Discrete Box-Constrained Minimax Classifier for Uncertain and Imbalanced Class Proportions

Cyprien Gilet, Susana Barbosa, Lionel Fillatre

Discrete Box-Constrained Minimax Classifier for Uncertain and Imbalanced Class Proportions

Cyprien Gilet, Susana Barbosa, Lionel Fillatre

Abstract—The goal of this paper is to build a supervised classifier addressing the following difficulties which commonly appear in safety-critical applications: imbalanced datasets, uncertain class proportions, dependencies between some features, presence of both numeric and categorical features, and arbitrary loss functions provided by experts. Many works have shown that discretizing the numeric features is relevant for dealing with mixed attributes. Thus, we develop a novel minimax classifier algorithm, designed for processing discrete or discretized features, which addresses all the previously mentioned issues. The usual minimax criterion derives from the computation of the class proportions which maximize the empirical Bayes risk over the probabilistic simplex. However, it can be potentially too pessimistic when the least favorable priors appear unrealistic or its risk of error becomes too high. In this case, under the assumption that the experts are able to provide independent bounds on some class proportions, our approach takes into account these constraints to decrease the risk of error. The resulting box-constrained minimax classifier appears as a trade-off between the discrete Bayes classifier and the usual minimax classifier.

Index Terms—Minimax Classifier, Γ-Minimax Classifier, Imbalanced datasets, Uncertain Class Proportions, Discrete Bayes Classifier, Histogram Rule, Bayesian Robustness.

1 Introduction

The task of supervised classification is becoming increasingly promising in several real applications such as medical diagnosis, condition monitoring, or fraud detection. However, the context of such applications often presents the following difficulties. Firstly, the class proportions (the priors) are generally imbalanced and may evolve in time for unknown reasons. Secondly, we generally have to work with both numeric and categorical features (mixed attributes), for which many of them present dependencies. Finally, we often have to take into account a specific loss function, provided by the experts of the application domain, in order to penalize differently the class classification errors. This introduction will discuss the main aspects of our work: 1) the difficulties to learn a classifier with imbalanced datasets, 2) the robustness of classifiers to uncertain class proportions, 3) the drawbacks of the usual minimax criterion, and 4) the difficulty to process mixed attributes. Then, it will be concluded by summarizing the contributions.

1.1 Working with imbalanced datasets

The common objective in the supervised classification task is to minimize the empirical global risk of errors, based on a set of labeled learning samples [1], [2]. This global risk of classification errors is the weighted sum of the class-conditional risks with respect to the associated class proportions [3]. Hence, when the training set is imbalanced, i.e. the classes are not equally represented, most of the classifiers essentially focus on the dominating classes containing the largest number of occurrences, and tend to underestimate the least represented ones [4], [5], [6], [7]. In other words, minimizing the empirical global risk leads the classifier to minimize the class-conditional risks of the dominating classes. A minority class with just a small number of occurrences will tend to have a large class-conditional risk.

A common approach to deal with imbalanced datasets is to balance the data by resampling the training set [4], [5]. However, this approach introduces a bias since the actual state of nature remains imbalanced. As shown in this paper, this bias increases linearly as the gap between the balanced class proportions and the actual class proportions increases.

An other common approach is the cost sensitive learning [4], [5], [8], [9] which aims at optimizing the cost of class classification errors in order to counterbalance the number of occurrences of each class. However, in our context, this approach presents two drawbacks: i) it modifies the loss function provided by the experts and ii) these costs are generally difficult to tune.

In our context, a relevant approach for working with imbalanced datasets is to fit the classifier by minimizing the maximum of the class-conditional risks. The resulting decision rule is called a minimax classifier. It derives from the computation of the least favorable priors which maximize the minimum empirical global risk of error [3], [10], [11]. These least favorable priors are generally difficult to calculate as underlined in [12], [13] and [11]. A pioneering work on the minimax criterion in the field of machine learning is [14]. This work studies the generalization error of a minimax classifier but does not provide any method to compute it. In [15], the authors proposed the Minimum Error Minimax Probability Machine for the task of binary classification only, the extension to multiple classes is difficult. This method is very close to [16]. The Support Vector Machine (SVM) decision rule can also be tuned for the

- *C. Gilet and L. Fillatre are with University Côte d'Azur, CNRS, I3S laboratory, Sophia-Antipolis, France.*
  E-mail: gilet@i3s.unice.fr, and lionel.fillatre@i3s.unice.fr
- *S. Barbosa is with University Côte d'Azur, CNRS, IPMC laboratory, Sophia-Antipolis, France.*
  E-mail: sudocarmo@gmail.com

Preprint.

minimax classification [17]. The study proposed in [17] is limited to the linear classifiers (using or not a feature mapping) and to the classification problems between only two classes. In [18], the authors proposed an approach which fits a decision rule by learning the probability distribution which minimizes the worst-case of misclassification over a set of distributions centered at the empirical distribution. When the class-conditional distributions of the training set belong to a known parametric family of probability distributions, the competitive minimax approach can be an alternative solution [19]. Finally, in [20], the authors proposed a fixed-point algorithm based on generalized entropy and strict sense Bayesian loss functions. To estimate the least-favorable priors, this approach alternates a resampling step of the learning set with an evaluation step of the class-conditional risk. However, the fixed-point algorithm needs the minimax rule to be an equalizer rule. We can show that this assumption is not always satisfied when considering discrete features. Moreover, when the training dataset is too small or highly imbalanced, it is not possible to resample the dataset with respect to some priors which demand too many random samples from the classes which contain initially just a few samples.

1.2 Working with uncertain class proportions

In many application fields like medicine, the distribution of the priors can change in time because of unknown reasons (for example unknown causal effects). We generally do not know when these changes may occur. This is an important issue since the overall risk of error evolves linearly when some changes in the class proportions occur [3], [10]. Nowadays, this drawback is more and more discussed in the Machine Learning field [21], [22], and the task of considering robust classifiers with respect to uncertainty in the priors distributions is becoming necessary. In the literature, this task is generally called Bayesian Robustness [11].

In addition to its relevance for working with imbalanced datasets, the minimax classifier is also designed to address the issue of uncertain class proportions [3], [10], [11], [23]. By minimizing the maximum of the class-conditional risks, we expect the overall risk to become almost constant for any prior. It is then robust to any changes in the priors. This asserts that a minimax classifier is relevant in our context.

1.3 Possible drawback regarding the minimax criterion

Although the minimax criterion is suitable for addressing the issues regarding the class proportions, this approach appears sometimes too pessimistic as discussed in [11], [24]. This drawback occurs when the least favorable priors seem unrealistic (i.e. too far from the actual state of nature), and the global risk of error becomes too high. In order to alleviate this drawback when it occurs, a solution is to consider a set Γ of reasonable or realistic prior distributions, which leads to the Γ-minimax criterion [11]. The calculation of a Γ-minimax classifier is difficult. Currently, no algorithm exist to calculate it in a general way.

In this paper, we consider Γ as a box-constraint on the priors. The main asset of considering such a box-constraint stems from the fact that the experts of the application domain can easily and rationally build it, by providing some independent bounds on each class proportion. For example, in the medical field, it may be reasonable to bound the maximum frequency of a given disease. To our knowledge, the approach of taking into account independent bounds on the priors has not been studied yet for addressing the minimax criterion drawback. This novel decision rule is called the "Box-constrained minimax classifier".

1.4 Working with both numeric and discrete features

The task of dealing with both numeric and categorical attributes is difficult for reaching optimal results. For computing a minimax classifier, we need to well estimate the joint distribution of the input features in each class. However, in the presence of mixed attributes, and due to the curse of dimensionality [10], [25], this estimation is quite difficult. In such a case, a weakening solution would be to consider the naïve approach of estimating the marginal distribution of each feature independently. But, as previously mentioned, this hypothesis is not acceptable since we want to take into account the dependencies between the features. Hence, a reasonable approach is to discretize the numeric attributes. It allows us to constrain the joint distribution of the features to be categorical, which simplifies its estimation.

Moreover, in the literature, many works have shown that the discretization of the numeric features generally leads to accurate results [26], [27], [28], [29], [30], with strong analytic properties. For example, in the case of binary classification with respect to the $L_{0\text{-}1}$ loss function, the true error rate of the histogram rule which minimizes the risk of error on a discrete training set can be computed exactly [31], [32], [33]. In our context, all these benefits encourage us to discretize the numeric features.

1.5 Contributions

The contributions of the paper are the following. Firstly, we introduce a specific Γ-minimax classifier, called the box-constrained minimax classifier, which takes into account some independent bounds on each class proportion. Secondly, we extend the calculation of the binary histogram rule established in [31], [32], [33] to the case of multiple classes with respect to any positive loss function. We propose a theoretical study of the minimum achievable risk of error in the case of discrete features, called the discrete empirical Bayes risk, as a function of the class proportions. We show that this is a non-differentiable concave multivariate piecewise affine function over the probabilistic simplex. Thirdly, we propose a projected-subgradient-based algorithm which computes the box-constrained minimax classifier in the case of discrete features. This algorithm searches for the priors which maximize the minimum risk of errors over the box-constrained simplex. We establish the convergence of this algorithm. It must be noted that this algorithm can also be used to compute the usual unconstrained minimax classifier, which is still challenging in general. Fourthly, we show that this algorithm can be coupled with a discretization process, like the k-means algorithm, to compute the box-constrained classifier in the context of mixed attributes.

This paper is in the field of Γ-minimaxity and Bayesian robustness for supervised classification tasks. It generalizes

our opening works published in the proceedings [34], [35], and presents more contents with all the proofs and new experiments.

The paper is organized as follows. Section 2 introduces the box-constrained minimax classifier concept. Section 3 studies the discrete empirical Bayes risk. In section 4, we derive the algorithm to compute the discrete box-constrained minimax classifier. In section 5, we show how to easily and accurately discretize databases containing both numeric and categorical features with the k-means algorithm. We then provide a rigorous experiment procedure to compare our novel classifier with other usual classifiers faced with the issues of imbalanced and uncertain class proportions. These experiments are based on six real databases coming from different application fields. Section 6 concludes the paper. The appendices support the main results of the paper, including the mathematical proofs.

2 Box-Constrained Minimax Classifier

Given $K \geq 2$ the number of classes, let $\mathcal{Y} = \{1, \ldots, K\}$ be the set of class labels and $\hat{\mathcal{Y}} = \mathcal{Y}$ the predicted labels. Let $\mathcal{X}$ be the space of all feature values. Let $L : \mathcal{Y} \times \hat{\mathcal{Y}} \to [0, +\infty)$ be the loss function such that, for all $(k, l) \in \mathcal{Y} \times \hat{\mathcal{Y}}$, $L(k, l) := L_{kl}$ corresponds to the loss, or the cost, of predicting the class l whereas the real class is k. For example, the $L_{0\text{-}1}$ loss function is defined by $L_{kk} = 0$ and $L_{kl} = 1$ when $k \neq l$. Given a multiset $\{(Y_i, X_i), i \in \mathcal{I}\}$ containing a number m of labeled learning samples, the task of supervised classification [1], [2], [10] is to learn a decision rule $\delta : \mathcal{X} \to \hat{\mathcal{Y}}$ which assigns each sample $i \in \mathcal{I}$ to a class $\hat{Y}_i \in \hat{\mathcal{Y}}$ from its feature vector $X_i := [X_{i1}, \ldots, X_{id}] \in \mathcal{X}$ composed of d observed features, and such that δ minimizes the empirical risk

$$\hat{r}(\delta) = \frac{1}{m} \sum_{i \in \mathcal{I}} L(Y_i, \delta(X_i)). \tag{1}$$

As explained in [3], this risk can be written as

$$\hat{r}\left(\delta_{\hat{\pi}}\right) = \sum_{k \in \mathcal{Y}} \hat{\pi}_k \hat{R}_k\left(\delta_{\hat{\pi}}\right). \tag{2}$$

Here and in the following, $\hat{\pi} := [\hat{\pi}_1, \ldots, \hat{\pi}_K]$ corresponds to the class proportions of the training set, such that for all $k \in \mathcal{Y}$, $\hat{\pi}_k = \frac{1}{m} \sum_{i \in \mathcal{I}} \mathbb{1}_{\{Y_i = k\}}$, where $\mathbb{1}_{\{Y_i = k\}}$ denotes the indicator function of the event $Y_i = k$. Moreover, in (2), $\hat{R}_k\left(\delta_{\hat{\pi}}\right)$ corresponds to the empirical class-conditional risk associated to class k, defined by

$$\hat{R}_k\left(\delta_{\hat{\pi}}\right) := \sum_{l \in \hat{\mathcal{Y}}} L_{kl}\, \hat{\mathbb{P}}(\delta_{\hat{\pi}}(X_i) = l \mid Y_i = k). \tag{3}$$

Here, $\hat{\mathbb{P}}(\delta_{\hat{\pi}}(X_i) = l \mid Y_i = k)$ denotes the empirical probability for the classifier $\delta_{\hat{\pi}}$ to assign the class l given that the true class is k. Note that in (2) and (3), the notation $\delta_{\hat{\pi}}$ means that the decision rule δ was fitted under the priors $\hat{\pi}$. More generally, we will use the notation δ_π to denote that the decision rule δ was fitted when considering the priors π, for any π in the K-dimensional probabilistic simplex $\mathbb{S}$ defined by $\mathbb{S} := \{\pi \in [0,1]^K : \sum_{k=1}^K \pi_k = 1\}$. In the following, $\Delta := \{\delta : \mathcal{X} \to \hat{\mathcal{Y}}\}$ denotes the set of all possible classifiers.

