Asymptotically Minimax Robust Hypothesis Testing

Gökhan Gül, Member, IEEE
Technische Universität Darmstadt, 64283, Darmstadt, Germany
E-mail: ggul@spg.tu-darmstadt.de

Abstract

The design of asymptotically minimax robust hypothesis testing is formalized for the Bayesian and Neyman-Pearson tests of Type I and II. The uncertainty classes based on the KL-divergence, $\alpha$-divergence, symmetrized $\alpha$-divergence, total variation distance, as well as the band model, moment classes and p-point classes are considered. It is shown with a counterexample that minimax robust tests do not always exist. Implications between single sample-, all-sample- and asymptotic minimax robustness are derived. Existence and uniqueness of asymptotically minimax robust tests are proven using Sion’s minimax theorem and the Karush-Kuhn-Tucker multipliers. The least favorable distributions and the corresponding robust likelihood ratio functions are derived in parametric forms, which can then be determined by solving a system of equations. The proposed theory proves that Dabak’s design does not produce any asymptotically minimax robust test. A generalization of the theory to multiple-, decentralized- and sequential hypothesis testing is discussed. The derivations are evaluated, examplified and applied to spectrum sensing.

Index Terms

Distributed detection, hypothesis testing, robustness, least favorable distributions, minimax optimization, multiple hypothesis testing, sequential probability ratio test, spectrum sensing.

I. INTRODUCTION

The detection of the events of interest is fundamental to many practical applications such as radar, sonar, digital communications or seismology [1]. Formally, any real world example of binary decision making can be modeled by a binary hypothesis test, where under each hypothesis
$\mathcal{H}_j$, the received data $y = (y_1, \ldots, y_n)$ follows a particular probability distribution $F_j$, $j \in \{0, 1\}$. The goal is to design a decision rule $\delta$ which assigns every $y$ either to $\mathcal{H}_0$ or $\mathcal{H}_1$, such that a predefined objective function, for example the error probability, is minimised. Clearly, the performance of $\delta$ strictly depends on the correctness of the assumption that the received data $y$ indeed follows the distributions $F_0$ or $F_1$. However, such an assumption is too strict and often does not hold in practice, for example in digital communications over wireless channels subject to severe time varying interference with a few dominant interferers.

In such a scenario, a classical way of dealing with the deviations from the model assumptions is via parametric modeling. The parameters of the statistical model can be estimated on the fly using some robust estimation techniques, e.g. M-estimators [2], [3]. However, it is implicitly assumed that the shape of the distributions is still perfectly known. Additionally, the changes in the distributions should be slow enough to be able to accurately update the estimates. These assumptions are invalid for non-Gaussian and statistically time varying applications [4], [5] and therefore necessitate considering non-parametric approaches [6].

Non-parametric tests, for example the sign test or the Wilcoxon test, are widely used in practice because they make only mild assumptions on the nominal distributions, are cheap to implement and their performance is acceptable for a variety of detection problems [6]. Their main drawback is that when compared to an optimum detector, their performance can be far away from being satisfactory, especially if there is some a priori knowledge available about the nominal distributions. Moreover, non-parametric tests can be obtained as the limiting tests of the minimax robust test [7]. Therefore, it is reasonable to design minimax robust tests and resort to non-parametric tests if and only if the limiting case holds. Such a design also provides flexibility to determine the detector complexity and to trade-off robustness versus performance considering the available knowledge about the nominal distributions.

Minimax robust hypothesis testing is a generalized version of the composite hypothesis testing, in which under each hypothesis $\mathcal{H}_j$, all possible probability distributions $G_j$, that may be the true distribution of the received data $y$, are stacked together in a so called uncertainty class $\mathcal{G}_j$. The choice of the uncertainty classes is usually application dependent and most common choices are either model based, e.g. $\epsilon$-contamination model, or distance based, where every $G_j$ is a member of $\mathcal{G}_j$ if $D(F_j, G_j) \leq \epsilon_j$, for some suitable distance $D$ and a robustness parameter $\epsilon_j$. Consequently, the ultimate goal of the designer is to find a minimax robust decision rule $\hat{\delta}$.
and a pair of least favorable distributions (LFDs) $(\hat{G}_0, \hat{G}_1) \in \mathcal{G}_0 \times \mathcal{G}_1$ such that a predefined objective function is mini-maximised (or maxi-minimized). Under some mild conditions, such a design provides the most powerful test in a well defined minimax sense, i.e. a robust test which provides the best guaranteeable detection performance irrespective of uncertainties imposed on the statistical model.

Although minimax robust tests are highly preferred for applications, where reliable detection is of utmost importance, unfortunately such tests do not always exist over the set of deterministic decision rules. Existence of minimax robust tests are completely determined by the choice of uncertainty sets. In case a minimax robust test does not exist over deterministic decision rules, it may still exist over the set of randomized decision rules. The problem with such a design is that the designed test is minimax robust only for a single sample and cannot be extended to multiple samples while maintaining the minimax robustness. In the presence of multiple samples and absence of minimax robust tests, probably the best option is to consider asymptotically minimax robust tests, which minimise the asymptotic decrease rates of error probabilities. As will be shown in the following the asymptotically minimax test is also minimax robust if in fact a minimax robust test exists and is unique. In summary, minimax robust tests can be broadly classified into three categories:

1) **All-sample** minimax robust tests (over deterministic decision rules) [8], [9].

2) **Single-sample** minimax robust tests (over randomized decision rules) [10], [11], [12].

3) **Asymptotically** minimax robust tests (over deterministic decision rules)[13], [14].

in terms of the number of samples available for testing. Another way of categorization is through the choice of

1) **Uncertainty classes** defined on some probability space [7].

A. Related work

The earliest work in robust hypothesis testing is attributed to P. J. Huber, who published a robust version of the probability ratio test for the $\epsilon$-contamination and total variation classes of probability distributions on 1965 [8]. Huber derived the least favorable distributions and showed that the clipped likelihood ratio test was the minimax robust test for both uncertainty classes. The conclusions of this work was later extended by Huber and Strassen to a larger class, which includes five different classes as special cases [15]. The largest classes known, for which a
minimax robust test exists and is a version of \( \hat{g}_1/\hat{g}_0 = d\hat{G}_1/d\hat{G}_0 \), are the 2-alternating capacities \[9\]. All aforementioned works \[8, 15\] and \[9\] are all-sample minimax robust, and it was shown by \[16\] that such tests do not always exist, for example when the uncertainty classes are built with respect to the KL-divergence.

Clipped likelihood ratio tests (CLRTs) resulting from the uncertainty classes in \[8\] and \[15\] are widely used in practice, especially to deal with outliers. However, the models leading to CLRTs may be unrealistic for many applications, e.g., a signal value which tends to infinity, and occurs infinitely often is, almost never seen, but such a scenario is fully considered by the models in \[8\] and \[15\]. This was first observed by Dabak and Johnson, who suggested that eliminating such distributions could lead to a smoothed uncertainty model, which may be better suited for practical applications, where modeling errors is of interest. Based on this idea, they considered the KL-divergence as the distance to build the uncertainty classes and derived the corresponding robust test for the asymptotic case, i.e. as the number of measurements tends to infinite \[13\]. Under several assumptions, Levy showed that a single sample minimax robust test could be designed for the same uncertainty model, if the error minimizing decision rules are allowed to be randomized. Considering a similar approach all the assumptions made by Levy were later removed \[12\]. The shortcomings of the model with the KL-divergence is that both the distance as well as the a-priori probabilities of the nominal test are not selectable \[12\]. Replacing the KL-divergence with the \( \alpha \)-divergence these two final constraints were also removed in \[11\]. Surprisingly, Dabak and Johnson’s asymptotically robust test was different from Levy’s minimax robust test, which was also different from the CLRT. Even more interestingly, for the whole \( \alpha \)-divergence neighborhood and for any a-priori probabilities of the hypotheses, the corresponding minimax robust test was a censored likelihood ratio test, with a well defined randomizing function \( \hat{\delta} \) \[11\].

As mentioned before, single-sample minimax robust tests cannot be extended straightforwardly to multiple samples while maintaining the minimax robustness property. Additionally, all aforementioned designs do not allow incorporating approximately known positions, shapes or statistics of the actual probability distributions into the considered model. With this motivation, several other uncertainty models have been proposed in the literature. One approach is that the uncertainty classes can fully be defined in terms of the statistics of the actual distributions, such as the moments \[14\]. Another approach is to consider the p-point classes, which allow designation of
the desired amount of area to the non-overlapping sub-sets of the domain of density functions \[17\]. The band models, which was first proposed by Kassam \[18\] and later revised by Fauß et. al. \[19\], on the other hand, enable the assignment of the approximate shape and location to the actual distributions.

All aforementioned works in the field of robust hypothesis testing are theoretical. There are also application oriented works, for example \[20\], where Huber’s clipped likelihood ratio test is applied to robust detection of a known signal in nearly Gaussian noise. These results are later strengthened for a known signal in contaminated non-Gaussian noise \[4\]. Robust detection of stochastic signals for Gaussian signal and Gaussian mixture noise is also studied for small and large samples sizes \[21\]. The first paper studying p-point classes was also application oriented \[5\].

B. Motivation

The motivation of this paper can be stated as follows:

1) The derivations in \[13\], \[22\], which are later summarized in \[23\], do not yield asymptotically minimax robust (Neyman-Pearson) tests. This requires derivations and analysis, which lead to minimax robustness.

2) The theoretical designs for the asymptotic case consider the NP formulations by default, probably because they result in simpler solutions \[13\], \[14\]. However, by Chernoff \[24\], it is well known that the NP tests have the worst error exponents. Therefore, it is necessary to obtain the asymptotically minimax robust tests for the fastest decay rate of the error probability.

C. Summary of the paper and its contributions

In this paper, the design of asymptotically minimax robust binary hypothesis tests is studied for various uncertainty classes. The existence and uniqueness of minimax robust tests are proven in general. Considering the Karush-Kuhn-Tucker (KKT) approach, the least favorable distributions and the robust likelihood ratio functions (LRFs) are derived in parametric forms, which can be made explicit by solving a set of non-linear equations. The results are extended to decentralized detection, where several sensors are available and to sequential detection, where on-the-fly
detection is of interest. In the sequel, the contributions of this paper together with their relation to prior works are summarized.

1) Existence and uniqueness of minimax robust tests are analysed in general, and it is shown that any minimax robust test can be designed via solving

$$\min_{u \in (0,1)} \max_{(g_0, g_1) \in \mathcal{G}_0 \times \mathcal{G}_1} \int g_1^u g_0^{1-u} d\mu.$$ 

The corresponding test is all-sample minimax robust if there exists one and is unique, otherwise the test is asymptotically minimax robust.

2) For the KL-divergence neighborhood, the LFDs of the asymptotically minimax robust -NP tests- of Type I and II are obtained in parametric forms. The parameters of LFDs can be found by solving four non-linear coupled equations and the corresponding test is different from the ones derived in [13], [22], [23].

3) For uncertainty classes based on the KL-divergence, α-divergence, symmetrized α-divergence, total variation distance as well as the band- and ϵ-contamination models, the LFDs of the (asymptotically) minimax robust -rate minimizing- tests are obtained in parametric forms. For moment classes and p-point classes, the design of minimax robust tests is defined as a convex optimization problem. The derivations regarding the total variation distance generalize the ones obtained earlier by [8], via allowing non equal robustness parameters $\epsilon_0 \neq \epsilon_1$ to be chosen. Moreover, the related analytical designs explain the choice of distributions and necessary parameters made by Huber and Kassam. Notice that in [8] and [18], the tests are heuristically designed, i.e. the LFDs are some trial versions, which in turn yielded minimax robust tests. Additionally, it is proven that two special cases of the band model give rise to the clipped likelihood ratio tests, which are shown to be single sample minimax robust.

4) The theory derived is extended to multiple-, decentralized- and sequential hypothesis testing. A new criteria for the minimax robustness of sequential tests is proposed [7].

D. Outline of the paper

The rest of the paper is organized as follows. In Section II, a brief overview of the fundamental concepts in minimax robust hypothesis testing is given. In Section III, single-sample, all-sample and asymptotic minimax robustness are defined, and existence and uniqueness of minimax robust
tests are explained and exemplified. In Section [IV] the equations formulating asymptotic minimax robustness are derived, saddle value condition is characterized and the problem statement is made. In Section [V] the least favorable distributions and asymptotically minimax robust tests are obtained for various uncertainty classes. In Section [VI] generalizations of the theory to multiple-, decentralized and sequential detection are discussed. In Section [VII] simulations are performed to evaluate, exemplify and apply the theory to spectrum sensing. Finally in Section [VIII] the paper is concluded.

II. FUNDAMENTALS OF MINIMAX ROBUST HYPOTHESIS TESTING

Let \((\mathcal{Y}, \mathcal{A})\) be a measurable space, where \(\mathcal{A}\) is a Borel \(\sigma\)-algebra, and \(Y = (Y_1, \ldots, Y_n)\) be a vector of independent and identically distributed (i.i.d.) random variables (r.v.s) taking values on \(\mathcal{Y}^n\). Under the hypothesis \(H_j\), the nominal and the actual distributions of the r.v. \(Y\) and \(Y = Y_1\) are commonly denoted by \(F_j\) and \(G_j\), respectively. Furthermore, let \(f_j\) and \(g_j\) be the density functions of the distributions \(F_j\) and \(G_j\) with respect to a dominating measure \(\mu\).

The actual distribution \(G_j\) of the r.v. \(Y\) under \(H_j\) is not known exactly but it is defined to be a member of the classes of distributions \(\mathscr{G}_j\). Given the nominal distributions \(F_j\), the uncertainty classes can be defined as

\[
\mathscr{G}_j = \{G_j : D(G_j, F_j) \leq \epsilon_j\}, \quad j \in \{0, 1\},
\]

where every \(g_j\) is at least \(\epsilon_j > 0\) close to the nominal density \(f_j\), with respect to a certain distance \(D\), for example the \(f\)-divergence,

\[
D_f(G_j, F_j) = \int_Y f \left( \frac{g_j}{f_j} \right) f_j d\mu, \quad j \in \{0, 1\}, \tag{1}
\]

where \(f\) is a convex function satisfying \(f(1) = 0\). In (1), it is implicitly assumed that \(G_j \succ F_j\), hence \(D_f\) is a smooth distance [13], [11].

Consider the binary hypothesis testing problem,

\[
\mathcal{H}_0 : Y \sim G_0 \\
\mathcal{H}_1 : Y \sim G_1
\]

with two possibly composite hypotheses. The goal is to find a decision rule \(\delta : \mathcal{Y} \to [0, 1]\) which assigns every observation \(y\) either to \(\mathcal{H}_0\) or \(\mathcal{H}_1\). In fact, any choice of \(\delta\) yields the false alarm
probability

\[ P_F(\delta, g_0) = \int_{\mathcal{Y}} \delta g_0 d\mu \]

miss detection probability

\[ P_M(\delta, g_1) = \int_{\mathcal{Y}} (1 - \delta) g_1 d\mu \]

and the overall error probability

\[ P_E(\delta, g_0, g_1) = P(H_0) P_F(\delta, g_0) + P(H_1) P_M(\delta, g_1). \]

The optimality of \( \delta \in \Delta \) depends on two conditions:

**Case 1**: \( \epsilon_0 = \epsilon_1 = 0 \).

In this case, the likelihood ratio test

\[ \delta(y) = \begin{cases} 
0, & l(y) < t \\
\kappa(y), & l(y) = t \\
1, & l(y) > t 
\end{cases} \]

for some threshold \( t = P(H_0)/P(H_1) \) and a function \( \kappa : \mathcal{Y} \rightarrow [0, 1] \) is optimum in the sense that it minimizes the overall error probability both in Bayes and Neyman-Pearson sense, where \( l(y) = f_1(y)/f_0(y) \) is the nominal likelihood ratio corresponding to the observation \( y \).

**Case 2**: Either \( \epsilon_0 > 0 \) or \( \epsilon_1 > 0 \), or both \( \epsilon_0 > 0 \) and \( \epsilon_1 > 0 \).

In this case, at least one of the hypotheses is composite and the interest is in finding a decision rule which minimizes the error probability for the worst case scenario. In principle, such a test may be obtained by solving the minimax optimization problem,

\[
\sup_{(g_0, g_1) \in \mathcal{G}_0 \times \mathcal{G}_1} \min_{\delta \in \Delta} P_E(\delta, g_0, g_1) = \min_{\delta \in \Delta} \sup_{(g_0, g_1) \in \mathcal{G}_0 \times \mathcal{G}_1} P_E(\delta, g_0, g_1). \tag{2}
\]

A solution to (2) with the least favorable densities \( \hat{g}_0, \hat{g}_1 \) and the robust decision rule \( \hat{\delta} \) implies a saddle value, i.e.

\[ P_E(\hat{\delta}, \hat{g}_0, \hat{g}_1) \geq P_E(\hat{\delta}, \hat{g}_0, \hat{g}_1) \geq P_E(\hat{\delta}, g_0, g_1), \tag{3} \]

which in turn implies

\[ P_F(\hat{\delta}, \hat{g}_0) \geq P_F(\hat{\delta}, g_0), \]
\[ P_M(\hat{\delta}, \hat{g}_1) \geq P_M(\hat{\delta}, g_1), \]

since \( P_E \) is distinct in \( g_0 \) and \( g_1 \).
III. SINGLE-SAMPLE, ALL-SAMPLE AND ASYMPTOTIC MINIMAX ROBUSTNESS

In the previous section, all formulations are given for a single random variable \( Y \), which may in fact be also multi-variable. If observations are repeated, we have a vector of random variables \( Y = (Y_1, \ldots, Y_n) \), where each r.v. \( Y_k \) is subjected to uncertainties. The uncertainty classes associated with the r.v. \( Y_k \) may in general be dependent on \( k \), however \( Y = Y_k \) in distribution for all \( k \) will be assumed here. For repeated observations, either we may be interested in fixed sample size or asymptotic minimax robustness. The strategy to obtain a fixed sample size minimax robust test is to solve the minimax optimization problem (2) for a single sample and then extend it to multiple samples via

\[
l(y) = \frac{dG_1(y)}{dG_0(y)} = \prod_{k=1}^{n} \frac{g_1(y_k)}{g_0(y_k)}.
\]

In general it is not mandatory to solve (2) analytically, rather one may also propose a solution and then see if it satisfies the saddle value condition (3), which was followed by Huber [8].

