Constructive Minimax Classification of Discrete Observations with Arbitrary Loss Function

Lionel Fillatre

Constructive Minimax Classification of Discrete Observations with Arbitrary Loss Function

Lionel Fillatre

Université Côte d'Azur, CNRS, I3S Laboratory
CS 40121 - 06903 Sophia Antipolis CEDEX, France
E-mail: lionel.fillatre@i3s.unice.fr

Abstract

This paper develops a multihypothesis testing framework for calculating numerically the optimal minimax test with discrete observations and an arbitrary loss function. Discrete observations are common in data processing and make tractable the calculation of the minimax test. Each hypothesis is both associated to a parameter defining the distribution of the observations and to an action which describes the decision to take when the hypothesis is true. The loss function measures the gap between the parameters and the actions. The minimax test minimizes the maximum classification risk. It is the solution of a finite linear programming problem which gives the worst case classification risk and the worst case prior distribution. The minimax test equalizes the classification risks whose prior probabilities are strictly positive. The minimax framework is applied to vector channel decoding which consists in classifying some codewords transmitted on a binary asymmetric channel. The Hamming metric is used to measure the number of differences between the emitted codeword and the decoded one.

Keywords: Multiple hypothesis testing, statistical classification, minimax test, linear programming

1. Introduction

The problem of classifying discrete distributions often appears in engineering applications, including pattern recognition with discrete-valued data [1, 2], sensor network with quantized observations [3, 4] image processing [5, 6, 7], and channel decoding [8, 9] among others. The goal of this work is to decide between K hypotheses $\mathcal{H}_1, \ldots, \mathcal{H}_K$ where the Probability Mass Function (pmf) of the observed data $\boldsymbol{x}$ depends on the known value of a certain parameter given the hypothesis. A decision error is measured with an arbitrary loss function which depends both on the true hypothesis and the chosen one. We assume that the prior probabilities of the hypotheses are unknown. This is a classical assumption when the prior knowledge of observations is insufficient.

1.1. Minimax Classification

Contrary to a purely Bayesian criterion which needs a complete statistical description of the problem [10], the minimax criterion is well adapted to classification problems where the probability of each hypothesis is unknown. This criterion consists in minimizing the largest probability to make a classification error. The optimal test consists in choosing the maximum of weighted likelihood functions. The weights are generally very difficult to calculate [11], even in some simple cases. Furthermore, the minimax test may satisfy the equalization property, i.e., the worst classification errors are all equal, which is quite interesting in practice.

There are two main trends in literature to design minimax test. The first trend consists in calculating analytically the minimax test. On the first hand, the minimax test is studied in a general setting [10, 11]. Although there is a vast literature, it is still difficult to find an algorithm which calculates the minimax test for a specific situation. For instance, the famous book [10] does not describe any algorithm to compute a minimax test. On the second hand, the minimax test is often established for a specific issue [12, 13, 14, 15, 16] but the algorithm can not be easily extended to an other observation model.

The second trend consists in computing numerically the minimax test [17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]. The paper [17] is certainly the first to use of programming techniques for testing two composite hypotheses based on discrete random variables. The LP approach was already implicit in [18] with real-valued observations. It is shown in [19] how to use the simplex method for calculating minimax decisions functions. Duality theory was first used in [20] and for the general case of minimax tests in [21]. The results are extended to the more general class of most stringent tests in [22]. The paper [23] introduced a framework where the theory of infinite LP is applicable. The survey [24] gives an overview of these pioneering approaches. All the above mentioned papers are focused on the classification of only two hypotheses.

In the case of several hypotheses, the work [26] is devoted to solving general minimax problems by iteration methods. To solve a decision problem with an arbitrary loss function, numerical solutions involving nonlinear optimization to obtain the least favorable distribution have been studied in [27, 31, 32]. Obtaining the least favorable distribution simplifies the problem but it does not provide necessarily an equalizer minimax test, even when it

exists. Indeed, in case of discrete observations, the equalization of the classification risks needs the randomization of the test, which is not obtained when only the least favorable distribution is computed. The work [25] showed that the theory of infinite LP can be exploited for multiple hypotheses testing problems but they do not propose any constructive algorithm to solve the problem. The case of several hypotheses is closely related to the problem of minimax estimation [33], except that the parameter space is generally not finite but continuous and compact in case of minimax estimation.

1.2. Discrete Observations and Finite Linear Programming

This paper is in favor of a "discretize-then-optimize" approach, i.e., the case of discrete observations can be interpreted as the discretization, or quantization, of continuous real observations. Discrete observations often occur in signal processing applications where the quantization of continuous values is necessary [34]. Digital communications and image processing are some fields where quantization is crucially important [35, 36] to limit the size of the storage or to describe a digital content with only a few features. Wireless sensor networks are characterized by limited resources, such as energy and communication bandwidth. One way to save energy is to limit the data transmitted in the network by using quantized data [3, 37, 38]. More generally, the approach studied in this paper can be easily applied to any signal processing applications where data quantization is of interest. The way the data are quantized is out of the scope of this work.

Discrete observations naturally involve finite Linear Programming (LP). In fact, as described in Section 2, the decision function is then a vector of reals which makes possible the construction of a finite LP problem to com-

pute the minimax test and the worst prior distribution. After discretization, the standard LP problem [39] can be solved using techniques for large-scale LP [40], e.g., interior point methods, Dantzig-Wolfe decomposition, etc. On the contrary, continuous observations lead to infinite LP because the decision function generally belongs to an infinite dimensional space, as shown for instance in [22]. Infinite LP has the advantage to fit a general case but it is numerically difficult to solve as shown for example in [41, 42]. The main way to numerically solve an infinite LP consists in discretizing the problem or to discretize the solution of the problem if it is known. Alternative approaches for solving infinite LP consist in approximating the initial problem by a sequence of LP problems with finite dimensional spaces [42]. The main drawback of this alternative "optimize-then-discretize" approach would be to develop ad-hoc optimization algorithms.

1.3. Contributions of the Paper

The approach proposed in this paper is based on [22] where the author solves a LP problem to calculate the minimax test between only two hypotheses (binary classification). The paper [22] does not consider any loss function; only the probability of misclassification is studied. It deals with hypotheses which can be composite, i.e., each hypothesis may refer to an infinite number of statistical models. It is focused on continuous observations and it studies the minimax test as the solution of an infinite LP problem. It also proves a weak duality theorem between the primal infinite LP problem and its dual. The solution of the dual LP problem gives the worst case distribution of the minimax test. All the results proposed in [22] are theoretical and no algorithm is proposed, or can be easily derived, to compute the min-

imax test. The case of discrete observations is just very briefly introduced as a motivation of the general study. This paper extends [22] to the multiple hypothesis framework (K-ary classification with $K \geq 3$) and to an arbitrary loss function, i.e., it considers that the classification risk between a couple of hypotheses can change with respect to the involved couple of hypotheses. It only considers simple hypothesis: each hypothesis refers to only one statistical model. It is focused on discrete observations in order to make tractable the computation of the minimax test.

The first contribution of this paper is the design of a minimax classification test between multiple hypotheses as the solution of a finite LP problem, called the primal problem, when the observations are discrete and the loss function is arbitrary. This contribution is summarized in Theorem 2. The explicit calculation of the randomized minimax test makes it possible to equalize the classification risks, which is discussed in Corollary 1. This equalization of the classification risks is generally not fulfilled by a Bayes test because it depends on the worst case distribution.

The second contribution is the computation of the worst case distribution, also called the least favorable prior, which is obtained as the solution of the dual LP problem. The minimax test is then expressed as the maximum of weighted likelihood functions, i.e., it is a Bayesian test associated to the worst case weights. This contribution is summarized in Theorem 3. The calculation is very accurate since there is no need of a stopping criterion to halt the algorithm.

Finally, the minimax test is applied to noisy channel decoding. The Hamming metric is used to measure the number of differences between the

emitted codeword and the decoded one. Assuming that the channel and the codebook are known but not the probabilities of each codeword, it is shown that the minimax test outperforms the conventional Maximum Likelihood (ML) decoder, also known as the Multiple Generalized Likelihood Ratio Test (MGLRT), which assumes an uniform prior over the codebook. The ML decoder is clearly suboptimal in case of the binary asymmetric channel when the prior distribution of the codewords is not uniform. It should be noted that the optimality of the minimax test is non-asymptotic and it is different from the random coding sense usually employed in channel decoding.

1.4. Organization of the Paper

The paper is organized as follows. Section 2 describes the statistical framework, including the presentation of the minimax criterion and the LP problem whose solution is the minimax test. Section 3 studies the solution of the LP problems, both the primal and the dual ones, which lead to the minimax test closed-form expression and the worst case distribution of the hypotheses. Section 4 shows the relevance and efficiency of the proposed test for noisy channel decoding. Finally, Section 5 concludes this paper.