2.1 Reminds on the Minimax classifier principle

Let $\{(Y'_i, X'_i), i \in \mathcal{I}'\}$ be the multiset containing a number m' of test samples satisfying the unknown class proportions $\pi' = [\pi'_1, \ldots, \pi'_K]$. The classifier $\delta_{\hat{\pi}}$ fitted with the samples $\{(Y_i, X_i), i \in \mathcal{I}\}$ is then used to predict the classes Y'_i of the test samples $i \in \mathcal{I}'$ from their associated features $X'_i \in \mathcal{X}$. As described in [3], the risk of misclassification with respect to the classifier $\delta_{\hat{\pi}}$ and as a function of π' is defined by

$$\hat{r}\left(\pi', \delta_{\hat{\pi}}\right) = \sum_{k \in \mathcal{Y}} \pi'_k \hat{R}_k\left(\delta_{\hat{\pi}}\right). \tag{4}$$

Fig. 1, left, illustrates the risk $\hat{r}\left(\pi', \delta_{\hat{\pi}}\right)$ for $K = 2$. In this case, it can be rewritten as

$$\hat{r}\left(\pi', \delta_{\hat{\pi}}\right) = \pi'_1 \left(\hat{R}_1\left(\delta_{\hat{\pi}}\right) - \hat{R}_2\left(\delta_{\hat{\pi}}\right)\right) + \hat{R}_2\left(\delta_{\hat{\pi}}\right). \tag{5}$$

It is then clear that $\hat{r}\left(\pi', \delta_{\hat{\pi}}\right)$ is a linear function of π'_1. It is easy to verify that the maximum value of $\hat{r}\left(\pi', \delta_{\hat{\pi}}\right)$ is $M(\delta_{\hat{\pi}}) := \max\{\hat{R}_1\left(\delta_{\hat{\pi}}\right), \hat{R}_2\left(\delta_{\hat{\pi}}\right)\}$. Since $M(\delta_{\hat{\pi}})$ is larger than $\hat{r}\left(\pi', \delta_{\hat{\pi}}\right)$, it involves that the risk of the classifier can change significantly when π' differs from $\hat{\pi}$.

More generally, for $K \geq 2$ classes, the maximum risk which can be attained by a classifier when π' is unknown is $M(\delta_{\hat{\pi}}) := \max\{\hat{R}_1\left(\delta_{\hat{\pi}}\right), \ldots, \hat{R}_K\left(\delta_{\hat{\pi}}\right)\}$. Hence, a solution to make a decision rule $\delta_{\hat{\pi}}$ robust with respect to the class proportions π' is to fit $\delta_{\hat{\pi}}$ by minimizing $M(\delta_{\hat{\pi}})$. As explained in [3], this minimax problem is equivalent to consider the following optimization problem:

$$\delta^B_{\hat{\pi}} = \operatorname*{argmin}_{\delta \in \Delta} \max_{\pi \in \mathbb{S}} \hat{r}(\pi, \delta_\pi) = \operatorname*{argmin}_{\delta \in \Delta} \max_{\pi \in \mathbb{S}} \hat{r}(\delta_\pi). \tag{6}$$

In [36], the famous Minimax Theorem establishes that

$$\min_{\delta \in \Delta} \max_{\pi \in \mathbb{S}} \hat{r}(\delta_\pi) = \max_{\pi \in \mathbb{S}} \min_{\delta \in \Delta} \hat{r}(\delta_\pi). \tag{7}$$

In our case, dealing only with discrete features involves that the set of possible classifiers Δ is finite. Looking at the proof of the Minimax theorem in [36] shows immediately that the Minimax theorem holds when Δ is finite. In the following, let us define

$$\delta^B_\pi := \arg\min_{\delta \in \Delta} \hat{r}(\delta_\pi) \tag{8}$$

the optimal Bayes classifier associated to a given prior $\pi \in \mathbb{S}$. Hence, according to (7), provided that we can calculate δ^B_π for any $\pi \in \mathbb{S}$, the optimization problem (6) is equivalent to compute the least favorable priors

$$\bar{\pi} := \arg\max_{\pi \in \mathbb{S}} \hat{r}(\delta^B_\pi), \tag{9}$$

so that the minimax classifier $\delta^B_{\hat{\pi}}$ solution of (6) is given by (8) when considering the prior (9).

2.2 Benefits of the Box-constrained minimax classifier

Sometimes, the minimax classifier appears too pessimistic in the case where the experts consider that the least favorable priors $\bar{\pi}$ are unrealistic (i.e., $\bar{\pi}$ is too far from $\hat{\pi}$), and that the global risk of errors associated to $\delta^B_{\hat{\pi}}$ is too high [11]. In such a case, a solution is to shrink the class proportions constraint, based on the knowledge, or the interest, of the experts from the application domain.

For example in Fig. 1, right, let concider that the proportions of class 1 are uncertain but bounded between

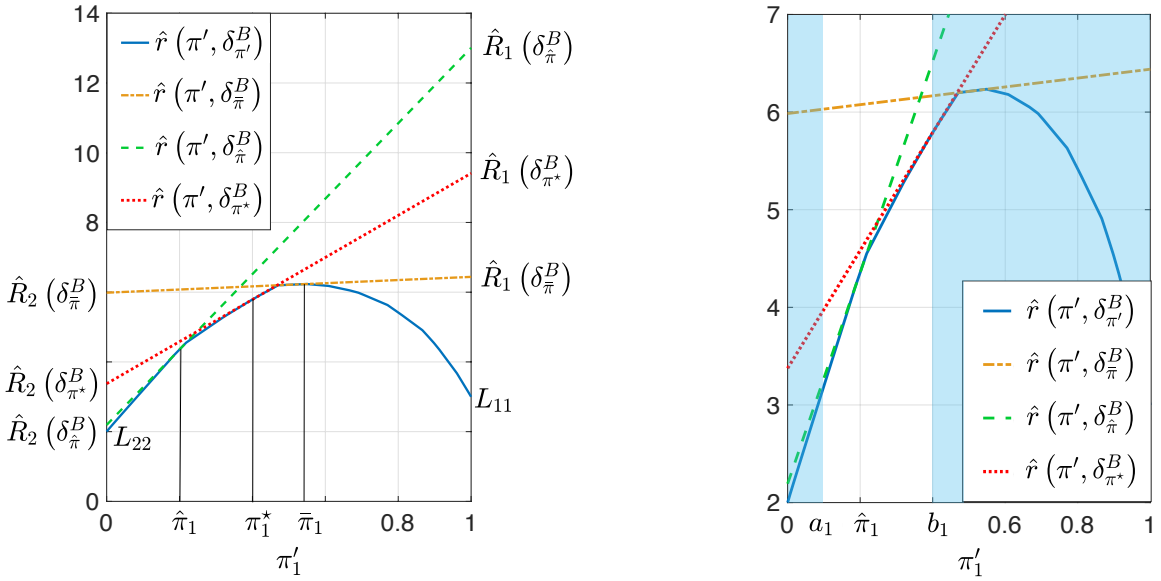

Fig. 1. Comparison between the empirical Bayes classifier $\delta^B_{\hat{\pi}}$, the minimax classifier $\delta^B_{\bar{\pi}}$ and the box-constrained minimax classifier $\delta^B_{\pi^\star}$. These results come from a synthetic dataset for which $K = 2$ classes. The generation of this dataset is detailed in Appendix A.

$a_1 = 0.1$ and $b_1 = 0.4$. If we look at the point b_1, it is clear that the classifier $\delta^B_{\bar{\pi}}$ fitted on the class proportions $\bar{\pi}_1$ of the training set is very far from the minimum empirical Bayes risk $\hat{r}\left(\pi', \delta^B_{\pi'}\right)$. The minimax classifier $\delta^B_{\bar{\pi}}$ is more robust and the box-constrained minimax classifier $\delta^B_{\pi^\star}$ has no loss. If we look now at the point a_1, the minimax classifier is disappointing but the loss of the box-constrained minimax classifier is still acceptable. In other words, the box-constrained minimax classifier seems to provide us with a reasonable trade-off between the global loss of performance, the minimization of the maximum of the class conditional risks, and the robustness to the prior change, based on the knowledge, or the interest, of the experts from the application domain. To our knowledge, the concept of box-constrained minimax classifier has not been studied yet.

More generally for $K \geq 2$ classes, in the case where we bound each class proportion π_k independently between $[a_k, b_k]_{k \in \mathcal{Y}}$, we set up the box-constraint

$$\mathbb{B} := \{\pi \in \mathbb{R}^K : \forall k \in \mathcal{Y},\ 0 \leq a_k \leq \pi_k \leq b_k \leq 1\}, \quad (10)$$

which results that we consider the box-constrained simplex

$$\mathbb{U} := \mathbb{S} \cap \mathbb{B}. \quad (11)$$

Hence, to compute the box-constrained minimax classifier with respect to $\mathbb{U}$, we therefore consider the minimax problem

$$\delta^B_{\pi^\star} = \arg\min_{\delta \in \Delta} \max_{\pi \in \mathbb{U}} \hat{r}(\delta_\pi).$$

And according to (7), provided that we can calculate δ^B_π for any $\pi \in \mathbb{U}$, this problem leads to the optimization problem

$$\pi^\star = \arg\max_{\pi \in \mathbb{U}} \hat{r}(\delta^B_\pi). \quad (12)$$

Remark 1. *It is worth noting that the minimax classifier $\delta^B_{\bar{\pi}}$ is a particular case of the box-constrained minimax classifier $\delta^B_{\pi^\star}$. Indeed, the least favorable priors $\bar{\pi}$ are still accessible when considering $\mathbb{B} = [0, 1]^K$, so that $\mathbb{U} = \mathbb{S}$ and $\pi^\star = \bar{\pi}$.*

3 Discrete Empirical Bayes Risk

Let consider that all the features are discrete, or beforehand discretized. In [32], [33], Dougherty et al established strong results concerning the discrete classification task in the case of $K = 2$ classes and when considering the $L_{0\text{-}1}$ loss function. Among these results, they calculate the histogram rule which minimizes the risk (1) on the training set. In this section, we extend the calculation of the discrete empirical Bayes classifier for $K \geq 2$ classes and when considering any positive loss function L. We then study its associated global risk of errors as a function of the priors over the simplex $\mathbb{S}$.

3.1 Empirical Bayes risk for the training set prior

For all $k \in \mathcal{Y}$, let $\mathcal{I}_k = \{i \in \mathcal{I} : Y_i = k\}$ be the set of learning samples from the class k, and $m_k = |\mathcal{I}_k|$ the number of samples in $\mathcal{I}_k$. Thus with these notations and in link with (3), we can write

$$\hat{\mathbb{P}}(\delta_{\bar{\pi}}(X_i) = l \mid Y_i = k) = \frac{1}{m_k} \sum_{i \in \mathcal{I}_k} \mathbb{1}_{\{\delta_{\bar{\pi}}(X_i) = l\}}. \quad (13)$$

Since each feature X_{ij} is discrete, it takes on a finite number of values t_j. It follows that the feature vector $X_i := [X_{i1}, \ldots, X_{id}]$ takes on a finite number of values in the finite set $\mathcal{X} = \{x_1, \ldots, x_T\}$ where $T = \prod_{j=1}^d t_j$. Each vector x_t can be interpreted as a "profile vector" which characterizes the samples. Let us note $\mathcal{T} = \{1, \ldots, T\}$ the set of indices. Let us define for all $k \in \mathcal{Y}$ and for all $t \in \mathcal{T}$,

$$\hat{p}_{kt} := \frac{1}{m_k} \sum_{i \in \mathcal{I}_k} \mathbb{1}_{\{X_i = x_t\}} \quad (14)$$

the probability estimate of observing the features profile $x_t \in \mathcal{X}$ given that the class label is k. In the context of statistical hypothesis testing theory, [37] calculates the risk of a statistical test with discrete inputs. In the next lemma, we extend this calculation to the empirical risk of a classifier $\delta \in \Delta$ with discrete features in the context of machine learning.

Lemma 1. *Given a classifier $\delta \in \Delta$, its associated empirical risk on the training dataset is given by*

$$\hat{r}\left(\delta_{\hat{\pi}}\right) = \sum_{t\in\mathcal{T}}\sum_{k\in\mathcal{Y}}\sum_{l\in\hat{\mathcal{Y}}} L_{kl}\,\hat{\pi}_k\,\hat{p}_{kt}\,\mathbb{1}_{\{\delta_{\hat{\pi}}(x_t)=l\}}. \tag{15}$$

Proof. The proof is detailed in Appendix C.1. □

According to Lemma 1, the performance of any classifier δ fitted on the learning dataset depends only on the loss function L, the probabilities $\hat{p}_{kt}$, and the priors $\hat{\pi}_k$. In this sense, the set of values $\{\hat{p}_{kt}, \hat{\pi}_k\}$ can be viewed as an exhaustive statistics of the training dataset. In the following theorem, we extend the calculation of the empirical discrete Bayes classifier [32], in our general context of $K \geq 2$ classes and when considering any positive loss function L.