In the following definition single sample minimax robustness is described. It is then extended to multiple samples. Existence and uniqueness of single sample and all sample minimax robust tests are mentioned in this section and that of the asymptotic minimax robust test more in details in the next sections.

**Definition III.1** (Single-sample minimax robustness). Let \( \hat{l} = \hat{g}_1/\hat{g}_0 \) be the robust likelihood ratio function. Then,

\[
G_0 \left[ \hat{l}(Y) < t \right] \geq \hat{G}_0 \left[ \hat{l}(Y) < t \right],
\]

\[
G_1 \left[ \hat{l}(Y) < t \right] \leq \hat{G}_1 \left[ \hat{l}(Y) < t \right]
\]

for all \( t \in \mathbb{R}_{\geq 0} \) and \( (G_0, G_1) \in \mathcal{G}_0 \times \mathcal{G}_1 \).

There is a tight connection between minimizing a distance and finding a solution to (4). This will be stated with the following theorem.

**Theorem III.1.** Over all \( (G_0, G_1) \in \mathcal{G}_0 \times \mathcal{G}_1 \),

\[
(\hat{G}_0, \hat{G}_1) \in \mathcal{G}_0 \times \mathcal{G}_1 \text{ satisfies (4)} \iff (\hat{G}_0, \hat{G}_1) \in \mathcal{G}_0 \times \mathcal{G}_1 \text{ minimizes } D_f
\]
for all twice differentiable convex functions $f$, where the $f$-divergence can alternatively be written as

$$D_f(G_0, G_1) = \int_0^\infty \left( \min(1, t) - \int_Y \min(g_0, tg_1) \, d\mu \right) \, d\mu_f(t)$$

with

$$\mu_f(a, b) = \partial^+ f(b) - \partial^+ f(a), \quad 0 < a < b < \infty$$

where $\partial^+$ denotes the right derivative operator.

Proof: Implication (5) was proven in [9, Section 6] for

$$D_f(G_0, G_1) := D_{f^*}(G_0, G_1) = \int_Y f \left( \frac{g_0}{g_0 + g_1} \right) (g_0 + g_1) \, d\mu.$$

For every $f^*$, the normalization $f^{**}(x) = f^*(x) - f^*(1/2)$ leads to $f^{**}(1/2) = 0$, where $f^{**}$ is also twice differentiable and convex. In [25, Equation 8], it was shown that $D_f = D_{f^{**}}$, where $D_f$ is the regular definition of the $f$-divergence given by (1). Hence, (5) follows, and the claim is proven. An alternative definition of $D_f$ given by (6) can be found in [26, Equation 10].

Depending on the definition of the uncertainty classes, $\mathcal{G}_0$ and $\mathcal{G}_1$, a minimax robust test satisfying Huber’s definition may or may not exit. For example it exits for the $\epsilon$-contamination classes of distributions [8] and it may not exist as given with the following example.

Example III.2. Let $U^n = \mathcal{U}(0, \frac{1}{n})$ be the uniform measure on the unit interval, where $\mathcal{G}_0$ corresponds to those with odd $n$ and $\mathcal{G}_1$ contains those with even $n$. For this choice of the uncertainty classes, there exists no $(\hat{G}_0, \hat{G}_1) \in \mathcal{G}_0 \times \mathcal{G}_1$ satisfying the inequalities in (4).

Proof: Let $f(x) = x^u$, where $D_f$ is to be maximized, since $f(x) = -x^u + 1$ with $f(1) = 0$ is an $f$-divergence for all $u \in (0, 1)$. There are two cases to be considered:

Case 1: $m > n$.

$$D_f(G_0, G_1) = \int_Y g_1^u g_0^{1-u} \, d\mu = \int_0^{1/m} m^u n^{1-u} \, dy = \left( \frac{n}{m} \right)^{1-u}$$

Case 1: $m < n$.

$$D_f(G_0, G_1) = \int_Y g_1^u g_0^{1-u} \, d\mu = \int_0^{1/n} m^u n^{1-u} \, dy = \left( \frac{m}{n} \right)^u$$

For every $u$, if $m > n$, then $n$ should be maximized and $m$ should be minimized and similarly, if $m < n$, then $n$ should be minimized and $m$ should be maximized such that $D_f$ is maximum.
This implies \(|m - n|\) should be minimized for both cases, which holds with \(\hat{m} = 1\). In this case neither of the conditions
\[
\left( \frac{n}{n + 1} \right)^{1-u} \geq \left( \frac{m}{m + 1} \right)^u \\
\left( \frac{m}{m + 1} \right)^u \geq \left( \frac{n}{n + 1} \right)^{1-u}
\]

is true for all \(u \in (0, 1)\). Hence no \((\hat{G}_0, \hat{G}_1) \in \mathcal{G}_0 \times \mathcal{G}_1\) minimizes all \(f\)-divergences. As a consequence of Theorem III.1 no pair is least favorable satisfying (4).

If the uncertainty classes do not allow a minimax robust test to exist for all thresholds, it may still be possible to obtain a minimax robust test for a unique decision rule, if randomized decision rules are considered, see [16]. The information in the randomization is lost by the multiplication of the likelihood ratios therefore straightforward extension of the test to multiple observations is not minimax robust. However, if a single sample minimax robust test exists, then it is also all-sample minimax robust. In order to prove this, let us first consider the following remark and lemma.

Remark III.1. Let \(X\) and \(Y\) be two random variables defined on the same measurable space \((\mathcal{Y}, \mathcal{A})\), having cumulative distribution functions \(P_X\) and \(P_Y\), respectively. \(X\) is called stochastically larger than \(Y\), i.e. \(X \succeq Y\), if \(P_Y(x) \geq P_X(x)\) for all \(x\).

Lemma III.3. Let \(X_1, X_2, Y_1\) and \(Y_2\) be four random variables on \((\mathcal{Y}, \mathcal{A})\), out of which \(X_1\) and \(X_2\), and \(Y_1\) and \(Y_2\) are independent. If \(X_1 \succeq Y_1\) and \(X_2 \succeq Y_2\), then \(X_1 + X_2 \succeq Y_1 + Y_2\).

Proof: From Remark III.1 we have \(P_{Y_1}(x) \geq P_{X_1}(x)\) and \(P_{Y_2}(x) \geq P_{X_2}(x)\) for all \(x\). Hence,
\[
P_{Y_1+Y_2}(z) = \int_{-\infty}^{+\infty} P_{Y_1}(z-x)dP_{Y_2}(x) \geq \int_{-\infty}^{+\infty} P_{X_1}(z-x)dP_{Y_2}(x) \\
\quad = \int \int_{x+y \leq z} dP_{X_1}(x)dP_{Y_2}(y) = \int_{-\infty}^{+\infty} P_{Y_2}(z-y)dP_{X_1}(y) \\
\quad \geq \int_{-\infty}^{+\infty} P_{X_2}(z-y)dP_{X_1}(y) = P_{X_1+X_2}(z).
\]

Proposition III.4 (All-sample minimax robustness). Minimax robustness defined by (4) for a single sample extends to multiple samples straightforwardly.
Proof: Let \( X_k = \ln \hat{g}_1 / \hat{g}_0(Y_k) \) for every \( k \). Then, from Remark III.1 we have \( \hat{X}_k^0 \succeq X_k^0 \). Applying Lemma III.3 to every pair of r.v.s \((X_k^0, \hat{X}_k^0)\) and \((X_{k+1}^0, \hat{X}_{k+1}^0)\) and iterating this process, we have \( \sum_{k=1}^n \hat{X}_k^0 \succeq \sum_{k=1}^n X_k^0 \), which in turn implies
\[
G_0 \left[ \sum_{k=1}^n X_k \leq t \right] \geq \hat{G}_0 \left[ \sum_{k=1}^n X_k \leq t \right].
\] (7)
The proof for the second inequality is similar and therefore omitted. ■

As to the uniqueness of the tests satisfying (4), all tests designed so far are known to be unique [15], [9].

If the uncertainty classes do not allow the existence of an all sample minimax robust test it is still possible to design the test for the case, where the total number of samples is large enough, and ideally for \( n \to \infty \). Such tests are called asymptotically minimax robust and can concretely be stated with the following definition.

**Definition III.2** (Asymptotic minimax robustness). Let \( S_n(Y) = \frac{1}{n} \ln \hat{l}(Y) \). Then, the tests satisfying
\[
\lim_{n \to \infty} \frac{1}{n} \ln G_0 \left[ S_n(Y) > t \right] \leq \lim_{n \to \infty} \frac{1}{n} \ln \hat{G}_0 \left[ S_n(Y) > t \right]
\]
\[
\lim_{n \to \infty} \frac{1}{n} \ln G_1 \left[ S_n(Y) \leq t \right] \leq \lim_{n \to \infty} \frac{1}{n} \ln \hat{G}_1 \left[ S_n(Y) \leq t \right]
\]
for some specific \( t \) and for all \((G_0, G_1) \in \mathcal{G}_0 \times \mathcal{G}_1\) are called asymptotically minimax robust.

The connection between asymptotic and all sample minimax robustness can be given with the following proposition.

**Proposition III.5.** For some uncertainty classes \( \mathcal{G}_0 \) and \( \mathcal{G}_1 \),

\[
\text{All sample minimax robustness } \iff \text{Asymptotically minimax robustness}
\]

if all sample and asymptotically minimax robust tests exit and are unique.

Proof: \( \implies \) is trivially true, and \( \impliedby \) is also true since both tests exist and are unique. ■

Proposition III.5 implies that finding an asymptotically minimax robust test is the best that can be done and the name *asymptotically* should not be misleading. Existence and uniqueness of asymptotically minimax robust tests will be discussed more in details in the next section.
IV. Formulation of Asymptotically Minimax Robustness

For many practical applications, there are more than a few samples and if the uncertainty model of interest does not allow an all-sample minimax robust test to exist, an asymptotic design is necessary. In the sequel, asymptotic minimax robustness will be formalized with first deriving the minimax equations, then, stating the existence of a saddle value and last, with a concrete definition of the problem.

A. Derivation of the Rate Functions

Large deviations theory can be used to derive the equations from where asymptotically minimax robust tests can be obtained. Let us consider the following theorem by Cramér [27].

Theorem IV.1. Let $X = (X_1, \ldots, X_n)$ be a vector of i.i.d. random variables, $S_n(X) = \frac{1}{n} \sum_{k=1}^{n} X_k$ be their average and $M_{X_1}(u) := E[e^{uX_1}] < \infty$ be the moment generating function of the r.v. $X_1$. Then, for all $t > E[X_1]$,

$$
\lim_{n \to \infty} \frac{1}{n} \ln P[S_n(X) > t] = -I(t)
$$

where the rate function

$$
I(t) = \sup_{u \in R} (tu - \ln M_{X_1}(u))
$$

(8)

is the Legendre transform of the log-moment generating function.

A proof of Theorem [IV.1] can be found in [23, pp. 108-111].

Remark IV.1. Theorem [IV.1] implies

$$
\lim_{n \to \infty} \frac{1}{n} \ln P[S_n(X) \leq t] = -I'(t)
$$

for all $t < E[X_1]$.

A proof of Remark [IV.1] can be found in [7, p. 84].

Let $X_k = \ln \hat{l}(Y_k)$, where $Y_k$ is distributed as $G_j \in \mathcal{G}_j$ under $\mathcal{H}_j$. Then, for all

$$
E_{G_0}[\ln \hat{l}(Y_1)] < t < E_{G_1}[\ln \hat{l}(Y_1)]
$$

(9)

from Theorem [IV.1] and Remark [IV.1]

$$
\lim_{n \to \infty} \frac{1}{n} \ln G_0[S_n(X) > t] = -I_0(t)
$$

$$
\lim_{n \to \infty} \frac{1}{n} \ln G_1[S_n(X) \leq t] = -I_1(t)
$$

(10)
where the rate functions are given by

\[ I_j(t) = \sup_{u \in \mathbb{R}} \left( tu - \ln M_{Y_1}^j(u) \right), \quad j \in \{0, 1\}, \quad (11) \]

with the moment generating functions

\[ M_{Y_1}^j(u) = \mathbb{E}_{G_j} \left[ \exp \left( u \ln \hat{l}(Y_1) \right) \right], \quad j \in \{0, 1\}. \quad (12) \]

For the selected robust likelihood ratio function \( \hat{l} \), the asymptotic decrease rates of false alarm and miss detection probabilities can be found from (11) for any threshold \( t \) satisfying (9). Of particular interest is the optimal value of \( t \) which maximizes the asymptotic decrease rate of the error probability. This problem was first solved by Chernoff [24]. In the following, a simple derivation will be made.

**B. Optimum Threshold**

Let the error probability be

\[ P_E(n, t) = P_0 P_F(n, t) + (1 - P_0) P_M(n, t) \]

for \( n \) samples and the threshold \( t \). From (10) one can write

\[ P_E(n, t) \approx C_F(n) P_0 \exp (-n I_0(t)) + C_M(n) (1 - P_0) \exp (-n I_1(t)), \]

where \( C_F \) and \( C_M \) satisfy

\[ \lim_{n \to \infty} \frac{1}{n} \ln C_F(n) = \lim_{n \to \infty} \frac{1}{n} \ln C_M(n). \]

Hence, the exponential decay rate of the error probability is governed by \( I_0 \) and \( I_1 \).

From [7, Remark. 5.2.2.], \( I_0 \) and \( I_1 \) are increasing and decreasing functions of \( u \), respectively. Let \( h_j : u \mapsto t \) be the mapping between maximizing \( u \) and \( t \) in (8). It is easy to see that \( h_j \) is increasing because it is the derivative of a convex function \( \ln M_{X_1}^j(u) \) [23, p. 77]. Hence, \( I_0(t) = I_0(h_0(u)) \) and \( I_1(t) = I_1(h_1(u)) \) are also increasing and decreasing functions respectively, as

\[ \frac{\partial I_0(h_0(u))}{\partial u} = I_0'(h_0(u)) h_0'(u) \geq 0, \]

\[ \frac{\partial I_1(h_1(u))}{\partial u} = I_1'(h_1(u)) h_1'(u) \leq 0. \]

December 12, 2017
Since $I_0$ and $I_1$ are increasing and decreasing functions, and the error probability $P_E$ decreases exponentially with $I_0$ and $I_1$, the maximum decay rate can be obtained via solving

$$\arg \lim_{n \to \infty} \min_t P_E(n, t) = \arg \max_t \min \{I_0(t), I_1(t)\}. $$

Furthermore, as $M_{Y_1}(u) = M_{Y_1}^0(u+1)$ for $G_j := \hat{G}_j$ together with (11) implies $I_1(t) = I_0(t) - t$, it is true that $I_0(0) = I_1(0)$ and one can write

$$I_m(t) = \min \{I_0(t), I_1(t)\} = \begin{cases} I_0(t), & t < 0 \\ I_1(t), & t > 0 \\ I_0(0) = I_1(0), & t = 0 \end{cases}. \quad (13)$$

Hence, we have

$$\arg \sup_t I_m(t) = 0. \quad (14)$$

Notice that if $G_j = \hat{G}_j$ a.e. is not true, we do not necessarily have (13) and (14).

