the domain $\Theta^{(\Lambda)}$ of g, and hence $\Lambda_{\theta_1} = 0$. Applying (b) above to this case we get $\alpha_\theta^{(\theta)} \cdot (z_\theta + \Lambda_\theta) \leqq \alpha_{\theta_1}^{(\theta)} z_{\theta_1}$, which is the conclusion (*) we wanted to reach. To finish the proof, we need to take care of a) and b):

ad a): If for $\theta \in \Theta^{(\Lambda)}$ and $\theta_1 = g(\theta)$ we can show that

$$\alpha_\theta^{(\theta)} \cdot (z_\theta + \Lambda_\theta) < \alpha_{\theta_1}^{(\theta_1)} \cdot (z_{\theta_1} + \Lambda_{\theta_1}),$$

then loops in sequences of iterates obviously cannot occur. Proof of this inequality: since $\theta \in \Theta^{(\Lambda)}$ we have $\Lambda_\theta > 0$, hence:

$$0 < \alpha_\theta^{(\theta)} \cdot (z_\theta + \Lambda_\theta) \leqq \alpha_{\theta_1}^{(\theta)} \cdot (z_{\theta_1} + \Lambda_{\theta_1}).$$

This implies $z_{\theta_1} + \Lambda_{\theta_1} > 0$. Since $\theta_1 \neq \theta$ we also have $\alpha_{\theta_1}^{(\theta)} < \alpha_{\theta_1}^{(\theta_1)}$, and hence:

$$\alpha_\theta^{(\theta)} \cdot (z_\theta + \Lambda_\theta) \leqq \alpha_{\theta_1}^{(\theta)} \cdot (z_{\theta_1} + \Lambda_{\theta_1}) < \alpha_{\theta_1}^{(\theta_1)} \cdot (z_{\theta_1} + \Lambda_{\theta_1}).$$

ad b): We will use induction: Assume for $m \geqq 0$ *that* $\theta_1 = g^m(\theta)$ exists and

$$\alpha_\theta^{(\theta)} \cdot (z_\theta + \Lambda_\theta) \leqq \alpha_{\theta_1}^{(\theta)} \cdot (z_{\theta_1} + \Lambda_{\theta_1}).$$

(This is trivially true for $m = 0$.) We show that it also holds for $m + 1$, if $\theta_1 \in \Theta^{(\Lambda)}$. Let then be $\theta_2 = g(\theta_1) = g^{m+1}(\Theta)$, for which we get:

$$\alpha_{\theta_1}^{(\theta_1)} \cdot (z_{\theta_1} + \Lambda_{\theta_1}) \leqq \alpha_{\theta_2}^{(\theta_1)} \cdot (z_{\theta_2} + \Lambda_{\theta_2}).$$

Combining this with the induction assumption, we obtain:

$$\alpha_{\theta}^{(\theta)} \cdot \alpha_{\theta_1}^{(\theta_1)} \cdot (z_{\theta}+\Lambda_{\theta}) \leqq \alpha_{\theta_2}^{(\theta_1)} \cdot \alpha_{\theta_1}^{(\theta)} \cdot (z_{\theta_2}+\Lambda_{\theta_2}).$$

To the right hand side we may apply 6.2.1), which gives

$$\alpha_{\theta}^{(\theta)} \cdot \alpha_{\theta_1}^{(\theta_1)} \cdot (z_{\theta}+\Lambda_{\theta}) \leqq \alpha_{\theta_1}^{(\theta_1)} \cdot \alpha_{\theta_2}^{(\theta)} \cdot (z_{\theta_2}+\Lambda_{\theta_2}),$$

and hence the required inequality follows:

$$\alpha_{\theta}^{(\theta)} \cdot (z_{\theta}+\Lambda_{\theta}) \leqq \alpha_{\theta_2}^{(\theta)} \cdot (z_{\theta_2}+\Lambda_{\theta_2}). \quad \square$$

Proposition 6.4. *The primal function $c \to k_c^{\lambda}(z)$ is lower differentiable at the supremum thresholds, and the lower partial derivatives are:*

$$D_{c;\,-e^{(\theta)}}\, k_c^{\lambda}(z) = \frac{1}{\alpha_{\theta}^{(\theta)}} \cdot \left(-k^{\alpha^{(\theta)}}(z) + \sum_{\theta_1 \neq \theta} \alpha_{\theta_1}^{(\theta)} z_{\theta_1}\right).$$