Theorem 1. *The empirical Bayes classifier $\delta^B_{\hat{\pi}}$, which minimizes the empirical risk (15) over Δ, is given by*

$$\delta^B_{\hat{\pi}} : X_i \mapsto \arg\min_{l\in\hat{\mathcal{Y}}} \sum_{t\in\mathcal{T}}\sum_{k\in\mathcal{Y}} L_{kl}\,\hat{\pi}_k\,\hat{p}_{kt}\,\mathbb{1}_{\{X_i=x_t\}}. \tag{16}$$

Its associated empirical risk is $\hat{r}\left(\delta^B_{\hat{\pi}}\right) = \sum_{k\in\mathcal{Y}} \hat{\pi}_k \hat{R}_k\left(\delta^B_{\hat{\pi}}\right)$, where for all $k \in \mathcal{Y}$,

$$\hat{R}_k\left(\delta^B_{\hat{\pi}}\right) = \sum_{t\in\mathcal{T}}\sum_{l\in\hat{\mathcal{Y}}} L_{kl}\,\hat{p}_{kt}\,\mathbb{1}_{\{\lambda_{lt}=\min_{q\in\hat{\mathcal{Y}}}\lambda_{qt}\}}, \tag{17}$$

with for all $l \in \hat{\mathcal{Y}}$ and all $t \in \mathcal{T}$, $\lambda_{lt} = \sum_{k\in\mathcal{Y}} L_{kl}\,\hat{\pi}_k\,\hat{p}_{kt}$.

Proof. The proof is detailed in Appendix C.2. □

According to Theorem 1, the empirical Bayes classifier $\delta^B_{\hat{\pi}}$ outperforms, on the training set, any more advanced classifiers. Let us note that this classifier is non-naïve, it takes into account all the possible dependencies between the features since we do not make any assumptions of independence between the attributes for calculating it.

3.2 Empirical Bayes risk extended to any prior

Since we can only consider the samples from the training set, the probabilities $\hat{p}_{kt}$ defined in (14) are assumed to be estimated once for all. Indeed, the statistical estimation theory [38] has established that the estimates $\hat{p}_{kt}$ correspond to the maximum likelihood estimates of the true probabilities p_{kt} for all couples $(k,t) \in \mathcal{Y}\times\mathcal{T}$. By estimating these probabilities with the full training set, we get the best unbiased estimate with the smallest variance. This paper assumes that these class-conditional probabilities are representative of the test set. However, as explained in Section 2, we can not be confident in the class proportions estimate $\hat{\pi}_k$. They are probably biased by the data collection, and the priors can change in time for unknown reasons. Thus, the empirical Bayes risk must be viewed as a function of the class proportions.

From Theorem 1 and keeping unchanged the class-conditional probabilities $\hat{p}_{kt}$, it follows that the empirical Bayes classifier (8) associated to any prior $\pi \in \mathbb{S}$ is given by

$$\delta^B_{\pi} : X_i \mapsto \arg\min_{l\in\hat{\mathcal{Y}}} \sum_{t\in\mathcal{T}}\sum_{k\in\mathcal{Y}} L_{kl}\,\pi_k\,\hat{p}_{kt}\,\mathbb{1}_{\{X_i=x_t\}}. \tag{18}$$

Moreover, the associated minimum empirical Bayes risk $\hat{r}\left(\delta^B_{\pi}\right)$ extended to any prior $\pi \in \mathbb{S}$ is given by the function $V : \mathbb{S} \to [0,1]$ defined by

$$V(\pi) := \hat{r}\left(\delta^B_{\pi}\right) = \sum_{k\in\mathcal{Y}} \pi_k \hat{R}_k\left(\delta^B_{\pi}\right), \tag{19}$$

where for all $k \in \mathcal{Y}$,

$$\hat{R}_k\left(\delta^B_{\pi}\right) = \sum_{t\in\mathcal{T}}\sum_{l\in\hat{\mathcal{Y}}} L_{kl}\,\hat{p}_{kt}\,\mathbb{1}_{\{\lambda_{lt}=\min_{q\in\hat{\mathcal{Y}}}\lambda_{qt}\}}, \tag{20}$$

with for all $l \in \hat{\mathcal{Y}}$ and all $t \in \mathcal{T}$, $\lambda_{lt} = \sum_{k\in\mathcal{Y}} L_{kl}\,\pi_k\,\hat{p}_{kt}$. The function $V : \pi \mapsto V(\pi)$ gives the minimum value of the empirical Bayes risk when the class proportions are π and the class-conditional probabilities $\hat{p}_{kt}$ remain unchanged. In other words, a classifier can be said robust to the priors if its risk remains very close to $V(\pi)$ whatever the value of $\pi \in \mathbb{S}$.

It is well known in the literature [3], [10] that the Bayes risk, as a function of the priors, is concave over the probabilistic simplex $\mathbb{S}$. The following proposition shows that this result holds when considering the empirical Bayes risk (19). Let us note that all the results are given for $\pi \in \mathbb{S}$, but they also hold over the box-constrained probabilistic simplex $\mathbb{U}$ since $\mathbb{U} \subset \mathbb{S}$.

Proposition 1. *The empirical Bayes risk $V : \pi \mapsto V(\pi)$ is concave over the probabilistic simplex $\mathbb{S}$.*

Proof. The proof is detailed in Appendix C.3. □

Then, the following proposition and its corollary study the non-differentiability of V over $\mathbb{S}$.

Proposition 2. *The empirical Bayes risk $V : \pi \mapsto V(\pi)$ is a multivariate piecewise affine function over $\mathbb{S}$ with a finite number of pieces.*

Proof. The proof is detailed in Appendix C.4. □

Corollary 1. *If the following condition*

$$\exists\,(\pi,\pi',k) \in \mathbb{S}\times\mathbb{S}\times\mathcal{Y} : \hat{R}_k\left(\delta^B_{\pi}\right) \neq \hat{R}_k\left(\delta^B_{\pi'}\right) \tag{21}$$

is satisfied, then V is non-differentiable over the simplex $\mathbb{S}$.

Proof. The proof is detailed in Appendix C.5. □

Note that the condition (21) is most likely achievable. Otherwise, each class conditional risk would remain equal whatever the prior. And if the condition (21) is not satisfied, it results that V is affine over $\mathbb{S}$.

4 Computation of the Box-Constrained Minimax Classifier

In order to compute our box-constrained minimax classifier, according to (12) and when considering (19), our objective is to solve the following optimization problem

$$\pi^\star = \arg\max_{\pi\in\mathbb{U}} V(\pi). \tag{22}$$

Since $V : \pi \mapsto V(\pi)$ is in general non-differentiable provided that the condition (21) is satisfied, it is necessary to develop an optimization algorithm adapted to both the non-differentiability of V and the domain $\mathbb{U}$.

4.1 Optimization procedure and convergence

In order to compute the least favorable priors $\pi^\star$ which maximize V over $\mathbb{U}$ in the general case where V is non-differentiable, we propose to use a projected subgradient algorithm based on [39] and following the scheme

$$\pi^{(n+1)} = \mathrm{P}_{\mathbb{U}}\left(\pi^{(n)} + \frac{\gamma_n}{\eta_n}\, g^{(n)}\right). \tag{23}$$

In (23), at each iteration $n \geq 1$, $g^{(n)}$ denotes a subgradient of V at the point $\pi^{(n)}$, γ_n denotes the subgradient step, $\eta_n = \max\{1, \|g^{(n)}\|_2\}$, and $\mathrm{P}_{\mathbb{U}}$ denotes the exact projection onto the box-constrained simplex $\mathbb{U}$. Let us note that this algorithm also holds in the particular case where the condition (21) is not satisfied, i.e. when the function V is affine over $\mathbb{U}$. The following lemma gives a subgradient of the target function V.

Lemma 2. *Given $\pi \in \mathbb{U}$, the vector composed by all the class-conditional risks $\hat{R}\left(\delta_\pi^B\right) := \left[\hat{R}_1\left(\delta_\pi^B\right), \ldots, \hat{R}_K\left(\delta_\pi^B\right)\right]$ is a subgradient of V at the point π.*

Proof. The proof is detailed in Appendix C.6. □

In the following, we choose $g^{(n)} = \hat{R}\left(\delta^B_{\pi^{(n)}}\right)$ at each iteration $n \geq 1$ in (23). The following theorem establishes the convergence of the iterates (23) to $\pi^\star$.

Theorem 2. *When considering $g^{(n)} = \hat{R}\left(\delta^B_{\pi^{(n)}}\right)$ and any sequence of steps $(\gamma_n)_{n\geq 1}$ satisfying*

$$\inf_{n\geq 1} \gamma_n > 0, \quad \sum_{n=1}^{+\infty} \gamma_n^2 < +\infty, \quad \sum_{n=1}^{+\infty} \gamma_n = +\infty, \tag{24}$$

the sequence of iterates (23) converges strongly to a solution $\pi^\star$ of (22), whatever the initialization $\pi^{(1)} \in \mathbb{S}$.

Proof. The proof is a consequence of Theorem 1 in [39]. Here we have the strong convergence since $\pi^{(n)}$ belongs to a finite dimensional space. □

It is worth noting that when the empirical Bayes risk V is not constantly equal to zero over $\mathbb{S}$, the subgradient $\hat{R}\left(\delta^B_{\pi^\star}\right)$ at the box-constrained minimax optimum cannot vanish, otherwise the associated risk $V(\pi^\star)$ would be null too due to (19). And this would be a contradiction with the fact that $\pi^\star$ is solution of (22). Hence, in this general case, the sequence (23) is infinite, and we need to consider a stopping criterion. To this aim, we propose to follow the reasoning in [40] which leads to the following corollary.

Corollary 2. *At the iteration $N \geq 1$,*

$$\left| \max_{n\leq N} \left\{ V\left(\pi^{(n)}\right)\right\} - V(\pi^\star) \right| \;\leq\; \varphi(N),$$

with

$$\varphi(N) := \max\left\{1, \sqrt{\sum_{k=1}^{K}\left[\sum_{l=1}^{K} L_{kl}\right]^2}\right\} \frac{\rho^2 + \sum_{n=1}^{N}\gamma_n^2}{2\sum_{n=1}^{N}\gamma_n} \tag{25}$$

and where ρ is a constant satisfying $\|\pi^{(1)} - \pi^\star\|_2 \leq \rho$.

Proof. The proof is summarized in Appendix C.7. □

In practice we can choose $\rho^2 = K$ since all the proportions belong to the probabilistic simplex. Since (25) converges to 0 as $N \to \infty$, we can choose a small tolerance $\varepsilon > 0$ as a stopping criterion: we fix ε and, then, we compute $N = N_\varepsilon$ such that the bound in (25) is smaller than ε.

When considering the sequence of iterates (23), we need to compute the exact projection onto the box-constrained probabilistic simplex $\mathbb{U}$ at each iteration n. To this aim, we propose to consider the algorithm provided by [41], which computes the exact projection onto polyhedral sets in Hilbert spaces. In Appendix B, we show how to apply this projection to our box-constrained simplex $\mathbb{U}$. Let us note that in the case where we are interested in computing the minimax classifier $\delta_{\hat{\pi}}^B$, we have $\mathbb{U} = \mathbb{S}$ (see Remark 1), and we can perform the projection onto $\mathbb{S}$ using the algorithm provided by [42], or its faster version [43], for which the complexity is lower.

4.2 Box-constrained minimax classifier Algorithm

The procedure for computing the box-constrained minimax classifier $\delta^B_{\pi^\star}$ is summarized in the step by step Algorithm 1. In practice, we choose the sequence of steps $(\gamma_n)_{n\geq 1} = 1/n$ which satisfies (24). Let us note that our approach does not need to resample the training set at each iteration n. Indeed, the uses of $\pi^{(n)}$ and $\pi^\star$ is only analytic, which allows to involve the entire information provided in the training set for computing our minimax classifier.

Algorithm 1 Box-constrained minimax classifier

Input: $(Y_i, X_i)_{i\in\mathcal{I}}$, K, N.
Compute $\pi^{(1)} = \hat{\pi}$
Compute the $\hat{p}_{kt}$'s as described in (14).
$r^\star \leftarrow 0$, $\pi^\star \leftarrow \pi^{(1)}$
for $n = 1$ **to** N **do**
  for $k = 1$ **to** K **do**
    $g_k^{(n)} \leftarrow \hat{R}_k\left(\delta^B_{\pi^{(n)}}\right)$ see (20)
  end for
  $r^{(n)} = \sum_{k=1}^K \pi_k^{(n)} g_k^{(n)}$ see (19)
  if $r^{(n)} > r^\star$ **then**
    $r^\star \leftarrow r^{(n)}$, $\pi^\star \leftarrow \pi^{(n)}$
  end if
  $\gamma_n \leftarrow 1/n$, $\eta_n \leftarrow \max\{1, \|g^{(n)}\|_2\}$
  $\pi^{(n+1)} \leftarrow \mathrm{P}_{\mathbb{U}}\left(\pi^{(n)} + \gamma_n\, g^{(n)}/\eta_n\right)$
end for
Output: $r^\star$, $\pi^\star$ and $\delta^B_{\pi^\star}$ provided by (18) with $\pi = \pi^\star$.

5 Numerical experiments

For illustrating the interest of our box-constrained minimax classifier, we applied our algorithm to 6 real databases [44], [45], [46], [47], [48], [49], coming from different application domains, and presenting the previously mentioned issues. These databases present different levels of difficulty, depending on the number of classes, the class proportions, the loss function, the number of features and number of samples. A detailed description of all these databases is available in Supplementary Material. An overview of the main characteristics of each database is given in Table 1, and their associated class proportions $\hat{\pi}$ are provided in Fig. 3.

TABLE 1
Overview on each database. Among the d features, d_n corresponds to the number of numeric features. Moreover, $Quad$ denotes the quadratic loss function, such that for all $(k,l) \in \mathcal{Y} \times \hat{\mathcal{Y}}$, $L_{kl} = (k-l)^2$. Finally, Stl denotes the loss function provided by the experts of the application domain [47], such that $L_{12} = 10$, $L_{21} = 500$, and $L_{11} = L_{22} = 0$.