### C. Minimax Equations

The minimax robust test that is intended to be designed is a likelihood ratio test with $\hat{t} = \hat{g}_1/\hat{g}_0$, for which the worst case data samples are also obtained from $\hat{g}_0$ and $\hat{g}_1$. As a result, $t = 0$ can safely be selected as the optimum threshold. Hence, using the equalities,

1) $t = 0$
2) $- \inf(f) = \sup(-f)$ for $f < 0$
3) $\inf(-f) = - \sup(f)$

the minimax equations can be obtained from (11) and (12) as

$$\hat{g}_0 = \arg \sup_{g_0 \in \mathcal{G}_0} \left( \inf_{u_0 \in \mathbb{R}} \ln \int_Y \left( \frac{\hat{g}_1}{\hat{g}_0} \right)^{u_0} g_0 d\mu \right),$$

$$\hat{g}_1 = \arg \sup_{g_1 \in \mathcal{G}_1} \left( \inf_{u_1 \in \mathbb{R}} \ln \int_Y \left( \frac{\hat{g}_1}{\hat{g}_0} \right)^{u_1} g_1 d\mu \right). \quad (15)$$

In their current forms, these two coupled equations are mathematically intractable, especially if $\mathcal{G}_0$ and $\mathcal{G}_1$ are infinite sets. A solution found for (15) with the least favorable densities $\hat{g}_0$ and $\hat{g}_1$ implies

$$C_0(\hat{g}_0, \hat{g}_1; u_0) = \inf_{u_0 \in \mathbb{R}} \ln \int_Y \hat{g}_1^{u_0} \hat{g}_0^{1-u_0} d\mu,$$

$$C_1(\hat{g}_0, \hat{g}_1; u_1) = \inf_{u_1 \in \mathbb{R}} \ln \int_Y \hat{g}_1^{1+u_1} \hat{g}_0^{-u_1} d\mu. \quad (16)$$

December 12, 2017
The following definition and remark will be used in the rest of the paper.

**Definition IV.1.**

\[ D_u(G_0, G_1) = \int_Y g_1^u g_0^{1-u} d\mu \]

is denoted as the \( u \)-divergence.

**Remark IV.2.** Some properties of the \( u \)-divergence are listed in Table I. While the points 1. and 4.-7. are trivially correct, 2. and 3. follow from the properties of the \( \alpha \)-divergence using 7. [28].

We can now prove the following proposition.

**Proposition IV.2.** For every \((\hat{g}_0, \hat{g}_1) \in \mathcal{G}_0 \times \mathcal{G}_1\) the results of both optimization problems in (16) are the same with \( u_0 = -u_1 \), where \( u_0 \in [0, 1] \), i.e. \( C_0(\hat{g}_0, \hat{g}_1; u_0) = C_1(\hat{g}_0, \hat{g}_1; -u_1) \).

**Proof:** Application of 6. in Table I to (16) implies a unique minimizing \( u_0 \), which lies in [0, 1]. Due to same reasoning, one can see that the minimizing \( u_1^* = -u_1 \) also lies in [0, 1]. Since \(-C_0\) is the Chernoff distance which is symmetric [23, p. 82], we have

\[ C_0(\hat{g}_0, \hat{g}_1; u_0) = C_0(\hat{g}_1, \hat{g}_0; u_0) = C_1(\hat{g}_0, \hat{g}_1; u_1^*). \]

Proposition IV.2 implies that both optimization problems are equivalent and have the same result for \( u_0 = -u_1 \). Hence, it is sufficient to consider only one of them. Eventually, the problem
to be solved reduces to

\[(\hat{g}_0, \hat{g}_1) = \arg \sup_{(g_0, g_1) \in \mathcal{G}_0 \times \mathcal{G}_1} \inf_u \ln D_u(G_0, G_1). \tag{17}\]

Equation 17 can further be simplified to

\[(\hat{g}_0, \hat{g}_1) = \arg \sup_{(g_0, g_1) \in \mathcal{G}_0 \times \mathcal{G}_1} \inf_u D_u(G_0, G_1) \tag{18}\]

because

\[\frac{\partial \ln D_u}{\partial u} = 0 \implies \frac{\partial D_u}{\partial u} = 0\]

and \(\ln\) is an increasing mapping from \([0, 1]\) to \(\mathbb{R}_{\leq 0}\), cf. 5. in Table I, i.e. for every \(u, (G_0, G_1)\) and \((G_0^*, G_1^*)\),

\[\ln D_u(G_0, G_1) < \ln D_u(G_0^*, G_1^*) \implies D_u(G_0, G_1) < D_u(G_0^*, G_1^*).\]

D. Saddle Value Condition

In this section existence of a saddle value, hence a solution to the minimax optimization problem in (18), is discussed. Uniqueness condition depends on the choice of the uncertainty classes and will be discussed in the next section. In general existence of a saddle value is described by a solution to

\[\min_u \sup_{(g_0, g_1) \in \mathcal{G}_0 \times \mathcal{G}_1} D_u(G_0, G_1) = \sup_{(g_0, g_1) \in \mathcal{G}_0 \times \mathcal{G}_1} \min_u D_u(G_0, G_1). \tag{19}\]

**Theorem IV.3** (Application of Sion’s minimax theorem \([29]\)). A solution to (19) exists if the following conditions hold:

- The objective function \(D_u\) is real valued, upper semi-continuous and quasi-concave on \(\mathcal{G}_0 \times \mathcal{G}_1\) for all \(u \in [0, 1]\).
- The objective function \(D_u\) is lower semi-continuous and quasi-convex on \([0, 1]\) for all \((g_0, g_1) \in \mathcal{G}_0 \times \mathcal{G}_1\).
- \([0, 1]\) is a compact convex subset of a linear topological space.
- \(\mathcal{G}_0 \times \mathcal{G}_1\) is a convex subset of a linear topological space.

**Proof:** The objective function is real valued, continuous in \(u\) and \((g_0, g_1)\), jointly concave on \(\mathcal{G}_0 \times \mathcal{G}_1\) for all \(u \in [0, 1]\), and convex on \([0, 1]\) for all \((g_0, g_1) \in \mathcal{G}_0 \times \mathcal{G}_1\), see 2. and 3. in Table I. The set \([0, 1]\) is trivially convex and is closed and bounded, hence compact with respect
to the standard topology by Heine-Borel theorem \cite[Theorem 2.41]{30}. Finally, \(\mathcal{G}_0\) and \(\mathcal{G}_1\) are convex sets, since \(D_f\) is a convex distance. As a result \(\mathcal{G}_0 \times \mathcal{G}_1\) is also convex.

Existence of a saddle value also implies
\[
D_u(\hat{G}_0, \hat{G}_1) \geq D_{u^*}(\hat{G}_0, \hat{G}_1) \geq D_{u^*}(G_0, G_1),
\]
where \(u^*\) is the minimizing \(u\), and \(\hat{G}_0\) and \(\hat{G}_1\) are the least favorable distributions.

### E. Problem Statement

From (20), given \((\hat{G}_0, \hat{G}_1)\), the objective function \(D_u\) needs to be minimized over \(u\) and given the minimizing \(u^*\), \(D_{u^*}\) needs to be maximized over \((G_0, G_1)\). This can compactly be written as

Maximization:
\[
\hat{g}_0 = \arg \sup_{g_0 \in \mathcal{G}_0} D_u(G_0, G_1) \quad \text{s.t.} \quad g_0 > 0, \quad \Upsilon(g_0) = \int_Y g_0 \, d\mu = 1
\]
\[
\hat{g}_1 = \arg \sup_{g_1 \in \mathcal{G}_1} D_u(G_0, G_1) \quad \text{s.t.} \quad g_1 > 0, \quad \Upsilon(g_1) = \int_Y g_1 \, d\mu = 1
\]

Minimization:
\[
u^* = \arg \min_{u \in [0,1]} D_u(\hat{G}_0, \hat{G}_1).
\]

The maximization stage involves two separate constrained optimization problems, which are coupled. The following minimization problem can be solved once \(\hat{g}_0\) and \(\hat{g}_1\) are derived as functions of \(u\).

### V. Least Favorable Distributions and Asymptotically Minimax Robust Tests

In this section LFDs and the asymptotically minimax robust tests are derived for various uncertainty classes considering the minimax optimization problem given by (21). Additionally, the asymptotic NP tests are also derived. Complete derivations are carried out for the Kullback-Leibler (KL)-divergence neighborhood, and similar steps are skipped for the sake of clarity when the same procedure is repeated for the \(\alpha\)- and the symmetrized \(\alpha\)-divergences. The derivations also include the uncertainty classes based on the total variation distance as well as the band model, moment classes, and p-point classes.
A. KL-divergence Neighborhood

The KL-divergence is a special case of the \( f \)-divergence, \( D_f \), with \( f(x) = x \ln x \), i.e.

\[
D_{KL}(G_j, F_j) = \int_Y \ln \left( \frac{g_j}{f_j} \right) g_j d\mu.
\]

It is the classical information divergence [31] and was used in earlier works as the distance to create the uncertainty classes [13], [10]. It is a smooth distance, hence suitable to deal with modeling errors [12]. The uncertainty classes are obtained as before, i.e. by defining

\[
\mathcal{G}_j = \{ G_j : D_{KL}(G_j, F_j) \leq \epsilon_j \}, \quad j \in \{0, 1\}. \tag{22}
\]

1) Rate Minimizing Tests: Asymptotically minimax robust tests for the KL-divergence neighborhood can be stated with the following theorem.

**Theorem V.1.** For the uncertainty classes given by (22), the LFDs

\[
\begin{align*}
\hat{g}_0 &= \exp \left[ -\frac{\lambda_0 - \mu_0}{\lambda_0} \right] \exp \left[ \frac{(1 - u)(\hat{g}_1/\hat{g}_0)^u}{\lambda_0} \right] f_0, \\
\hat{g}_1 &= \exp \left[ -\frac{\lambda_1 - \mu_1}{\lambda_1} \right] \exp \left[ \frac{u(\hat{g}_1/\hat{g}_0)^{-1+u}}{\lambda_1} \right] f_1,
\end{align*}
\]

with the robust likelihood ratio function

\[
\frac{\hat{g}_1}{\hat{g}_0} = \exp \left[ -\frac{\mu_1 + \mu_0}{\lambda_1} \right] \exp \left[ \frac{u(\hat{g}_1/\hat{g}_0)^{-1+u} + (-1 + u)(\hat{g}_1/\hat{g}_0)^u}{\lambda_1} \right] l \tag{23}
\]

provide a unique solution to (21). Moreover, the Lagrangian parameters \( \lambda_0 \) and \( \lambda_1 \), hence \( \mu_0 \) and \( \mu_1 \), can be obtained by solving

\[
\begin{align*}
\int_Y r_0 \ln (r_0/s_0) f_0 d\mu/s_0 &= \epsilon_0, \\
\int_Y r_1 \ln (r_1/s_1) f_1 d\mu/s_1 &= \epsilon_1,
\end{align*}
\]

\[
\hat{g}_1/\hat{g}_0 = (r_1s_0)/(r_0s_1), \tag{24}
\]

where

\[
\begin{align*}
s_0(\lambda_0) &= \int_Y r_0(\lambda_0, \hat{g}_1/\hat{g}_0) f_0 d\mu = \int_Y \exp \left[ \frac{(1 - u)(\hat{g}_1/\hat{g}_0)^u}{\lambda_0} \right] f_0 d\mu = \exp \left[ \frac{\lambda_0 + \mu_0}{\lambda_0} \right], \\
s_1(\lambda_1) &= \int_Y r_1(\lambda_1, \hat{g}_1/\hat{g}_0) f_1 d\mu = \int_Y \exp \left[ \frac{u(\hat{g}_1/\hat{g}_0)^{-1+u}}{\lambda_1} \right] f_1 d\mu = \exp \left[ \frac{\lambda_1 + \mu_1}{\lambda_1} \right].
\end{align*}
\]
A proof of Theorem [V.1] is given in three stages. In the maximization stage, Karush-Kuhn-Tucker (KKT) multipliers are used to determine the parametric forms of the LFDs, $\hat{g}_0$ and $\hat{g}_1$. In the minimization stage, the minimizing $u$ and the uniqueness conditions are established. Finally, in the last stage, a set of equations are derived from where the Lagrangian parameters can be obtained.

Proof:

a) Maximization: Consider the Lagrangian

$$L_0(g_0, g_1, \lambda_0, \mu_0) = D_u(G_0, G_1) + \lambda_0(\epsilon_0 - D_{\text{KL}}(G_0, F_0)) + \mu_0(1 - \Upsilon(g_0)),$$  \hspace{1cm} (25)

where $\lambda_0$ and $\mu_0$ are the KKT multipliers. A solution to (25) can uniquely be determined, in case all KKT conditions are met [32, Chapter 5], because $L_0$ is a strictly concave functional of $g_0$, as $\frac{\partial^2 L_0}{\partial g_0^2} < 0$ for every $\lambda_0 > 0$. Writing (25) explicitly, it follows that

$$L_0(g_0, g_1, \lambda_0, \mu_0) = \int_Y \left[ \left( \frac{g_1}{g_0} \right)^u - \lambda_0 \ln \left( \frac{g_0}{f_0} \right) - \mu_0 \right] g_0 d\mu + \lambda_0 \epsilon_0 + \mu_0.$$  \hspace{1cm} (26)

Imposing the stationarity condition of KKT multipliers and hereby taking the Gâteaux’s derivative of Equation (26) in the direction of $\psi_0$, yields

$$\int_Y \left[ (1 - u) \left( \frac{g_1}{g_0} \right)^u - \lambda_0 \ln \left( \frac{g_0}{f_0} \right) - \lambda_0 - \mu_0 \right] \psi_0 d\mu,$$

which implies

$$(1 - u)g_0^{-u}g_1^u - \lambda_0 \ln g_0 = \lambda_0 + \mu_0 - \lambda_0 \ln f_0,$$  \hspace{1cm} (27)

since $\psi_0$ is an arbitrary function. Similarly, writing the Lagrangian $L_1$ for $D_{\text{KL}}(G_1, F_1)$ and the Lagrangian parameters $\lambda_1$ and $\mu_1$, and taking the Gâteaux’s derivative with respect to $g_1$ in the direction of $\psi_1$ leads to

$$ug_1^{-1+u}g_0^{1-u} - \lambda_1 \ln g_1 = \lambda_1 + \mu_1 - \lambda_1 \ln f_1.$$  \hspace{1cm} (28)

From (27) and (28) the least favorable densities can be obtained as a functional of the robust likelihood ratio function $\hat{g}_1/\hat{g}_0$ and the nominal densities $f_0$ and $f_1$ as given in Theorem [V.1]. The second Lagrangian $L_1$ is similarly strictly concave for every $\lambda_1 > 0$. 

December 12, 2017  
DRAFT
b) Minimization: If the LFDs are known, finding the minimizing \( u \) is a simple convex optimization problem. Because \( \inf \) and \( \int \) can be interchanged
\[
(\hat{g}_0, \hat{g}_1) = \arg \sup_{(g_0, g_1) \in \mathcal{G}_0 \times \mathcal{G}_1} \int_{Y} \left( \frac{g_1}{g_0} \right)^u g_0 d\mu,
\]
as the integrand is continuous both in \( u, g_0 \) and \( g_1 \), see 2. and 3. in Table I. Performing the minimization leads to
\[
u^* = \arg \int_{Y} \ln \left( \frac{g_1}{g_0} \right) \left( \frac{g_1}{g_0} \right)^u g_0 d\mu = 0
\]
and reformulation of the original problem as
\[
(\hat{g}_0, \hat{g}_1) = \arg \sup_{(g_0, g_1) \in \mathcal{G}_0 \times \mathcal{G}_1} \int_{Y} \left( \frac{g_1}{g_0} \right)^{u^*} g_0 d\mu.
\]
However, before performing the maximization, it is not possible to write \( u^* \) analytically as a function of \( g_0 \) and \( g_1 \) and insert into (31). Also, after the maximization the least favorable densities are some, not necessarily simple, functions of \( u \). In fact, the objective function is not necessarily convex after inserting \( g_0(u) \) and \( g_1(u) \) in (29). Therefore, if minimization is performed after the maximization, the condition \( (30) \) does not necessarily hold. In this case, performing maximization first is more suitable, because finding the optimum value of \( u \) in \([0, 1]\) can easily be performed.