Proof. Lower differentiability follows immediately from 6.3 and 5.14. The lower partial derivatives can be obtained from the remarks preceding 5.14, from which it follows that the θ-component of the largest element in the set $\{\Lambda \geqq 0 \,|\, \langle \tilde{W}(z+\Lambda), \Lambda\rangle = \langle c, \Lambda\rangle, \Lambda \geqq 0\}$:

$$\begin{aligned} D_{c;\,-e^{(\theta)}}\, k_c^{\lambda}(z) &= \frac{1}{\alpha_{\theta}^{(\theta)}} \cdot \max_{\theta_1} \alpha_{\theta_1}^{(\theta)} z_{\theta_1} - z_{\theta} \\ &= \frac{1}{\alpha_{\theta}^{(\theta)}} \cdot (\max_{\theta_1} \alpha_{\theta_1}^{(\theta)} z_{\theta_1} - \alpha_{\theta}^{(\theta)} z_{\theta}) \\ &= \frac{1}{\alpha_{\theta}^{(\theta)}} \cdot \left(-k^{\alpha^{(\theta)}}(z) + \sum_{\theta_1 \neq \theta} \alpha_{\theta_1}^{(\theta)} z_{\theta_1}\right) \end{aligned}$$

where the last equality can be obtained from the end of Sect. 2. □

Proposition 6.5. *The loss function $\tilde{W}_{\theta}^{\lambda}$ is robust under ε-contamination of (P_{θ}) if*

$$(1-\varepsilon) \cdot S^{(P)}(k^{\alpha^{(\theta)}}) + \varepsilon \cdot \frac{1}{n} \leqq \frac{1}{n} \cdot (1-\alpha_{\theta}^{(\theta)}) \qquad \forall\, \theta.$$

Proof. By 5.12 and in view of lower differentiability (6.4), the criterion is

$$S^{(P)}(D_{c;\,-e^{(\theta)}}\, k_c^{\lambda}) \geqq \frac{1}{n} \cdot \frac{\varepsilon}{1-\varepsilon} \qquad \forall\, \theta.$$

With 6.4 we rewrite the left hand side:

$$\begin{aligned} S^{(P)}(D_{c;\,-e^{(\theta)}}\, k_c^{\lambda}) &= \frac{1}{\alpha_{\theta}^{(\theta)}} \cdot S^{(P)}(-k^{\alpha^{(\theta)}}(z) + \sum_{\theta_1 \neq \theta} \alpha_{\theta_1}^{(\theta)} z_{\theta_1}) \\ &= \frac{1}{\alpha_{\theta}^{(\theta)}} \cdot \left(-S^{(P)}(k^{\alpha^{(\theta)}}) + \frac{1}{n} \cdot (1-\alpha_{\theta}^{(\theta)})\right) \end{aligned}$$

where we used the basic fact $S^{(P)}(z_\theta)=\frac{1}{n}$, see part I, Sect. 5. Simple calculation yields the assertion. □

After these technicalities, things fall in place since Theorem 5.18 is now applicable. We would like to interpret the results and eliminate the purely technical devices such as the mixing measures λ and the threshold derivatives $D_{c;\,-e^{(\theta)}}\,k_c^\lambda(z)$ from our formulations. We summarize as follows:

Theorem 6.6. *Let $\alpha^*=(\alpha_\theta^*)$ be a prior satisfying $\alpha_\theta^*>0\ \forall\,\theta$, and as before, let $\alpha^{(\theta)}=(1-\eta)\cdot\alpha^*+\eta\cdot e^{(\theta)}$ be the extremal contaminated prior which concentrates all contamination on the parameter value θ. If we have*

$$(1-\varepsilon)\cdot S^{(P)}(k^{\alpha^{(\theta)}})+\varepsilon\cdot\frac{1}{n}\leqq\frac{1}{n}\cdot(1-\alpha_\theta^{(\theta)})\qquad\forall\,\theta,$$

then there exists an experiment $Q_\theta=(1-\varepsilon)\cdot P_\theta+\varepsilon\cdot H_\theta$ which is simultaneously least favorable for all contaminated priors $\alpha=(1-\eta)\cdot\alpha^*+\eta\cdot\zeta$ (where $\zeta\in K$ is arbitrary). For these contaminated priors α we have:

$$s^{(v)}(k^\alpha)=(1-\varepsilon)\cdot S^{(P)}(k^\alpha)+\varepsilon\cdot\frac{1}{n}.$$

