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

Cyprien Gilet, Susana Barbosa, Lionel Fillatre

Abstract—This paper aims to build a supervised classifier for dealing with imbalanced datasets, uncertain class proportions, dependencies between features, the presence of both numeric and categorical features, and arbitrary loss functions. The Bayes classifier suffers when prior probability shifts occur between the training and testing sets. A solution is to look for an equalizer decision rule whose class-conditional risks are equal. Such a classifier corresponds to a minimax classifier when it maximizes the Bayes risk. We develop a novel box-constrained minimax classifier which takes into account some constraints on the priors to control the risk maximization. We analyze the empirical Bayes risk with respect to the box-constrained priors for discrete inputs. We show that this risk is a concave non-differentiable multivariate piecewise affine function. A projected subgradient algorithm is derived to maximize this empirical Bayes risk over the box-constrained simplex. Its convergence is established and its speed is bounded. The optimization algorithm is scalable when the number of classes is large. The robustness of our classifier is studied on diverse databases. Our classifier, jointly applied with a clustering algorithm to process mixed attributes, tends to balance the class-conditional risks while being not too pessimistic.

Index Terms—Minimax Classifier, Γ-Minimax Classifier, Imbalanced datasets, Uncertain Class Proportions, Prior Probability Shift, 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: First, 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), many of which present dependencies. Finally, we often have to take into account a specific loss function, provided by the experts in the relevant field, in order to unequally penalize the class classification errors.

1.1 Related work

The common objective in supervised classification tasks is to minimize the empirical global risk of errors, based on a set of labeled training samples as presented in [1], [2]. This global risk of classification errors is the weighted sum of the class-conditional risks with respect to the associated class proportions as shown in [3]. Such a classifier can be extremely sensitive to the class proportions when the classes are not perfectly separable. Classes are said perfectly separable when samples at any given location in the features space can come from only one possible class. Otherwise, the classes are not perfectly separable, and this issue often occurs in most real applications, like for example in medical field. When the classes are not perfectly separable and the training set is imbalanced, that is, the classes are unequally represented, most classifiers essentially focus on the dominating classes that contain the largest number of occurrences, and tend to underestimate the least represented ones as underlined in [4], [5], [6], [7], [8]. In other words, 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 as studied in [4], [5]. However, this approach introduces a bias since the actual state of nature remains imbalanced. Another common approach is cost-sensitive learning, studied in [4], [5], [9], [10], which aims to optimize the cost of class classification errors in order to counterbalance the number of occurrences of each class.

Furthermore, a classifier presenting imbalanced class-conditional risks becomes sensitive to prior probability shifts. Prior probability shift, as defined in [11], [12], occurs when the true state of nature can change in time due to unknown reasons, and the priors associated with the test samples differ from the class proportions observed in the training set. As discussed in [13], the sensitivity of a classifier to prior probability shifts is greater when the class-conditional risks are imbalanced. And the issue of prior probability shifts remains essential to resolve since the global risk of error is expected to evolve linearly when prior probability shifts occur as shown in [3], [14]. Since the early 2000s a new supervised classification field has emerged for addressing this issue of prior probability shifts, namely 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*

This paper corresponds to the version accepted (not published) in IEEE Transactions on Pattern Analysis and Machine Intelligence. Manuscript submitted on 1st April 2020; revised on 24th September 2020; revised on 1st December 2020; accepted on 4th December 2020. The official published version of this article is available on the IEEE Transactions on Pattern Analysis and Machine Intelligence Journal (DOI: 10.1109/TPAMI.2020.3046439).

quantification as studied in [13], [15], [16], [17], [18]. Using a training set, the task of quantification consists in estimating the class proportions of a test set in order to improve the classification performances associated with this given test sample. However, the main drawback of quantification approaches is that the task of classifying test instances is not performed individually but for the whole sample, which is not always possible for many real-world applications like medicine.

Hence, prior probability shifts and training with imbalanced datasets share a common trait, namely the sensitivity to unequal class-conditional risks. Equalizing the class-conditional risks is therefore essential to obtain a robust classifier. A famous and relevant approach for designing a robust classifier in the presence of imbalanced datasets and prior probability shifts is to fit the classifier by minimizing the maximum of the class-conditional risks as studied in [3], [14], [19], [20]. The resulting decision rule is called minimax classifier. Statistical decision theory in [21] shows that an equalizer Bayesian classifier, one whose class-conditional risks are all equal, is necessarily a minimax classifier. The minimax criterion is part of the field named *Bayesian Robustness* which characterizes the task of considering robust classifiers with respect to prior probability shifts, as mentioned in [19].

1.2 Motivation for the study

A pioneering article on the minimax criterion in the field of machine learning is [22]. This paper studies the generalization error of a minimax classifier but does not provide any method to compute it. In [23], the authors proposed the Minimum Error Minimax Probability Machine for the task of binary classification only, and the extension to multiple classes is difficult. This method is very close to the one described in [24]. The Support Vector Machine (SVM) decision rule can also be tuned for minimax classification as in [25]. The study proposed in [25] is limited to linear classifiers (either using or not using a feature mapping) and to the classification problems between only two classes. In [26], the authors proposed an approach that fits a decision rule by learning the probability distribution which minimizes the worst case of misclassification over a set of distributions centered on 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 as proposed in [27]. Finally, in [28], 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 that demand too many random occurrences from the classes containing initially just a few instances.