DATABASE	m	d	d_n	K	L
FRAMINGHAM [44]	3,658	15	8	2	$L_{0\text{-}1}$
DIABETES [45]	768	8	8	2	$L_{0\text{-}1}$
ABALONE [46]	4,177	8	7	5	$Quad$
SCANIA TRUCKS [47]	69,309	130	130	2	Stl
NASA PC3 [48]	1,563	37	36	2	$L_{0\text{-}1}$
SATELLITE [49]	5,100	36	36	2	$L_{0\text{-}1}$

5.1 Features discretization

In order to apply our algorithm, we need to discretize the numeric features. To this aim, many methods can be applied. As explained in [26], [27], we can use supervised discretization methods such as [50], [51], [52], or unsupervised methods such as the k-means algorithm [53]. For our experiments, after having compared many of these methods of discretization in terms of computation time, and their impact on the risk of misclassifiactions and on the generalization error, it resulted that the k-means algorithm was the most convenient and leaded to the most interesting results.

For each database, we therefore decided to quantize the features using the k-means algorithm with a number $T \geq K$ of centroids. In other words, each real feature vector $X_i \in \mathbb{R}^d$ composed of all the features was quantized with the index of the centroid closest to it, i.e., $Q(X_i) = j$ where $Q : \mathbb{R}^d \mapsto \{1, \ldots, T\}$ denotes the k-means quantizer and j is the index of the centroid of the cluster in which X_i belongs to. The choice of the number of centroids T is important since it has an impact on the generalization error. It was established from a 10-sub-fold cross-validation over the main training set, and such that the generalization error computed over the validation set, as a function of T, should not exceed the training error by more than $\varepsilon > 0$. An example of this procedure is given in Fig. 2.

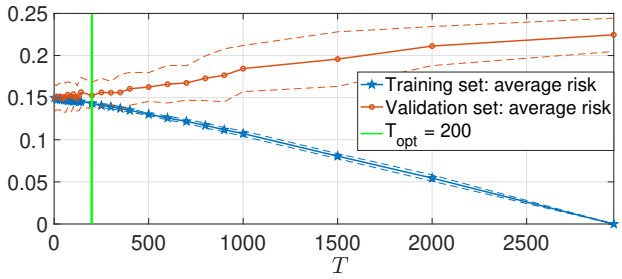

Fig. 2. Framingham database: choice of T from the training set in the first iteration of the 10-fold cross-validation procedure, and when considering $\varepsilon = 0.01$. The dashed curves show the standard-deviation around the mean of $\hat{r}\left(\hat{\pi}, \delta_{\hat{\pi}}^B\right)$.

5.2 Box-constraint generation

In practice, the Box-constraint can be established by experts by bounding independently some or all the priors. For our experiments, in order to illustrate the benefits of the box-constrained minimax classifier $\delta_{\pi^\star}^{\mathrm{B}}$ compared to the minimax classifier $\delta_{\bar{\pi}}^{\mathrm{B}}$ and the discrete Bayes classifier $\delta_{\hat{\pi}}^{\mathrm{B}}$, we consider a box-constraint $\mathbb{B}_\beta$ centered in $\hat{\pi}$, and such that, given $\beta \in [0,1]$,

$$\mathbb{B}_\beta = \left\{ \pi \in \mathbb{R}^K : \forall k \in \mathcal{Y},\ \hat{\pi}_k - \rho_\beta \leq \pi_k \leq \hat{\pi}_k + \rho_\beta \right\}, \quad (26)$$

with $\rho_\beta := \beta \left\| \hat{\pi} - \bar{\pi} \right\|_\infty = \beta \max_{k \in \mathcal{Y}} |\hat{\pi}_k - \bar{\pi}_k|$. Our box-constrained probabilistic simplex is therefore $\mathbb{U}_\beta = \mathbb{S} \cap \mathbb{B}_\beta$. Thus, when $\beta = 0$, $\mathbb{B}_0 = \{\hat{\pi}\}$, hence $\mathbb{U}_0 = \{\hat{\pi}\}$ and $\pi^\star = \hat{\pi}$. When $\beta = 1$, $\bar{\pi} \in \mathbb{B}_1$, hence $\bar{\pi} \in \mathbb{U}_1$ and $\pi^\star = \bar{\pi}$.

5.3 Procedures of the experiments

For each database, we performed a 10-fold cross-validation procedure and we applied our box-constrained minimax classifier $\delta_{\pi^\star}^{\mathrm{B}}$ with respect to the box constraint $\mathbb{B}_{0.5}$. We compared $\delta_{\pi^\star}^{\mathrm{B}}$ to the discrete Bayes classifier $\delta_{\hat{\pi}}^{\mathrm{B}}$ (16), the minimax classifier $\delta_{\bar{\pi}}^{\mathrm{B}}$, the Logistic Regression [54] denoted by $\delta_{\hat{\pi}}^{\mathrm{LR}}$, and the Random Forest [55] denoted by $\delta_{\hat{\pi}}^{\mathrm{RF}}$. We applied $\delta_{\hat{\pi}}^{\mathrm{LR}}$ and $\delta_{\hat{\pi}}^{\mathrm{RF}}$ to both the real datasets and the discretized datasets in order to evaluate the impact of the discretization. As explained in Remark 1, we computed the minimax classifier $\delta_{\bar{\pi}}^{B}$ and its associated least favorable priors $\bar{\pi}$ using our box-constrained minimax algorithm when considering $\mathbb{B} = [0,1]^K$. Let us denote $\Delta^E := \left\{\delta_{\bar{\pi}}^{\mathrm{LR}}, \delta_{\bar{\pi}}^{\mathrm{RF}}, \delta_{\hat{\pi}}^{\mathrm{B}}, \delta_{\pi^\star}^{\mathrm{B}}, \delta_{\bar{\pi}}^{\mathrm{B}}\right\} \subset \Delta$. For these experiments we evaluate each classifier on five different criteria.

We first measure the performance of each classifier by computing their global empirical risks (2) on both the training set and the test set of the cross-validation procedure.

The databases we are considering here are imbalanced, or highly imbalanced, which complicates the task of well classifying the samples from the classes with the smallest priors. For measuring the performance of each classifier $\delta \in \Delta^E$ on this difficult task, we compute $\max_{k \in \mathcal{Y}} \hat{R}_k(\delta)$ on both the training sets and the test sets, so that the smaller this criterion is, the more accurate the classifier δ appears for well classifying samples from the smallest classes.

In order to illustrate the fact that the minimax classifiers $\delta_{\pi^\star}^{\mathrm{B}}$ and $\delta_{\bar{\pi}}^{\mathrm{B}}$ aim at balancing as more as possible the class conditional risks with respect to the constraints $\mathbb{U}_\beta$ and $\mathbb{S}$, we moreover consider the criterion $\psi : \Delta \to \mathbb{R}^+$ such that

$$\psi(\delta) := \max_{k \in \mathcal{Y}} \hat{R}_k(\delta) - \min_{k \in \mathcal{Y}} \hat{R}_k(\delta), \quad (27)$$

In other words, the criterion ψ aims at measuring how equalizer a given classifier $\delta \in \Delta$ is.

In order to evaluate the robustness of each classifier when the class proportions are uncertain, i.e when π' differs from $\hat{\pi}$, we generated 1,000 random priors $\pi^{(s)}$, $s \in \{1, \ldots, 1000\}$, uniformly dispersed over the box-constrained simplex $\mathbb{U}_{0.5}$. To this aim, we uniformly generated a sequence of priors over $\mathbb{S}$ using the procedure [56], until that 1000 of them also satisfy the constraint $\mathbb{B}_{0.5}$. Then, for each repetition of the cross-validation procedure, we generated 1000 test subsets $\left\{ \left(Y_i', X_i'\right), i \in \mathcal{I}^{(s)} \right\}$ by randomly selecting samples from the full test fold set $\left\{ \left(Y_i', X_i'\right), i \in \mathcal{I}' \right\}$, and such that each test subsets satisfies one of the random priors $\pi^{(s)}$. Each fitted classifier $\delta \in \Delta^E$ was finally tested when considering all the 1000 random priors over $\mathbb{U}_{0.5}$. In order to measure the robustness of each classifier δ, we look at the boxplot of

TABLE 2
Results associated to each classifer $\delta \in \Delta^E$ and each database after the 10-folds cross-validation procedure. The notation δ (K) means that the classifier δ was performed on the discretized version of each database. The results are presented as $[\text{mean} \pm \text{std}]$. For each criterion and for each database, the green font characterizes the most efficient classifier, whereas the red font characterizes the classifier with the worst result. For each criterion, in order to get a better overview for comparing each classifier, we moreover computed the average rank of each decision rule δ based on their results associated to the 6 databases. Finally, the computing time criterion does not take into account the preprocessing task of discretizing the data. All these results were obtained on a MacBook Pro (Intel Core i7 processor [3.1 GHz], 16 GO of RAM), using MATLAB. The algorithms used to obtained these results are available at `https://github.com/cypgilet/discrete_box_constrained_minimax_classifier`.