In the previous section, the parametric forms of the LFDs are obtained uniquely as \( L_0 \) and \( L_1 \) are a strictly concave functionals of \( g_0 \) and \( g_1 \), respectively, as long as \( \lambda_0 \) and \( \lambda_1 \) are positive. Similarly, for every pair of density functions, \((g_0, g_1) \in \mathcal{G}_0 \times \mathcal{G}_1 \), \( D_u \) is strictly convex in \( u \in (0, 1) \). Hence, for \( \hat{g}_0 \) and \( \hat{g}_1 \) the minimizing \( u \) is unique as well.

c) Equation Solving: In order to obtain the parameters, given any choice of \( u \), originally there are four non-linear equations
\[
\Upsilon(g_0(\lambda_0, \mu_0, \lambda_1, \mu_1)) = 1
\]
\[
\Upsilon(g_1(\lambda_0, \mu_0, \lambda_1, \mu_1)) = 1
\]
\[
D_{KL}(g_0(\lambda_0, \mu_0, \lambda_1, \mu_1), f_0) = \epsilon_0
\]
\[
D_{KL}(g_1(\lambda_0, \mu_0, \lambda_1, \mu_1), f_1) = \epsilon_1
\]
which need to be solved together with (23). These five equations can be reduced to three without any loss of generality. From the first and second equations we have \( s_0 \) and \( s_1 \) as functionals
of \( r_0 \) and \( r_1 \), respectively, as given in Theorem V.1. Using \((r_0, s_0)\) and \((r_1, s_1)\) in the last two equations of (32) as well as in (23) the three equations given in (24) can be obtained. □

2) Neyman-Pearson Tests: The asymptotic NP tests are designed in such a way that one of the error exponents has the highest exponential decay rate, while the other (although not wanted) has the lowest. For Type-I NP-tests the threshold is chosen as \( t_0 = \lim_{\epsilon \to 0} E_{G_0}[\ln \hat{I}(Y_1)] + \epsilon \) such that \( P_F \) is asymptotically guaranteed to get below any \( \epsilon > 0 \), while \( P_M \) has the highest decay rate. Similarly, for Type-II NP-tests the threshold is chosen as \( t_1 = \lim_{\epsilon \to 0} E_{G_1}[\ln \hat{I}(Y_1)] - \epsilon \) such that \( P_M \) is asymptotically guaranteed to get below any \( \epsilon > 0 \), while \( P_F \) has the highest decay rate. Using the thresholds \( t_0 \) and \( t_1 \) in (11) and keeping in mind that \( I_1(t) = I_0(t) - t \), one can obtain for the Type-I NP-test \( I_0(t_0) = 0 \) and \( I_1(t_0) = -t_0 = D_{KL}(G_0, G_1) \) and for the Type-II NP-test, \( I_0(t_1) = t_1 = D_{KL}(G_1, G_0) \) and \( I_1(t_1) = 0 \). Hence, the minimax problem formulation becomes

\[
\begin{align*}
\text{Type-I NP-test:} & \quad \min_{(g_0, g_1) \in \mathcal{F}_0 \times \mathcal{G}_1} D_{KL}(G_0, G_1) \quad \text{s.t.} \quad g_0 > 0, \; g_1 > 0, \; \Upsilon(g_0) = 1, \; \Upsilon(g_1) = 1 \\
\text{Type-II NP-test:} & \quad \min_{(g_0, g_1) \in \mathcal{F}_0 \times \mathcal{G}_1} D_{KL}(G_1, G_0) \quad \text{s.t.} \quad g_0 > 0, \; g_1 > 0, \; \Upsilon(g_0) = 1, \; \Upsilon(g_1) = 1
\end{align*}
\]

(33)

A solution to the Type-I NP-test formulation can be stated with the following theorem.

**Theorem V.2.** The LFDs of the asymptotically minimax robust Type-I NP-test are given by

\[
\begin{align*}
\hat{g}_0 &= \left(1 + \frac{\lambda_1}{\lambda_0}\right)^{-\frac{1}{\lambda_0}} \exp \left[-1 - \frac{1 + \mu_0}{\lambda_0}\right] W \left[\lambda_0 \exp \left[-\frac{\lambda_1 - \mu_0 + \lambda_0 g_1}{\lambda_1 (1 + \lambda_0)} \right] l^{-\frac{\lambda_0}{1 + \lambda_0}}\right] f_0 \\
\hat{g}_1 &= \left(1 + \frac{\lambda_1}{\lambda_0}\right)^{-\frac{1}{\lambda_0}} \exp \left[-1 - \frac{1 + \mu_0}{\lambda_0}\right] W \left[\lambda_0 \exp \left[-\frac{\lambda_1 - \mu_0 + \lambda_0 g_1}{\lambda_1 (1 + \lambda_0)} \right] l^{-\frac{\lambda_0}{1 + \lambda_0}}\right] f_0
\end{align*}
\]

(34)

where \( W \) is the Lambert-W function.

**Proof:** The solution can again be obtained by KKT multipliers. Considering the Lagrangians

\[
\begin{align*}
L(g_0, g_1, \lambda_0, \mu_0) &= D_{KL}(G_0, G_1) + \lambda_0 (D_{KL}(G_0, F_0) - \epsilon_0) + \mu_0 (\Upsilon(g_0) - 1), \\
L(g_0, g_1, \lambda_1, \mu_1) &= D_{KL}(G_0, G_1) + \lambda_1 (D_{KL}(G_1, F_1) - \epsilon_1) + \mu_1 (\Upsilon(g_1) - 1),
\end{align*}
\]

(35)
and following the same steps as before, one can get, respectively,

\[ g_1 = \exp [1 + \lambda_0 + \mu_0] g_0^{1+\lambda_0} f_0^{-\lambda_0}, \quad (36) \]

\[ g_0 = g_1 (\mu_1 + \lambda_1 (1 + \ln (g_1/f_1))). \quad (37) \]

Solving (36) and (37) for \( g_0 \) and \( g_1 \), respectively, the least favorable densities of the asymptotically minimax robust Type-I NP-test can be obtained as given in Theorem V.2.

Remark V.1. The Type-II minimax robust NP-test can similarly be obtained either by following the same procedure for the objective function \( D_{KL}(G_1, G_0) \) or by considering the same equations given by (34). To accomplish the latter one, before the optimization \( \epsilon_0 \) and \( f_0 \) need to be interchanged by \( \epsilon_1 \) and \( f_1 \) respectively, and after obtaining the LFDs, \( \hat{g}_0 \) needs to be interchanged by \( \hat{g}_1 \), cf. (33). The related parameters can be obtained directly by solving the equations in (32) for the LFDs (34). As a side note both NP-tests are the limiting tests of the rate minimizing asymptotically minimax robust test as

\[ \max_{(g_0, g_1) \in \mathcal{G}_0 \times \mathcal{G}_1} D_u(G_0, G_1) \equiv \min_{(g_0, g_1) \in \mathcal{G}_0 \times \mathcal{G}_1} D_\alpha(G_0, G_1), \quad \forall u = \alpha \in (0, 1) \]

and \( D_\alpha(G_0, G_1) \) converges to \( D_{KL}(G_0, G_1) \) and \( D_{KL}(G_1, G_0) \), respectively, for \( \alpha \to 1 \) and \( \alpha \to 0 \) from 7. in Table I, see also Section V-B.

Interestingly, the robust likelihood ratio function is a nonlinear functional of only \( l \). Moreover, compared to the rate minimizing asymptotically minimax robust tests, the LFDs of their NP counterparts are given only as a functional of the nominal distributions, without coupling with \( \hat{g}_1/\hat{g}_0 \). This simplification, however, results in a complication of the closed form LFDs. Note that the problem formulation given by (33) differs from that of Dabak’s formulation [13], [22], see also [23, pp. 250-255], i.e. Dabak’s test is the result of a joint minimization of \( I_0(t_1) \) over all \( g_1 \in \mathcal{G}_1 \) and \( I_1(t_0) \) over all \( g_0 \in \mathcal{G}_0 \), hence it yields simpler analytic forms for the LFDs, but not an asymptotically minimax robust test, see Section VII. However, Dabak’s test is surprisingly asymptotically minimax robust for the expected number of samples of the sequential probability ratio test [7], [12].

B. \( \alpha \)-divergence Neighborhood

Another alternative for the choice of the uncertainty classes is the \( \alpha \)-divergence,

\[ D_\alpha(G_j, F_j) := \frac{1}{\alpha(1-\alpha)} \left( \int \left( (1-\alpha)f_j + \alpha g_j - g_j^{\alpha} f_j^{1-\alpha} \right) \text{d}\mu \right), \quad \alpha \in \mathbb{R} \setminus \{0, 1\} \]
which is a special case of the $f$-divergence with

$$ f(t) = \begin{cases} 
- \ln t, & \alpha = 0 \\
 t \ln t, & \alpha = 1 \\
 \frac{t^{\alpha-1}}{\alpha(\alpha-1)}, & \text{otherwise}
\end{cases} $$

The $\alpha-$divergence also includes various distances as special cases [28, p.1537], e.g. $D_{KL}$ as $\alpha \to 1$, and can be estimated from the data samples [33]. The LFDs resulting from the $\alpha-$ divergence neighborhood is stated with the following theorem.

**Theorem V.3.** The least favorable densities of the $\alpha-$divergence neighborhood can be given as

$$ \hat{g}_0 = \left( \frac{1 - \alpha}{\lambda_0} \left( \mu_0 - (1 - u) \left( \frac{\hat{g}_1}{\hat{g}_0} \right)^u \right) + 1 \right)^{\frac{1}{\alpha-1}} f_0, $$

$$ \hat{g}_1 = \left( \frac{1 - \alpha}{\lambda_1} \left( \mu_1 - u \left( \frac{\hat{g}_1}{\hat{g}_0} \right)^{-1+u} \right) + 1 \right)^{\frac{1}{\alpha-1}} f_1, $$

where

$$ \frac{\hat{g}_1}{\hat{g}_0} = \left( \frac{\frac{1 - \alpha}{\lambda_1} \left( \mu_1 - u \left( \frac{\hat{g}_1}{\hat{g}_0} \right)^{-1+u} \right) + 1 \right)^{\frac{1}{\alpha-1}} l. $$

**Proof:** The proof follows by using the same Lagrangian approach as before, i.e. by replacing $D_{KL}$ with $D_\alpha$ in (25) and performing the derivations.

The parameters are obtained similarly by solving four non-linear equations coupled with (39).

1) **Special Cases:** The LFDs in (38) can explicitly be written for some special choices of the parameters. For instance if $\alpha = 1/2$ and $u = 1/2$, the robust likelihood ratio function simplifies to

$$ \hat{l} = c_1 l + c_2 l^{1/2} + c_3 $$

where

$$ c_1 = \frac{2\lambda_0 \lambda_1 + \mu_0 \lambda_1}{4\lambda_0^2 \lambda_1^2 + \lambda_0^2 \mu_0^2 + 4\lambda_0^2 \lambda_1 \mu_0}, $$

$$ c_2 = \frac{\lambda_0(2\lambda_0 \lambda_1 + \mu_0 \lambda_1)}{4\lambda_0^2 \lambda_1^2 + \lambda_0^2 \mu_0^2 + 4\lambda_0^2 \lambda_1 \mu_0}, $$

$$ c_3 = \frac{0.25\lambda_0^2}{4\lambda_0^2 \lambda_1^2 + \lambda_0^2 \mu_0^2 + 4\lambda_0^2 \lambda_1 \mu_0}. $$
The robust likelihood ratio function \( \hat{l} \) is a functional of only \( l \) and for all practical purposes
\[
\hat{l} = c_0 + \sum_{k \in I} c_k l^k
\]
is a possible generalization where \( I \) is the set of indexes and \( c_k \) are the parameters, which can be solved for training data.

C. Symmetrized \( \alpha \)-divergence Neighborhood

The \( \alpha \)-divergence is not a symmetric distance in general, where \( \alpha = 1/2 \) is an exception. A symmetrized version of the \( \alpha \)-divergence, i.e.
\[
D_\alpha(G_j, F_j) := \frac{1}{\alpha(1-\alpha)} \left( \int_Y ((f_j^\alpha - g_j^\alpha)(f_j^{1-\alpha} - g_j^{1-\alpha})) d\mu \right), \quad \alpha \in \mathbb{R} \setminus \{0, 1\}
\]
can be obtained by
\[
\frac{1}{2} (D_\alpha(G_0, G_1) + D_\alpha(G_0, G_1)),
\]
aborting the scaling factor 1/2. Symmetrized \( \alpha \)-divergence is also an \( f \)-divergence \[21, 26\] including various other symmetrized divergences such as symmetric \( \chi^2 \)-squared divergence \[28\]. The LFDs resulting from the symmetrized \( \alpha \)-divergence neighborhood is stated with the following theorem.

**Theorem V.4.** The LFDs of the symmetrized \( \alpha \)-divergence neighborhood can be written in terms of two coupled equations
\[
\frac{\lambda_0}{1-\alpha} \left( \frac{\hat{g}_0}{f_0} \right)^{2\alpha-1} + \left( 1-u \right) \left( \frac{\hat{g}_1}{\hat{g}_0} \right)^u - \frac{\lambda_0}{\alpha(1-\alpha)} - \mu_0 \right) \left( \frac{\hat{g}_0}{f_0} \right)^\alpha + \frac{\lambda_0}{\alpha} = 0, \quad (40)
\]
\[
\frac{\lambda_1}{1-\alpha} \left( \frac{\hat{g}_1}{f_1} \right)^{2\alpha-1} + \left( u \left( \frac{\hat{g}_1}{\hat{g}_0} \right)^{u-1} - \frac{\lambda_1}{\alpha(1-\alpha)} - \mu_1 \right) \left( \frac{\hat{g}_1}{f_1} \right)^\alpha + \frac{\lambda_1}{\alpha} = 0. \quad (41)
\]

**Proof:** The proof follows by using the same Lagrangian procedure as before. \[\Box\]

In general, (40) and (41) need to be solved jointly with four non-linear equations obtained from the Lagrangian constraints in order to determine the parameters. It is however possible to reduce the total number of equations to five if \( \alpha \) is given. The idea is to solve the equations such that \( \hat{g}_0/f_0 = h_0(\hat{g}_1/\hat{g}_0) \) and \( \hat{g}_1/f_1 = h_1(\hat{g}_1/\hat{g}_0) \), respectively, hence, \( \hat{g}_1/\hat{g}_0 = (h_1/h_0)l \) is the coupling equation, where \( h_0 \) and \( h_1 \) are some functions.
D. Total Variation Neighborhood

The total variation neighborhood is defined as

\[ \mathcal{G}_j = \{ G_j : D_{TV}(G_j, F_j) \leq \epsilon_j \}, \quad j \in \{0, 1\}, \]

where

\[ D_{TV}(G_j, F_j) = \frac{1}{2} \int_Y |g_j - f_j| \, d\mu. \]

The LFDs and the corresponding minimax robust test for the uncertainty classes created by the total variation neighborhood were found earlier by Huber [8]. However, the design approach is heuristic, many choices of the parameters and/or functions are unknown and the test is obtained under the assumption that \( \epsilon_0 = \epsilon_1 \). Since asymptotic minimax robustness implies all sample minimax robustness, if both tests exist and are unique, the minimax robust test resulting from the total variation neighborhood can also be analytically derived following the same design procedure introduced before. The following theorem substantiate this claim.

**Theorem V.5.** For the total variation neighborhood, the robust LRF is given by

\[
\hat{g}_1 = \begin{cases} 
  t_l, & l < (k_0 t_l)/k_1, \\
  \frac{k_1 l}{k_0}, & (k_0 t_l)/k_1 \leq l \leq (k_0 t_u)/k_1, \\
  t_u, & l > (k_0 t_u)/k_1
\end{cases}
\]

Moreover, the LFDs can be chosen as

\[
\hat{g}_0 = \begin{cases} 
  (f_0 + f_1)/c_1, & l < (k_0 t_l)/k_1, \\
  k_0 f_0, & (k_0 t_l)/k_1 \leq l \leq (k_0 t_u)/k_1 \\
  (f_0 + f_1)/c_2, & l > (k_0 t_u)/k_1
\end{cases}
\]

and

\[
\hat{g}_1 = \begin{cases} 
  t_l(f_0 + f_1)/c_1, & l < (k_0 t_l)/k_1, \\
  k_1 f_1, & (k_0 t_l)/k_1 \leq l \leq (k_0 t_u)/k_1 \\
  t_u(f_0 + f_1)/c_2, & l > (k_0 t_u)/k_1
\end{cases}
\]

where

\[ c_1 = \frac{1}{k_0} + \frac{t_l}{k_1} \quad \text{and} \quad c_2 = \frac{1}{k_0} + \frac{t_u}{k_1} \]
Proof: Consider the Lagrangians,
\[ \begin{align*}
L_0(g_0, g_1, \lambda_0, \mu_0) &= D_u(G_0, G_1) + \lambda_0(D_{TV}(G_0, F_0) - \epsilon_0) + \mu_0(\Upsilon(G_0) - 1), \\
L_1(g_0, g_1, \lambda_1, \mu_1) &= D_u(G_0, G_1) + \lambda_1(D_{TV}(G_1, F_1) - \epsilon_1) + \mu_1(\Upsilon(G_1) - 1). 
\end{align*} \] (43)

There are three cases of interest:

Case 1: \( g_j = k_j f_j \)

Here, no derivatives are necessary and we simply have \( g_0 = k_0 f_0 \) and \( g_1 = k_1 f_1 \) a.e.