Proof. We have to verify the assumptions of 5.18. Measurability of the standard loss function and continuity of the concave function are no problem. Second, for every α in the interior of K, i.e., $\alpha_\theta>0\ \forall\,\theta$, the standard loss $\tilde{W}_\theta^\alpha(z)$ takes on its supremum α_θ at any z for which $z_\theta=0$. All elements of our η-contamination neighborhood are in the interior of K. Last, there remains continuity and robustness of the mixture standard loss function $\tilde{W}_\theta^\lambda$. Continuity is no problem according to the remarks at the beginning of this section. Robustness follows from Proposition 6.5. □

Although our point of view is the one of the theory of comparison of experiments which is characterized by focusing on risks rather than procedures, we will decode the theorem in terms of procedures. Recall from Sect. 2 that for any standard loss function such as $\tilde{W}_\theta^\alpha$ the Bayes procedure for an exact model (P_θ) is the vector of densities $p=(p_\theta)$. This amounts to nothing else than using the ordinary Bayes procedure for (P_θ), e.g., in the case of classification the rule based on maximal weighted densities: given the observation y, decide for a parameter θ which satisfies $\alpha_\theta\cdot p_\theta(y)=\max_{\theta_1}\alpha_{\theta_1}\cdot p_{\theta_1}(y)$, or equivalently:

$$p(y)\in\{z\in K\mid a_\theta\, z_\theta=\max_{\theta_1}\alpha_{\theta_1}\, z_{\theta_1}\}=B_\theta(\alpha)$$

in the "standard formulation". Optimality of p is expressed by

$$S^{(P)}(k^\alpha)=R((P),p,(\tilde{W}^\alpha)),$$

see part I, Sect. 5. (Values of standard measures on concave functions are just Bayes risks with regard to the corresponding loss functions.)

Now consider the procedure which decides a priori in favor of θ. Its Bayes risk is independent of the model and depends on the prior weights only:

$$R((v), \sigma \equiv \theta, (\tilde{W}^\alpha)) = \frac{1}{n} \cdot (1-\alpha_\theta).$$

Assume that the density vector p does not boil down to an a priori decision, which means in the "standard formulation": *image*$(p) \not\subset A_\theta(\alpha)$ a.s. $\forall\,\theta$. The upper Bayes risk of p under ε-contamination becomes:

$$\begin{aligned} R((v), p, (\tilde{W}^\alpha)) &= (1-\varepsilon) \cdot R((P), p, (\tilde{W}^\alpha)) + \varepsilon \cdot \frac{1}{n} \sum_\theta \sup_y \tilde{W}_\theta^\alpha(p(y)) \\ &= (1-\varepsilon) \cdot S^{(P)}(k^\alpha) + \varepsilon \cdot \frac{1}{n}. \end{aligned}$$

With this in mind, we see that the inequality

$$(1-\varepsilon) \cdot S^{(P)}(k^\alpha) + \varepsilon \cdot \frac{1}{n} \leqq \frac{1}{n} \cdot (1-\alpha_\theta)$$

holds true iff:

1) the optimal procedure p based on (P_θ) does not degenerate to an a priori decision for θ, and

2) the procedure p yields a better or not worse upper Bayes risk under ε-contamination than deciding a priori for θ.

The assumption of Theorem 6.6 says that p must not break down in favor of θ in the sense of 1) and 2) even if the prior α^* is η-contaminated in favor of θ.

References

1. Chow, S.-N., Hale, J.K.: Methods of bifurcation theory. Series of Comprehensive Studies in Mathematics **251**. Berlin-Heidelberg-New York: Springer 1982
2. Huber, P.J.: A robust version of the probability ratio test. Annals of Mathematical Statistics **36**, 1753–1758 (1965)
3. Huber, P.J.: Robust confidence limits. Z. Wahrscheinlichkeitstheorie Verw. Gebiete **10**, 269–278 (1968)
4. Huber, P.J., Strassen, V.: Minimax tests and the Neyman-Pearson lemma for capacities. Ann. Stat. **1**, 251–263 (1973)
5. Holmes, B.H.: A course on optimization and best approximation. Lecture Notes in Mathematics **257**. Berlin-Heidelberg-New York: Springer 1972
6. Luenberger, D.G.: Optimization by vector space methods. New York-London-Sydney-Toronto: John Wiley 1969
7. Le Cam, L.: Sufficiency and approximate sufficiency. Ann. Math. Statist. **35**, 1419–1499 (1964)
8. Parthasarathy, K.R.: Probability measures on metric spaces. New York-San Francisco-London: Academic Press 1967
9. Smart, D.R.: Fixed point theorems. Cambridge Tracts in Mathematics **66**. Cambridge-London-New York: Cambridge University Press 1974

Received June 9, 1980; in revised form November 10, 1984