Criteria	Databases	Classifiers						
		$\delta^{LR}_{\hat{\pi}}$	$\delta^{RF}_{\hat{\pi}}$	$\delta^{LR}_{\hat{\pi}}$ (K)	$\delta^{RF}_{\hat{\pi}}$ (K)	$\delta^{B}_{\hat{\pi}}$ (K)	$\delta^{B}_{\pi^\star}$ (K)	$\delta^{B}_{\bar{\pi}}$ (K)
Training $\hat{r}(\hat{\pi}, \delta)$	Framingham	$\mathbf{0.14 \pm 0.00}$	$\mathbf{0.14 \pm 0.01}$	0.15 ± 0.00	0.15 ± 0.00	0.15 ± 0.00	0.17 ± 0.01	0.33 ± 0.01
	Diabetes	0.22 ± 0.01	$\mathbf{0.20 \pm 0.02}$	0.35 ± 0.00	0.23 ± 0.01	0.23 ± 0.01	0.23 ± 0.01	0.25 ± 0.01
	Abalone	0.34 ± 0.01	$\mathbf{0.29 \pm 0.02}$	0.59 ± 0.01	0.34 ± 0.02	0.32 ± 0.02	0.47 ± 0.04	0.54 ± 0.06
	Scania Trucks	2.09 ± 0.04	1.28 ± 0.38	5.88 ± 0.08	4.44 ± 0.41	$\mathbf{0.95 \pm 0.04}$	2.78 ± 0.59	4.48 ± 1.48
	NASA pc3	$\mathbf{0.09 \pm 0.00}$	0.10 ± 0.00	0.10 ± 0.00	0.10 ± 0.00	$\mathbf{0.09 \pm 0.00}$	0.13 ± 0.01	0.25 ± 0.01
	Satellite	$\mathbf{0.004 \pm 0.0}$	0.006 ± 0.0	0.014 ± 0.0	0.006 ± 0.0	0.007 ± 0.0	0.016 ± 0.0	0.033 ± 0.01
Classifier Average Rank		1.83	**1.50**	4.33	2.83	2.00	3.67	4.83
Test $\hat{r}(\pi', \delta)$	Framingham	$\mathbf{0.15 \pm 0.02}$	$\mathbf{0.15 \pm 0.01}$	$\mathbf{0.15 \pm 0.01}$	$\mathbf{0.15 \pm 0.02}$	$\mathbf{0.15 \pm 0.02}$	0.20 ± 0.02	0.36 ± 0.03
	Diabetes	$\mathbf{0.22 \pm 0.05}$	0.24 ± 0.04	0.35 ± 0.05	0.30 ± 0.03	0.27 ± 0.05	0.27 ± 0.06	0.28 ± 0.05
	Abalone	$\mathbf{0.34 \pm 0.04}$	0.35 ± 0.04	0.59 ± 0.05	0.37 ± 0.05	0.36 ± 0.03	0.52 ± 0.04	0.59 ± 0.08
	Scania Trucks	2.28 ± 0.41	2.09 ± 0.42	5.88 ± 0.74	4.46 ± 0.63	$\mathbf{1.10 \pm 0.12}$	2.94 ± 0.56	4.63 ± 1.44
	NASA pc3	0.10 ± 0.02	0.10 ± 0.02	$\mathbf{0.10 \pm 0.02}$	$\mathbf{0.10 \pm 0.02}$	0.12 ± 0.02	0.18 ± 0.02	0.29 ± 0.04
	Satellite	0.007 ± 0.0	$\mathbf{0.006 \pm 0.00}$	0.014 ± 0.0	0.012 ± 0.0	0.009 ± 0.0	0.020 ± 0.0	0.035 ± 0.01
Classifier Average Rank		**1.67**	**1.67**	4.17	3.50	2.33	4.00	5.17
Training $\max_{k \in \hat{\mathcal{Y}}} \hat{R}_k(\delta)$	Framingham	0.91 ± 0.01	0.91 ± 0.04	1.00 ± 0.00	0.99 ± 0.01	0.92 ± 0.02	0.69 ± 0.04	$\mathbf{0.33 \pm 0.01}$
	Diabetes	0.42 ± 0.01	0.40 ± 0.07	1.00 ± 0.00	0.68 ± 0.12	0.45 ± 0.03	0.34 ± 0.04	$\mathbf{0.26 \pm 0.02}$
	Abalone	3.24 ± 0.16	3.47 ± 0.16	9.00 ± 0.00	3.33 ± 0.40	3.12 ± 0.42	0.74 ± 0.17	$\mathbf{0.61 \pm 0.15}$
	Scania Trucks	177 ± 3	108 ± 32	500 ± 0	376 ± 33	39 ± 6	8 ± 2	$\mathbf{5 \pm 1}$
	NASA pc3	0.79 ± 0.02	0.97 ± 0.05	1.00 ± 0.00	1.00 ± 0.00	0.88 ± 0.03	0.51 ± 0.04	$\mathbf{0.25 \pm 0.01}$
	Satellite	0.21 ± 0.02	0.37 ± 0.02	1.00 ± 0.00	0.80 ± 0.21	0.41 ± 0.10	0.13 ± 0.09	$\mathbf{0.03 \pm 0.01}$
Classifier Average Rank		3.50	4.00	6.50	5.50	3.83	2.00	**1.00**
Test $\max_{k \in \hat{\mathcal{Y}}} \hat{R}_k(\delta)$	Framingham	0.92 ± 0.04	0.94 ± 0.03	1.00 ± 0.00	1.00 ± 0.00	0.94 ± 0.04	0.78 ± 0.05	$\mathbf{0.46 \pm 0.05}$
	Diabetes	0.43 ± 0.09	0.48 ± 0.08	1.00 ± 0.00	0.73 ± 0.08	0.51 ± 0.10	0.40 ± 0.09	$\mathbf{0.32 \pm 0.06}$
	Abalone	3.18 ± 1.12	3.45 ± 1.21	9.00 ± 0.00	3.74 ± 0.93	4.08 ± 0.73	2.11 ± 0.98	$\mathbf{2.01 \pm 1.00}$
	Scania Trucks	192 ± 23	177 ± 30	500 ± 0	378 ± 37	51 ± 13	23 ± 8	$\mathbf{19 \pm 10}$
	NASA pc3	0.81 ± 0.10	0.98 ± 0.04	1.00 ± 0.00	1.00 ± 0.00	0.97 ± 0.05	0.77 ± 0.11	$\mathbf{0.48 \pm 0.12}$
	Satellite	0.33 ± 0.12	0.38 ± 0.20	1.00 ± 0.00	0.84 ± 0.15	0.52 ± 0.19	0.37 ± 0.25	$\mathbf{0.24 \pm 0.19}$
Classifier Average Rank		3.17	4.17	6.50	5.67	4.50	2.17	**1.00**
Training $\psi(\delta)$	Framingham	0.91 ± 0.01	0.90 ± 0.05	1.00 ± 0.00	0.99 ± 0.01	0.91 ± 0.03	0.61 ± 0.06	$\mathbf{0.01 \pm 0.01}$
	Diabetes	0.31 ± 0.01	0.31 ± 0.07	1.00 ± 0.00	0.62 ± 0.13	0.35 ± 0.04	0.17 ± 0.07	$\mathbf{0.02 \pm 0.01}$
	Abalone	3.12 ± 0.16	3.37 ± 0.15	9.00 ± 0.00	3.20 ± 0.40	2.96 ± 0.42	0.35 ± 0.12	$\mathbf{0.15 \pm 0.10}$
	Scania Trucks	177 ± 3	108 ± 32	500 ± 0	376 ± 33	38 ± 6	5 ± 2	$\mathbf{2 \pm 2}$
	NASA pc3	0.78 ± 0.02	0.97 ± 0.05	1.00 ± 0.00	1.00 ± 0.00	0.87 ± 0.03	0.43 ± 0.05	$\mathbf{0.01 \pm 0.01}$
	Satellite	0.21 ± 0.02	0.37 ± 0.02	1.00 ± 0.00	0.80 ± 0.21	0.41 ± 0.10	0.11 ± 0.09	$\mathbf{0.01 \pm 0.00}$
Classifier Average Rank		3.67	4.17	6.50	5.50	3.83	2.00	**1.00**
Test $\psi(\delta)$	Framingham	0.91 ± 0.04	0.93 ± 0.04	1.00 ± 0.00	1.00 ± 0.00	0.93 ± 0.05	0.68 ± 0.07	$\mathbf{0.11 \pm 0.06}$
	Diabetes	0.31 ± 0.10	0.36 ± 0.10	1.00 ± 0.00	0.66 ± 0.09	0.38 ± 0.10	0.19 ± 0.09	$\mathbf{0.06 \pm 0.06}$
	Abalone	3.06 ± 1.13	3.32 ± 1.22	9.00 ± 0.00	3.60 ± 0.93	3.90 ± 0.73	1.71 ± 0.96	$\mathbf{1.59 \pm 1.04}$
	Scania Trucks	192 ± 23	177 ± 30	500 ± 0	378 ± 37	51 ± 13	20 ± 9	$\mathbf{15 \pm 11}$
	NASA pc3	0.79 ± 0.10	0.98 ± 0.04	1.00 ± 0.00	1.00 ± 0.00	0.95 ± 0.05	0.65 ± 0.11	$\mathbf{0.22 \pm 0.13}$
	Satellite	0.32 ± 0.12	0.38 ± 0.20	1.00 ± 0.00	0.84 ± 0.15	0.52 ± 0.19	0.36 ± 0.25	$\mathbf{0.22 \pm 0.18}$
Classifier Average Rank		3.17	4.17	6.50	5.67	4.50	2.17	**1.00**
Training Time (s)	Framingham	0.20 ± 0.05	142 ± 56	0.12 ± 0.03	140 ± 24	$\mathbf{0.01 \pm 0.00}$	0.67 ± 0.14	0.35 ± 0.07
	Diabetes	0.05 ± 0.04	94 ± 44	0.04 ± 0.01	151 ± 57	$\mathbf{0.01 \pm 0.00}$	0.53 ± 0.12	0.21 ± 0.05
	Abalone	6.74 ± 1.95	156 ± 66	0.21 ± 0.05	76 ± 14	$\mathbf{0.03 \pm 0.01}$	9.21 ± 0.79	1.01 ± 0.23
	Scania Trucks	636 ± 1	288 ± 14	4.72 ± 0.18	4.36 ± 0.42	$\mathbf{0.10 \pm 0.03}$	1.25 ± 0.16	0.50 ± 0.16
	NASA pc3	3.50 ± 0.23	98 ± 42	0.06 ± 0.02	131 ± 32	$\mathbf{0.01 \pm 0.00}$	0.66 ± 0.12	0.33 ± 0.01
	Satellite	5.48 ± 2.33	101 ± 36	0.24 ± 0.06	209 ± 82	$\mathbf{0.02 \pm 0.01}$	0.96 ± 0.34	0.63 ± 0.34
Classifier Average Rank		4.50	6.33	2.50	6.17	**1.00**	4.33	3.17

$[\hat{r}(\pi^{(1)}, \delta), \ldots, \hat{r}(\pi^{(1000)}, \delta)]$, which allows to both evaluate the dispersion and the values of the risks $\hat{r}(\pi^{(s)}, \delta)$, $s \in \{1, \ldots, 1000\}$. In this experiment, we set that each one of the 1000 test subsets contains a number $m^{(s)} \approx m \cdot \min\{\hat{\pi}_1, \ldots, \hat{\pi}_K\}/10$ of observations. For ensuring statistically significant results, we need $m^{(s)}$ to be large enough, and we set that $m^{(s)}$ should be greater than 50 observations. We can therefore only consider the Framingham database and the Scania Trucks database for this criterion. And it results that each one of the 1000 test subsets contains around 54 samples for the Framingham database and 81 samples for the Scania Truck database.

5.4 Results

The class proportions $\pi^\star$ and $\bar{\pi}$ computed for each database are summarized in Fig. 3. As observed for the databases Abalone, Scania Trucks and Satellite, it is important to note that the least favorable priors $\bar{\pi}$, which are the most

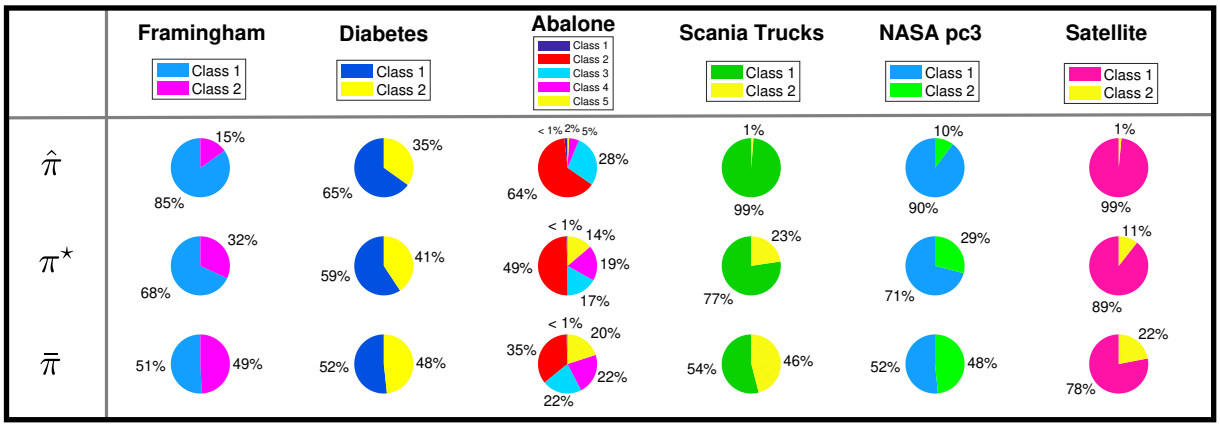

Fig. 3. Pie plots corresponding to the class proportions $\hat{\pi}$, $\pi^\star$ and $\bar{\pi}$ associated to each databases. These results correspond to the average of the computed priors at each iteration of the 10-folds cross-validation procedure.

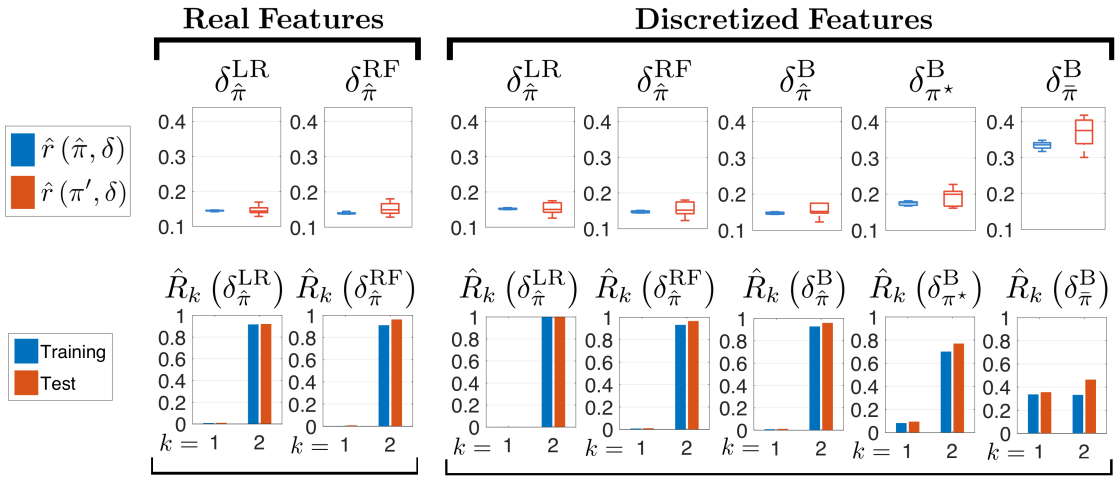

Fig. 4. Framingham database: Comparison of the risks of misclassification after the 10-fold cross-validation procedure for which the class proportions π' of the test set were similar to $\hat{\pi}$. On the top, the boxplots (training versus test) illustrate the dispersion of the global risks of misclassification. On the bottom, the barplots correspond to the average class-conditional risk associated to each classifier.

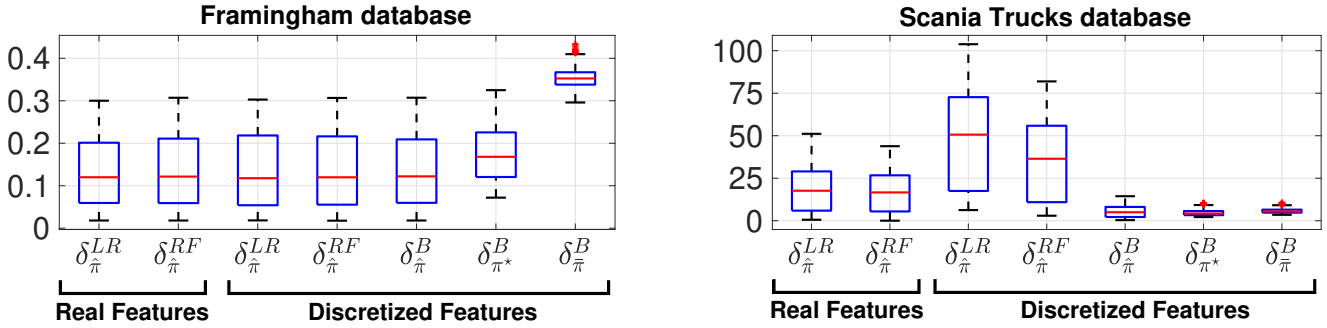

Fig. 5. Evaluation of the robustness of each classifier when $\pi' = \pi^{(s)}$ changes over $\mathbb{U}_{0.5}$. Here, $\hat{r}(\pi^{(s)}, \delta)$ corresponds to the 10-fold cross-validation average risk associated to the test set satisfying the priors $\pi^{(s)} \in \mathbb{U}_{0.5}$, $s \in \{1, \ldots, 1000\}$.

designed for equalizing the class conditional risks, are not always balanced. This illustrates the fact that the common solution mentioned in the state of the art, which aims at re-sampling the training set for satisfying the balanced class proportions $\hat{\pi} = [1/K, \ldots, 1/K]$, can be not optimal.