Case 2: \( g_j < k_j f_j \)

Taking the Gâteaux’s derivatives of the Lagrangians, \( L_0 \) and \( L_1 \), respectively, leads to
\[ \begin{align*}
\int ((1 - u)(g_1/g_0)^u + \mu_0 - \lambda_0)\psi d\mu &= 0, \\
\int (u(g_1/g_0)^{u-1} + \mu_1 - \lambda_1)\psi d\mu &= 0. 
\end{align*} \] (44)

Case 3: \( g_j > k_j f_j \)

Similarly we get
\[ \begin{align*}
\int ((1 - u)(g_1/g_0)^u + \mu_0 + \lambda_0)\psi d\mu &= 0, \\
\int (u(g_1/g_0)^{u-1} + \mu_1 + \lambda_1)\psi d\mu &= 0. 
\end{align*} \] (45)

Since Case 2 and Case 3 cannot jointly coexist, from these three cases, at most three different disjoint sets can be defined:

\[ \begin{align*}
A_1 &= \{ y : g_0 = k_0 f_0, g_1 > k_1 f_1 \} \equiv \{ y : g_1 = k_1 f_1, g_0 < k_0 f_0 \} \equiv \{ y : g_1 > k_1 f_1, g_0 < k_0 f_0 \}, \\
A_2 &= \{ y : g_0 = k_0 f_0, g_1 = k_1 f_1 \} , \\
A_3 &= \{ y : g_0 = k_0 f_0, g_1 < k_1 f_1 \} \equiv \{ y : g_1 = k_1 f_1, g_0 > k_0 f_0 \} \equiv \{ y : g_1 < k_1 f_1, g_0 > k_0 f_0 \} .
\end{align*} \] (46)

Solving the equations from (44) and (45) together with Case 1, we get
\[ \frac{\dot{g}_1}{g_0} = \begin{cases} 
\frac{u(-\lambda_0 + \mu_0)}{(1-u)(\lambda_1 + \mu_1)}, & A_1 \\
\frac{k_1 f_1}{k_0 f_0}, & A_2 \\
\frac{u(-\lambda_0 - \mu_0)}{(1-u)(\lambda_1 - \mu_1)}, & A_3
\end{cases} \] (47)
where

\[ u = \frac{\ln \left( \frac{\lambda_0 + \mu_0}{-\lambda_0 + \mu_0} \right)}{\ln \left( \frac{\lambda_0 + \mu_0}{-\lambda_0 + \mu_0} \right) + \ln \left( \frac{-\lambda_1 + \mu_1}{\lambda_1 + \mu_1} \right)} . \]

The robust LRF given in Theorem V.5 is then immediate by using (46) in (47).

Clearly, the minimax robust test must be unique. However, the Lagrangian approach considered imposes no constraints on the choice of the LFDs as long as the LFDs yield the robust likelihood ratio function given by (42). As a result, the LFDs can be chosen as given in Theorem V.5. In order \( \hat{g}_0 \) and \( \hat{g}_1 \) to be continuous, the limits from the left and right should agree for two meeting points of the piece-wise defined functions. This implies:

\[
\begin{align*}
\frac{k_0 t_1}{k_1} & \mapsto f_1 = f_0 \frac{k_0 t_1}{k_1} \quad \text{and} \quad k_0 f_0 = \frac{f_0 + f_1}{c_1} \implies c_1 = \frac{1}{k_0} + \frac{t_1}{k_1}, \\
\frac{k_0 t_u}{k_1} & \mapsto f_1 = f_0 \frac{k_0 t_u}{k_1} \quad \text{and} \quad k_0 f_0 = \frac{f_0 + f_1}{c_2} \implies c_2 = \frac{1}{k_0} + \frac{t_u}{k_1}.
\end{align*}
\]

Hence, the proof is completed.

Remark V.2. In Theorem V.5, \( \hat{g}_0 \) and \( \hat{g}_1 \) are obtained in four parameters. The parameters can again be determined by imposing the four constraints defined by (43), cf. (32). These results generalize Huber’s results allowing the robustness parameters to be chosen without the restriction of \( \epsilon_0 = \epsilon_1 \) [8]. Moreover, it is clear why the robust test is unique, the densities are not necessarily and the parameters \( c_1 \) are \( c_2 \) are chosen as such. Additionally, the robust likelihood ratio test is independent of \( u \). This result is in line with Theorem III.1.

E. Band Model

So far, the nominal distributions have been assumed to be known or could roughly be determined before constructing the uncertainty classes. In fact, depending on the application, the nominal distributions may also be unknown; for example only a partial statistics of the data samples may be available [14], [17] or the shape of the actual distributions may lie within a given band [18]. These cases will be studied here and in the following two sections.

The band classes were first proposed by Kassam [18] and later revisited by Faüß et. al. [19]. They are given by the uncertainty classes

\[ G_j = \{ g_j \in \mathcal{M} : g_j^L \leq g_j \leq g_j^U \} \tag{48} \]
where \( \mathcal{M} \) is the set of all density functions on \( \mathcal{Y} \), and \( g^L_j \) and \( g^U_j \) are non-negative lower and upper bounding functions such that \( \mathcal{G}_0 \) and \( \mathcal{G}_1 \) are nonempty sets. This implies
\[
\int_{\mathcal{Y}} g^L_j d\mu \leq 1 \leq \int_{\mathcal{Y}} g^U_j d\mu, \quad j \in \{0, 1\}.
\]
Moreover, \( g^L_j \) and \( g^U_j \) should be chosen such that \( g_0 \) and \( g_1 \) are distinct density functions, if not \( \mathcal{G}_0 \cap \mathcal{G}_1 \neq \emptyset \) and minimax hypothesis testing is not possible. Theoretically, there are two main reasons to study band models. First, the band models are not equivalent to distance based uncertainty classes introduced so far. Because, distribution functions which are not absolutely continuous with respect to nominal distributions can belong to the band model, while this is not possible for the \( f \)-divergence based uncertainty classes. This result also includes the total variation distance since for any chosen total variation based uncertainty class, the band model should accept \( g^L_j = 0 \) and \( g^U_j = \infty \). Otherwise, there are density functions of type \( \alpha g_j + (1 - \alpha)\delta_x \), where \( \delta_x \) is a dirac delta function at \( y = x \), which belong to the total variation based uncertainty classes but not to the band model. On the other hand, choosing \( g^L_j = 0 \) and \( g^U_j = \infty \) defines the set of all density functions on \( \mathcal{Y} \), which is definitely not produced by the total variation distance unless \( \epsilon_j \) are infinite.

The second reason to consider the band models is that the band models are in general capacity classes, however, whether they are two alternating has been unclear [17]. Therefore, the theory introduced by Huber was not directly applied to band models [9]. In fact, Huber has never defined the LFDs explicitly in [9], i.e. the LFDs are the distributions which maximize a version of the \( f \)-divergence over all distributions belonging to the related uncertainty classes.

Practically, the main motivation behind considering band models is that for some applications the density functions estimated from the training data are expressed as lying within a confidence interval and for these applications the band classes are the natural uncertainty model [2], [13]. The asymptotically minimax robust test and least favorable distributions arising from the band model can similarly be obtained as before. Consider the Lagrangians:
\[
L_0(g_0, g_1, \lambda_0, \theta_0, \mu_0) = D_u(G_0, G_1) + \lambda_0(g_0 - g^L_0) + \nu_0(g^U_0 - g_0) + \mu_0(\Upsilon(g_0) - 1),
\]
\[
L_1(g_0, g_1, \lambda_1, \theta_1, \mu_1) = D_u(G_0, G_1) + \lambda_1(g_1 - g^L_1) + \nu_1(g^U_1 - g_1) + \mu_1(\Upsilon(g_1) - 1),
\]
where \( \lambda_j \) and \( \mu_j \) are scalar and \( \nu_j \) are functional Langrangian multipliers. Taking the Gateux derivatives of the Lagrangians, at the direction of unit area integrable functions \( \psi_0 \) and \( \psi_1 \),
respectively, leads to

\[
\frac{\partial L_0}{\partial g_0} = \int \left( (1 - u) \left( \frac{g_1}{g_0} \right)^u + \lambda_0 - \nu_0 + \mu_0 \right) \psi_0 \, d\mu = 0, \\
\frac{\partial L_1}{\partial g_1} = \int \left( u \left( \frac{g_1}{g_0} \right)^{u-1} + \lambda_1 - \nu_1 + \mu_1 \right) \psi_1 \, d\mu = 0,
\]

(49)

Three cases can separately be investigated.

**Case 1:** $g_0^U = \infty$ and $g_1^U = \infty$ (no upper bounding functions).

In this case we have $\nu_0 = 0$ and $\nu_1 = 0$ everywhere, and hence, no constraints regarding the upper bounding functions are in effect. There are four conditions regarding the Lagrangians

- $L_0$:
  - $g_0 = g_0^L$ on $A_0$ and $g_0 > g_0^L$ on $\mathcal{Y}\setminus A_0$,
  - $L_1$:
  - $g_1 = g_1^L$ on $A_1$ and $g_1 > g_1^L$ on $\mathcal{Y}\setminus A_1$.

The integrals in (49) are defined for $g_0 > g_0^L$ and $g_1 > g_1^L$, respectively. Since $\nu_0 = 0$ and $\nu_1 = 0$ everywhere, with the assumption that $\lambda_j$ are constant functions, it is the case that

\[
\frac{g_1}{g_0} = \frac{1}{k_2} \quad \text{on} \quad \bar{A}_0 = \mathcal{Y}\setminus A_0 = \{ y : g_0 > g_0^L \},
\]

\[
\frac{g_1}{g_0} = k_1 \quad \text{on} \quad \bar{A}_1 = \mathcal{Y}\setminus A_1 = \{ y : g_1 > g_1^L \},
\]

(50)

where $k_1$ and $k_2$ are some positive constants.

**Theorem V.6.** From (50), it follows that the LFDs and the corresponding likelihood ratio function are unique and given by

\[
\hat{g}_0 = \begin{cases} 
  g_0^L, & y \in A_0 \\
  k_2 g_1^L, & y \in \bar{A}_0 
\end{cases}, \quad \hat{g}_1 = \begin{cases} 
  g_1^L, & y \in A_1 \\
  k_1 g_0^L, & y \in \bar{A}_1 
\end{cases},
\]

(51)

and

\[
\frac{\hat{g}_1}{\hat{g}_0} = \begin{cases} 
  \frac{1}{k_2}, & y \in \bar{A}_0 \cap A_1 \\
  \frac{g_1^L}{g_0^L}, & y \in A_0 \cap A_1 \\
  k_1, & y \in A_0 \cap \bar{A}_1 
\end{cases}.
\]

**Proof:** The claim follows from the conditions:

1. The sets $A_0$, $A_1$, $\bar{A}_0$ and $\bar{A}_1$ are all non-empty.
2. The set $\bar{A}_0 \cap \bar{A}_1$ is empty.
3. On $\bar{A}_0$ and $\bar{A}_1$, respectively, we have $\hat{g}_0 = k_2g_1^L$ and $\hat{g}_1 = k_1g_0^L$.

1. The sets $\bar{A}_0$ and $\bar{A}_1$ are trivially non-empty. If not, we have $\int_Y \hat{g}_0 = \int_Y g_0^L d\mu < 1$ and $\int_Y \hat{g}_1 d\mu = \int_Y g_1^L d\mu < 1$, which are contradictions with the fact that $\hat{g}_0$ and $\hat{g}_1$ are density functions. The set $A_0$ is also non-empty and this can be shown again with contradiction. Assume that $A_0$ is empty. In this case, $A_1$ can either be empty or non-empty. Assume that $A_1$ is also empty. Then, by (50), we necessarily have $\hat{g}_0 = \hat{g}_1$ a.e., which is excluded by a suitable choice of $g_0^L$ and $g_1^L$. Therefore, $A_1$ is non-empty. If $A_1$ is non-empty, then we must have $\hat{g}_0 = k_2g_1^L$ on $\bar{A}_0$. If not, $\hat{g}_1/\hat{g}_0$ will not be a constant function on $\bar{A}_0 \cap A_1$, which is non-empty since $\bar{A}_0 = \mathcal{Y}$. This again yields a contradiction with (50). Since $\hat{g}_0 = k_2g_1^L$ is defined on $\mathcal{Y}$, in order to satisfy (50), $\hat{g}_1$ must also be $g_1^L$ on $\bar{A}_1$. Hence, we have $\hat{g}_1 = g_1^L$ a.e. which is again a contradiction with the fact that $\int_Y \hat{g}_1 = 1$. Therefore, $A_0$ is non-empty. A similar analysis shows that $A_1$ is also non-empty.

2. The set $\bar{A}_0 \cap \bar{A}_1$ is empty. If not, from (51) and (50) we have
\[
\frac{\hat{g}_1}{\hat{g}_0} = \frac{k_1g_0^L}{k_2g_1^L} = k_1 = \frac{1}{k_2}.
\]
This implies $(\bar{A}_0 \cap A_1) \cup (A_0 \cap \bar{A}_1) = \mathcal{Y}$, hence, both $\bar{A}_0 \cap \bar{A}_1$ and $A_0 \cap A_1$ are empty sets. Since, $A_0 \cap A_1$ is non-empty, we have a contradiction, hence, $\bar{A}_0 \cap \bar{A}_1$ must be empty.

3. The set $A_0 \cap A_1$ is non-empty. If not, $A_0$ and $A_1$ are disjoint sets. This implies at least non-empty $\bar{A}_0 \cap A_1$ and $A_0 \cap \bar{A}_1$ and at most additionally non-empty $\bar{A}_0 \cap \bar{A}_1$. Non-empty $\bar{A}_0 \cap \bar{A}_1$ implies $\hat{g}_1/\hat{g}_0 = k$ a.e on $\mathcal{Y}$, see (52), and this is impossible, unless $k = 1$. If only $\bar{A}_0 \cap A_1$ and $A_0 \cap \bar{A}_1$ are non-empty, i.e. if $\bar{A}_0 \cap \bar{A}_1$ and $A_0 \cap A_1$ are empty, hence, $(\bar{A}_0 \cap A_1) \cup (A_0 \cap \bar{A}_1) = \mathcal{Y}$, we have $A_0 = \bar{A}_1$ and $A_1 = \bar{A}_0$ together with $A_0 \cup A_1 = \mathcal{Y}$. This is possible if and only if $k = k_1 = 1/k_2$, because
\[
\bar{A}_0 \cap A_1 = \{g_0 > g_0^L, g_1 = g_1^L\} = \{1/k_2 = g_1/g_0 < g_1^L/g_0^L\},
\]
\[
A_0 \cap \bar{A}_1 = \{g_1 > g_1^L, g_0 = g_0^L\} = \{k_1 = g_1/g_0 > g_1^L/g_0^L\}.
\]
The condition $k = k_1 = 1/k_2$ also implies $\hat{g}_1/\hat{g}_0 = 1$ a.e on $\mathcal{Y}$, which is avoided by suitable choices of $g_0^L$ and $g_1^L$. Hence, $A_0 \cap A_1$ cannot be empty.

The sets $\bar{A}_0 \cap A_1$ and $A_0 \cap \bar{A}_1$ are both non-empty. From $A_0 \cap A_1 \neq \emptyset$, there are four cases $A_0 \subset A_1$, $A_1 \subset A_0$, $A_0 = A_1$, or $A_0 \setminus A_1$ and $A_1 \setminus A_0$ are both non-empty. The first three conditions imply either non-empty $\bar{A}_0 \cap \bar{A}_1$, or $A_0 = \mathcal{Y}$, $A_1 = \emptyset$ or both. The first condition is
a contradiction with (52) and the other three imply \( \hat{g}_j = g_j^L \) on \( \mathcal{Y} \), which is impossible, see (51). Therefore, we have non-empty \( A_0 \cap A_1 \) together with non-empty \( A_0 \setminus A_1 \) and \( A_1 \setminus A_0 \). This eventually implies non-empty \( \bar{A}_0 \cap A_1 \) and \( A_0 \cap \bar{A}_1 \).

It is known that \( \hat{g}_1 = g_1^L \) on \( A_1 \) and on \( \bar{A}_0 \cap A_1 \) we have \( \hat{g}_1/\hat{g}_0 = 1/k_2 \). Hence, on \( \bar{A}_0 \) we must have \( \hat{g}_0 = k_2 g_1^L \). Similarly, on \( \bar{A}_1 \) we have \( \hat{g}_1 = k_1 g_0^L \).