The following notations are used throughout the paper. The notation $X \sim p$ means that X follows the pmf p. The expectation of the function $f(X)$ with respect to the distribution of X is denoted $\mathbb{E}^X[f(X)]$. If $X \sim p_{\boldsymbol{\theta}}$ follows the distribution $p_{\boldsymbol{\theta}}$ parametrized by a vector $\boldsymbol{\theta}$, then the expectation is denoted $\mathbb{E}^X_{\boldsymbol{\theta}}[f(X)]$. Lower-case and upper-case letters are for scalar variables or random variables, bold lower-case letters for column vectors, bold upper-case letters for matrices and calligraphic upper-case letters or upper-case Greek letters for sets. Transposition, the transformation of columns into

rows in a vector $\boldsymbol{x}$, resp. a matrix $\boldsymbol{A}$, is denoted by $\boldsymbol{x}^\top$, resp. $\boldsymbol{A}^\top$.

2. Minimax Test for Discrete Distributions

This section presents the multiple hypotheses testing problem which consists in classifying a discrete random vector characterized by its pmf.

2.1. Randomized Classification Problem

Let X be a discrete random variable taking values $\boldsymbol{x}$ in a finite sample space $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ where $\boldsymbol{x}_i \in \mathcal{A}^n$, $\mathcal{A}$ is the coding alphabet of finite size and n is the size of $\boldsymbol{x}$. Assume that X has a pmf denoted $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ where $\boldsymbol{\theta} \in \mathbb{R}^p$ is a vector of parameters characterizing the distribution of X. It is assumed that there are K possible parameters, i.e., $\boldsymbol{\theta} \in \Theta = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$. It is then desirable to solve the multiple hypotheses testing problem between the statistical hypotheses $\mathcal{H}_1$,...,$\mathcal{H}_K$ where $\mathcal{H}_k$ is defined by

$$\mathcal{H}_k : \{X \sim p_{\boldsymbol{\theta}_k}\}. \tag{1}$$

The statistical decision problem between multiple hypotheses is stated as the triplet (Θ, Ψ, w). On the basis of the outcome of the experiment $X = \boldsymbol{x}$ where the pmf of X is $p_{\boldsymbol{\theta}}$ for $\boldsymbol{\theta} \in \Theta$, it is desirable to choose an action $d(\boldsymbol{x}) \in \Psi = \{\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_K\}$ where $d : \mathcal{X} \mapsto \Psi$ is the decision. The loss for state $\boldsymbol{\theta}$ and decision $d(X)$ is the positive random quantity $w(\boldsymbol{\theta}, d(X))$ where $w : \Theta \times \Psi \mapsto [0, +\infty)$ is the loss function and its expectation, called the risk function for state $\boldsymbol{\theta}$ and decision d, is

$$R(\boldsymbol{\theta}, d) = \mathbb{E}_{\boldsymbol{\theta}}^X[w(\boldsymbol{\theta}, d(X))]. \tag{2}$$

Typical loss functions are the $0-1$ loss function, i.e., $w(\boldsymbol{\theta}_j, \boldsymbol{\psi}_k) = 0$ if $j = k$ and 1 otherwise, and the quadratic loss function, .i.e., $w(\boldsymbol{\theta}_j, \boldsymbol{\psi}_k) = \| \boldsymbol{\theta}_j - \boldsymbol{\psi}_k \|^2$ where $\| \cdot \|$ denotes the Euclidean norm when $\boldsymbol{\theta}_j$ and $\boldsymbol{\psi}_k$ belong to the same vector space. Furthermore, this paper exploits randomized decision tests. The decision problem is then viewed as the triplet (Θ, Ψ^*, w) where Ψ^* denotes the set of all pmf defined over Ψ.

Definition 1. *A randomized test for testing K hypotheses $\mathcal{H}_1,\ldots,\mathcal{H}_K$ is any measurable mapping $\boldsymbol{\delta}(\boldsymbol{x}) : \mathcal{X} \mapsto \Psi^*$. The test function $\boldsymbol{\delta}(\boldsymbol{x})=(\delta_1(\boldsymbol{x}), \ldots, \delta_K(\boldsymbol{x}))$ satisfies $0 \leq \delta_k(\boldsymbol{x}) \leq 1$ for all $k = 1, \ldots, K$ and*

$$\sum_{k=1}^{K} \delta_k(\boldsymbol{x}) = 1, \ \ \forall \, \boldsymbol{x} \in \mathcal{X}. \tag{3}$$

The set of randomized tests is denoted $\mathcal{D}^*$.

The randomized decision function $\boldsymbol{\delta} \in \mathcal{D}^*$ chooses action $\boldsymbol{\psi}_i$ with the probability $\delta_i(\boldsymbol{x})$. Then, the average loss is

$$\mathbb{E}^{\boldsymbol{\delta}(\boldsymbol{x})}[w(\boldsymbol{\theta}, \boldsymbol{\delta}(\boldsymbol{x}))] = \sum_{j=1}^{K} \delta_j(\boldsymbol{x}) w(\boldsymbol{\theta}, \boldsymbol{\psi}_j). \tag{4}$$

The risk function $R(\boldsymbol{\theta}, \boldsymbol{\delta})$ becomes:

$$\begin{aligned} R(\boldsymbol{\theta}, \boldsymbol{\delta}) &= \mathbb{E}_{\boldsymbol{\theta}}^{X}[w(\boldsymbol{\theta}, \boldsymbol{\delta}(X))] = \sum_{i=1}^{m} p_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \mathbb{E}^{\boldsymbol{\delta}(\boldsymbol{x}_i)}[w(\boldsymbol{\theta}, \boldsymbol{\delta}(\boldsymbol{x}_i))] \\ &= \sum_{i=1}^{m} \sum_{j=1}^{K} p_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \delta_j(\boldsymbol{x}_i) w(\boldsymbol{\theta}, \boldsymbol{\psi}_j). \end{aligned} \tag{5}$$

Definition 2. *A randomized test $\boldsymbol{\delta}^* \in \mathcal{D}^*$ is a minimax test between the hypotheses $\mathcal{H}_1,\ldots,\mathcal{H}_K$ if*

$$\max_{\boldsymbol{\theta} \in \Theta} R(\boldsymbol{\theta}, \boldsymbol{\delta}^*) = \inf_{\boldsymbol{\delta} \in \mathcal{D}^*} \max_{\boldsymbol{\theta} \in \Theta} R(\boldsymbol{\theta}, \boldsymbol{\delta}). \tag{6}$$

The minimax test is a special case of the Bayes test whose definition is recalled hereafter. Define the K-dimensional unit simplex

$$\mathcal{S}_K = \{\boldsymbol{q} \in [0,1]^K : \sum_{j=1}^{K} q_j = 1\}. \tag{7}$$

Definition 3 (Bayes test)**.** *Let $\boldsymbol{q} \in \mathcal{S}_K$ be a prior distribution. Given $1 \leq j \leq K$, let $g_j^{(\boldsymbol{q})}(\boldsymbol{x})$ be the j-th weighted likelihood defined by*

$$g_j^{(\boldsymbol{q})}(\boldsymbol{x}) = \sum_{k=1}^{K} q_k p_{\boldsymbol{\theta}_k}(\boldsymbol{x}) w(\boldsymbol{\theta}_k, \boldsymbol{\psi}_j),\ \forall \boldsymbol{x} \in \mathcal{X}. \tag{8}$$

The Bayes test function, also called the weighted likelihood test function, $\boldsymbol{\delta}^{(\boldsymbol{q})}(\boldsymbol{x}) = (\delta_1^{(\boldsymbol{q})}(\boldsymbol{x}), \ldots, \delta_K^{(\boldsymbol{q})}(\boldsymbol{x}))$ is given by:

$$\delta_j^{(\boldsymbol{q})}(\boldsymbol{x}) = \begin{cases} 1 & \text{if } g_j^{(\boldsymbol{q})}(\boldsymbol{x}) < \min_{k \neq j}\{g_k^{(\boldsymbol{q})}(\boldsymbol{x})\}, \\ \eta_j(\boldsymbol{x}) & \text{if } g_j^{(\boldsymbol{q})}(\boldsymbol{x}) = \min_{k \neq j}\{g_k^{(\boldsymbol{q})}(\boldsymbol{x})\}, \\ 0 & \text{if } g_j^{(\boldsymbol{q})}(\boldsymbol{x}) > \min_{k \neq j}\{g_k^{(\boldsymbol{q})}(\boldsymbol{x})\}, \end{cases} \tag{9}$$

where $0 \leq \eta_j(\boldsymbol{x}) \leq 1$ for all $\boldsymbol{x}$.

Let $r(\boldsymbol{q}, \boldsymbol{\delta})$ be the Bayes risk, associated to prior distribution $\boldsymbol{q} \in \mathcal{S}_K$ and the test $\boldsymbol{\delta}$, defined by

$$\begin{aligned} r(\boldsymbol{q}, \boldsymbol{\delta}) &= \mathbb{E}^{\boldsymbol{\theta}, X} R(\boldsymbol{\theta}, \boldsymbol{\delta}) = \sum_{k=1}^{K} q_k R(\boldsymbol{\theta}_k, \boldsymbol{\delta}) \\ &= \sum_{k=1}^{K} \sum_{i=1}^{m} \sum_{j=1}^{K} q_k p_{\boldsymbol{\theta}_k}(\boldsymbol{x}_i) w(\boldsymbol{\theta}_k, \boldsymbol{\psi}_j) \delta_j(\boldsymbol{x}_i) \end{aligned} \tag{10}$$

where $\boldsymbol{\theta}$ is considered as a random variable such that $\boldsymbol{\theta} \sim \boldsymbol{q}$. The following theorem [10, section 2.9, Theorem 1] shows that the minimax test can be interpreted as a Bayes test associated to a least favorable prior distribution.