Basically, the minimax classifier derives from the computation of the least favorable priors which maximize the minimum empirical global risk of error over the probabilistic simplex as shown in [3], [14], [19]. These least favorable priors are generally difficult to compute as underlined in [19], [29], [30]. Simple algorithms are still required to compute the least favorable priors for any classification problems. Moreover, although the minimax criterion is suitable for addressing the issues regarding class proportions, this approach appears sometimes too pessimistic, as discussed in [19], [31]. This drawback occurs when the least favorable priors seem unrealistic 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 introduced in [19]. However, the calculation of a Γ-minimax classifier is difficult too. Currently, no algorithm exists to calculate it in a general way. Finally, it is difficult to achieve optimal results when dealing with both numeric and categorical attributes. To compute a minimax classifier, we need a good estimate of the joint distribution of the input features in each class. However, in the presence of mixed attributes, and due to the curse of dimensionality (as noted in [14], [32]), this estimation is quite difficult. In such a case, a weaker solution would be to consider the naïve approach of estimating the marginal distribution of each feature independently. But this hypothesis is not acceptable since we want to take into account the dependencies between the features. Thenceforth, a reasonable approach is to discretize the numeric attributes in order to reduce the complexity of the joint distribution estimation. Especially since many papers in the literature have shown that the discretization of the numeric features generally leads to accurate results, as in [33], [34], [35], [36], [37], 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 calculated exactly as in [38], [39], [40]. In our context, all these benefits encourage us to discretize the numeric features. For all these reasons, our motivation is to develop a Γ-minimax classifier adapted to discrete or discretized features which can be easily computed.

Our approach is especially relevant for applications with severe requirements on the global risk of classification errors and strongly imbalanced datasets including for example medical diagnosis [41], [42], image classification [43], fault detection and isolation [29] and fraud detection [44]. For these kinds of applications, it is crucially important to propose a simple and fast algorithm that can almost equalize the class-conditional risks whatever the training set at hand.

1.3 Contribution and organization of the paper

The contributions of the paper are the following. First, we introduce a specific Γ-minimax classifier, called the "Box-constrained minimax classifier", which takes into account some independent bounds on each class proportion. The main advantage of considering such a box-constraint stems from the fact that experts in the field of application 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 yet been studied to address the minimax criterion drawback. Secondly, 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 priors. 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 remains still challenging in general. Fourthly, we show that this algorithm can be coupled with a discretization process, such as the k-means algorithm, to compute the box-constrained minimax classifier in the context of mixed attributes. Finally, we carefully show the robustness of our box-constrained minimax classifier since it is an almost equalizer classifier. We train it on imbalanced datasets and test it on test sets with prior probability shifts.

This paper is part of the field of Γ-minimaxity and Bayesian robustness for supervised classification tasks. It generalizes our preliminary works published in [45], [46]. We introduced our minimax classifier without considering any box-constraint on the priors in [45]. We generalized [45] by introducing the concept of box-constraint on the priors in [46]. In this paper, we reinforce [46] by providing all the proofs and carefully analyzing and interpreting the numerical performance of our classifier. We consider many state-of-the-art approaches and compare them to our classifier on both simulated and real datasets. We exploit several datasets with different and complementary characteristics.

The paper is organized as follows. Section 2 introduces the box-constrained minimax classifier concept. Section 3 studies the discrete empirical Bayes risk and derives the algorithm to compute the discrete box-constrained minimax classifier. In section 4, we show how to easily and accurately discretize databases containing both numeric and categorical features with the k-means algorithm. We then carry out a rigorous experimental procedure to compare our novel classifier with other traditional classifiers that deal with the issues of imbalanced and uncertain class proportions. These experiments are based on seven real databases coming from different application fields. Section 5 concludes the paper. The appendices support the main results of the paper, and provide 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 class l when 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 $k \neq l$. We consider a multiset $\mathcal{S} = \{(Y_i, X_i), i \in \mathcal{I}\}$ containing a number m of labeled training samples where $\mathcal{I}$ is a finite set of indices. Let $\mathbb{1}_{\{Y_i=k\}}$ be the indicator function of the event $Y_i = k$. In the following, $\hat{\pi} := [\hat{\pi}_1, \ldots, \hat{\pi}_K]$ corresponds to the class proportions of the training set:

$$\hat{\pi}_k = \frac{1}{m} \sum_{i \in \mathcal{I}} \mathbb{1}_{\{Y_i=k\}}, \ \forall k \in \mathcal{Y}. \tag{1}$$

The task of supervised classification as defined in [1], [2], [14] is to learn a decision rule $\delta : \mathcal{X} \to \hat{\mathcal{Y}}$ which assigns each instance $i \in \mathcal{I}$ 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, such that δ minimizes the empirical risk

$$\hat{r}(\delta) = \frac{1}{m} \sum_{i \in \mathcal{I}} L(Y_i, \delta(X_i)). \tag{2}$$

In the following, we will use the notation δ_π to denote that the decision rule δ was fitted, by minimizing (2), with the priors π, for any π in the K-dimensional probability simplex $\mathbb{S}$ defined by $\mathbb{S} := \{\pi \in [0,1]^K : \sum_{k=1}^K \pi_k = 1\}$. Let $\hat{\mathbb{P}}(\delta_{\hat{\pi}}(X_i) = l \mid Y_i = k)$ denote the empirical probability for the classifier $\delta_{\hat{\pi}}$ to assign the class l given that the true class is k:

$$\hat{\mathbb{P}}(\delta_{\hat{\pi}}(X_i) = l \mid Y_i = k) \ = \ \frac{1}{m_k} \sum_{i \in \mathcal{I}_k} \mathbb{1}_{\{\delta_{\hat{\pi}}(X_i)=l\}}, \tag{3}$$

where $\mathcal{I}_k = \{i \in \mathcal{I} : Y_i = k\}$ be the set of training samples from class k and $m_k = |\mathcal{I}_k|$ is the number of instances in $\mathcal{I}_k$. As explained in [3], the risk (2) can be written as

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

where $\hat{R}_k\left(\delta_{\hat{\pi}}\right)$ is the empirical class-conditional risk associated with class k, defined by

$$\hat{R}_k\left(\delta_{\hat{\pi}}\right) \ := \ \sum_{l \in \hat{\mathcal{Y}}} L_{kl} \, \hat{\mathbb{P}}(\delta_{\hat{\pi}}(X_i) = l \mid Y_i = k). \tag{5}$$

In the following, $\Delta := \{\delta : \mathcal{X} \to \hat{\mathcal{Y}}\}$ denotes the set of all possible classifiers.