The results associated to each criterion previously introduced are presented in Table 2, and spotlights results on the Framingham database are illustrated in Fig. 4. In Table 2, in order to get a better overview of the results associated to each criterion, we computed the average rank of each decision rule $\delta \in \Delta^E$ based on the 6 databases.

Concerning the global risks $\hat{r}(\hat{\pi}, \delta)$ and $\hat{r}(\pi', \delta)$, we can observe that as theoretically established, the discrete Bayes classifier $\delta^{\mathrm{B}}_{\hat{\pi}}$ always gets the best training results compared to all the classifiers applied to the discretized datasets. Moreover, $\delta^{\mathrm{B}}_{\hat{\pi}}$ can well challenge the Logistic Regression and the Random Forest applied both to the real features, and sometimes $\delta^{\mathrm{B}}_{\hat{\pi}}$ can even outperform all the others classifiers

like for the Scania Trucks database. Finally, as we expected, the minimax classifier $\delta^{\mathrm{B}}_{\bar{\pi}}$ gets generally the worst results in terms of global risks, and $\delta^{\mathrm{B}}_{\pi^\star}$ appears as a trade-off between $\delta^{\mathrm{B}}_{\hat{\pi}}$ and $\delta^{\mathrm{B}}_{\bar{\pi}}$.

Now, if we look at the maximum of the class conditional risks, the minimax classifier $\delta^{\mathrm{B}}_{\bar{\pi}}$ gets the best results. This means that the task of well classifying the samples from the smallest classes is better performed using $\delta^{\mathrm{B}}_{\bar{\pi}}$ than using all the other classifiers, even those applied to the real features. Our box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$ generally holds the second rank for this criterion, and appears here again as a trade-off between $\delta^{\mathrm{B}}_{\hat{\pi}}$ and $\delta^{\mathrm{B}}_{\bar{\pi}}$. Finally, we can observe similar results concerning the criterion $\psi(\delta)$. In other words, $\delta^{\mathrm{B}}_{\pi^\star}$ allows to find a trade-off between $\delta^{\mathrm{B}}_{\hat{\pi}}$ and $\delta^{\mathrm{B}}_{\bar{\pi}}$ for equalizing the class conditional risks.

Now, if we look at Fig. 5, the minimax classifier $\delta^{B}_{\bar{\pi}}$ was the most robust when the class proportions of the 1000 test sets differed from $\hat{\pi}$ since its associated risks $\hat{r}(\pi^{(s)}, \delta)$, $s \in \{1, \ldots, 1000\}$ were the less dispersed. Concerning the Framigham database, $\delta^{B}_{\bar{\pi}}$ stays however still too pessimistic. But concerning the Scania Trucks database, $\delta^{B}_{\bar{\pi}}$ was much more satisfying than $\delta^{LR}_{\hat{\pi}}$, $\delta^{RF}_{\hat{\pi}}$ and $\delta^{B}_{\hat{\pi}}$ in terms of both dispersion and risk values. Our box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$ appears here again as a trade-off between $\delta^{B}_{\hat{\pi}}$ and $\delta^{B}_{\bar{\pi}}$ in terms of dispersions and risk values.

Finally, if we look at the processing training times, we observe that the fastest classifier was $\delta^{B}_{\hat{\pi}}$, which particularly outperformed the Logistic Regression and the Random Forest applied to both the real and discretized databases. The processing times of the two minimax classifiers $\delta^{B}_{\pi^\star}$ and $\delta^{B}_{\bar{\pi}}$ are pretty low. We can however observe that $\delta^{B}_{\bar{\pi}}$ is generally faster than $\delta^{B}_{\pi^\star}$. This difference comes from the fact that for computing $\delta^{B}_{\bar{\pi}}$, the projection onto $\mathbb{S}$ is performed using the algorithm provided by [42], whereas concerning $\delta^{B}_{\pi^\star}$, the procedure for projected onto $\mathbb{U}$ is more complex.

5.5 Impact of the Box-constraint radius

We have seen previously that the box-constrained minimax classifier $\delta^{B}_{\pi^\star}$ allows to find a trade-off between satisfying an acceptable global risk and equalizing the class-conditional risks, with respect to independent bounds on the priors. And this trade-off depends on the box constraint bounds.

For illustrating this fact on the Framingham database, we considered different box-constraints $\mathbb{B}_\beta$ by changing the radius ρ_β in (26). When β ranges from 0 to 1, we increase the radius ρ_β of $\mathbb{B}_\beta$ until that $\bar{\pi}$ belongs to $\mathbb{U}_\beta$. Hence, as illustrated in Fig. 6, the more ρ_β increases, the more equalizer $\delta^{B}_{\pi^\star}$ becomes, then the more accurate $\delta^{B}_{\pi^\star}$ becomes for well classifying the samples from the smallest classes. However, the more ρ_β increases, the more pessimistic $\delta^{B}_{\pi^\star}$ becomes since $V(\pi^\star)$ converges to $V(\bar{\pi})$.

Hence, if necessary, the experts can easily tighten or spread the box-constraint bounds in order to find an acceptable trade-off.

6 Conclusion and Discussions

This paper proposes a box-constrained minimax classifier which fits in the field of Γ-minimaxity and Bayesian robustness for supervised classification tasks. Our approach

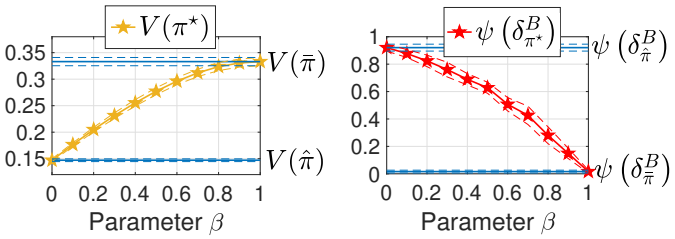

Fig. 6. Framingham database: Impact of the box-constraint radius on $\delta^{\mathrm{B}}_{\pi^\star}$ when β increases from 0 to 1 in (26), after a 10-fold cross-validation procedure. Results are presented as $\mathrm{mean} \pm \mathrm{std}$.

aims at addressing the issues of imbalanced datasets and uncertain class proportions, for multiple classes, when considering any positive loss function. The box-constraint can be conveniently defined by experts in the application field. Our method allows to find a trade-off between minimizing the maximum of the class conditional risks and satisfying an acceptable global risk of errors.

Our algorithm does not assume independence between features. To compute our minimax classifier, we beforehand need to discretize the numeric features, which allows us to calculate and model the empirical discrete non-naïve Bayes risk over the simplex. The performance of our classifier depends on the features discretization. We have seen that using the k-means algorithm leads to accurate results.

Future work will be devoted to fit our algorithm for learning a minimax regret classifier [11], [24], to study the generalization error of our minimax classifier, and to improve the computation time of the exact projection onto the box-constrained simplex, which would be preferable for dealing with databases containing a large number of classes.

Acknowledgments

The authors would like to thank Nicolas Glaichenhaus for his contributions and his help in this project, and the Provence-Alpes-Côte d'Azur region for its financial support.

References

[1] V. Vapnik, "An overview of statistical learning theory," *IEEE transactions on Neural Networks*, vol. 10 5, pp. 988–99, 1999.
[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer-Verlag New York, 2009.
[3] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. Springer-Verlag New York, 1994.
[4] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1263–1284, 2009.
[5] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, pp. 429–449, 2002.
[6] C. Elkan, "The foundations of cost-sensitive learning," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, 2001, pp. 973–978.
[7] Q. Dong, S. Gong, and X. Zhu, "Imbalanced deep learning by minority class incremental rectification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
[8] B. Ávila Pires, C. Szepesvari, and M. Ghavamzadeh, "Cost-sensitive multiclass classification risk bounds," in *Proceedings of the 30th International Conference on Machine Learning*, 2013.
[9] C. Drummond and R. C. Holte, "C4.5, class imbalance, and cost sensitivity: Why under-sampling beats oversampling," *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Datasets*, 2003.
[10] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. John Wiley and Sons, 2000.

[11] J. O. Berger, *Statistical decision theory and Bayesian analysis; 2nd ed.*, ser. Springer Series in Statistics. New York: Springer, 1985.

[12] L. Fillatre and I. Nikiforov, "Asymptotically uniformly minimax detection and isolation in network monitoring," *IEEE Transactions on Signal Processing*, vol. 60, no. 7, pp. 3357–3371, 2012.

[13] L. Fillatre, "Constructive minimax classification of discrete observations with arbitrary loss function," *Signal Processing*, vol. 141, pp. 322–330, 2017.

[14] A. Cannon, J. Howse, D. Hush, and C. Scovel, "Learning with the Neyman-Pearson and min-max criteria," *Los Alamos National Laboratory, Tech. Rep. LA-UR*, pp. 02–2951, 2002.

[15] H. Kaizhu, Y. Haiqin, K. Irwin, R. L. Michael, and L. Chan, "The minimum error minimax probability machine," *Journal of Machine Learning Research*, pp. 1253–1286, 2004.

[16] H. Kaizhu, Y. Haiqin, K. Irwin, and R. L. Michael, "Imbalanced learning with a biased minimax probability machine," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 4, pp. 913–923, Aug 2006.

[17] M. A. Davenport, R. G. Baraniuk, and C. D. Scott, "Tuning support vector machines for minimax and Neyman-Pearson classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 10, pp. 1888–1898, 2010.

[18] F. Farnia and D. Tse, "A minimax approach to supervised learning," in *Advances in NIPS 29*, 2016, pp. 4240–4248.

[19] M. Feder and N. Merhav, "Universal composite hypothesis testing: A competitive minimax approach," *IEEE Transactions on information theory*, vol. 48, no. 6, pp. 1504–1517, 2002.

[20] A. Guerrero-Curieses, R. Alaiz-Rodriguez, and J. Cid-Sueiro, "A fixed-point algorithm to minimax learning with neural networks," *IEEE Transactions on Systems, Man and Cybernetics, Part C, Applications and Reviews*, vol. 34, no. 4, pp. 383–392, Nov 2004.

[21] "Machine learning for health (workshop at neurips 2019): What makes machine learning in medicine different?" https://nips.cc/Conferences/2019/Schedule?showEvent=13162.

[22] "Machine learning for computational biology and health (tutorial at neurips 2019," https://nips.cc/Conferences/2019/Schedule?showEvent=13210.

[23] M. Yablon and J. T. Chu, "Approximations of bayes and minimax risks and the least favorable distribution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 35–40, 1982.

[24] R. Alaiz-Rodríguez, A. Guerrero-Curieses, and J. Cid-Sueiro, "Minimax regret classifier for imprecise class distributions," *Journal of Machine Learning Research*, vol. 8, pp. 103–130, Jan 2007.

[25] G. V. TRUNK, "A problem of dimensionality: A simple example," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1979.

[26] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," *International Conference on Machine Learning*, 1995.

[27] L. Peng, W. Qing, and G. Yujia, "Study on comparison of discretization methods," *IEEE, International Conference on Artificial Intelligence and Computational Intelligence*, pp. 380–384, 2009.

[28] Y. Yang and G. I. Webb, "Discretization for naive-bayes learning: managing discretization bias and variance," *Machine Learning*, vol. 74, no. 1, pp. 39–74, Jan 2009.

[29] S. García, J. Luengo, and F. Herrera, "Tutorial on practical tips of the most influential data preprocessing algorithms in data mining," *Knowledge-Based Systems*, vol. 98, pp. 1–29, 2016.

[30] J. L. Lustgarten, V. Gopalakrishnan, H. Grover, and S. Visweswaran, "Improving classification performance with discretization on biomedical datsets," *AMIA 2008 Symposium Proceedings*, pp. 445–449, 2008.

[31] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, 2nd ed. Springer-Verlag New York, 1996.

[32] U. Braga-Neto and E. R. Dougherty, "Exact performance of error estimators for discrete classifiers," *Elsevier Pattern Recognition*, vol. 38, no. 11, pp. 1799–1814, 2005.

[33] L. A. Dalton and E. R. Dougherty, "Bayesian minimum mean-square error estimation for classification error - part i: Definition and the bayesian mmse error estimator for discrete classification," *IEEE Transactions on Signal Processing*, vol. 59, pp. 115–129, 2011.

[34] C. Gilet, S. Barbosa, and L. Fillatre, "Minimax classifier with box constraint on the priors," in *Machine Learning for Health (ML4H) at NeurIPS 2019*. Proceedings of Machine Learning Research, 2019.

[35] C. Gilet and L. Fillatre, "Anomaly detection with discrete minimax classifier for imbalanced datasets or uncertain class proportions," in *World Congress on Condition Monitoring 2019*. Springer, 2019.

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

[37] M. Schlesinger and V. Hlavác, *Ten Lectures on Statistical and Structural Pattern Recognition*, 1st ed. Springer Netherlands, 2002.

[38] C. R. Rao, *Linear Statistical Inference and its Applications*. Wiley, 1973.

[39] Y. I. Alber, A. N. Iusem, and M. V. Solodov, "On the projected subgradient method for nonsmooth convex optimization in a hilbert space," *Mathematical Programming*, vol. 81, pp. 23–35, 1998.

[40] S. Boyd, L. Xiao, and A. Mutapcic, "Lecture notes: Subgradient methods, stanford university," 2003, uRL: http://web.mit.edu/6.976/www/notes/subgrad_method.pdf.