Theorem 1 (The Minimax Theorem)**.** *For a given decision problem* (Θ, Ψ^*, w)*, there exists a minimax test* $\boldsymbol{\delta}^*$ *and a least favorable distribution* $\boldsymbol{q}^*$*, also called the worst case prior, such that*

$$\inf_{\boldsymbol{\delta}\in\mathcal{D}^*} \sup_{\boldsymbol{q}\in\mathcal{S}_K} r(\boldsymbol{q},\boldsymbol{\delta}) = \sup_{\boldsymbol{q}\in\mathcal{S}_K} \inf_{\boldsymbol{\delta}\in\mathcal{D}^*} r(\boldsymbol{q},\boldsymbol{\delta}) = r(\boldsymbol{q}^*,\boldsymbol{\delta}^*). \tag{11}$$

The test $\boldsymbol{\delta}^*$ *is Bayes with respect to* $\boldsymbol{q}^*$*.*

Theorem 1 establishes the existence of the minimax test. Furthermore, let $\boldsymbol{\delta}^{(\boldsymbol{q})}(\boldsymbol{x})$ be the Bayes test which minimizes the Bayes risk when $\boldsymbol{q}$ is given, i.e., $r(\boldsymbol{q},\boldsymbol{\delta}^{(\boldsymbol{q})}) = \inf_{\boldsymbol{\delta}\in\mathcal{D}^*} r(\boldsymbol{q},\boldsymbol{\delta})$. Theorem 1 shows that the minimax test achieves the maximum Bayes risk $\sup_{\boldsymbol{q}\in\mathcal{S}_K} r(\boldsymbol{q},\boldsymbol{\delta}^{(\boldsymbol{q})})$.

2.2. Primal and Dual Linear Programming Problems

The following theorem shows that the calculation of the minimax test is stated as the solution of a LP problem. To the best of our knowledge, although the proof is straightforward, it was not established in a previous paper.

Theorem 2. *The test* $\boldsymbol{\delta}^*(\boldsymbol{x})$ *is a minimax test for testing* $\mathcal{H}_1,\ldots,\mathcal{H}_K$ *in the class* $\mathcal{D}^*$ *if and only if there exists a number* $\gamma^* \geq 0$ *such that* $(\boldsymbol{\delta}^*,\gamma^*)$ *is a solution of the LP problem where the linear form* $b(\boldsymbol{\delta},\gamma) = \gamma$ *has to be minimized among the couples* $(\boldsymbol{\delta},\gamma)$ *of the class* $\mathcal{K} \subset \Psi^* \times \mathbb{R}^+$ *that is defined*

by the following restrictions:

$$\delta_i(\boldsymbol{x}) \geq 0, \qquad \forall \boldsymbol{x} \in \mathcal{X}, \forall 1 \leq i \leq K, \tag{12}$$

$$-\delta_i(\boldsymbol{x}) \geq -1, \qquad \forall \boldsymbol{x} \in \mathcal{X}, \forall 1 \leq i \leq K, \tag{13}$$

$$\sum_{i=1}^{K} \delta_i(\boldsymbol{x}) \geq 1, \qquad \forall \boldsymbol{x} \in \mathcal{X}, \tag{14}$$

$$-\sum_{i=1}^{K} \delta_i(\boldsymbol{x}) \geq -1, \qquad \forall \boldsymbol{x} \in \mathcal{X}, \tag{15}$$

$$\gamma - R(\boldsymbol{\theta}_k, \boldsymbol{\delta}) \geq 0, \qquad \forall k \in 1, \ldots, K, \tag{16}$$

where $R(\boldsymbol{\theta}_k, \boldsymbol{\delta})$, given in (5), is a linear function of $\boldsymbol{\delta}$.

Proof. If $\boldsymbol{\delta}^*(\boldsymbol{x})$ is a minimax test with the maximum loss $\gamma^* = \max_{1 \leq k \leq K} R(\boldsymbol{\theta}_k, \boldsymbol{\delta}^*) \geq 0$, then the pair $(\boldsymbol{\delta}^*, \gamma^*)$ satisfies the restrictions (12)-(16). Furthermore, if the test $\boldsymbol{\delta}^*$ is minimax, then an other pair $(\boldsymbol{\delta}, \gamma)$ can not satisfy these restrictions unless $\gamma \geq \gamma^*$.

If $(\boldsymbol{\delta}^*, \gamma^*)$ is a solution to the LP problem, then $\boldsymbol{\delta}^* \in \mathcal{D}^*$ on account of (12)-(15). Moreover, $\max_{1 \leq k \leq K} R(\boldsymbol{\theta}_k, \boldsymbol{\delta}^*) \leq \gamma^*$ on account of (16). It is obvious that the optimum solution should satisfy $\max_{1 \leq k \leq K} R(\boldsymbol{\theta}_k, \boldsymbol{\delta}^*) = \gamma^*$ since γ^* is minimized. If $\boldsymbol{\delta}^*$ is not a minimax test, then there exists $\boldsymbol{\delta}' \in \mathcal{D}^*$ with $0 \leq \max_{1 \leq k \leq K} R(\boldsymbol{\theta}_k, \boldsymbol{\delta}') = \gamma' < \gamma$ and hence there exists a feasible element $(\boldsymbol{\delta}', \gamma') \in \mathcal{K}$ with $\gamma' < \gamma$, which contradicts the fact that $(\boldsymbol{\delta}^*, \gamma^*)$ is a solution to the LP problem. □

All the restrictions in Theorem 2 are given under the form of inequalities but, if necessary, the restrictions (14) and (15) can be merged and rewritten as the equality contraint (3). The LP problem given in Theorem 2 is easily stated in matrix form. Specifically, let $w_{j,k} = w(\boldsymbol{\theta}_j, \boldsymbol{\psi}_k)$ and $p_{i,j}$ be the

probability

$$p_{i,j} = p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_i). \tag{17}$$

The randomized test $\boldsymbol{\delta}(\boldsymbol{x})$ is defined by the matrix $\Delta = [\delta_{i,j}]$ where $\delta_{i,j} = \delta_j(\boldsymbol{x}_i)$ for all $1 \leq i \leq m$ and $1 \leq j \leq K$. The risk function in (5) is

$$R(\boldsymbol{\theta}_k, \boldsymbol{\delta}) = \sum_{i=1}^{m} \sum_{j=1}^{K} p_{i,k}\, w_{k,j}\, \delta_{i,j} \tag{18}$$

for all $\boldsymbol{\theta}_k$. The vectorization of the $K \times m$ matrix $\Delta = [\delta_{i,j}]$, denoted by $\mathrm{vec}(\Delta)$ or $\mathrm{vec}(\delta_{i,j})$, is the $Km \times 1$ column vector obtained by stacking the columns of the matrix Δ on top of one another:

$$\mathrm{vec}(\Delta) = (\delta_{1,1}, \ldots, \delta_{m,1}, \ldots, \delta_{1,K}, \ldots, \delta_{m,K})^\top. \tag{19}$$

Then, it is aimed to solve the minimax LP problem MLP1:

$$\min_{\boldsymbol{y} \geq 0} \quad \boldsymbol{b}^\top \boldsymbol{y} = \gamma \tag{20}$$

$$\text{s.t.} \quad \boldsymbol{A}^\top \boldsymbol{y} \geq \boldsymbol{c} \tag{21}$$

where the vector $\boldsymbol{y}$ is given by

$$\boldsymbol{y} = (\mathrm{vec}(\Delta)^\top, \gamma)^\top \in \mathbb{R}^{Km+1}, \tag{22}$$

the vector $\boldsymbol{b}$ associated to the linear form is

$$\boldsymbol{b} = (0, \ldots, 0, 1)^\top \in \mathbb{R}^{Km+1}, \tag{23}$$

the second member $\boldsymbol{c} \in \mathbb{R}^{m(K+2)+K}$ is

$$\boldsymbol{c} = (-\mathbf{1}_{1\times mK},\, \mathbf{1}_{1\times m},\, -\mathbf{1}_{1\times m},\, \mathbf{0}_{1\times K})^\top, \tag{24}$$

and the class of feasible solutions is characterized by the $(m(K+2)+K) \times (mK+1)$ matrix $\boldsymbol{A}$ given by

$$\boldsymbol{A}^\top = \begin{bmatrix} -\mathbf{I}_{mK} & \mathbf{0}_{mK\times 1} \\ \mathbf{1}_{1\times K} \otimes \mathbf{I}_m & \mathbf{0}_{m\times 1} \\ -\mathbf{1}_{1\times K} \otimes \mathbf{I}_m & \mathbf{0}_{m\times 1} \\ -\boldsymbol{M} & \mathbf{1}_{K\times 1} \end{bmatrix}. \tag{25}$$