**Corollary V.7.** The parameters should satisfy \( k_1 < 1/k_2 \), hence,

\[
A_0 \cap A_1 = \{k_1 \leq g_1^L/g_0^L \leq 1/k_2\}.
\]

Moreover,

\[
A_0 = \{g_1^L/g_0^L < 1/k_2\}, \quad A_1 = \{g_1^L/g_0^L > k_1\}.
\]

**Proof:** \( k_1 = 1/k_2 \) implies empty \( A_0 \cap A_1 \), which is impossible, and \( k_1 > 1/k_2 \) implies non-empty \( (\bar{A}_0 \cap A_1) \cap (A_0 \cap \bar{A}_1) \), which in turn implies \( k_1 = 1/k_2 \), another contradiction. Therefore, we have \( k_1 < 1/k_2 \). Accordingly, the sets \( A_0 \) and \( A_1 \) can be written as

\[
A_0 = (A_0 \cap A_1) \cup (A_0 \cap \bar{A}_1) = \{k_1 \leq g_1^L/g_0^L \leq 1/k_2\} \cup \{k_1 > g_1^L/g_0^L\} = \{g_1^L/g_0^L \leq 1/k_2\},
\]

\[
A_1 = (A_0 \cap A_1) \cup (\bar{A}_0 \cap A_1) = \{k_1 \leq g_1^L/g_0^L \leq 1/k_2\} \cup \{1/k_2 < g_1^L/g_0^L\} = \{g_1^L/g_0^L \geq k_1\}.
\]

**Remark V.3.** Let \( t_u = 1/k_2 \), \( t_l = k_1 \) and \( l = g_1^L/g_0^L \). Then, the LFDs and the robust LRF can be rewritten as

\[
\hat{g}_0 = \begin{cases} 
    g_0^L, & l \leq t_u \\
    1/t_u g_1^L, & l > t_u
\end{cases}, \quad \hat{g}_1 = \begin{cases} 
    g_1^L, & l \geq t_l \\
    t_l g_0^L, & l < t_l
\end{cases},
\]

(53)

and

\[
\frac{\hat{g}_1}{\hat{g}_0} = \begin{cases} 
    t_u, & l > t_u \\
    t_l, & t_l \leq l \leq t_u \\
    t_l, & l < t_l
\end{cases}.
\]

(54)

The lower bounding function constraints are satisfied automatically. Because, on \( \{l \leq t_u\} \) and \( \{l \geq t_l\} \), \( \hat{g}_j \geq g_j^L \) holds with equality, and on \( \{l > t_u\} \) and \( \{l < t_l\} \), we necessarily have
\[ \hat{g}_0 = \frac{1}{t_u} g_1^L \geq g_0^L \text{ and } \hat{g}_1 = t_l g_0^L \geq g_1^L, \text{ respectively, as } l = \frac{g_1^L}{g_0^L}. \] The density function constraints are satisfied by solving

\[
\int_{t < t_u} g_0^L \, d\mu + \frac{1}{t_u} \int_{t > t_u} g_1^L \, d\mu = 1, \\
\int_{t > t_l} g_1^L \, d\mu + t_l \int_{t < t_l} g_0^L \, d\mu = 1. \tag{55}
\]

**Remark V.4.** Letting \( g_j^L = (1 - \epsilon_j) f_j \), the band model reduces to the \( \epsilon \)-contamination model

\[ G_{\epsilon}^{-j} = \{ g_j : g_j = (1 - \epsilon_j) f_j + \epsilon_j h, h \in \mathcal{M} \} \]

with the nominal distributions \( f_j \), where \( 0 \leq \epsilon_j < 1 \) \[18\]. By Huber, the equations in (55) have unique solutions and the LFDs in (53) are single sample minimax robust \[8\]. From Theorem \[III.1\], single sample minimax robust LFDs minimize all \( f \)-divergences, hence they also maximize all \( u \)-divergences. This proves that choosing \( \lambda_j \) as scalars, which has been made to simplify the derivations, is a correct assumption. By (55), it is also implied that the parameters \( t_l \) and \( t_u \) are only dependent on \( g_0^L \) and \( g_1^L \), i.e. they are independent of the choice of \( u \). This is in accordance with Theorem \[III.1\].

**Case 2:** \( g_0^L = 0 \) and \( g_1^L = 0 \) (no lower bounding functions).

Assume that \( \lambda_0 = 0 \) and \( \lambda_1 = 0 \) everywhere, and hence, no constraints regarding the upper bounding functions are in effect. Similarly, the positivity constraints are also not imposed as before because, as it can be seen later, the density functions automatically satisfy these constraints. In this case, there are four conditions regarding the Lagrangians:

\[ L_0 : \quad g_0 = g_0^U \text{ on } A_0 \quad \text{and} \quad g_0 < g_0^U \text{ on } \mathcal{Y} \setminus A_0, \]

\[ L_1 : \quad g_1 = g_1^U \text{ on } A_1 \quad \text{and} \quad g_1 < g_1^U \text{ on } \mathcal{Y} \setminus A_1. \tag{56} \]

The integrals in (49) are defined for \( g_0 < g_0^U \) and \( g_1 < g_1^U \), respectively. Since \( \lambda_0 = 0 \) and \( \lambda_1 = 0 \) everywhere, and with the assumption that \( \nu_j \) are constant functions, it is the case that

\[ \frac{g_1}{g_0} = \frac{1}{k_2} \text{ on } \bar{A}_0 = \mathcal{Y} \setminus A_0 = \{ y : g_0 < g_0^U \}, \]

\[ \frac{g_1}{g_0} = k_1 \text{ on } \bar{A}_1 = \mathcal{Y} \setminus A_1 = \{ y : g_1 < g_1^U \}, \tag{57} \]

where \( k_1 \) and \( k_2 \) are some positive constants.
Theorem V.8. Let $t_1 = 1/k_2$, $t_u = k_1$ and $l = g_1^U / g_0^U$. It follows that the LFDs and the corresponding LRF are unique and given by

$$\hat{g}_0 = \begin{cases} 
    g_0^U, & l \geq t_1 \\
    1/t_1 g_1^U, & l < t_1 
\end{cases}, \quad \hat{g}_1 = \begin{cases} 
    g_1^U, & l \leq t_u \\
    t_u g_0^U, & l > t_u 
\end{cases},$$

and

$$\frac{\hat{g}_1}{\hat{g}_0} = \begin{cases} 
    t_1, & l < t_1 \\
    t_u, & l > t_u \\
    l, & t_1 \leq l \leq t_u 
\end{cases}.$$

Moreover, all the Lagrangian constraints are satisfied and in particular the LFDs are obtained by solving

$$\int_{l \geq t_1} g_0^U \, d\mu + \frac{1}{t_1} \int_{l < t_1} g_1^U \, d\mu = 1,$$

$$\int_{l \leq t_u} g_1^U \, d\mu + t_U \int_{l > t_u} g_0^U \, d\mu = 1.$$

Proof: The definition of the sets $A_j$, their intersections, their relation to $l$, $k_1$ and $k_2$, and the fact that $k_1 > 1/k_2$ trivially follow from the same line of arguments used in Theorem V.6 and Corollary V.7 by considering (56) and (57). The lower bounding function constraints are automatically satisfied as $\hat{g}_0$ and $\hat{g}_1$ are non-negative functions. The upper bounding function constraints are also satisfied in the same way as explained in Case 1. The LFDs are obtained by unit density function constraints.

Theorem V.9. The LFDs in Theorem V.8 are single sample minimax robust, i.e.

$$G_0 \left[ l < t \right] \geq \hat{G}_0 \left[ l < t \right],$$

$$G_1 \left[ l < t \right] \leq \hat{G}_1 \left[ l < t \right]$$

for all $t \in \mathbb{R}_{\geq 0}$ and $(G_0, G_1) \in \mathcal{G}_0 \times \mathcal{G}_1$, and/hence, $\nu_j$ can be chosen as constant functions.

Proof: For $g_0^U = 0$ and $g_1^U = 0$ the band model (48) can equivalently be written as

$$\mathcal{G}_j^+ = \left\{ g_j : g_j = (1 + \varepsilon_j) f_j - \varepsilon_j h, h \in \mathcal{M} \right\}$$
where $\varepsilon_j > 0$, and $f_j$ are the nominal density functions. For any $g_j \in \mathcal{G}_j$, if $t > t_u$, the event $A = [\hat{t} < t]$ has a full probability and if $t \leq t_1$, it has a null probability. Therefore, (60) holds trivially for these cases. For $t_1 < t \leq t_u$, we have

$$G_1(A) = (1 + \varepsilon_1)F_1(A) - \varepsilon_1 h \leq (1 + \varepsilon_1)F_1(A) = \hat{G}_1(A)$$
$$G_0(A) = (1 + \varepsilon_0)F_0(A) - \varepsilon_0 h \geq (1 + \varepsilon_0)F_0(A) - \varepsilon_0 = 1 - (1 + \varepsilon_0)(1 - F_0(A))$$
$$= 1 - (1 + \varepsilon_0)F_0(\bar{A}) = 1 - G_0^U(\bar{A}) = \hat{G}_0(A).$$

Hence, $\hat{g}_0$ and $\hat{g}_1$ are single sample minimax robust. Moreover, by the virtue of Theorem III.1, single sample minimax robust LFDs minimize all $f$-divergences, accordingly they also maximize all $u$-divergences. This proves that choosing $\nu_j$ as constant functions, which was made to simplify the derivations, was a correct assumption.

\begin{itemize}
  \item \textbf{Case 3:} $g_j^L < g_j < g_j^U$ (the general case).
\end{itemize}

The uncertainty classes for the general case are obtained by the intersection of the two contamination neighborhoods

$$\mathcal{G}_j = \mathcal{G}_j^c \cap \mathcal{G}_j^c.$$

There are six conditions regarding the Lagrangians

$$L_0 : \quad g_0 = g_0^L \quad \text{on} \quad A_0, \quad g_0 = g_0^U \quad \text{on} \quad A_1 \quad \text{and} \quad g_0^L < g_0 < g_0^U \quad \text{on} \quad A_2,$$

$$L_1 : \quad g_1 = g_1^L \quad \text{on} \quad A_3, \quad g_1 = g_1^U \quad \text{on} \quad A_4 \quad \text{and} \quad g_1^L < g_1 < g_1^U \quad \text{on} \quad A_5. \quad (61)$$

The integrals in (49) are defined for $g_0^L < g_0 < g_0^U$ and $g_1^L < g_1 < g_1^U$, respectively. With the assumption that both $\lambda_j$ and $\nu_j$ are constant functions, it is the case that

$$\frac{g_1}{g_0} = k_2 \quad \text{on} \quad A_2 = \{ y : g_0^L < g_0 < g_0^U \},$$

$$\frac{g_1}{g_0} = k_1 \quad \text{on} \quad A_5 = \{ y : g_1^L < g_1 < g_1^U \}. \quad (62)$$

where $k_1$ and $k_2$ are some positive constants.

\textbf{Theorem V.10.} Assume that both $\lambda_j$ and $\nu_j$ are constant functions. Then, there are at least three
different asymptotically minimax robust LRFs,

Type I: \[ \begin{align*}
\hat{g}_1 \mapsto g_0 & \begin{cases}
g^U_1 \geq g^L_0, & g^U_1/g^L_0 \leq k_2 \\
_k \geq g^U_1/g^L_0 > k_2 > g^U_1/g^L_0 & \\
g^U_1 \leq g^L_0, & k_1 < g^U_1/g^L_0 \leq k_1 \\
g^U_1 > g^L_0, & g^U_1/g^L_0 > g^U_1/g^L_0 \geq k_1 \\
\end{cases}
\end{align*} \]

Type II: \[ \begin{align*}
\hat{g}_1 \mapsto g_0 & \begin{cases}
g^U_1 \geq g^L_0, & g^U_1/g^L_0 \leq k_1 \\
_k \geq g^U_1/g^L_0 > k_1 > g^U_1/g^L_0 & \\
g^U_1 \leq g^L_0, & k_1 < g^U_1/g^L_0 \leq k_1 \\
g^U_1 > g^L_0, & g^U_1/g^L_0 > g^U_1/g^L_0 \geq k_1 \\
\end{cases}
\end{align*} \]

Type III: \[ \begin{align*}
\hat{g}_1 \mapsto g_0 & \begin{cases}
g^U_1 \geq g^L_0, & g^U_1/g^L_0 \leq k_1 \\
_k \geq g^U_1/g^L_0 > k_1 > g^U_1/g^L_0 & \\
g^U_1 \leq g^L_0, & k_1 < g^U_1/g^L_0 \leq k_1 \\
g^U_1 > g^L_0, & g^U_1/g^L_0 > g^U_1/g^L_0 \geq k_1 \\
\end{cases}
\end{align*} \]

with the corresponding pairs of LFDs, respectively,

\[ \hat{g}_0 = \begin{cases}
g^L_0, & g^U_1/g^L_0 \leq k_2 \\
_k g^U_1, & g^U_1/g^L_0 > k_2 \geq g^U_1/g^L_0 & \\
g^L_0, & g^U_1/g^L_0 \geq k_2 \\
\end{cases} \]

\[ \hat{g}_1 = \begin{cases}
g^L_1, & g^L_1/g^L_0 \geq k_1 \\
_k g^U_1, & g^U_1/g^L_0 > k_1 > g^U_1/g^L_0 & \\
g^L_1, & g^U_1/g^L_0 \leq k_1 \\
\end{cases} \]
Moreover, Type I and Type III LRFs tend to clipped likelihood ratio functions given by (54) and (59) with the corresponding LFDs defined in (53) and (58).

Proof: From (61) and (62), LFDs can be written as

\[
\hat{g}_0 = \begin{cases} 
  g^L_0, & g^L_1 / g^L_0 \leq k_2 \\
  \frac{1}{k^2} g^L_1, & g^L_1 / g^L_0 > k_2 \geq g^L_1 / g^L_0
\end{cases}, \quad \hat{g}_1 = \begin{cases} 
  g^U_1, & g^U_1 / g^L_0 \geq k_1 \\
  k_1 g^L_0, & g^U_1 / g^L_0 > k_1 > g^L_1 / g^L_0
\end{cases}
\]

Let \( \hat{g}_0 = \frac{1}{k^2} g^L_1 \) on \( A_2 \) and \( \hat{g}_1 = k_1 g^L_1 \) on \( A_5 \). Then,

\[
A_1 \cap A_5, \quad A_2 \cap A_4, \quad \text{and} \quad A_2 \cap A_5
\]

are all empty sets, because their existence contradicts with (62). Accordingly, the robust LRF can implicitly be written as

\[
\frac{\hat{g}_1}{\hat{g}_0} = \begin{cases} 
  \frac{g^U_1 / g^L_0}{k^2}, & A_0 \cap A_4 \\
  k_1, & A_0 \cap A_5 \\
  g^L_1 / g^L_0, & A_0 \cap A_3 \\
  \frac{g^U_1 / g^L_0}{k^2}, & A_1 \cap A_4 \\
  k_2, & A_2 \cap A_3 \\
  g^L_1 / g^L_0, & A_1 \cap A_3
\end{cases}
\]

Furthermore, from (61) and (62) we have

\[
A_0 \cap A_5 = \{g^L_1 < g_1 < g^U_0, \ g_0 = g^L_0\} = \{g^L_1 / g^L_0 < k_1 = g_1 / g_0 < g^U_1 / g^L_0\},
\]

\[
A_2 \cap A_3 = \{g^L_0 < g_0 < g^U_1, \ g_1 = g^L_1\} = \{g^L_1 / g^L_0 < k_2 = g_1 / g_0 < g^U_1 / g^L_0\}.
\]

The empty sets in (64) imply \( A_2 \subset A_3 \) and \( A_5 \subset A_0 \), which in turn imply \( A_5 = A_0 \cap A_5 \) and \( A_2 = A_2 \cap A_3 \). Accordingly, \( A_2 \) and \( A_5 \) can also be made explicit in (63). The sets \( A_0, A_1 \) and \( A_2 \) are disjoint, as well as the sets \( A_3, A_4 \) and \( A_5 \). On \( A_2 \) we have \( g^L_1 / k_2 < g^U_0 \) and due to
continuity $\frac{1}{k_2}g_1^L = g_0^U$ at least on a single point. It is also at most on a single point, if not $A_1$ and $A_2$ are not disjoint. For $A_1$, the only choice left is then $A_1 = \{g_1^L/k_2 \geq g_0^U\}$. Similarly, i.e. considering $g_0^L < g_1^L/k_2$ on $A_2$ etc., we have $A_0 = \{g_1^L \leq g_1^L/k_2\}$. Performing the same analysis over $A_2 \cap A_3$, leads to the explicit definition of the sets $A_3$, $A_4$ and $A_5$. This implies that $A_1 \cap A_4$ is an empty set. Hence, $\hat{g}_0$, $\hat{g}_1$ and $\hat{g}_1/\hat{g}_0$ follow as defined in Theorem V.10, Type III. Following the same line of arguments for the cases $\hat{g}_0 = \frac{1}{k_1}g_1^U$ on $A_2$ and $\hat{g}_1 = k_2g_1^U$ on $A_5$ we have

$$A_1 \cap A_5 = \{g_1^L < g_1 < g_1^U, g_0 = g_0^U\} = \{g_1^L/g_0^U < k_1 = g_1/g_0 < g_1^U/g_0^U\},$$

$$A_2 \cap A_4 = \{g_0^L < g_0 < g_0^U, g_1 = g_1^U\} = \{g_1^U/g_0^U < k_2 = g_1/g_0 < g_1^U/g_0^U\},$$

in the places of $A_0 \cap A_5$ and $A_2 \cap A_3$, respectively, empty $A_0 \cap A_3$, and the explicit definition of the sets $A_j$, which leads to the robust LRF Type I and the corresponding LFDs. The robust LRF of Type II is a special case arising from merging the middle three regions of the robust LRFs of Type I or III. Clipped likelihood ratio functions are obtained again from the robust LRF of Type I and II for $k_1$ small enough and $k_2$ large enough, and $k_1$ large enough and $k_2$ small enough, respectively. This implies empty $A_0 \cap A_4$ and $A_1 \cap A_3$.

**Remark V.5.** Three different types of LFDs given in Theorem V.10 were first proposed in [18] without any details about how they were obtained. Here, the robust LRFs and the corresponding LFDs have been derived analytically with the assumption that optimum Lagrangian parameters $\lambda_j$ and $\nu_j$ are constant functions. The correctness of these assumptions is due to Theorem 1 and 2 in [18], which show that these pairs of LFDs are minimax robust and minimize all $f$-divergences.

There are two different cases of consideration. If the type of LRFs are/can be known, it may be preferable to obtain the LFDs by solving the unit area density function equations in $k_1$ and $k_2$. However, if this knowledge is unavailable, (21) may need to be solved by a convex optimization method for the uncertainty classes given by (48), which introduce linear constraints. To do this, the densities may first be sampled, and hence discretized. For any integration, a numerical integration method can be adopted, for instance the trapezoidal integration.