[41] K. E. Rutkowski, "Closed-form expressions for projectors onto polyhedral sets in hilbert spaces," *SIAM Journal on Optimization*, vol. 27, pp. 1758–1771, 2017.

[42] L. Condat, "Fast projection onto the simplex and the ℓ_1 ball," *Mathematical Programming*, vol. 158, no. 1, pp. 575–585, 2016.

[43] G. Perez, M. Barlaud, L. Fillatre, and J.-C. Régin, "A filtered bucket-clustering method for projection onto the simplex and the ℓ_1 ball," *Mathematical Programming*, 2019.

[44] B. University, the National Heart Lung, and B. Institute, "The framingham heart study," From 1948, downloaded data: https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset.

[45] R. S. Johannes, "Using the adap learning algorithm to forecast the onset of diabetes mellitus," *Johns Hopkins APL Technical Digest*, vol. 10, pp. 262–266, 1988.

[46] N. Warwick, J, T. L. Sellers, T. Simon, R, C. Andrew, J, F. Wes, B, and T. M. R. Laboratories., "The population biology of abalone (haliotis species) in tasmania. 1, blacklip abalone (h. rubra) from the north coast and the islands of bass strait," *Sea Fisheries Division, Technical Report*, no. 48, 1994.

[47] S. C. AB, "Aps failure at scania trucks data set," 2016. [Online]. Available: https://www.kaggle.com/uciml/aps-failure-at-scania-trucks-data-set

[48] J. Sayyad Shirabad and T. Menzies, "Pc3 software defect prediction," *The PROMISE Repository of Software Engineering Databases, School of Information Technology and Engineering, University of Ottawa, Canada.*, 2005.

[49] "Satellite database," https://www.openml.org/d/40900.

[50] R. Kerber, "Chimerge: Discretization of numeric attributes," *AAAI-92 Proceedings*, pp. 123–127, 1992.

[51] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," *IEEE, International Conference on tools with Artificial Intelligence*, 1995.

[52] A. L. Kurgan and K. J. Cios, "Caim discretization algorithm," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, pp. 145–153, 2004.

[53] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.

[54] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd ed. Chapman & Hall, 1990.

[55] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[56] W. J. Reed, "Random points in a simplex," *Pacific J. Math.*, vol. 54, no. 2, pp. 183–198, 1974.

[57] N. I. of Diabetes, Digestive, and K. Diseases, "Pima indians diabetes database," 1988. [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database

[58] T. Menzies, J. Greenwald, and A. Frank, "Data mining static code attributes to learn defect predictors," *IEEE Transactions on Software Engineering*, vol. 32, 2007.

[59] M. Shepperd, Q. Song, Z. Sun, and C. Mair, "Data quality: Some comments on the nasa software defect datasets," *IEEE Transactions on Software Engineering*, vol. 39, 2013.

[60] T. McCabe, "A complexity measure," *IEEE Transactions on Software Engineering*, vol. 32, 1976.

[61] L. Halstead, "Elements of software science," *Elsevier*, 1977.

Cyprien Gilet received the M.Sc. degree in applied mathematics (Master MIGS, Dijon, France) in 2017, and he is currently PhD student in Machine Learning at the University of Côte d'Azur in the I3S laboratory ("Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis"). The subject of his thesis is to develop a new mathematical algorithm addressing the issues that commonly appear in Machine Learning for Health, in order to help the physicians in the patients' diagnosis. His thesis is in close collaboration with the IPMC laboratory ("Institut de Pharmacologie Moléculaire et Cellulaire") in Sophia-Antipolis.

Susana Barbosa is a biologist with a postgraduate course in applied mathematics for biological sciences from Nova University in Lisbon in 2006 and a PhD degree in tropical medicine from University of Liverpool in 2012. From 2013 to 2016 she has worked as a post doctoral researcher at the University of São Paulo in Brazil. Since 2017 she is a post doctoral researcher at the Institut de Pharmacologie Moléculaire et Cellulaire in Sophia-Antipolis. Her current interests include epidemiology, molecular psychiatry, machine learning and deep learning.

Lionel Fillatre received the M.Sc. degree in decision and information engineering and the Ph.D. degree in systems optimization from the Troyes University of Technology (UTT), France, in 2001 and 2004, respectively.

From 2005 to 2007, he worked at Télécom Bretagne, Brest, France. From 2007 to 2012, he was an Associate Professor at the Systems Modelling and Dependability Laboratory, UTT. Since 2012, he is a full Professor at the University Côte d'Azur in the I3S laboratory ("Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis").

His current research interests include statistical decision theory, machine learning, deep learning, signal and image processing, and bio-inspired processing.

Discrete Box-Constrained Minimax Classifier for Uncertain and Imbalanced Class Proportions
Cyprien Gilet, Susana Barbosa, Lionel Fillatre

Appendices

APPENDIX A
SYNTHETIC DATASET FOR FIGURE 1

The results presented in Fig. 1 come from a synthetic dataset. This dataset was generated as follow. We considered $K = 2$ classes and $d = 3$ features. We generated $m = 20,000$ samples such that for each sample $i \in \mathcal{I}$, $Y_i \sim \text{Cat}(K, \hat{\pi})$ with $\hat{\pi} = [0.2, 0.8]$. The categorical distribution, which is denoted as $\text{Cat}(K, \hat{\pi})$, is a discrete distribution with support $\{1, \dots, K\}$ such that the probability of output k is $\hat{\pi}_k$. For all $j \in \{1, \dots, d\}$, we generated the features X_{ij} as follow:

$$X_{ij} \;=\; \mathbb{1}_{\{Y_i=1\}} U_i \;+\; \mathbb{1}_{\{Y_i=2\}} V_i \,,$$

with $U_i \sim \mathcal{N}(\mu_{1j}, \sigma_{1j})$ and $V_i \sim \mathcal{N}(\mu_{2j}, \sigma_{2j})$ where

$$\mu = \begin{bmatrix} 37.5 & 6.5 & 19 \\ 39 & 7 & 20 \end{bmatrix}, \quad \sigma = \begin{bmatrix} 1 & 1.5 & 1.2 \\ 2 & 0.8 & 2 \end{bmatrix}.$$

The univariate normal distribution with mean μ and standard-deviation σ is denoted $\mathcal{N}(\mu, \sigma)$. We then discretized each feature $j \in \{1, \dots, d\}$ into 6 uniform bins over $[\min_{i \in I} X_{ij}, \max_{i \in I} X_{ij}]$. Finally, we considered the following loss function L such that $L_{11} = 3$, $L_{12} = 15$, $L_{21} = 25$, $L_{22} = 2$.

APPENDIX B
PROJECTION ONTO THE CONSTRAINT $\mathbb{U}$

Let us remind that $\mathbb{U} = \mathbb{S} \cap \mathbb{B}$, where $\mathbb{B} := \{\pi \in \mathbb{R}^K : \forall k = 1, \dots, K,\ 0 \le a_k \le \pi_k \le b_k \le 1\}$. Let us define for all $i \in \{1, \dots, 2K + 2\}$

$$U_i = \begin{cases} \{\pi \in \mathbb{R}^K : \langle \pi, e_i \rangle \le b_i\} & \text{if } i \in \{1, \dots, K\} \\ \{\pi \in \mathbb{R}^K : \langle \pi, -e_{(i-K)} \rangle \le -a_i\} & \\ & \text{if } i \in \{K+1, \dots, 2K\} \\ \{\pi \in \mathbb{R}^K : \langle \pi, \mathbf{1}_K \rangle \le 1\} & \text{if } i = 2K+1 \\ \{\pi \in \mathbb{R}^K : \langle \pi, -\mathbf{1}_K \rangle \le -1\} & \text{if } i = 2K+2 \end{cases}$$

where, for all $k \in \{1, \dots, K\}$, $e_k \in \mathbb{R}^K$ is the indicator vector with 1 in coordinate k, and $\mathbf{1}_K \in \mathbb{R}^K$ is the vector fully composed of ones. We therefore can write $\mathbb{U}$ as

$$\mathbb{U} \;=\; \bigcap_{i=1}^{2K+2} U_i. \tag{28}$$

In [41], the author proposes an algorithm to compute the exact projection onto polyhedral sets in Hilbert spaces, which is the case of our box-constrained simplex (28).

APPENDIX C
PROOFS OF THE PAPER

C.1 Proof of Lemma 1

From (2), (3), (13) and (14) it follows that:

$$\begin{aligned} \hat{r}(\delta_{\hat{\pi}}) \;&=\; \sum_{k \in \mathcal{Y}} \sum_{l \in \hat{\mathcal{Y}}} L_{kl}\, \hat{\pi}_k\, \hat{\mathbb{P}}(\delta_{\hat{\pi}}(X_i) = l \mid Y_i = k) \\ &=\; \sum_{k \in \mathcal{Y}} \sum_{l \in \hat{\mathcal{Y}}} L_{kl}\, \hat{\pi}_k\, \frac{1}{m_k} \sum_{i \in \mathcal{I}_k} \mathbb{1}_{\{\delta_{\hat{\pi}}(X_i) = l\}}. \end{aligned}$$

The indicator function in the last equation can be rewritten as

$$\mathbb{1}_{\{\delta_{\hat{\pi}}(X_i) = l\}} = \sum_{t \in \mathcal{T}} \mathbb{1}_{\{\delta_{\hat{\pi}}(x_t) = l\}}\, \mathbb{1}_{\{X_i = x_t\}}.$$

Hence, we finally get:

$$\begin{aligned} \hat{r}(\delta_{\hat{\pi}}) \;&=\; \sum_{t \in \mathcal{T}} \sum_{k \in \mathcal{Y}} \sum_{l \in \hat{\mathcal{Y}}} \mathbb{1}_{\{\delta_{\hat{\pi}}(x_t) = l\}}\, L_{kl}\, \hat{\pi}_k\, \frac{1}{m_k} \sum_{i \in \mathcal{I}_k} \mathbb{1}_{\{X_i = x_t\}} \\ &=\; \sum_{t \in \mathcal{T}} \sum_{k \in \mathcal{Y}} \sum_{l \in \hat{\mathcal{Y}}} \mathbb{1}_{\{\delta_{\hat{\pi}}(x_t) = l\}}\, L_{kl}\, \hat{\pi}_k\, \hat{p}_{kt}\,. \end{aligned}$$

□

C.2 Proof of Theorem 1

Let $\delta \in \Delta$, let $t \in \mathcal{T}$, and let $h_t = \underset{l \in \hat{\mathcal{Y}}}{\text{argmin}} \sum_{k \in \mathcal{Y}} L_{kl}\, \hat{\pi}_k\, \hat{p}_{kt}$,

$$\begin{aligned} &\sum_{l \in \hat{\mathcal{Y}}} \sum_{k \in \mathcal{Y}} L_{kl}\, \hat{\pi}_k\, \hat{p}_{kt}\, \mathbb{1}_{\{\delta(x_t) = l\}} \\ &\qquad\qquad \ge\; \sum_{k \in \mathcal{Y}} L_{kh_t}\, \hat{\pi}_k\, \hat{p}_{kt} \sum_{l \in \hat{\mathcal{Y}}} \mathbb{1}_{\{\delta(x_t) = l\}} \\ &\qquad\qquad \ge\; \sum_{k \in \mathcal{Y}} L_{kh_t}\, \hat{\pi}_k\, \hat{p}_{kt}. \end{aligned}$$

The last inequality can be rewritten as

$$\begin{aligned} &\sum_{l \in \hat{\mathcal{Y}}} \sum_{k \in \mathcal{Y}} L_{kl}\, \hat{\pi}_k\, \hat{p}_{kt}\, \mathbb{1}_{\{\delta(x_t) = l\}} \\ &\ge \sum_{l \in \hat{\mathcal{Y}}} \sum_{k \in \mathcal{Y}} L_{kl} \hat{\pi}_k \hat{p}_{kt} \mathbb{1}_{\{\sum_{k \in \mathcal{Y}} L_{kl}\, \hat{\pi}_k \hat{p}_{kt} = \min_{q \in \hat{\mathcal{Y}}} \sum_{k \in \mathcal{Y}} L_{kq} \hat{\pi}_k \hat{p}_{kt}\}} \\ &\ge \sum_{l \in \hat{\mathcal{Y}}} \sum_{k \in \mathcal{Y}} L_{kl}\, \hat{\pi}_k\, \hat{p}_{kt}\, \mathbb{1}_{\{\lambda_{lt} = \min_{q \in \hat{\mathcal{Y}}} \lambda_{qt}\}}, \end{aligned}$$

where for all $(q, t) \in \hat{\mathcal{Y}} \times \mathcal{T}$, $\lambda_{qt} = \sum_{k \in \mathcal{Y}} L_{kq}\, \hat{\pi}_k\, \hat{p}_{kt}$. Hence, from (15), and for all $\delta \in \Delta$, we get

$$\hat{r}(\delta_{\hat{\pi}}) \;\ge\; \sum_{t \in \mathcal{T}} \sum_{l \in \hat{\mathcal{Y}}} \sum_{k \in \mathcal{Y}} L_{kl}\, \hat{\pi}_k\, \hat{p}_{kt}\, \mathbb{1}_{\{\lambda_{lt} = \min_{q \in \hat{\mathcal{Y}}} \lambda_{qt}\}}. \tag{29}$$

It follows that (29) is a lower bound of the empirical Bayes risk. It is straightforward to verify that the decision rule