The matrix $\boldsymbol{M} = \mathbf{Q}^\top(\boldsymbol{W} \otimes \mathbf{I}_m)$ of size $K \times mK$ models the K restrictions (16) within the matrix $\boldsymbol{A}$. It is detailed as follows:

$$\boldsymbol{M} = \begin{bmatrix} p_{1,1}\, w_{1,1} & \dots & p_{m,1}\, w_{1,1} & p_{1,1}\, w_{1,2} & \dots \\ p_{1,2}\, w_{2,1} & \dots & p_{m,2}\, w_{2,1} & p_{1,2}\, w_{2,2} & \dots \\ \vdots & \vdots & \vdots & \vdots & \dots \\ p_{1,K}\, w_{K,1} & \dots & p_{m,K}\, w_{K,1} & p_{1,K}\, w_{K,2} & \dots \end{bmatrix}$$
$$\begin{bmatrix} \dots & p_{m,1}\, w_{1,2} & \dots & p_{1,1}\, w_{1,K} & \dots & p_{m,1}\, w_{1,K} \\ \dots & p_{m,2}\, w_{2,2} & \dots & p_{1,2}\, w_{2,K} & \dots & p_{m,2}\, w_{2,K} \\ \dots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \dots & p_{m,K}\, w_{K,2} & \dots & p_{1,K}\, w_{K,K} & \dots & p_{m,K}\, w_{K,K} \end{bmatrix} \tag{26}$$

where $\boldsymbol{W} = [w_{j,k}]_{1\le j,k\le K}$ is the loss matrix,

$$\mathbf{Q} = \sum_{i=1}^{K} (\boldsymbol{f}_i \boldsymbol{f}_i^\top) \otimes (\mathbf{P}\boldsymbol{f}_i) = \begin{bmatrix} \boldsymbol{p}_1 & \mathbf{0}_{m\times 1} & \dots & \mathbf{0}_{m\times 1} \\ \mathbf{0}_{m\times 1} & \boldsymbol{p}_2 & \dots & \mathbf{0}_{m\times 1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{m\times 1} & \mathbf{0}_{m\times 1} & \dots & \boldsymbol{p}_K \end{bmatrix}, \tag{27}$$

$\boldsymbol{p}_j$ is the j-th pmf viewed as the vector

$$\boldsymbol{p}_j = (p_{1,j}, p_{2,j}, \dots, p_{m,j})^\top, \tag{28}$$

and $\mathbf{P}$ is the $m \times K$ matrix containing all the pmfs:

$$\mathbf{P} = \Big[\boldsymbol{p}_1 \quad \boldsymbol{p}_2 \quad \dots \quad \boldsymbol{p}_K \Big]. \tag{29}$$

The symbol $\otimes$ denotes the Kronecker product, $\boldsymbol{f}_i$ denotes the ith unit vector of $\mathbb{R}^K$ (its ith component is one, all others zero) and $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. The symbol $\mathbf{1}_{m\times n}$, respectively $\mathbf{0}_{m\times n}$, denotes the $m \times n$ matrix with all entries one, respectively zero.

We consider also the dual minimax LP problem MLP2:

$$\max_{\boldsymbol{z}\geq 0} \quad \boldsymbol{c}^\top \boldsymbol{z} \tag{30}$$

$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{z} \leq \boldsymbol{b} \tag{31}$$

The vector $\boldsymbol{z} \in \mathbb{R}^{m(K+2)+K}$ is decomposed into four subvectors $\boldsymbol{v} \in \mathbb{R}^{mK}$, $\boldsymbol{\lambda} \in \mathbb{R}^m$, $\boldsymbol{\mu} \in \mathbb{R}^m$ and $\boldsymbol{q} \in \mathbb{R}^K$ such that

$$\boldsymbol{z} = (v_{1,1}, \dots, v_{m,K}, \lambda_1, \dots, \lambda_m, \mu_1, \dots, \mu_m, q_1, \dots, q_K) = (\boldsymbol{v}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{q}). \tag{32}$$

With these notations, the $mK + 1$ restrictions $\boldsymbol{A}\boldsymbol{z} \leq \boldsymbol{b}$ can be detailed as:

$$-v_{i,j} + \lambda_i - \mu_i - \sum_{k=1}^{K} q_k p_{i,k} w_{k,j} \leq 0 \tag{33}$$

for all $1 \leq i \leq m$, $1 \leq j \leq K$ and

$$\sum_{k=1}^{K} q_k \leq 1. \tag{34}$$

The class of feasible solutions satisfying $\boldsymbol{z} \geq 0$ and (31) is denoted $\mathcal{L}$.

3. Solution of the Linear Programming Problem

This section first describes the theoretical solution of the LP problem, then it underlines how this solution can be computed in practice.

3.1. Theoretical Solution

The following well-known lemma is recalled and adapted to our statistical problem in order to underline the meaning of each variable and to make understandable the proof of Theorem 3.

Lemma 1. *If $\boldsymbol{y} \in \mathcal{K}$ is feasible for MLP1 and $\boldsymbol{z} \in \mathcal{L}$ is feasible for MLP2,*

$$\boldsymbol{c}^\top \boldsymbol{z} \leq \boldsymbol{b}^\top \boldsymbol{y}. \tag{35}$$

Proof.

$$\begin{aligned}
\boldsymbol{c}^\top \boldsymbol{z} &= -\sum_{i=1}^{m}\sum_{j=1}^{K} v_{i,j} + \sum_{i=1}^{m} \lambda_i - \sum_{i=1}^{m} \mu_i && (36)\\
&\leq -\sum_{i=1}^{m}\sum_{j=1}^{K} \delta_{i,j} v_{i,j} + \sum_{i=1}^{m} (\lambda_i - \mu_i) \sum_{j=1}^{K} \delta_{i,j} && (37)\\
&\leq -\sum_{i=1}^{m}\sum_{j=1}^{K} \delta_{i,j} v_{i,j} + \sum_{i=1}^{m} (\lambda_i - \mu_i) \sum_{j=1}^{K} \delta_{i,j} \\
&\quad + \sum_{k=1}^{K} q_k (\gamma - \sum_{i=1}^{m}\sum_{j=1}^{K} p_{i,k} w_{k,j} \delta_{i,j}) && (38)
\end{aligned}$$

where inequality (37) comes from (12), (13), (14), (15) and inequality (38) comes from (16), (18) and $q_k \geq 0$ for all k. It follows that

$$\boldsymbol{c}^\top \boldsymbol{z} \leq \sum_{i=1}^{m}\sum_{j=1}^{K} \delta_{i,j} (-v_{i,j} + \lambda_i - \mu_i - \sum_{k=1}^{K} q_k p_{i,k} w_{k,j}) + \gamma \sum_{k=1}^{K} q_k. \tag{39}$$

The restrictions (33) and (34) yield $\boldsymbol{c}^\top \boldsymbol{z} \leq \gamma = \boldsymbol{b}^\top \boldsymbol{y}$. □

The following well-known lemma [39] is just recalled.

Lemma 2. *If $\boldsymbol{y}^* \in \mathcal{K}$ is feasible for MLP1 and $\boldsymbol{z}^* \in \mathcal{L}$ is feasible for MLP2 such that*

$$\boldsymbol{c}^\top \boldsymbol{z}^* = \boldsymbol{b}^\top \boldsymbol{y}^* \tag{40}$$

then $(\boldsymbol{y}^*, \boldsymbol{z}^*)$ constitutes a pair of optimal solutions for MLP1 and MLP2.

The following lemma establishes that MLP1 and MLP2 admit a pair of optimal solutions $(\boldsymbol{y}^*, \boldsymbol{z}^*)$.

Lemma 3. *There exist an optimal solution $\boldsymbol{y}^* \in \mathcal{K}$ for MLP1 and an optimal solution $\boldsymbol{z}^* \in \mathcal{L}$ for MLP2 such that (40) is satisfied.*

Proof. According to Theorem 1, the minimax test exists so the primal problem MLP1 has a finite optimal solution $\boldsymbol{y}^*$. Due to the strong duality property of LP [39, chap. 4], the dual problem MLP2 has also an optimal solution $\boldsymbol{z}^*$ and (40) is satisfied. □

The following theorem, based on Lemma 3, is the main result of this paper. It shows that the worst case distribution is given as the solution of the dual LP problem.