2.1 Empirical Bayes risk for the training set prior

Let us consider that each feature X_{ij} is discrete or beforehand discretized and 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 instances. Let $\mathcal{T} = \{1, \ldots, T\}$ be the set of indices. Then let us define for all $k \in \mathcal{Y}$ and for all $t \in \mathcal{T}$,

$$\hat{p}_{kt} \ := \ \frac{1}{m_k} \sum_{i \in \mathcal{I}_k} \mathbb{1}_{\{X_i=x_t\}}, \tag{6}$$

the probability estimate of observing the feature profile $x_t \in \mathcal{X}$ given that the class label is k. In the context of statistical hypothesis testing theory, [47] 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*

$$\hat{r}\left(\delta_{\hat{\pi}}\right) = \sum_{t\in\mathcal{T}}\sum_{k\in\mathcal{Y}}\sum_{l\in\hat{\mathcal{Y}}} L_{kl}\,\hat{\pi}_k\,\hat{p}_{kt}\,\mathbb{1}_{\{\delta_{\hat{\pi}}(x_t)=l\}}. \tag{7}$$

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

According to Lemma 1, the performance of any classifier δ fitted on the training 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 collectively exhaustive of the training dataset. The following theorem precises the discrete empirical Bayes classifier for $K \geq 2$ classes and any positive loss function L.

Theorem 1. *The empirical Bayes classifier $\delta^B_{\hat{\pi}}$, which minimizes the empirical risk (7) over Δ, is given by*

$$\delta^B_{\hat{\pi}} : X_i \mapsto \underset{l\in\hat{\mathcal{Y}}}{\arg\min} \sum_{t\in\mathcal{T}}\sum_{k\in\mathcal{Y}} L_{kl}\,\hat{\pi}_k\,\hat{p}_{kt}\,\mathbb{1}_{\{X_i=x_t\}}. \tag{8}$$

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 $k \in \mathcal{Y}$,

$$\hat{R}_k\left(\delta^B_{\hat{\pi}}\right) = \sum_{t\in\mathcal{T}}\sum_{l\in\hat{\mathcal{Y}}} L_{kl}\,\hat{p}_{kt}\,\mathbb{1}_{\{\lambda_{lt}=\min_{q\in\hat{\mathcal{Y}}}\lambda_{qt}\}}, \tag{9}$$

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. We 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.

2.2 Background on the minimax classifier principle

Let $\mathcal{S}' = \{(Y'_i, X'_i), i \in \mathcal{I}'\}$ be a multiset, where $\mathcal{I}'$ is a finite set of indices, containing a number m' of test samples satisfying the unknown class proportions $\pi' = [\pi'_1, \ldots, \pi'_K]$. The classifier $\delta_{\hat{\pi}}$ fitted using the training set $\mathcal{S}$ 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

$$\hat{r}\left(\pi', \delta_{\hat{\pi}}\right) = \sum_{k\in\mathcal{Y}} \pi'_k \hat{R}_k\left(\delta_{\hat{\pi}}\right). \tag{10}$$

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

$$\hat{r}\left(\pi', \delta_{\hat{\pi}}\right) = \pi'_1\left(\hat{R}_1\left(\delta_{\hat{\pi}}\right) - \hat{R}_2\left(\delta_{\hat{\pi}}\right)\right) + \hat{R}_2\left(\delta_{\hat{\pi}}\right). \tag{11}$$

It is 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 implies 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 that can be attained by a classifier when π' shifts over the simplex 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 ensure a decision rule $\delta_{\hat{\pi}}$ is 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 considering the following optimization problem:

$$\delta^B_{\hat{\pi}} = \underset{\delta\in\Delta}{\operatorname{argmin}} \max_{\pi\in\mathbb{S}} \hat{r}(\pi, \delta_\pi) = \underset{\delta\in\Delta}{\operatorname{argmin}} \max_{\pi\in\mathbb{S}} \hat{r}(\delta_\pi). \tag{12}$$

The upper index B in (12) means that $\delta^B_{\hat{\pi}}$ is a Bayes classifier. The famous Minimax Theorem in [48] establishes that

$$\min_{\delta\in\Delta}\max_{\pi\in\mathbb{S}} \hat{r}(\delta_\pi) = \max_{\pi\in\mathbb{S}}\min_{\delta\in\Delta} \hat{r}(\delta_\pi). \tag{13}$$

In our case, dealing only with discrete features entails that the set of possible classifiers Δ is finite. Looking at the proof of the Minimax theorem in [48] shows immediately that the Minimax theorem holds when Δ is finite. In the following, let us define

$$\delta^B_{\pi} := \underset{\delta\in\Delta}{\arg\min}\, \hat{r}(\delta_\pi) \tag{14}$$

the optimal Bayes classifier associated with any given priors $\pi \in \mathbb{S}$. Hence, according to (13), provided that we can calculate δ^B_π for any $\pi \in \mathbb{S}$, the optimization problem (12) is equivalent to computing the least favorable priors

$$\bar{\pi} := \underset{\pi\in\mathbb{S}}{\arg\max}\, \hat{r}\left(\delta^B_\pi\right), \tag{15}$$

so that the minimax classifier $\delta^B_{\hat{\pi}}$ solution of (12) is given by (14) when considering the priors (15).