**F. Moment Classes**

The motivation behind modeling the uncertainties through moment classes is that the most common approach to partial statistical modeling is through moments, typically mean and cor-
relation. Among others, moment classes have been considered in applications to finance [34], admission control [35] and queueing theory [36]. The moment classes, which was originally introduced in [14], can be generalized as

$$\mathcal{G}_j = \left\{ g_j \in \mathcal{M} : c_{j,0}^k \leq \int_{\mathcal{Y}} h_j^k(Y) g_j d\mu \leq c_{j,1}^k \right\}$$

(66)

where $h_j^k$ are real valued continuous functions and $c_{j,0}^k$ and $c_{j,1}^k$, $k \in \{1, \ldots, K\}$, are some constants. The constants and the functions should be chosen such that $\mathcal{G}_0 \cap \mathcal{G}_1 = \emptyset$. The original version of the moment classes have been studied for asymptotically minimax robust NP-tests in [14]. Here, its extended version is studied for rate minimizing asymptotically minimax robust tests, by replacing the constraints in Section [V-E] by the ones above (66), see Section [VI].

G. P-point Classes

The partial information available for the robust hypothesis testing may also be in the form of masses which are assigned to every non-overlapping subsets of $\mathcal{Y}$. Such classes are called p-point classes and have been used in robust detection [5], [37], rate-distortion [38] and robust smoothing problems [39]. The original definition of p-point classes can be extended covering a more general case as follows:

$$\mathcal{G}_j = \left\{ g_j \in \mathcal{M} : c_{j,0}^k \leq G_j(A_j^k) \leq c_{j,1}^k \right\}$$

(67)

where $A_j^k \in \mathcal{A}$ are some disjoint subsets of $\mathcal{Y}$. The robust designs considering the p-point classes [5], [37], are application dependent. Using the same theory and techniques, i.e. by replacing the constraints in problem definition given in Section [IV-E] with the ones defined by (67), more general designs can be made. Moreover, a hybrid model which combines p-point classes with moment classes can also be of interest for engineering applications.

VI. GENERALIZATIONS

There are possible generalizations of the theory introduced to multiple-, decentralized-, and sequential hypothesis testing.
A. Multiple Hypothesis Testing

In many practical applications, there are more than two hypothesis being tested \cite{40}, \cite{41}. For testing $K$ hypothesis, an optimum decision rule requires $\binom{K}{2}$ binary hypothesis tests, such that the overall error probability, e.g.

$$P_E = \sum_{i \neq j} P(H_i | H_j) P(H_j)$$

is minimized \cite{1}. This implies that asymptotically minimax robust multiple hypothesis testing can be obtained by constructing $\binom{K}{2}$ asymptotically minimax robust binary hypothesis tests using the same theory introduced so far.

B. Decentralized Hypothesis Testing

Asymptotically minimax robust tests satisfy the saddle value conditions asymptotically. For parallel sensor networks, it is known that if every sensor satisfies the saddle value conditions, the whole sensor network is minimax robust under some mild conditions \cite{16}. For serial networks, in order to guarantee error free asymptotic detection, the likelihood ratio function should not be bounded \cite{42}, \cite{43}. This is true for most of the designs considered here, however, the mismatch cases need to be studied separately. For more general networks it is possible that the likelihood ratio tests are not optimum \cite{44}, p. 331. For such networks, throughout designs maybe required.

C. Sequential Probability Ratio Test

Sequential probability ratio test (SPRT) was proposed by Wald as an alternative to classical hypothesis testing \cite{45}. Decisions are made on the fly if and only if a certain level of confidence is reached, and a new data sample is collected, otherwise. Therefore, not only the false alarm and miss detection probabilities but also the expected number of samples under each hypothesis need to be considered for the minimax robustness. Furthermore, the computations must be exact and Wald’s approximations cannot be considered \cite{12}.

Let $S_n = \sum_{k=1}^{n} \ln \hat{l}(Y_k)$, where $Y_k \sim G_j$ under $H_j$, and $0 < t_l < 1$ and $1 < t_u < \infty$ be the lower and upper thresholds of the SPRT, respectively. Furthermore, let the stopping time of the stochastic process be denoted by

$$\tau = \min\{n \geq 1 : S_n \geq \ln t_u \text{ or } S_n \leq \ln t_l\}.$$
Then, according to Huber, a sequential statistical test is minimax robust if

\[ P_F(t_l, t_u, \hat{G}_0) \geq P_F(t_l, t_u, G_0), \]
\[ P_M(t_l, t_u, \hat{G}_1) \geq P_M(t_l, t_u, G_1), \] \hspace{1cm} (68)

and

\[ E_{\hat{G}_0}[\tau(t_l, t_u)] \geq E_{G_0}[\tau(t_l, t_u)], \]
\[ E_{\hat{G}_1}[\tau(t_l, t_u)] \geq E_{G_1}[\tau(t_l, t_u)], \] \hspace{1cm} (69)

for all \((g_0, g_1) \in \mathcal{G}_0 \times \mathcal{G}_1\) and for all \((t_l, t_u)\). However, it is known that no known minimax robust hypothesis test satisfies both (67) and (68) [7]. Moreover, it is not clear why both of these conditions must hold, because the related objective functions are not compatible, i.e. a test satisfying one does not imply satisfying the other. In engineering applications, the requirement is to have the quickest detection with lowest probability of error on average. Based on this statement, a new, and a more practically oriented definition of minimax robust sequential hypothesis testing can be made.

**Definition VI.1.** Let \( t_u = t_l^c, \ c > 0, \) and

\[ h_{G_0} : E_{G_0}[\tau(t_l, t_u)] \mapsto P_F(t_l, t_u, G_0), \]
\[ h_{G_1} : E_{G_1}[\tau(t_l, t_u)] \mapsto P_M(t_l, t_u, G_1). \]

Then, a sequential probability ratio test is minimax robust if

\[ h_{\hat{G}_0}(n(c)) \geq h_{G_0}(n(c)), \ \forall c, n, \forall g_0 \in \mathcal{G}_0 \]
\[ h_{\hat{G}_1}(n(c)) \geq h_{G_1}(n(c)), \ \forall c, n, \forall g_1 \in \mathcal{G}_1 \] \hspace{1cm} (70)

where \( n \) is a common notation for \( E_{G_0} \) and \( E_{G_1} \). The test is called asymptotically minimax robust if for a specific \( c \), and for \( n \to \infty \), (70) holds.

Existence of the tests satisfying the inequalities in (70) is an open problem. Some examples of the asymptotically minimax robust tests extended to sequential tests are shown in Section VII.
TABLE II

ROBUST TESTS USED IN SIMULATIONS FOR COMPARISON

| Acronym | Description |
|---------|-------------|
| (a)-test | Asymptotically minimax robust test |
| (a')-test | Dabak’s asymptotically robust test [13] |
| (h)-test | Huber’s clipped likelihood ratio test [8] |
| (m)-test | Minimax robust test for modeling errors [12] |
| (n)-test | Nominal test |

TABLE III

PAIR OF NOMINAL DISTRIBUTIONS USED IN THE EXAMPLES

| Acronym | Under $\mathcal{H}_0$ | Under $\mathcal{H}_1$ |
|---------|----------------------|----------------------|
| $d_1$   | $\mathcal{N}(-1,1)$   | $\mathcal{N}(1,1)$   |
| $d_2$   | $\mathcal{N}(-1,1)$   | $\mathcal{N}(1,4)$   |
| $d_3$   | $\mathcal{L}(0,1)$    | $f_\mathcal{L}(y)(\sin(2\pi y) + 1)$ |

VII. SIMULATIONS

In this section, the theoretical findings are evaluated, exemplified, and applied to spectrum sensing, as an example of a signal processing application. Observations are assumed to be real valued for all uncertainty models. In general, the presented theory allows single observations to be vector valued. However, this extension is straightforward, hence, is not simulated for the sake of simplicity. For solving all systems of equations damped Newton’s method [46] and for all convex optimization problems interior point methods [47] are used. To make the simulations transparent and easily repeatable the parameter values are explicitly stated. Five different tests are considered as tabulated in Table II. The notation $|b|_{a}$ stands for testing with the (a)-test while the data samples are obtained from the LFDs of the (b)-test.

A. Theoretical Examples

In all theoretical examples, the nominal distributions listed in Table III are considered. The notation $\mathcal{N}(\mu, \sigma^2)$ stands for the Gaussian distribution with mean $\mu$ and variance $\sigma^2$ whereas $\mathcal{L}(0,1)$ denotes the standard Laplace distribution with the respective parameters. The density...
functions are similarly denoted by $f_N$ and $f_L$, respectively. In the following, the least favorable distributions, robust likelihood ratio functions, parameters of the equations, and (non)-convexity of $D_u$ are illustrated.

1) LFDs and Robust LRFs: Comparative simulations are required in order to get the intuition about how robustness is achieved depending on the choice of the distance. Consider the pair of distributions denoted by $d_1$ in Table III and let the robustness parameters be $\epsilon_0 = \epsilon_1 = 0.1$. 

Fig. 1. The nominal distributions denoted by $d_1$ and the corresponding LFDs for $\epsilon_0 = \epsilon_1 = 0.1$, where $u = 0.5$ for all $\alpha$.

Fig. 2. Robust and nominal LRFs found for the nominal distributions denoted by $d_1$ and $\epsilon_0 = \epsilon_1 = 0.1$, including the symmetric case $\alpha_s$. 

December 12, 2017
For this setup, Figure 1 illustrates the LFDs together with the nominal distributions for the KL-divergence ($\alpha \to 1$) as well as for various $\alpha$-divergences. Symmetrized $\alpha$-divergence is not included for the sake of clarity. There are two observations from this example:

- The LFDs are non-Gaussian (not visible but verified by means of curve fitting).
- The variance of the LFDs are decreasing as $\alpha$ increases.

In Figure 2 the corresponding likelihood ratio functions are depicted, including the symmetrized $\alpha = 2$-divergence, denoted by $\alpha_s = 2$. For $\alpha = 0.1$ and $\alpha = 0.5$ there is a strong amplification of the likelihood ratios for larger observations and clipping for smaller observations (not well visible) in order to achieve asymptotic robustness. Moreover, there is no recognizable difference between the LRFs of $\alpha = 2$ and $\alpha_s = 2$.

The pair of nominal distributions $d_1$ are symmetric about the origin and asymmetric nominals are known to complicate the solution of the non-linear equations [12]. Additionally, the LRFs can be visualized in a reduced range, focusing more on the clipping range rather than the range of amplification, to be complimentary to the previous example. In this regard, the pair of nominal distributions denoted by $d_2$ in Table III are considered. Figure 3 illustrates the LFDs together with the nominals whereas Figure 4 shows the corresponding robust LRFs for $\epsilon_0 = \epsilon_1 = 0.1$. For larger observations in absolute value, there is huge amplification (not well visible), whereas for the smaller observations there is no hard clipping as in the previous example. The difference between $\alpha = 2$ and the symmetrized $\alpha_s = 2$ divergences is now visible.

The nominal LRFs are either increasing, or first decreasing and then increasing, respectively, for the pair of distributions $d_1$ and $d_2$. It is possible to construct an example for which the nominal LRF is repeatedly increasing and decreasing. This case both confirms the solvability of the related non-linear equations and serves as an example for the convexity analysis in the next section. Let the nominal distributions be denoted by $d_3$ as given in Table III. Furthermore, let $\epsilon_0 = \epsilon_1 = 0.05$, as the nominal distributions are now closer to each other. For this setup, Figure 5 and Figure 6 illustrate the LFDs together with the nominals and the robust LRFs, respectively, for the KL-divergence neighborhood. Similar to the previous examples, the nominal LRFs which are smaller than 1 are amplified and those larger than 1 are attenuated.

2) Convexity of $D_u$ and the Lagrangian Parameters: It was mentioned earlier that $D_u$ is convex in $u$, for a fixed pair of distributions. However, it is not necessarily convex if for every $u$ the distribution functions are possibly different. This is especially the case when one considers
Fig. 3. The nominal distributions denoted by \( d_2 \) and the corresponding LFDs for \( \epsilon_0 = \epsilon_1 = 0.1 \).

Fig. 4. Robust and nominal LRFs found for the nominal distributions denoted by \( d_2 \) and \( \epsilon_0 = \epsilon_1 = 0.1 \), including the symmetric case \( \alpha_s \). The optimum values of \( u \) are 0.95, 0.67, 0.56, 0.59, 0.61, respectively, from \( \alpha = 0.1 \) to \( \alpha_s = 2 \).

the LFDs which are found as a function of \( u \), cf. Section IV. In order to see whether the convexity arguments still hold in general, \( D_u \) is plotted for the pair of nominal distributions \( d_1 \), \( d_2 \) and \( d_3 \), when the distance is the KL-divergence and additionally for \( d_2 \), when the distance is the \( \alpha = 0.1 \)-divergence. The robustness parameters are the same as in the previous simulations. Figure 7 illustrates the outcome of this simulation, which proves the existence of distances (i.e. \( \alpha = 0.1 \)) for which \( D_u \) is not necessarily convex, although it may not possibly be the case for
Fig. 5. The nominal distributions denoted by $d_3$ and the corresponding LFDs for $\epsilon_0 = \epsilon_1 = 0.05$, where $u = 0.46$.

Fig. 6. Robust and nominal LRFs found for the nominal distributions denoted by $d_3$ and $\epsilon_0 = \epsilon_1 = 0.05$.

the KL-divergence.

The LFDs are obtained by solving a system of non-linear equations for every choice of $u$. These parameters can be depicted so that the results can easily be verified by others. In this example again the KL-divergence neighborhood is considered with $\epsilon_0 = \epsilon_1 = 0.1$. In Figure 8 the KKT parameters $\lambda_0$, $\lambda_1$, $\mu_0$ and $\mu_1$ are illustrated for the pairs of nominal distributions denoted by $d_1$ and $d_2$. For both examples, the KKT parameters follow similar paths.
Fig. 7. The $u$-divergence as a function of $u$ for the LFDs obtained for various pairs of distributions as well as uncertainty classes.

Fig. 8. KKT parameters for two pairs of nominal distributions and the KL-divergence neighborhood with $\epsilon_0 = \epsilon_1 = 0.1$.

3) Asymptotically Minimax Robust NP-tests: Dabak’s test is neither minimax robust nor asymptotically minimax robust as it was shown in Section IV, V-A. This result can be demonstrated with an example. Let the nominal distributions be given as in Table III with $d_2$ and let $\epsilon_0 = \epsilon_1 = 0.01$. The rate functions $I_0$ and $I_1$ are of interest for two cases; $|a_1|$ and $|a^*|$, i.e. when the test is asymptotically minimax robust Type-I NP-test and Dabak’s test, respectively. Figures 9 and 10 illustrate the rate functions $I_0$ and $I_1$. Both in Figure 9 and Figure 10 the performance of the $(a^*)$-test is degraded by the data samples obtained from the LFDs of the
(a)-test in comparison to those obtained from the LFDs of the (a*-)test.

4) Band Model: Asymptotically minimax robust tests arising from the band model can similarly be simulated. Consider the lower bounding functions

\[ g_0^L(y) = (1 - \epsilon)f_N(y; -1, 4), \quad g_1^L(y) = (1 - \epsilon)f_N(y; 1, 4), \]

where the contamination ratio is chosen to be \( \epsilon = 0.2 \). Furthermore, let the upper bounding functions be

\[ g_0^U(y) = (1 + \epsilon)f_N(y; -1, 4), \quad g_1^U(y) = (1 + \epsilon)f_N(y; 1, 4), \]

with the parameters \( \epsilon = 0.2 \) (Type-I), \( \epsilon = 0.5 \) (Type-II), \( \epsilon = 1.5 \) (Type-III) or \( \epsilon = 19 \) (Type-III), simulating three different types of robust LRFs resulting from the band model, cf. Section V-E. For this setup, and excluding \( \epsilon = 19 \) for the sake of clarity, Figure 11 illustrates the corresponding LFDs together with the bounding functions. For \( \epsilon = 1.5 \), the LFDs are overlapping around \( y = 0 \), leading to \( \hat{l} = 1 \). This type of overlapping has previously been reported by [12] for single sample minimax robust tests obtained from the KL-divergence neighborhood. However, the test in [12] is not minimax robust unless a well defined randomized decision rule is used.

In Figure 12, the corresponding robust likelihood ratio functions are illustrated. Increasing \( \epsilon \) transforms the corresponding robust LRF from Type-I to Type-II and then to Type-III. Further increasing \( \epsilon \), i.e. when \( \epsilon = 19 \), the robust LRF tends to a clipped likelihood ratio test, which is the limiting LRF stated in Section V-E Case 3, and also derived in Case 1. The robust LRFs can take different shapes depending on the bounding functions. Similar patterns were stated with equations in [18] and also observed in [19].