Theorem 3 (Minimax test)**.** *Let $\boldsymbol{y}^* = (\boldsymbol{\delta}^*, \gamma^*) \in \mathcal{K}$ and $\boldsymbol{z}^* = (\boldsymbol{v}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*, \boldsymbol{q}^*) \in \mathcal{L}$ be some solutions of, respectively, MLP1 and MLP2. For all $1 \leq i \leq m$ and $1 \leq j \leq K$, let $g^*_{i,j}$ be the discrete decision function:*

$$g^*_{i,j} = \sum_{k=1}^{K} q^*_k \, p_{i,k} \, w_{k,j} \tag{41}$$

where q^*_k is given in $\boldsymbol{q}^*$. The minimax test $\boldsymbol{\delta}^* = [\delta^*_{i,j}]$ given in $\boldsymbol{y}^*$ satisfies

$$\delta^*_{i,j} = \begin{cases} 1 & \text{if} \quad g^*_{i,j} < \min_{k \neq j} g^*_{i,k}, \\ 0 & \text{if} \quad g^*_{i,j} > \min_{k \neq j} g^*_{i,k}. \end{cases} \tag{42}$$

In case of equality between at least two discrete decision functions, i.e., $g^*_{i,j} = \min_{k \neq j} g^*_{i,k}$*, the minimax test is using a tie-breaking decision rule based on the values* $0 \leq \delta^*_{i,j} \leq 1$ *given by the solution of MLP1. The distribution* $\boldsymbol{q}^*$*, given by the solution of MLP2, is the worst prior distribution associated to the minimax test. Futhermore, the test* $\boldsymbol{\delta}^*$ *satisfies*

$$\max_{1 \leq k \leq K} R(\boldsymbol{\theta}_k, \boldsymbol{\delta}^*) = \gamma^*. \tag{43}$$

Proof. The proof is inspired from [22] but it contains two main technical differences: we consider i) more than two hypotheses and ii) an arbitrary loss function. The optimal solutions $\boldsymbol{y}^*$ and $\boldsymbol{z}^*$ are given in Lemma 3. The equality (40) holds if and only if equality holds everywhere in the proof of Lemma 1. Hence, the optimal pairs of solution $(\boldsymbol{y}^*, \boldsymbol{z}^*)$ satisfies

i) $\delta^*_{i,j} = 1$ when $v^*_{i,j} > 0$ from (37),

ii) $\gamma^* - \sum_{i=1}^{m} \sum_{j=1}^{K} p_{i,k} w_{k,j} \delta^*_{i,j} = 0$ when $q^*_k > 0$ from (38),

iii) $\sum_{j=1}^{K} \delta^*_{i,j} = 1$ for all $1 \leq i \leq m$ from (37),

iv) $\delta^*_{i,j} = 0$ when $-v^*_{i,j} + \lambda^*_i - \mu^*_i - \sum_{k=1}^{K} q^*_k p_{i,k} w_{k,j} < 0$ from (39),

v) $\sum_{k=1}^{K} q^*_k = 1$ when $\gamma^* > 0$ from (39).

According to (33), we have $v^*_{i,j} \geq \lambda^*_i - \mu^*_i - g^*_{i,j}$. If $\lambda^*_i - \mu^*_i + g^*_{i,j} \leq 0$, it is necessary that $v^*_{i,j} = 0$ to maximize $\boldsymbol{c}^\top \boldsymbol{z}$ since $\boldsymbol{c}^\top \boldsymbol{z}$ decreases as $v_{i,j}$ increases as shown in (36). Hence, in order to maximize $\boldsymbol{c}^\top \boldsymbol{z}$, it is clear that

$$v^*_{i,j} = \max\left\{0, \lambda^*_i - \mu^*_i - g^*_{i,j}\right\} \tag{44}$$

where $g^*_{i,j}$ is given in (41). To derive the closed form of the test, let us consider the family of points $\{v^*_{i,1}, \ldots, v^*_{i,K}\}$ for a given index i. Two cases can occur as described hereafter.

Case 1: there exists j such that $v^*_{i,j} > 0$. This involves that $\delta^*_{i,j} = 1$ according to i). Hence, for all $\ell \neq j$, $\delta^*_{i,\ell} = 0$ since $\sum_{j=1}^K \delta^*_{i,j} = 1$. This imposes that $v^*_{i,\ell} = 0$ according to i). This also involves that j is the unique index such that $v^*_{i,j} > 0$. Since $v^*_{i,\ell} = 0$, it follows from (44) that $\lambda^*_i - \mu^*_i \leq g^*_{i,\ell}$. On the contrary, $v^*_{i,j} > 0$ involves that $g^*_{i,j} < \lambda^*_i - \mu^*_i$. Hence, it follows that

$$g^*_{i,j} = \min_{1 \leq \ell \leq K} g^*_{i,\ell} < \min_{1 \leq \ell \neq j \leq K} g^*_{i,\ell}. \tag{45}$$

Case 2: $v^*_{i,j} = 0$ for all j. It is impossible that $\delta^*_{i,j} = 0$ for all j since the constraint iii) should be satisfied. Hence, according to constraint iv), there exist some coefficients $j_1, \ldots, j_t$ such that $g^*_{i,j_\ell} = \sum_{k=1}^K q^*_k p_{i,k} w_{k,j_\ell} = \lambda^*_i - \mu^*_i = \varrho^*_i$ for all ℓ. The indices j such that $\delta^*_{i,j} = 0$ satisfy $g^*_{i,j} > \varrho^*_i$ according to iv). It follows that

$$g^*_{i,j_\ell} = \min_{1 \leq j \leq K} g^*_{i,j} \tag{46}$$

for all ℓ. Hence, $\delta_{i,j} = 0$ for all $j \notin \{j_1, \ldots, j_t\}$ and $\delta_{i,j_\ell} > 0$ for all ℓ such that $\sum_{\ell=1}^t \delta^*_{i,j_\ell} = 1$. □

It should be noted that the decision function (41) of the minimax test corresponds to the decision functions of the Bayes test in Definition 3 for the worst prior distribution $\boldsymbol{q}^*$. In case of a tie-break between some discrete decision functions, as underlined in Theorem 3, the dual problem does not precise the value the decision function for the tie-break; the solution of the primal problem is then crucial. The tie-breaking decision rule is required to obtain an equalizer test as shown in the next corollary.

Corollary 1 (Equalizer test)**.** *Let* $\boldsymbol{y}^* = (\boldsymbol{\delta}^*, \gamma^*) \in \mathcal{K}$ *be a solution of MLP1 and* $\boldsymbol{q}^*$ *the worst prior associated to* $\boldsymbol{\delta}^*$*. The optimal minimax test* $\boldsymbol{\delta}^*$ *is an equalizer test almost everywhere, i.e., it satisfies* $R(\boldsymbol{\theta}_k, \boldsymbol{\delta}^*) = \gamma^*$ *for all* $q_k^* > 0$ *and* $R(\boldsymbol{\theta}_k, \boldsymbol{\delta}^*) \leq \gamma^*$ *when* $q_k^* = 0$*.*

Proof. It is a consequence of ii) in the proof of Theorem 3. □

The equalization property of the minimax test is very interesting in practice. It must be noted that, in general, the randomization is necessary to satisfy this property. Under mild assumptions, we can show that the Bayes risk $r(\boldsymbol{q}^*, \boldsymbol{\delta}^*)$ of the minimax test $\boldsymbol{\delta}^*$ can be achieved by a non-randomized Bayes test $\boldsymbol{\delta}^\star$ (see [10] for instance) but it does not involve that the non-randomized Bayes test equalizes the risk functions $R(\boldsymbol{\theta}_k, \boldsymbol{\delta}^\star)$ for all k such that $q^k > 0$.

3.2. Practical Aspects of the Solution

The minimax test can be computed in two different ways. The first way consists in computing the minimax test directly under the form $\Delta = [\delta_{i,j}]$. The test looks like a LookUp Table (LUT) where the discrete observation plays the role of the input and the output is the decision. The LUT form describes entirely the test but it is not easily interpretable in practice. The computation of the LUT $\Delta = [\delta_{i,j}]$ requires to solve explicitly the LP problem MLP1 (20)-(21). In case of a large number m of discrete values or a large number K of hypotheses, this large-scale optimization problem can be resource demanding and time consuming. Fortunately, as mentioned in Subsection 1.2, many numerical approaches and tools exist to deal with large-scale LP.

The second way consists in computing the minimax test under the form (41)-(42). This second form is simpler to use than the previous form for taking a decision and it avoids the LUT. Furthermore, this form of the test makes more interpretable the decision by highlighting the role of each element: the prior, the loss function and the pmf of the discrete observation. Unfortunately, this form needs to know the prior distribution coefficients q_k^* which have to be computed by solving the LP problem MLP2 (30)-(31). Hence, this second solution may also require to solve a large-scale optimization problem. Moreover, since the test values $\Delta = [\delta_{i,j}]$ are not explicitly computed, the resulting test is not necessarily able to deal with the tie-breaking decision rule (see the discussion after Corollary 1). Hence, the equalization property may not hold.

In practice, whatever the way to compute and to use the minimax test, it is recommended to exploit existing software libraries to solve the LP problem. The matrix form given by (23), (24) and (25) can be easily implemented. When the matrix and vectors are too large, an alternative solution is to compute directly the linear constraints (12)-(16) with an adequate programming library able to deal with large-scale optimization problem. It must be noted that the LUT $\Delta = [\delta_{i,j}]$ or the worst case distribution $\boldsymbol{q}^*$ are computed only once.