2.3 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 with $\delta^B_{\hat{\pi}}$ is too high as noted in [19]. In such a case, a solution is to shrink the class proportions constraint, based on the knowledge, or the focus of interest, of the experts in the application domain.

For example in Fig. 1, right, let us consider that the proportions of class 1 are uncertain but bounded between $a_1 = 0.1$ and $b_1 = 0.4$. If we look at the point b_1, it is clear that the classifier $\delta^B_{\hat{\pi}}$ fitted on the class proportions $\hat{\pi}_1$ of the training set is very far from the minimum empirical Bayes risk $\hat{r}\left(\pi', \delta^B_{\pi'}\right)$. The minimax classifier $\delta^B_{\hat{\pi}}$ 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 change of priors, based on the knowledge, or the interest, of the experts in the application domain. To our knowledge, the concept of box-constrained minimax classifiers 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

$$\mathbb{B} := \left\{\pi \in \mathbb{R}^K : \forall k \in \mathcal{Y},\ 0 \leq a_k \leq \pi_k \leq b_k \leq 1\right\}, \tag{16}$$

which results in the box-constrained simplex

$$\mathbb{U} := \mathbb{S} \cap \mathbb{B}. \tag{17}$$

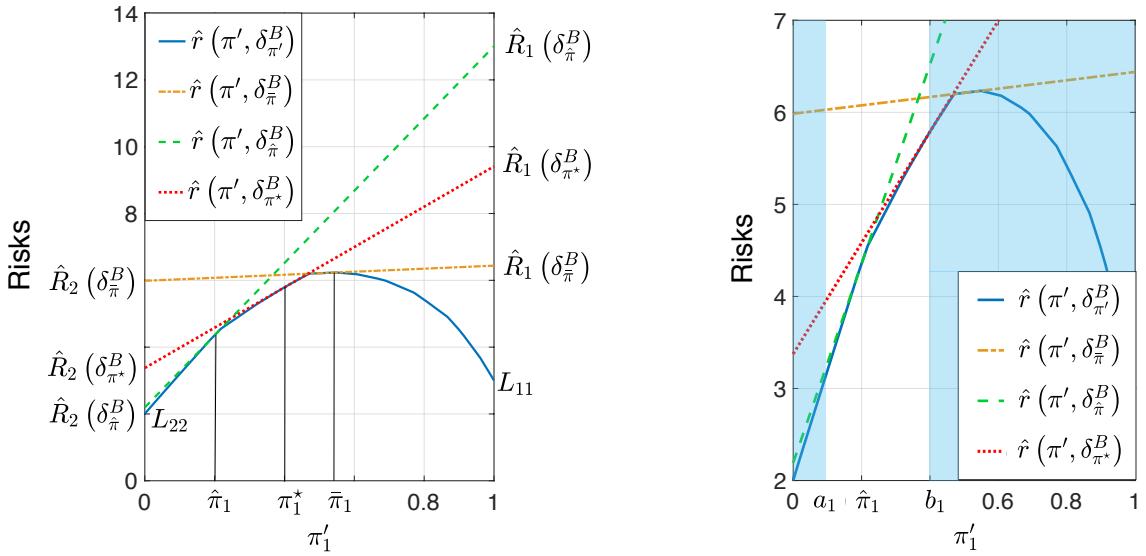


Fig. 1. Comparison between the empirical Bayes classifier $\delta^B_{\hat{\pi}}$, the minimax classifier $\delta^B_{\bar{\pi}}$ and the box-constrained minimax classifier $\delta^B_{\pi^\star}$. These results come from a synthetic dataset for which $K = 2$ classes. The generation of this dataset is detailed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

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

$$\pi^\star \;=\; \arg\max_{\pi\in\mathbb{U}} \hat{r}\left(\delta^B_\pi\right). \tag{18}$$

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 Computation of the Box-Constrained Minimax Classifier

Let us remind that all the features are discrete, or have been discretized. In [39], [40], [49], [50], the authors analyze the risk (2) for the discrete classification task, any number K of classes and an arbitrary loss function. These studies are limited to a given prior or a random prior with a given distribution. This section extends the study of the risk as a function of the priors over the simplex $\mathbb{S}$. This extension is necessary to compute the least favorable prior $\pi^\star$.

3.1 Empirical Bayes risk extended to any prior

Since we can only exploit the instances from the training set, the probabilities $\hat{p}_{kt}$ defined in (6) are assumed to be estimated once and for all. This is an usual assumption in the literature [13]. Statistical estimation theory in [51] has established that the estimates $\hat{p}_{kt}$ correspond to the maximum likelihood estimates of the true probabilities p_{kt} for all $(k,t) \in \mathcal{Y}\times\mathcal{T}$. By estimating these probabilities with the full training set, we get the best unbiased estimate with the smallest variance. This paper assumes that the class-conditional probabilities are representative of the test set.

However, as explained in Section 2, we cannot be confident in the class proportion estimates $\hat{\pi}_k$. Indeed, when the training set is imbalanced, these estimates $\hat{\pi}_k$ can lead to a biased Bayes classifier toward the most probable classes. Furthermore, the estimates $\hat{\pi}_k$ can be uncertain: i) the estimates are not informative about the true a priori distribution, or ii) the estimates, which are computed only one time, can not capture the priors probability shifts that can occur in time. Thus, the empirical Bayes risk must be viewed as a function of the priors. By this way, the performance of the classifier can be assessed whatever the class proportions are.