5) Moment Classes: The LFDs and robust LRFs arising from the moment classes can be exemplified by solving the convex optimization problem given in Section IV-E through replacing the constraints with the ones defined by the moment classes. Consider the constraints

\[ -2 \leq E_{G_0}[Y] \leq -0.5, \quad 0.5 \leq E_{G_1}[Y] \leq 2, \]
\[ 0 \leq E_{G_0}[Y^2] \leq 2, \quad 2 \leq E_{G_1}[Y^2] \leq 4, \]

defined over the first and second moments of the probability density functions. Figures 13 and 14 illustrate the LFDs and the corresponding robust LRF, respectively.
6) **P-point Classes:** Similarly, an example to the asymptotically minimax robust test arising from the p-point classes can be given. Consider the p-point classes defined by the constraints

\[ \int_{-5}^{3} g_0(y)dy \leq 0.3, \quad \int_{0}^{3} g_1(y)dy \geq 0.8. \]

Figures [15] and [16] illustrate the LFDs and the corresponding robust LRF, respectively.
Fig. 11. Three different pairs of LFDs arising from the band model together with the bounding functions for $\varepsilon \in \{1.2, 1.5, 2.5\}$.

Fig. 12. Three different types of Robust LRFs arising from the band model for $\varepsilon \in \{1.2, 1.5, 2.5, 19\}$ together with the nominal LRF.

B. Signal Processing Examples

The theory presented in this paper is applicable to any signal processing example, where robust detection of events is of interest. Let us consider spectrum sensing used in cognitive radio to allow unlicensed or secondary users to use spectrum holes that are not occupied by licensed or
primary users [48]. Presence of absence of a signal is formulated by a binary hypothesis test

\[ \mathcal{H}_0 : y[n] = w[n], \quad n \in \{1, \ldots, N\} \]
\[ \mathcal{H}_1 : y[n] = \theta x[n] + w[n], \quad n \in \{1, \ldots, N\} \]

where \( w[n] \) are noise samples, \( x[n] \) are unattenuated samples of the primary signal, \( \theta > 0 \) is the unknown channel gain and \( y[n] \) are the received signal samples. Both the primary signal samples \( x[n] \) and the noise samples \( w[n] \), which are independent of \( x[n] \), are i.i.d. standard Gaussian.
Fig. 15. Least favorable distributions arising from the p-point classes in the given example.

Fig. 16. The robust LRF arising from the p-point classes in the given example.

Under each hypothesis, it is assumed that the distribution of $Y$ may deviate from its nominal distribution by a factor of $\epsilon_0 = \epsilon_1 = 0.02$ with respect to the KL-divergence. Furthermore, the channel gain is assumed to be perfectly estimated as $\theta = \sqrt{3}$. The LFDs corresponding to the robust tests listed in Table II are found by solving the related equations, e.g. (24). The LFDs of the (h)-test are determined from the $\epsilon$-contamination neighborhood such that $D_{KL}(\hat{g}_0, f_0) = \epsilon_0$ and $D_{KL}(\hat{g}_1, f_1) = \epsilon_1$. In Figure 17 the ratio of the robust LRF to the nominal LRF for four different robust hypothesis testing schemes is depicted. Both similarities and differences can be observed, and in particular $\hat{l}/l$ is not integrable for the (m)-test.
For the aforementioned scenario, the goal is to evaluate the performance of the (a)-test, (a*)-test, (h)-test and the (n)-test under fixed sample size and sequential concepts for various statistics of the data samples.

1) Fixed Sample Size Test: For a fixed number of samples ranging from $n = 1$ to $n = 100$, the performance of the robust tests are evaluated with Monte-Carlo simulations for $10^6$ samples. The threshold of the fixed sample size test is set to $t = 1$. False alarm and miss detection probabilities of the (a)-test in comparison to that of the (a*)-test are illustrated in Figures 18 and 19 when the tested data samples are obtained from the LFDs of the robust tests listed in Table II. In Figures 20 and 21 similar experiments are repeated for the (h)-test in comparison to the (n)-test. The following conclusions can be made from these experiments.

- The (a)-test does not degrade its performance as the theory suggests.
- The (a*)-test degrades its performance for the data samples obtained from the LFDs of the (a)- and (h)-tests, cf. Figure 18
- The data samples obtained from the LFDs do not always yield worse results than that of the nominal test, cf. Figure 19
- Minimax robustness of the (h)-test for the $\epsilon$-contamination neighborhood is in trouble if indeed the uncertainties can be well modeled by the KL-divergence neighborhood, see Figure 20.
Fig. 18. False alarm probability as a function of the total number of samples for the asymptotically minimax robust test ((a)-test) in comparison to Dabak's test ((a')-test).

Fig. 19. Miss detection probability as a function of the total number of samples for the asymptotically minimax robust test ((a)-test) in comparison to Dabak's test ((a')-test).

2) Sequential Test: The performance of the sequential version of the (a)-test can similarly be evaluated. Of particular interest is whether the new definitions of the minimax robust sequential test in Section VI hold. The value of $c = 1$ is chosen so that $t_u = 1/t_l$. The SPRT is run for every $t_u \in \{0.01, 0.02, \ldots, 4\}$ assuming the same experimental setup used for fixed sample size tests. Figure 22 illustrates the false alarm and miss detection probabilities resulting from the sequential (a)-test as a function of $n$. In Figure 23 the same simulation is repeated for the pair
of nominal distributions denoted by $d_2$ in Table III with $\epsilon_0 = \epsilon_1 = 0.1$. According to both results, the sequential (a)-test does not degrade its performance as $n$ gets larger for any input data that is considered. It was also verified that the (a)-test is not asymptotically minimax robust both for the equations defined by (68) and (69).
Fig. 22. False alarm and miss detection probabilities resulting from the sequential (a)-test as a function of $n$ for the given signal processing example.

Fig. 23. False alarm and miss detection probabilities resulting from the sequential (a)-test as a function of $n$ for the nominal distributions denoted by $d_2$ and the KL-divergence neighborhood with $\epsilon_0 = \epsilon_1 = 0.1$.

VIII. CONCLUSIONS

Designing (asymptotically) minimax robust Bayesian hypothesis testing schemes for any reasonable construction of the uncertainty classes has been shown to be made via maximizing $D_u(G_0, G_1)$ over all $(\hat{g}_0, \hat{g}_1) \in \mathcal{G}_0 \times \mathcal{G}_1$ for the minimizing $u$. The uncertainty classes based on the KL-divergence, $\alpha$-divergence, symmetrized $\alpha$-divergence, the total variation distance, as well as the band model, moment classes and p-point classes have been considered. For the first
six classes, the KKT multipliers approach was employed to derive the LFDs and the minimax robust tests in parametric forms. The parameters can be obtained by means of solving non-linear systems of equations. For the latter two classes, the asymptotically minimax robust tests has been evaluated as a convex optimization problem. In addition to the Bayesian formulation, Neyman-Pearson versions of the minimax robust hypothesis testing schemes have been derived considering the KL-divergence neighborhood. The asymptotically minimax robust Neyman-Pearson tests of Type I and II correspond to a non-linear transformation of the nominal LRF, which involves the Lambert-W function. This result proved that Dabak’s test was not asymptotically minimax robust. Existence and uniqueness of a single sample deterministic minimax robust test imply existence and uniqueness of an all-sample minimax robust test. It was shown that single sample minimax robust tests may fail to exist even for a simple example. Moreover, randomized versions of single sample minimax robust tests, if they exist, do not lead to all sample minimax robust tests. These results inevitably motivate considering asymptotical designs. Although the term asymptotic means applicability of the theory for very large sample sizes, it was proven that the proposed approach finds at least the asymptotically minimax, and at best, if it exists, the all-sample minimax robust test. Therefore, in some sense, the asymptotic design may be called the best possible design in terms of minimax robustness. Generalizations of the theory include multiple-, decentralized- and sequential hypothesis testing. The asymptotic results straightforwardly apply to multiple hypothesis and decentralized detection with parallel sensor networks, while such an extension is not taken for granted for more general networks. Similarly, the sequential version of the (a)-test has been verified not to be asymptotically minimax robust according to Huber’s definitions. In engineering applications, the quickest detection for the lowest error probabilities is the major requirement. Based on this idea a novel characterization of minimax robustness for the SPRT was proposed. In order to evaluate theoretical findings and their applicability simulations have been performed. In particular, the LFDs and the robust LRFs were examplified for every pair of uncertainty classes. Although clipping was already known to result in minimax robustness, it was first observed here that asymptotic minimax robustness may require amplifying the nominal likelihood ratios by a great factor. Additionally, a large number of constraints may be required such that the LFDs resulting from the moment and p-point classes are smooth enough. Limited simulations supported
that the sequential (a)-test may be asymptotically minimax robust for this new definition, though, no general proofs were provided. A signal processing example from spectrum sensing shows the usefulness of the theory in practice, both for fixed sample size and sequential tests.

From this work the following questions are open:

- Is it true that if $f_0(-y) = f_1(y)$ for all $y$, then $u = 1/2$ is the minimizer for all/some pairs of uncertainty classes?
- Is it true that the pair $(\hat{g}_0, \hat{g}_1)$ maximizing $D_u$ for every $u \in \mathcal{Y}$ implies single sample, hence, all-sample minimax robustness?
- Are there uncertainty classes for which the sequential (a)-test is (asymptotically) minimax robust for the new definitions made?
- How should the asymptotic design look like if each random variable $Y_k$ is subjected to different uncertainties, or if $Y_k$ are not mutually independent?

**References**

[1] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory*. Prentice Hall PTR, Jan. 1998.

[2] P. J. Huber, *Robust statistics*. Wiley New York, 1981.

[3] S. Kassam, G. Moustakides, and J. Shin, “Robust detection of known signals in asymmetric noise,” *IEEE Transactions on Information Theory*, vol. 28, no. 1, pp. 84–91, January 1982.

[4] S. Kassam and J. Thomas, “Asymptotically robust detection of a known signal in contaminated non-Gaussian noise,” *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 22–26, January 1976.

[5] A. El-Sawy and V. VandeLinde, “Robust detection of known signals,” *IEEE Transactions on Information Theory*, vol. 23, no. 6, pp. 722–727, November 1977.

[6] J. D. Gibson and J. L. Melsa, *Introduction to nonparametric detection with applications*, ser. Mathematics in science and engineering. New York, San Francisco, London: Academic Press, 1975.

[7] G. Gül, *Robust and Distributed Hypothesis Testing*, ser. Lecture Notes in Electrical Engineering. Springer, 2017, vol. 414.

[8] P. J. Huber, “A robust version of the probability ratio test,” *Ann. Math. Statist.*, vol. 36, pp. 1753–1758, 1965.

[9] P. J. Huber and V. Strassen, “Minimax tests and the Neyman-Pearson lemma for capacities,” *Ann. Statistics*, vol. 1, pp. 251–263, 1973.

[10] B. C. Levy, “Robust hypothesis testing with a relative entropy tolerance,” *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 413–421, 2009.

[11] G. Gül and A. M. Zoubir, “Robust hypothesis testing with $\alpha$ -divergence,” *IEEE Transactions on Signal Processing*, vol. 64, no. 18, pp. 4737–4750, Sept 2016.

[12] G. Gül and A. M. Zoubir, “Minimax robust hypothesis testing,” *IEEE Transactions on Information Theory*, vol. PP, no. 99, pp. 1–1, 2017.
[13] A. G. Dabak and D. H. Johnson, “Geometrically based robust detection,” in Proceedings of the Conference on Information Sciences and Systems, Johns Hopkins University, Baltimore, MD, May 1994, pp. 73–77.

[14] C. Pandit and S. Meyn, “Worst-case large-deviation asymptotics with application to queueing and information theory,” Stochastic Processes and their Applications, vol. 116, no. 5, pp. 724 – 756, 2006.

[15] P. J. Huber and V. Strassen, “Robust confidence limits,” Z. Wahrscheinlichkeitstheorie und verw. Gebiete, vol. 10, pp. 269–278, 1968.

[16] G. Gül, “Minimax robust decentralized detection in parallel sensor networks,” Jan. [Online]. Available: https://arxiv.org/abs/1703.07670

[17] K. Vastola and H. Poor, “On the p-point uncertainty class (corresp.),” IEEE Transactions on Information Theory, vol. 30, no. 2, pp. 374–376, March 1984.

[18] S. Kassam, “Robust hypothesis testing for bounded classes of probability densities (corresp.),” IEEE Transactions on Information Theory, vol. 27, no. 2, pp. 242–247, March 1981.

[19] M. Fauss and A. M. Zoubir, “Old bands, new tracks-revisiting the band model for robust hypothesis testing,” IEEE Transactions on Signal Processing, vol. 64, no. 22, pp. 5875–5886, Nov 2016.

[20] R. Martin and S. Schwartz, “Robust detection of a known signal in nearly Gaussian noise,” IEEE Transactions on Information Theory, vol. 17, no. 1, pp. 50–56, Jan 1971.

[21] R. Martin and C. McGath, “Robust detection to stochastic signals (corresp.),” IEEE Transactions on Information Theory, vol. 20, no. 4, pp. 537–541, Jul 1974.

[22] A. G. Dabak, “A geometry for detection theory,” Ph.D. dissertation, Rice University, May 1993.

[23] B. C. Levy, Principles of Signal Detection and Parameter Estimation, 1st ed. Springer Publishing Company, Incorporated, 2008.

[24] H. Chernoff, “A measure of asymptotic efficiency for tests of a hypothesis based on the sums of observations,” Annals of Mathematical Statistics, vol. 23, pp. 409–507, 1952.

[25] F. Osterreicher and I. Vajda, “Statistical information and discrimination,” IEEE Transactions on Information Theory, vol. 39, no. 3, pp. 1036–1039, May 1993.

[26] A. Guntuboyina, S. Saha, and G. Schiebinger, “Sharp inequalities for f-divergences,” IEEE Transactions on Information Theory, vol. 60, no. 1, pp. 104–121, Jan 2014.

[27] H. Cramér, “Sur un nouveau théorème-limite de la théorie des probabilités,” Actualités Scientifiques et Industrielles, no. 736, Hermann Cie, Paris, 1938.

[28] A. Cichocki and S. Amari, “Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities,” Entropy, vol. 12, no. 6, pp. 1532–1568, 2010.

[29] M. Sion, “On general minimax theorems.” Pacific Journal of Mathematics, vol. 8, no. 1, pp. 171–176, 1958.

[30] W. Rudin, Principles of Mathematical Analysis, ser. International series in pure and applied mathematics. Paris: McGraw-Hill, 1976.

[31] F. Liese and I. Vajda, “On divergences and informations in statistics and information theory,” IEEE Trans. Information Theory, vol. 52, no. 10, pp. 4394–4412, 2006.

[32] D. Bertsekas, A. Nedić, and A. Ozdaglar, Convex Analysis and Optimization, ser. Athena Scientific optimization and computation series. Athena Scientific, 2003.

[33] B. Pczos and J. G. Schneider, “On the estimation of alpha-divergences,” in Proceedings of the Fourteenth International
[34] J. E. Smith, “Generalized chebychev inequalities: Theory and applications in decision analysis,” Oper. Res., vol. 43, no. 5, pp. 807–825, Oct. 1995.

[35] F. Brichet and A. Simonian, “Conservative Gaussian models applied to measurement-based admission control,” in Quality of Service, 1998. (IWQoS 98) 1998 Sixth International Workshop on, May 1998, pp. 68–71.

[36] M. A. Johnson and M. R. Taaffe, “An investigation of phase-distribution moment-matching algorithms for use in queueing models,” Queueing Systems, vol. 8, no. 1, pp. 129–147, Dec 1991.

[37] A. El-Sawy and V. VandeLinde, “Robust sequential detection of signals in noise,” IEEE Transactions on Information Theory, vol. 25, no. 3, pp. 346–353, May 1979.

[38] D. Sakrison, “The rate of a class of random processes,” IEEE Transactions on Information Theory, vol. 16, no. 1, pp. 10–16, January 1970.

[39] L. Cimini and S. Kassam, “Robust and quantized wiener filters for p-point spectral classes,” in 14th Conf. on Inform. Sciences and Systems, Princeton University, Princeton, NJ, 1980, pp. 314–319.

[40] A. K. Jain, R. P. W. Duin, and J. Mao, “Statistical pattern recognition: a review,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 4–37, Jan 2000.

[41] S. Dudoit, J. P. Shaffer, and J. C. Boldrick, “Multiple hypothesis testing in microarray experiments,” Statistical Science, vol. 18, no. 1, pp. 71–103, 2003.

[42] T. M. Cover, “Hypothesis testing with finite statistics,” Ann. Math. Statist., vol. 40, no. 3, pp. 828–835, 06 1969.

[43] M. E. Hellman and T. M. Cover, “Learning with finite memory,” Ann. Math. Statist., vol. 41, no. 3, pp. 765–782, 06 1970.

[44] J. N. Tsitsiklis, “Decentralized detection,” in In Advances in Statistical Signal Processing, JAI Press, 1993, pp. 297–344.

[45] A. Wald, “Sequential tests of statistical hypotheses,” The Annals of Mathematical Statistics, vol. 16, no. 2, pp. 117–186, 06 1945.

[46] D. Ralph, “Global convergence of damped newton’s method for nonsmooth equations via the path search,” Math. Oper. Res., vol. 19, no. 2, pp. 352–389, 1994.

[47] F. A. Potra and S. J. Wright, “Interior-point methods,” J. Comput. Appl. Math., vol. 124, no. 1-2, pp. 281–302, Dec 2000.

[48] Q. Zhao and B. M. Sadler, “A survey of dynamic spectrum access,” IEEE Signal Processing Magazine, vol. 24, no. 3, pp. 79–89, May 2007.