The next section presents a simple application where the samples are naturally discrete. In case of real observations, as mentioned in Subsection 1.2, the discrete samples can be obtained by quantizing the real observations. The parameter m obviously depends on the resolution of the quantizer. It is reasonable to maintain m as small as possible. From this way, the approach

proposed in this paper can be extended to any classification problems where the statistical distribution of the observations is exactly known for each possible hypothesis. The loss function does not have to satisfy strong assumptions except that it takes on only non-negative values. As discussed in [10], the loss function can play an important role in practice and it should be chosen carefully. When the worst case distribution is significantly different from the uniform distribution, it is expected that the minimax test will significantly outperform the MGLRT which remains the standard solution to solve this kind of decision problem. By definition, even if the prior distribution of the hypotheses is known, the minimax test will also outperform the Bayes test in terms of the maximum classification risk. Finally, since the minimax test is based on the worst case distribution, it is more robust to a possible prior distribution misspecification than the Bayes test.

4. Application to Noisy Channel Decoding

This section considers the problem of channel decoding which is a natural case of discrete observations.

4.1. Minimax Channel Decoding

The problem of symbol decoding is frequently encountered in coded communication. The standard solution is given by the ML when the channel is known [8]. A number of related works [43] replace the ML channel decoder by a Maximum-A-Posteriori (MAP) decoder that incorporates the statistics of the source fed to the channel encoder. When the channel is unknown, there exist universal decoders in the random coding sense [8, 44]. Specifically, the exponential decay rate of the average error probability of these

universal decoders, w.r.t. the ensemble of randomly chosen codes, is the same as that of the average error probability obtained by the optimum ML decoder. However, universality in the random-coding sense does not imply that for a specific code, a decoder can not be more powerful than the ML decoder. This example considers that the channel is known but the prior distribution of the input codewords is unknown, which is especially relevant for non-uniform sources [45, 46].

4.2. Problem Statement

We assume a given discrete memoryless channel whose transition probability matrix, $\pi = [\pi(\theta, x)]_{\theta,x\in\mathcal{A}}$, is known: $\pi(\theta, x)$, also denoted $\pi(x|\theta)$, is the probability of the channel producing the output symbol $x \in \mathcal{A}$ when the input is $\theta \in \mathcal{A}$. For simplicity, we assume that the output alphabet of the channel is the same as the input alphabet. We assume a given loss function, also called the fidelity criterion, $\Lambda : \mathcal{A}^2 \to [0, \infty)$, represented by a matrix $\Lambda = [\Lambda(\theta, x)]_{\theta,x\in\mathcal{A}}$, where $\Lambda(\theta, x)$ denotes the loss incurred by decoding the symbol θ with the symbol x. An example of such a loss function is the Hamming metric, i.e., $\Lambda(\theta, x) = 0$ when $\theta = x$, and $\Lambda(\theta, x) = 1$ otherwise.

Consider the vector channel $\pi^n(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{j=1}^n \pi(x_j|\theta_j)$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n) \in \mathcal{A}^n$ is the vector channel input and $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{A}^n$ is the observed vector channel output. A codebook $\Theta = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ of length n and rate R is a collection of $K = \lfloor 2^{nR} \rfloor$ vectors $\boldsymbol{\theta}_i = (\theta_{i,1}, \ldots, \theta_{i,n})$ in $\mathcal{A}^n$ which represent the set of messages to be transmitted across the channel. Here, $\lfloor a \rfloor$ denotes the integer part of the real value a. Upon transmitting one of the K messages $\boldsymbol{\theta}_i$, a vector $\boldsymbol{x}$ is received at the channel output, under the conditional pmf $\pi^n(\cdot|\boldsymbol{\theta}_i)$. The decoder, which observes $\boldsymbol{x} \in \mathcal{A}^n$, has to

decide which message $\boldsymbol{\theta} \in \Theta$ was truly transmitted. This is a typical instance of the hypothesis problem (1) where

$$p_{\boldsymbol{\theta}_i}(\boldsymbol{x}) = \pi^n(\boldsymbol{x}|\boldsymbol{\theta}_i) = \prod_{j=1}^{n} \pi(x_j|\theta_{i,j}) \tag{47}$$

for all $i \in \{1, \ldots, K\}$. It is assumed that the action set and the parameter set are equal, i.e., $\Psi = \Theta$. The loss function is then

$$w(\boldsymbol{\theta}, \boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \Lambda(\theta_i, x_i). \tag{48}$$

Hence, the randomized decision $\boldsymbol{\delta}(\boldsymbol{x}) \in \Theta$ can be interpreted as a n-block decoder which minimizes the average Hamming distortion between the channel input vector and the channel output decoded vector.

4.3. Numerical Results

The proposed example considers a Binary Asymmetric Channel (BAC) where $\mathcal{A} = \{0, 1\}$ and

$$\pi = \begin{bmatrix} 1 - \pi_0 & \pi_0 \\ \pi_1 & 1 - \pi_1 \end{bmatrix} \tag{49}$$

where $0 \leq \pi_0, \pi_1 \leq 1$. The case $\pi_0 = \pi_1$ corresponds the binary symmetric channel and the case $\pi_0 = 0$ to the Z-channel. Without loss of generality, we can restrict the values of the parameters π_0 and π_1 as follows:

$$0 \leq \pi_0 \leq \pi_1 \leq 1, \tag{50}$$

$$\pi_0 \leq 1 - \pi_0, \tag{51}$$

$$\pi_0 \leq 1 - \pi_1. \tag{52}$$

In fact, by permuting the columns of π, i.e., by flipping the channel outputs (change zero to one and one to zero), and/or by permuting its rows, i.e., by

flipping the channel inputs, we can easily obtain a matrix π such that $1-\pi_0$ is a maximum element of the matrix. Hence, $1-\pi_0 \geq \pi_1$ yields $\pi_0 \leq 1-\pi_1$ and $1-\pi_0 \geq 1-\pi_1$ yields $\pi_0 \leq \pi_1$. It follows that π_0 is necessary a minimum element of the matrix and all the restrictions (50), (51) and (52) are satisfied. Note that (51) can be simplified to $\pi_0 \leq 1/2$.

Without any loss of generality but to simplify the numerical experiment, we assume that the codebook is composed of all the words of length $n = 2$:

$$\Theta = \Big\{ \boldsymbol{\theta} = (\theta_1, \theta_2),\ \theta_i \in \{0,1\} \Big\} \tag{53}$$

such that $\boldsymbol{\theta}_1 = (0,0)$, $\boldsymbol{\theta}_2 = (0,1)$, $\boldsymbol{\theta}_3 = (1,0)$ and $\boldsymbol{\theta}_4 = (1,1)$. The pmf $p_{\boldsymbol{\theta}_i}(\boldsymbol{x})$ of hypothesis $\mathcal{H}_i$ is then given by (47) where the marginal pmf $\pi(x_j|\theta_{i,j})$ is

$$(1-\pi_0)^{(1-\theta_{i,j})(1-x_j)} \pi_0^{(1-\theta_{i,j})x_j} \pi_1^{\theta_{i,j}(1-x_j)} (1-\pi_1)^{\theta_{i,j}x_j}. \tag{54}$$

Without loss, we assume that $\boldsymbol{x}_j = \boldsymbol{\theta}_j$ for all $1 \leq j \leq K = 4$, i.e., $\mathcal{X} = \Theta$. Figure 1 shows $\max_{1\leq k\leq K} R(\boldsymbol{\theta}_k, \boldsymbol{\delta})$ for the ML decoder $\hat{\boldsymbol{\delta}}$ and the minimax decoder $\boldsymbol{\delta}^*$ for all the possible values of (π_0, π_1) satisfying the restrictions (50), (51) and (52). In other words, π_0 varies from 0 to 0.5 and π_1 varies from π_0 to $1-\pi_0$. This explains why the domain where $\max_{1\leq k\leq K} R(\boldsymbol{\theta}_k, \boldsymbol{\delta}) > 0$ has a triangular shape. The minimax test is calculated by solving the primal LP problem MLP1 with the simplex algorithm [39].