From Theorem 1, and keeping the class-conditional probabilities $\hat{p}_{kt}$ unchanged, it follows that the empirical Bayes classifier (14) associated with any prior $\pi \in \mathbb{S}$ is given by

$$\delta^B_\pi : X_i \mapsto \arg\min_{l\in\hat{\mathcal{Y}}} \sum_{t\in\mathcal{T}}\sum_{k\in\mathcal{Y}} L_{kl}\,\pi_k\,\hat{p}_{kt}\,\mathbb{1}_{\{X_i=x_t\}}. \tag{19}$$

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 $V : \mathbb{S} \to [0,+\infty)$ defined by

$$V(\pi) \;:=\; \hat{r}\left(\delta^B_\pi\right) \;=\; \sum_{k\in\mathcal{Y}} \pi_k \hat{R}_k\left(\delta^B_\pi\right), \tag{20}$$

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

$$\hat{R}_k\left(\delta^B_\pi\right) = \sum_{t\in\mathcal{T}}\sum_{l\in\hat{\mathcal{Y}}} L_{kl}\,\hat{p}_{kt}\,\mathbb{1}_{\{\lambda_{lt}=\min_{q\in\hat{\mathcal{Y}}}\lambda_{qt}\}}, \tag{21}$$

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 to be robust to priors probability shifts if its risk remains very close to $V(\pi)$ whatever the value of $\pi \in \mathbb{S}$.

It is well known in the literature, see [3], [14], that the Bayes risk, as a function of the priors, is concave over the probability simplex $\mathbb{S}$. The following proposition shows that this result holds when considering the empirical Bayes risk (20). Let us note that all the results are given for $\pi \in \mathbb{S}$, but

they also hold over the box-constrained probability simplex $\mathbb{U}$ since $\mathbb{U} \subset \mathbb{S}$.

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

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

Then, the following proposition and its corollary consider 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*

$$\exists\,(\pi, \pi', k) \in \mathbb{S} \times \mathbb{S} \times \mathcal{Y} : \hat{R}_k\left(\delta_{\pi}^{B}\right) \neq \hat{R}_k\left(\delta_{\pi'}^{B}\right) \tag{22}$$

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

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

Note that condition (22) is most likely achievable. Otherwise, each class-conditional risk would remain equal whatever the prior. And if condition (22) is not satisfied, it follows that V is affine over $\mathbb{S}$.

3.2 Optimization procedure and convergence

In order to compute our box-constrained minimax classifier, according to (18) and when considering (20), our objective is to solve the following optimization problem

$$\pi^{\star} \;=\; \arg\max_{\pi \in \mathbb{U}} V(\pi). \tag{23}$$

Since $V : \pi \mapsto V(\pi)$ is in general non-differentiable provided that condition (22) is satisfied, it is necessary to develop an optimization algorithm adapted to both the non-differentiability of V and the domain $\mathbb{U}$. To this aim, we propose to use a projected subgradient algorithm based on [52] and following the scheme

$$\pi^{(n+1)} = \mathrm{P}_{\mathbb{U}}\left(\pi^{(n)} + \frac{\gamma_n}{\eta_n}\, g^{(n)}\right), \tag{24}$$

where, 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 $\mathbb{U}$. We note that this algorithm also holds in the particular case where condition (22) is not satisfied, that is, 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 $g^{(n)} = \hat{R}\left(\delta_{\pi^{(n)}}^{B}\right)$ at each iteration $n \geq 1$ in (24). The following theorem establishes the convergence of the iterates (24) to $\pi^{\star}$.

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

$$\inf_{n\geq 1} \gamma_n > 0, \quad \sum_{n=1}^{+\infty} \gamma_n^2 < +\infty, \quad \sum_{n=1}^{+\infty} \gamma_n = +\infty, \tag{25}$$

the sequence of iterates (24) converges strongly to a solution $\pi^{\star}$ of (23), whatever the initialization $\pi^{(1)} \in \mathbb{S}$.

Proof. The proof is a consequence of Theorem 1 in [52]. Here we have 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 $\hat{R}\left(\delta_{\pi^{\star}}^{B}\right)$ at the box-constrained minimax optimum cannot vanish, otherwise the associated risk $V(\pi^{\star})$ would be null too due to (20). And this would be a contradiction with the fact that $\pi^{\star}$ is solution of (23). Hence, in this general case, the sequence (24) is infinite and we need to consider a stopping criterion. With this aim, we propose to follow the reasoning in [53] which leads to the following corollary.

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

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

with

$$\varphi(N) := \max\left\{1, \sqrt{\sum_{k=1}^{K}\left[\sum_{l=1}^{K} L_{kl}\right]^2}\right\} \frac{\rho^2 + \sum_{n=1}^{N}\gamma_n^2}{2\sum_{n=1}^{N}\gamma_n}, \tag{26}$$

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 probability simplex. Since (26) converges to 0 as $N \to \infty$, we can choose a small tolerance $\varepsilon > 0$ as a stopping criterion: we fix ε and then compute $N = N_{\varepsilon}$ such that the bound in (26) is smaller than ε.

When considering the sequence of iterates (24), we need to compute the exact projection onto the box-constrained probability simplex $\mathbb{U}$ at each iteration n. To this end, we propose to consider the algorithm provided in [54], 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 $\mathbb{U}$. We 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 algorithms provided in [55] or [56] for which the complexity are lower.