(16) achieves the lower bound (29). Hence, the classifier (16) minimizes (15), and its associated empirical Bayes risk is:

$$\hat{r}\left(\delta^B_{\hat{\pi}}\right) \;=\; \sum_{t\in\mathcal{T}}\sum_{l\in\hat{\mathcal{Y}}}\sum_{k\in\mathcal{Y}} L_{kl}\,\hat{\pi}_k\,\hat{p}_{kt}\,\mathbb{1}_{\{\lambda_{lt}=\min_{q\in\hat{\mathcal{Y}}}\lambda_{qt}\}}. \quad (30)$$

From (2) and (30), we finally identify the empirical class-conditional risk of class $k\in\mathcal{Y}$ as (17). □

C.3 Proof of Proposition 1

Let $\alpha\in[0,1]$ and let consider the class proportions $\pi,\pi',\pi''\in\mathbb{S}$ such that $\pi''=\alpha\pi+(1-\alpha)\pi'$. Thus,

$$\begin{aligned}
V(\pi'') \;&=\; \hat{r}\left(\delta^B_{\pi''}\right) \;=\; \sum_{k\in\mathcal{Y}}\pi''_k\hat{R}_k\left(\delta^B_{\pi''}\right)\\
&=\; \alpha\sum_{k\in\mathcal{Y}}\pi_k\hat{R}_k\left(\delta^B_{\pi''}\right)+(1-\alpha)\sum_{k\in\mathcal{Y}}\pi'_k\hat{R}_k\left(\delta^B_{\pi''}\right)\\
&=\; \alpha\,\hat{r}\left(\pi,\delta^B_{\pi''}\right)+(1-\alpha)\,\hat{r}\left(\pi',\delta^B_{\pi''}\right)\\
&\geq\; \alpha\,\hat{r}\left(\pi,\delta^B_{\pi}\right)+(1-\alpha)\,\hat{r}\left(\pi',\delta^B_{\pi'}\right)\\
&\geq\; \alpha\,\hat{r}\left(\delta^B_{\pi}\right)+(1-\alpha)\,\hat{r}\left(\delta^B_{\pi'}\right)\\
&\geq\; \alpha\,V(\pi)+(1-\alpha)\,V(\pi').
\end{aligned}$$

This shows that V is concave over $\mathbb{S}$. □

C.4 Proof of Proposition 2

Let us consider the equivalence relation $\mathcal{R}$ over the simplex $\mathbb{S}$ such that for all $(\pi,\pi')\in\mathbb{S}\times\mathbb{S}$,

$$\begin{aligned}
\pi\mathcal{R}\pi' \quad\Longleftrightarrow\quad & \forall(l,t)\in\hat{\mathcal{Y}}\times\mathcal{T},\\
& \mathbb{1}_{\{\lambda_{lt}=\min_{q\in\mathcal{Y}}\lambda_{qt}\}}=\mathbb{1}_{\left\{\lambda'_{lt}=\min_{q\in\mathcal{Y}}\lambda'_{qt}\right\}},
\end{aligned}$$

with

$$\lambda_{lt}=\sum_{k\in\mathcal{Y}}L_{kl}\,\pi_k\,\hat{p}_{kt}\quad\text{and}\quad\lambda'_{lt}=\sum_{k\in\mathcal{Y}}L_{kl}\,\pi'_k\,\hat{p}_{kt}.$$

Let $\pi\in\mathbb{S}$, and let $[\pi]\subset\mathbb{S}$ denote the equivalence class to which π belongs. Thus, according to (20), for all $k\in\mathcal{Y}$, there exists a constant $\alpha_k\geq 0$ such that for all $\pi'\in[\pi]$, $\hat{R}_k\left(\delta^B_{\pi'}\right)=\alpha_k$. Then, by considering $\alpha=[\alpha_1,\ldots,\alpha_K]$ and according to (19) we have for all $\pi'\in[\pi]$, $V(\pi')=\sum_{k=1}^K\pi'_k\alpha_k$, which shows that V is affine over $[\pi]$. Since the set of equivalence classes is a partition of the simplex $\mathbb{S}$, V is piecewise affine over $\mathbb{S}$.

Moreover, we can show that $\pi'\in[\pi]$ if and only if $\delta^B_{\pi'}(x_t)=\delta^B_{\pi}(x_t)$ for all $t\in\mathcal{T}$. Thus, by denoting $\mathbb{S}/\mathcal{R}$ the quotient set of $\mathbb{S}$, there exists an injection $\varphi:\mathbb{S}/\mathcal{R}\to\mathcal{Y}^{\mathcal{X}}$. Hence $|\mathbb{S}/\mathcal{R}|\leq|\mathcal{Y}|^{|\mathcal{X}|}=K^T$. It follows that the number of pieces composing V is finite. □

C.5 Proof of Corollary 1

Let us suppose that there exist $\pi,\pi'\in\mathbb{S}$ and $k\in\mathcal{Y}$ such that $\hat{R}_k\left(\delta^B_{\pi}\right)\neq\hat{R}_k\left(\delta^B_{\pi'}\right)$. Then, from the proof of Proposition 2, V is at least composed of two affine pieces since it is impossible to have a single equivalence class. Hence, V is non-differentiable over the intersections of these pieces. □

C.6 Proof of Lemma 2

Let us remind that, for a concave function $f:\mathbb{R}^K\to\mathbb{R}$, g is a subgradient of f at point $u\in\mathbb{R}^K$ if g satisfies $f(v)\leq f(u)+\langle v-u,g\rangle$ for all $v\in\mathbb{R}^K$. Here, $\langle a,b\rangle$ denotes the dot product between the vectors a and b. In our case, given $\pi\in\mathbb{U}$, let consider $\pi'\in\mathbb{U}$. Denoting $\hat{R}\left(\delta^B_{\pi}\right)$ the vector $\hat{R}\left(\delta^B_{\pi}\right):=\left[\hat{R}_1\left(\delta^B_{\pi}\right),\ldots,\hat{R}_K\left(\delta^B_{\pi}\right)\right]$ of all class-conditional risks, we get:

$$\begin{aligned}
&V(\pi)+\left\langle\pi'-\pi,\hat{R}\left(\delta^B_{\pi}\right)\right\rangle\\
&\qquad=\;\sum_{k\in\mathcal{Y}}\pi_k\,\hat{R}_k\left(\delta^B_{\pi}\right)+\sum_{k\in\mathcal{Y}}(\pi'_k-\pi_k)\,\hat{R}_k\left(\delta^B_{\pi}\right)\\
&\qquad=\;\sum_{k\in\mathcal{Y}}\pi'_k\hat{R}_k\left(\delta^B_{\pi}\right)\\
&\qquad\geq\;\hat{r}\left(\pi',\delta^B_{\pi'}\right)=\hat{r}\left(\delta^B_{\pi'}\right)=V(\pi').
\end{aligned}$$

This inequality holds for any $\pi'\in\mathbb{U}$, hence the result. □

C.7 Proof of Corollary 2

Following the reasoning in [40] when considering the subgradient definition associated to a concave function, we can show that at the iteration $N\geq 1$

$$\begin{aligned}
&V(\pi^\star)-\max_{n\leq N}\left\{V\left(\pi^{(n)}\right)\right\}\\
&\qquad\leq\;\frac{\left\|\pi^{(1)}-\pi^\star\right\|_2^2+\sum_{n=1}^N\frac{\gamma_n^2}{\eta_n^2}\left\|g^{(n)}\right\|_2^2}{2\sum_{n=1}^N\frac{\gamma_n}{\eta_n}}. \quad (31)
\end{aligned}$$

Since $\eta_n=\max\left\{1,\left\|g^{(n)}\right\|_2\right\}$, we can moreover show that

$$\sum_{n=1}^N\frac{\gamma_n^2}{\eta_n^2}\left\|g^{(n)}\right\|_2^2\;\leq\;\sum_{n=1}^N\gamma_n^2. \quad (32)$$

Since at each iteration we choose $g^{(n)}=\hat{R}\left(\delta^B_{\pi^{(n)}}\right)$, we have

$$\begin{aligned}
\left\|g^{(n)}\right\|_2\;&=\;\sqrt{\sum_{k=1}^K\left[\hat{R}_k\left(\delta^B_{\pi^{(n)}}\right)\right]^2}\\
&=\;\sqrt{\sum_{k=1}^K\left[\sum_{l=1}^K L_{kl}\,\hat{\mathbb{P}}\left(\delta^B_{\pi^{(n)}}(X_i)=l\mid Y_i=k\right)\right]^2}\\
&\leq\;\sqrt{\sum_{k=1}^K\left[\sum_{l=1}^K L_{kl}\right]^2}.
\end{aligned}$$

It follows that for all $n\in\{1,\ldots,N\}$, $\eta_n\leq\max\{1,h(L)\}$, with

$$h(L):=\sqrt{\sum_{k=1}^K\left[\sum_{l=1}^K L_{kl}\right]^2}.$$

Hence we have,

$$\sum_{n=1}^N\frac{\gamma_n}{\eta_n}\;\geq\;\frac{1}{\max\{1,h(L)\}}\sum_{n=1}^N\gamma_n. \quad (33)$$

From (31), (32) and (33), we finally get (25). □

Discrete Box-Constrained Minimax Classifier for Uncertain and Imbalanced Class Proportions
Cyprien Gilet, Susana Barbosa, Lionel Fillatre

Supplementary Material

Databases descriptions

Framingham Heart database: This database comes from the Framingham Heart study [44], and contains the clinical observations of 3,658 individuals (after removing individuals with missing values) who have been followed for 10 years. The objective of the Framingham study was to predict the development of a Coronary Heart Disease (CHD) within 10 years based on $d = 15$ observed features measured at inclusion. We therefore have $K = 2$ classes, with class 2 corresponding to individuals who have developed a CHD, and class 1 corresponding to the others. Among the 15 features, 7 are categorical (*sex, education, smoking status, previous history of stroke, diabetes, hypertension, antihypertensive treatment*) and 8 are numeric (*age, number of cigarettes per day, cholesterol levels, systolic blood pressure, diastolic blood pressure, heart rate, body mass index (BMI), glycemia*). The dataset is imbalanced: $\hat{\pi} = [0.85, 0.15]$, which means that 15% of the individuals have developed a CHD within 10 years. For this database, we considered the $L_{0\text{-}1}$ loss function.

Diabetes prediction database: Another example of machine learning application in medicine field is to predict the onset diabetes based on diagnostic measurements. We consider here the database studied in [45] which was originally provided by the National Institute of Diabetes and Digestive and Kidney Diseases, and available at [57]. This database contains the measurements of 8 clinical and biological features (*Number of times pregnant, Plasma glucose concentration, Diastolic blood pressure, Triceps skin fold thickness, 2-Hour serum insulin, BMI, Diabetes pedigree function, Age*) for 768 patients. We have $K = 2$ classes, where the class 2 corresponds to the patients who were tested positive for diabetes. The class proportions of this dataset are $\hat{\pi} = [0.65, 0.35]$. For this database, we considered the $L_{0\text{-}1}$ loss function.

Abalone database: The Abalone dataset contains the physical measurements of 4,177 abalones from Tasmania [46]. This dataset is composed of 8 features (1 categorical and 7 numerical) from which the objective is predict the age of each abalone. The initial ages to predict ranged from 1 to 29. For this experiment, we decided to rather consider $K = 5$ classes $\{A_1, A_2, A_3, A_4, A_5\}$ associated to the age groups $\{[\leq 4], [5, 10], [11, 15], [16, 20], [\geq 21]\}$ satisfying the class proportions $\hat{\pi} = [0.02, 0.64, 0.28, 0.05, 0.01]$. These classes are imbalanced. For this database we considered the quadratic loss function: for all $(k, l) \in \mathcal{Y} \times \hat{\mathcal{Y}}$, $L_{kl} = (k-l)^2$, so that the more the predicted class is far from the true class, the more important this error is.

APS Failure Trucks database: This real condition monitoring database [47] focuses on Air Pressure System (APS) used for various functions in Scania trucks such as braking and gear changes. Measurements of a specific APS component were collected from heavy Scania trucks in everyday usage. The goal is to predict a potential failure of this component. We therefore consider $K = 2$ classes where the class 1 corresponds to the APS without failures, and class 2 to the defect APS components. For this database, the costs of class misclassifications were given by experts:

$$L = \begin{bmatrix} 0 & 10 \\ 500 & 0 \end{bmatrix}, \tag{34}$$

so that the cost of predicting a non-existing failure is $\$10$, while the cost of missing a failure is $\$500$. After removing missing values, the database contains the measurements of 69,309 samples for which 68,494 do not present any failure and 815 present a failure. Hence, the class proportions $\hat{\pi} = [0.9882, 0.0118]$ are highly imbalanced, which highly complicates the task of predicting a failure. Finally, each sample is described by $d = 130$ numeric and categorical anonymized features.

NASA pc3 software database: The purpose of this database is to detect certain defects in a flight software of a satellite in Earth orbit [48], [58]. More details on this database and on this task are given in [58] and [59]. For our experiments, we downloaded the data at `https://www.openml.org/d/1050`. This database is composed by 1,563 samples and 37 attributes measured with McCabe [60] and Halstead [61] "module"-based metrics. We have $K = 2$ classes, where the class 2 corresponds to the defect programs. The class proportions $\hat{\pi} = [0.8976, 0.1024]$ are imbalanced, which complicates the task of detecting defect programs. For this database, we considered the $L_{0\text{-}1}$ loss function.

Satellite database: We consider another real highly imbalanced database, downloaded at `https://www.openml.org/d/40900`, for which the motivation is to classify images of soil taken from a satellite into $K = 2$ classes satisfying the class proportions $\hat{\pi} = [0.9853, 0.0147]$. This database is composed by 5,100 samples and 36 attributes, and we considered the $L_{0\text{-}1}$ loss function.