To compare more precisely the minimax decoder to the ML one, let us observe Figure 2 where the crossover probability $\pi_0 = 0.1$ is fixed and π_1 is varying from π_0 to $1-\pi_0$. We can see the four risk functions $R(\boldsymbol{\theta}_k, \boldsymbol{\delta})$ for all codewords $\boldsymbol{\theta}_k \in \Theta$ and for both the decoders, $\hat{\boldsymbol{\delta}}$ and $\boldsymbol{\delta}^*$. It can be noted that the minimax decoder is an equalizer test, as discussed in Corollary 1, contrary to the ML decoder. We can see that $R(\boldsymbol{\theta}_k, \hat{\boldsymbol{\delta}}) < R(\boldsymbol{\theta}_k, \boldsymbol{\delta}^*)$ for $k \in \{1,2,3\}$ but

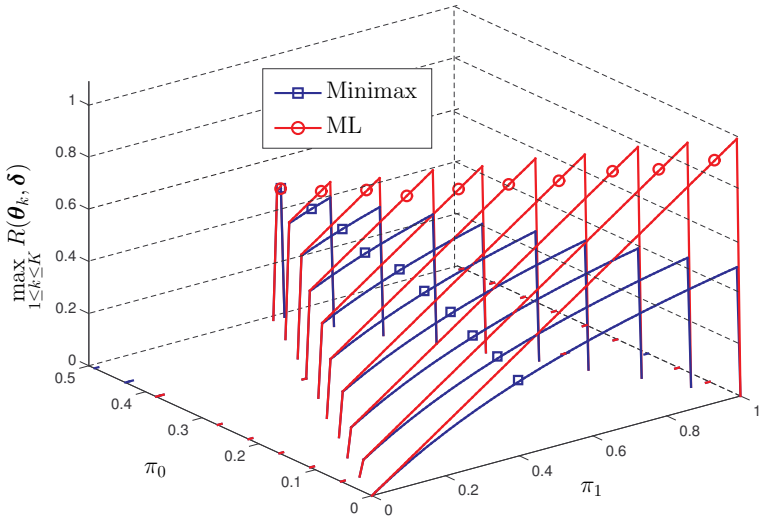

Figure 1: The maximum risk $\max_{1\leq k\leq K} R(\boldsymbol{\theta}_k, \boldsymbol{\delta})$ for the minimax decoder $\boldsymbol{\delta}^*$ (blue square) and the ML decoder $\hat{\boldsymbol{\delta}}$ (red circle) as a function of (π_0, π_1).

$R(\boldsymbol{\theta}_4, \hat{\boldsymbol{\delta}}) > R(\boldsymbol{\theta}_4, \boldsymbol{\delta}^*)$. This explains why the maximum risk is smaller for the minimax decoder. For example, when $\pi_1 = 1 - \pi_0 = 0.9$, the transmission of the symbol $1 \in \mathcal{A}$ is rarely correct. Hence, the transmission of the codeword $\boldsymbol{\theta}_4 = (1, 1)$ is often decoded as $\boldsymbol{\theta}_1 = (0, 0)$ by the ML decoder. This worst case of decoding error leads to a large risk function $R(\boldsymbol{\theta}_4, \hat{\boldsymbol{\delta}})$. Contrary to the ML decoder, the minimax decoder is automatically tuned to take into account this worst case. This tuning is done by using the worst case prior distribution $\boldsymbol{q}^*$ described in Theorem 3. It is interesting to note that the ML decoder is an equalizer decoder when $\pi_0 = \pi_1$ or $\pi_0 = 1 - \pi_1$. This behavior is explained in the following paragraph.

The worst case prior distribution is shown in Figure 3 as a function of π_1. We can see that the worst prior converges to the uniform prior as π_1 is increasing. Let us explain this behavior. Let $\pi^n_{\pi_1} = [\pi^n(\boldsymbol{x}_j|\boldsymbol{\theta}_i)]_{1\leq i,j\leq K}$ be

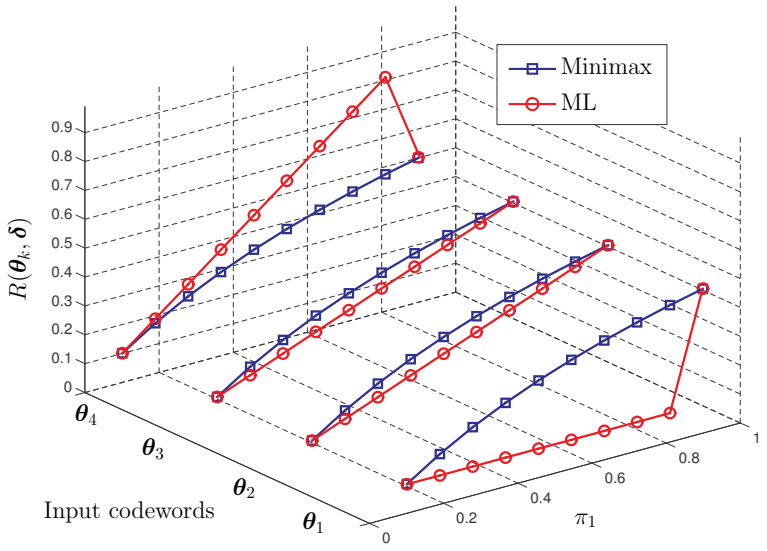

Figure 2: The decoding risks $R(\boldsymbol{\theta}_k, \boldsymbol{\delta})$ for the minimax decoder $\boldsymbol{\delta}^*$ (blue square) and the ML decoder $\hat{\boldsymbol{\delta}}$ (red circle) as a function of $\boldsymbol{\theta}_k$ and π_1 when $\pi_0 = 0.1$.

the vector channel matrix when $\pi_0 = 0.1$ and π_1 is a free parameter. When $\pi_1 = 0.1$, the matrix $\pi^n_{0.1}$, given by

$$\pi^n_{0.1} = \begin{bmatrix} 0.81 & 0.09 & 0.09 & 0.01 \\ 0.09 & 0.81 & 0.01 & 0.09 \\ 0.09 & 0.01 & 0.81 & 0.09 \\ 0.01 & 0.09 & 0.09 & 0.81 \end{bmatrix}, \tag{55}$$

is symmetric. Hence, the decoding problem is invariant with respect to the permutation of the input codewords [10]. This involves that the worst case prior is uniform. Consequently, the minimax decoder and the ML one are

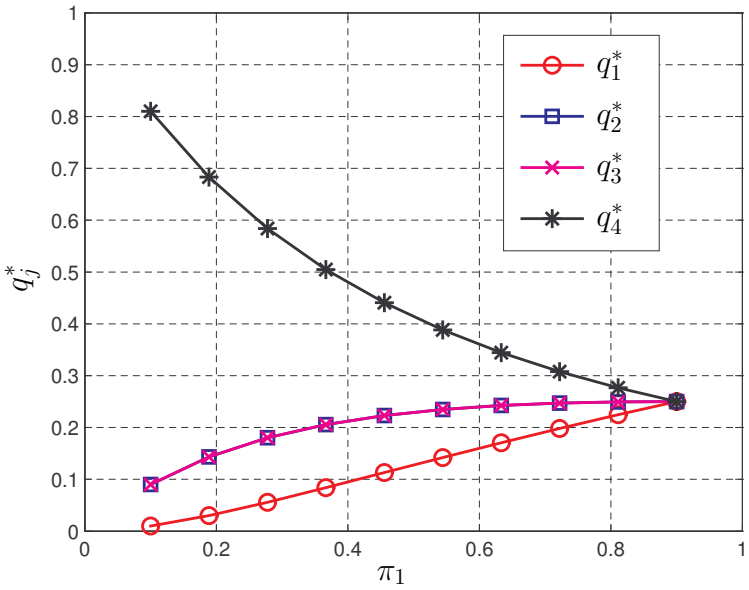

Figure 3: The worst case prior probabilities as a function of π_1 when $\pi_0 = 0.1$.

equivalent. When $\pi_1 = 0.9$, the matrix $\pi_{0.9}^n$ is given by

$$\pi_{0.9}^n = \begin{bmatrix} 0.81 & 0.09 & 0.09 & 0.01 \\ 0.81 & 0.09 & 0.09 & 0.01 \\ 0.81 & 0.09 & 0.09 & 0.01 \\ 0.81 & 0.09 & 0.09 & 0.01 \end{bmatrix}. \tag{56}$$

The rows of the matrix are the same. Hence, the decoding problem is invariant with respect to the permutation of the input codewords. This involves that the worst case prior is also uniform. When $0.1 < \pi_1 < 0.9$, the matrix $\pi_{\pi_1}^n$ has no special properties and the uniform prior is not necessarily the optimal one. In this situation, the minimax decoder clearly outperforms the ML one in the minimax sense. It is interesting to note that the worst case distribution is not unique: when $\pi_1 = 0.1$, the minimax worst prior is shown

in Figure 3 and it is clearly different from the uniform prior associated to the ML decoder which is also a worst prior in this case.

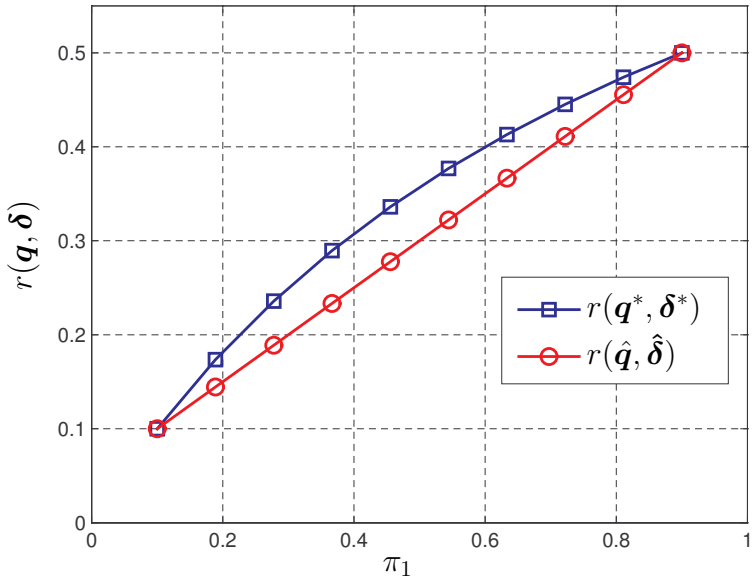

Figure 4: The Bayes risk $r(\boldsymbol{q},\boldsymbol{\delta})$ for the minimax decoder $\boldsymbol{\delta}^*$ and the ML decoder $\hat{\boldsymbol{\delta}}$ as a function of π_1 when $\pi_0 = 0.1$.