3.3 Box-constrained minimax classifier algorithm

The procedure for computing the box-constrained minimax classifier $\delta_{\pi^{\star}}^{B}$ is summarized step by step in Algorithm 1. In practice, we can choose the sequence of steps $(\gamma_n)_{n\geq 1} = 1/n$ which satisfies (25). Let us note that our approach does not need to resample the training set at each iteration n. Indeed, $\pi^{(n)}$ and $\pi^{\star}$ are used only analytically, which enables us to include all the information provided in the training set for computing our minimax classifier.

A Matlab version and a Python version of our algorithm are available at [57].

Algorithm 1 Box-constrained minimax classifier

Input: $(Y_i, X_i)_{i\in\mathcal{I}}$, K, N.
Compute $\pi^{(1)} = \hat{\pi}$
Compute the $\hat{p}_{kt}$ values, given by (6).
$r^\star \leftarrow 0$, $\pi^\star \leftarrow \pi^{(1)}$
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 (21)
  end for
  $r^{(n)} = \sum_{k=1}^K \pi_k^{(n)} g_k^{(n)}$ see (20)
  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 (19) with $\pi = \pi^\star$.

4 Numerical Experiments

In this section, we illustrate the interest of our box-constrained minimax classifier on one syhthetic database described in Appendix A.2, and on six real ones described in [58], [59], [60], [61], [62], [63], 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 the number of instances. 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 in [61], such that $L_{12} = 10$, $L_{21} = 500$, and $L_{11} = L_{22} = 0$.

Database	m	d	d_n	K	L
Synthetic	10,000	2	2	3	$L_{0\text{-}1}$
Framingham [58]	3,658	15	8	2	$L_{0\text{-}1}$
Diabetes [59]	768	8	8	2	$L_{0\text{-}1}$
Abalone [60]	4,177	8	7	5	$Quad$
Scania Trucks [61]	69,309	130	130	2	Stl
NASA pc3 [62]	1,563	37	36	2	$L_{0\text{-}1}$
Satellite [63]	5,100	36	36	2	$L_{0\text{-}1}$

4.1 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 [33], [34], we can use supervised discretization methods such as [64], [65], [66], or unsupervised methods such as the k-means algorithm in [67]. 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. By discretizing the features space using the Kmeans algorithm, our approach considers that the instances belonging to the same cluster may have similar behavior and assigns them to the same class. This philosophy is closely related to the clustering of bandits based approaches for which the objective is to identify clusters of users so that the users belonging to the same cluster are supposed to have similar behavior, which allows to improve the contents recommendation based on the payoffs computed in each cluster [68], [69].

The choice of the number of centroids T is important since it has an impact on the generalization error. It was established from a 4-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.

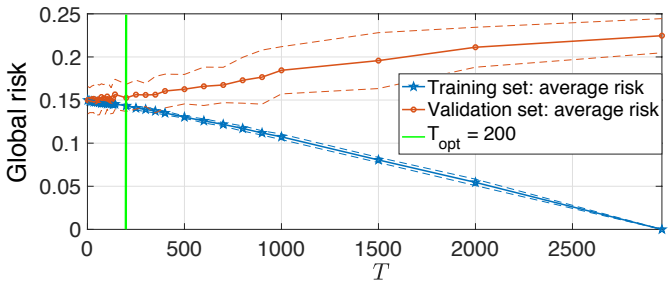


Fig. 2. Framingham database: choice of T from the training set in the first iteration of the 4-fold cross-validation procedure, and when considering $\varepsilon = 0.01$. The dashed curves show the standard-deviation around the mean of $\hat{r}\left(\hat{\pi}, \delta^B_{\hat{\pi}}\right)$.

4.2 Box-constraint generation

In practice, the experts of the application domain can establish the Box-constraint by bounding independently some or all the priors. Concerning the synthetic database, we set the box-constraint as

$$\mathbb{B} := \{\pi \in \mathbb{R}^3 : 0.6 \leq \pi_1 \leq 1, 0.1 \leq \pi_2 \leq 0.25, \\ 0 \leq \pi_3 \leq 0.1\}. \quad (27)$$

Regarding the real databases, in order to illustrate the benefits of the box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$ compared to the minimax classifier $\delta^{\mathrm{B}}_{\hat{\pi}}$ and the discrete Bayes classifier $\delta^{\mathrm{B}}_{\hat{\pi}}$, we consider a box-constraint $\mathbb{B}_\beta$ centered in $\hat{\pi}$, and such that, given $\beta \in [0,1]$,

$$\mathbb{B}_\beta = \left\{\pi \in \mathbb{R}^K : \forall k \in \mathcal{Y},\ \hat{\pi}_k - \rho_\beta \leq \pi_k \leq \hat{\pi}_k + \rho_\beta\right\}, \quad (28)$$

with $\rho_\beta := \beta\,\|\hat{\pi} - \bar{\pi}\|_\infty = \beta\,\max_{k\in\mathcal{Y}} |\hat{\pi}_k - \bar{\pi}_k|$. 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 $\pi^\star = \bar{\pi}$.