Let $\boldsymbol{q}^*_{\pi_1}$ be the worst prior when $\pi_0 = 0.1$ and π_1 is a free parameter. Let $\hat{\boldsymbol{q}}$ be the uniform prior. The mean risks, i.e., the Bayes risks $r(\boldsymbol{q}^*,\boldsymbol{\delta}^*)$ and $r(\hat{\boldsymbol{q}},\hat{\boldsymbol{\delta}})$, of both the decoders are shown in Figure 4. The Bayes risk of the ML decoder is always smaller than the Bayes risk of the minimax decoder, which is a direct consequence of the fact that the minimax test is associated to the largest Bayes risk (see Theorem 1). Hence, it should be noted that the minimax decoder minimizes the maximum decoding risk but it does not outperform the ML decoder in the mean decoding error sense.

5. Conclusion

This paper proposes an algorithm to calculate the minimax test with an arbitrary loss function between simple hypotheses by solving a linear programming problem. As a by-product, the solution of the dual problem gives the worst case distribution. The minimax test is applied to the problem of binary asymmetric channel decoding when the prior distribution of the codewords is unknown and the loss function is the Hamming metric. Contrary to the Bayes test, the minimax test equalizes the decoding risks.

References

[1] A. Jain, R. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (1) (2000) 4–37.

[2] H. Tang, B. Yin, Y. Sun, Y. Hu, 3D face recognition using local binary patterns, Signal Processing 93 (8) (2013) 2190–2198.

[3] J.-J. Xiao, A. Ribeiro, Z.-Q. Luo, G. B. Giannakis, Distributed compression-estimation using wireless sensor networks, IEEE Signal Process. Mag. 23 (4) (2006) 27–41.

[4] S. Talarico, N. A. Schmid, M. Alkhweldi, M. C. Valenti, Distributed estimation of a parametric field: Algorithms and performance analysis, IEEE Trans. Signal Process. 62 (5) (2014) 1041–1053.

[5] A. Cheddad, J. Condell, K. Curran, P. Mc Kevitt, Digital image steganography: Survey and analysis of current methods, Signal Processing 90 (3) (2010) 727–752.

[6] L. Fillatre, Adaptive steganalysis of least significant bit replacement in grayscale natural images, IEEE Trans. Signal Process. 60 (2) (2012) 556–569.

[7] R. Davarzani, S. Mozaffari, K. Yaghmaie, Scale- and rotation-invariant texture description with improved local binary pattern features, Signal Processing 111 (2015) 274–293.

[8] I. Csiszár, J. Körner, Information Theory: Coding Theorems for Discrete Memoryless Systems, Academic, New York, 1981.

[9] J. G. Proakis, Digital communications, McGraw-Hill, New York, USA, 1983.

[10] T. Ferguson, Mathematical Statistics : A Decision Theoretic Approach, Academic Press, 1967.

[11] E. L. Lehmann, J. P. Romano, Testing statistical hypotheses, 3rd Edition, Springer Texts in Statistics, Springer, New York, 2005.

[12] J. Stuller, Generalized likelihood signal resolution, IEEE Trans. Inf. Theory 21 (3) (1975) 276–282.

[13] C. W. Helstrom, Minimax detection of signals with unknown parameters, Signal Processing 27 (2) (1992) 145–159.

[14] B. Baygün, A. O. Hero, Optimal simultaneous detection and estimation under a false alarm constraint, IEEE Trans. Inf. Theory 41 (3) (1995) 688–703.

[15] L. Fillatre, I. Nikiforov, Asymptotically Uniformly Minimax Detection and Isolation in Network Monitoring, IEEE Trans. Signal Process. 60 (7) (2012) 3357–3371.

[16] G. H. Jajamovich, A. Tajer, X. Wang, Minimax-optimal hypothesis testing with estimation-dependent costs, IEEE Trans. Signal Process. 60 (12) (2012) 6151–6165.

[17] E. W. Barankin, On systems of linear equations, with applications to linear programming and the theory of tests of statistical hypotheses, University of California publications in statistics 1 (8) (1951) 161–214.

[18] G. B. Dantzig, A. Wald, On the Fundamental Lemma of Neyman and Pearson, The Annals of Mathematical Statistics 22 (1) (1951) 87–93.

[19] L. Weiss, Statistical Decision Theory, McGraw-Hill series in probability and statistics, McGraw-Hill, 1961.

[20] H. Witting, Mathematische Statistik: eine Einführung in Theorie und Methoden, Leitfäden der Angewandten Mathematik und Mechanik LAMM, B.G. Teubner, 1966.

[21] O. Krafft, H. Witting, Optimale tests und ungünstigste verteilungen, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 7 (4) (1967) 289–302.

[22] W. Schaafsma, Most stringent and maximin tests as solutions of linear programming problems, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 14 (4) (1970) 290–307.

[23] V. Baumann, Eine parameterfreie theorie der ungünstigsten verteilungen für das testen von hypothesen, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 11 (1) (1968) 41–60.

[24] O. Krafft, Programming methods in statistics and probability theory, in: J. Rosen, O. Mangasarian, K. Ritter (Eds.), Nonlinear Programming, Academic Press, 1970, pp. 425–446.

[25] O. Krafft, N. Schmitz, A symmetrical multiple decision problem and linear programming, Operations Research Verfahren 7 (1970) 126–149.

[26] W. Nelson, Minimax solution of statistical decision problems by iteration, Ann. Math. Statist. 37 (6) (1966) 1643–1657.

[27] C.-I. Chang, L. Davisson, Two iterative algorithms for finding minimax solutions, IEEE Trans. Inf. Theory 36 (1) (1990) 126–140.

[28] B. Baygün, A. O. Hero, An iterative solution to the min-max simultaneous detection and estimation problem, in: Proc. of the IEEE Workshop on Statistical Signal and Array Processing, 1996, pp. 8–11.

[29] R. F. Noubiap, W. Seidel, An efficient algorithm for constructing Γ-minimax tests for finite parameter spaces, Computational Statistics & Data Analysis 36 (2) (2001) 145–161.

[30] A. Goldenshluger, A. Juditsky, A. Nemirovski, Hypothesis testing by convex optimization, Electronic Journal of Statistics 9 (2) (2015) 1645–1712.

[31] P. J. Kempthorne, Numerical specification of discrete least favorable prior distributions, SIAM Journal on Scientific and Statistical Computing 8 (2) (1987) 171–184.

[32] J. C. Preisig, A minmax approach to adaptive matched field processing in an uncertain propagation environment, IEEE Trans. Signal Process. 42 (6) (1994) 1305–1316.

[33] Y. N. Levinbook, State estimation: A decision theoretic approach, Ph.D. thesis, University of Florida (2007).

[34] R. Gray, D. Neuhoff, Quantization, IEEE Trans. Inf. Theory 44 (1998) 2325–2384.

[35] S. S. Selvi, A. Makur, Image vector quantization with variable dimension blocks and edge preserving cost function, Signal Processing 83 (8) (2003) 1823–1826.

[36] R. Dianat, F. Marvasti, P. Azmi, S. Talebi, New vector quantization-based techniques for reducing the effect of channel noise in image transmission, Signal Processing 84 (11) (2004) 2153–2163.

[37] H. He, P. K. Varshney, Fusing censored dependent data for distributed detection, IEEE Trans. Signal Process. 63 (16) (2015) 4385–4395.

[38] Z. Li, P.-J. Chung, B. Mulgrew, Distributed target localization using quantized received signal strength, Signal Processing 134 (2017) 214–223.

[39] D. Bertsimas, J. Tsitsiklis, Introduction to linear optimization, Athena Scientific, 1997.

[40] R. K. Martin, Large scale linear and integer optimization: a unified approach, Springer United States, Boston, 1999.

[41] E. J. Anderson, P. Nash, Linear Programming in Infinite-Dimensional Spaces, Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons, 1987.

[42] J. Nocedal, S. J. Wright, Numerical Optimization, Springer, 2006.

[43] L. Bahl, J. Cocke, F. Jelinek, J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate (corresp.), IEEE Trans. Inf. Theory 20 (2) (1974) 284–287.

[44] N. Merhav, Universal decoding for arbitrary channels relative to a given class of decoding metrics, IEEE Trans. Inf. Theory 59 (9) (2013) 5566–5576.

[45] T. M. Cover, J. A. Thomas, Elements of information theory, Wiley series in telecommunications, J. Wiley & Sons, New York, Chichester, Brisbane, 1991.

[46] F. Cabarcas, R. D. Souza, J. Garcia-Frias, Turbo coding of strongly nonuniform memoryless sources with unequal energy allocation and PAM signaling, IEEE Trans. Signal Process. 54 (5) (2006) 1942–1946.