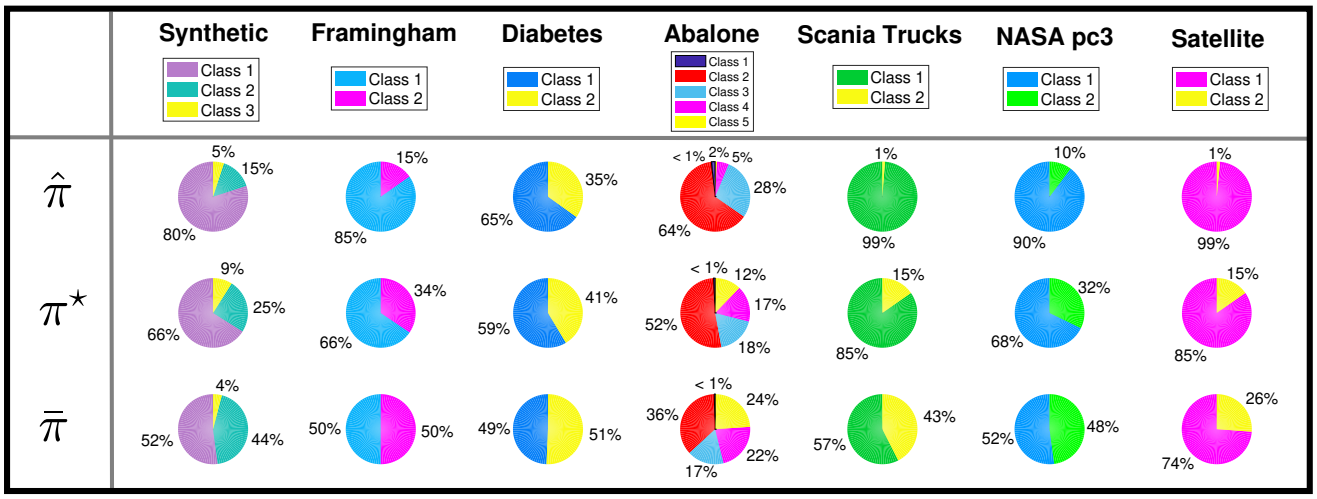


Fig. 3. Pie plots corresponding to the priors $\hat{\pi}$, $\pi^\star$ and $\bar{\pi}$ associated with each databases. These results correspond to the average of the computed priors at each iteration of the 4-folds cross-validation procedure.

4.3 Procedures of the experiments

For each database, we performed a 4-fold cross-validation procedure and we applied our box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$ with respect to their associated box-constraint. We set to 4 the number of folds in order to keep large enough test folds for each database. Concerning the synthetic database, we considered the box (27). Concerning the real databases, we considered the boxes $\mathbb{B}_{0.6}$ for the Framingham, Diabetes, Scania Trucks, NASA pc3 and Satellite databases, and $\mathbb{B}_{0.4}$ for the Abalone database.

4.3.1 *Flowchart of the Box-constrained minimax classifier*

In subsection 4.1, we presented a preprocessing approach for discretizing the numeric features in order to apply our discrete box-constrained minimax criterion. On the following, we will consider the box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$ as the assembly of this features discretization step with the box-constrained minimax computation described in Algorithm 1. The flowchart presented in Fig. 4 illustrates how these two procedures are assembled. Furthermore, as explained in Remark 1, the usual minimax classifier $\delta^{B}_{\bar{\pi}}$ is a particular case of the box-constrained minimax classifier when $\mathbb{B} = [0,1]^K$.

4.3.2 *Classifiers considered for the experiments*

We compared our box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$ to the Logistic Regression [70] denoted by $\delta^{\mathrm{LR}}_{\hat{\pi}}$, the K-Nearest-Neighbors denoted by $\delta^{\mathrm{NN}}_{\hat{\pi}}$, the discrete Bayes classifier $\delta^{\mathrm{B}}_{\hat{\pi}}$ (8), the minimax classifier $\delta^{\mathrm{B}}_{\bar{\pi}}$.

We moreover compare our new algorithm with three common approaches adapted for dealing with imbalanced datasets: the Weighted Logistic Regression denoted by $\delta^{\mathrm{WLR}}_{\hat{\pi}}$, the Weighted Random Forest denoted by $\delta^{\mathrm{WRF}}_{\hat{\pi}}$, Weighted K-Nearest-Neighbors denoted by $\delta^{\mathrm{WNN}}_{\hat{\pi}}$. The Weighted Logistic Regression and the Weighted Random Forests are fitted by considering class-weights inversely proportional to class frequencies and using the algorithms provided by Scikit-Learn [71]. Concerning the Weighted K-Nearest-Neighbors we use the approach provided by [72]. This approach attributes the class for which the sum of the class-weighted instances is the maximum in the neighborhood, where in each class $k \in \mathcal{Y}$, the associated class-weight is $w_k = 1 - \hat{\pi}_k$.

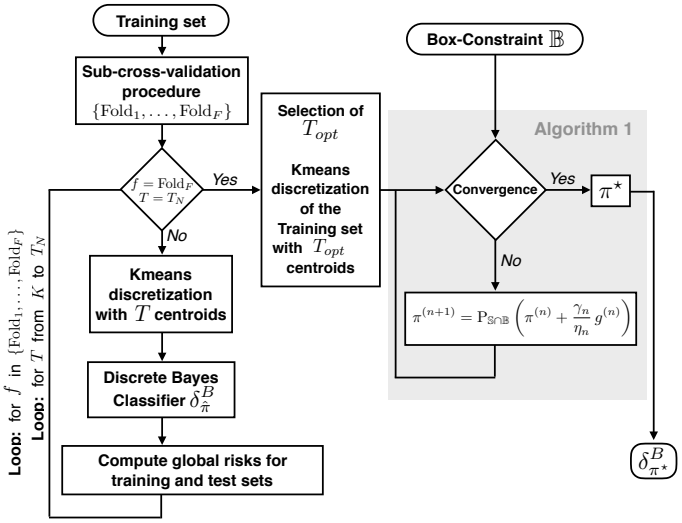


Fig. 4. Flowchart of the box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$ which includes both the discretization and the training steps.

We finally consider two quantification approaches designed for dealing with prior probability shifts. To this aim, we applied the discrete Bayes classifier (19) associated to the class proportions estimated beforehand on the test sets with the adjusted count approach described in [13], [15], [16], and with a more advanced method based on energy distance given in [73]. In the following, these two adjusted classifiers will be respectively denoted as δ^{AC} and δ^{epc}. The set of all these classifiers is denoted $\Delta^E := \{\delta^{\mathrm{LR}}_{\hat{\pi}}, \delta^{\mathrm{RF}}_{\hat{\pi}}, \delta^{\mathrm{NN}}_{\hat{\pi}}, \delta^{\mathrm{WLR}}_{\hat{\pi}}, \delta^{\mathrm{WRF}}_{\hat{\pi}}, \delta^{\mathrm{WNN}}_{\hat{\pi}}, \delta^{\mathrm{B}}_{\hat{\pi}}, \delta^{\mathrm{B}}_{\pi^\star}, \delta^{\mathrm{B}}_{\bar{\pi}}, \delta^{\mathrm{AC}}, \delta^{\mathrm{epc}}\}$.

4.3.3 *Criteria of comparisons*

For these experiments we evaluate each classifier on five different criteria during a common cross-validation procedure.

At each iteration of the cross-validation procedure, we first compare the global risk (4) associated with each classi-

fier $\delta \in \Delta^E$ on both the training set $\mathcal{S} = \{(Y_i, X_i), i \in \mathcal{I}\}$ and the test set $\mathcal{S}' = \{(Y_i', X_i'), i \in \mathcal{I}'\}$.

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 $\delta^{\mathrm{B}}_{\pi^\star}$ and $\delta^{\mathrm{B}}_{\bar{\pi}}$ 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 $\psi : \Delta^E \to \mathbb{R}^+$ such that

$$\psi(\delta) \;:=\; \max_{k\in\mathcal{Y}} \hat{R}_k(\delta) - \min_{k\in\mathcal{Y}} \hat{R}_k(\delta). \tag{29}$$

In other words, the criterion ψ aims to measure how equalizer a given classifier $\delta \in \Delta^E$ is.

In order to evaluate the robustness of each classifier when the class proportions are uncertain and prior probability shifts occur, we generated 100 random priors $\pi^{(s)}$, $s \in \{1, \ldots, 100\}$, uniformly dispersed over the box-constrained simplex $\mathbb{U}$. To this aim, we uniformly generated a sequence of priors over the simplex $\mathbb{S}$ using the procedure in [74], until that 100 of them also satisfy the constraint $\mathbb{U}$. Then, for each repetition of the cross-validation procedure, we generated 100 test subsets $\mathcal{S}^{(s)} = \left\{(Y_i', X_i'), i \in \mathcal{I}^{(s)}\right\}$ by randomly selecting instances from the full test fold set $\mathcal{S}'$, and such that each test subset $\mathcal{S}^{(s)}$ satisfies one of the random priors $\pi^{(s)}$. Each classifier $\delta \in \Delta^E$ was finally tested when considering all the 100 random priors over $\mathbb{U}$. In order to measure the robustness of each classifier δ, we look at the boxplot of $[\hat{r}(\pi^{(1)}, \delta), \ldots, \hat{r}(\pi^{(100)}, \delta)]$, which allows to both evaluate the dispersion and the values of the risks $\hat{r}(\pi^{(s)}, \delta)$, $s \in \{1, \ldots, 100\}$.

4.4 Results

In this subsection, we first present a detailed description of the results associated with the synthetic database, and we then summarize the results associated with the real ones.

4.4.1 *Results associated with the synthetic database*

The values of the global risks $\hat{r}(\hat{\pi}, \delta)$ and $\hat{r}(\pi', \delta)$ associated respectively with the training and test sets are given in Fig. 7. We can observe that the discrete Bayes classifier $\delta^{\mathrm{B}}_{\hat{\pi}}$ applied on the discretized database can well challenge with the Logistic Regression $\delta^{\mathrm{LR}}_{\hat{\pi}}$ and the K-Nearest-Neighbors $\delta^{\mathrm{NN}}_{\hat{\pi}}$ applied both to the real features. This confirms that the discretization of the features has a negligible impact. Fig. 5 shows the samples of the synthetic dataset. It is clear that the class 2 is the most difficult class to discriminate: class 2 represents just 15% of the dataset and it overlaps significantly the samples of class 1 which represent 80% of the dataset. We can note that the classifiers not tuned for imbalance datasets have some difficulties for well classifying the samples of class 2.

The priors $\pi^\star$ and $\bar{\pi}$ computed with our minimax algorithm for this synthetic database are given in Fig. 3. It is important to note that the least favorable priors $\bar{\pi}$, which are

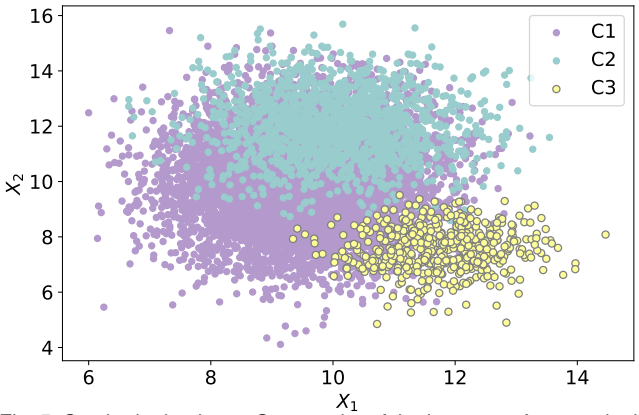


Fig. 5. Synthetic database: Scatter plot of the instances from each class {C1, C3, C3}. The database generation is described in Appendix A.2.

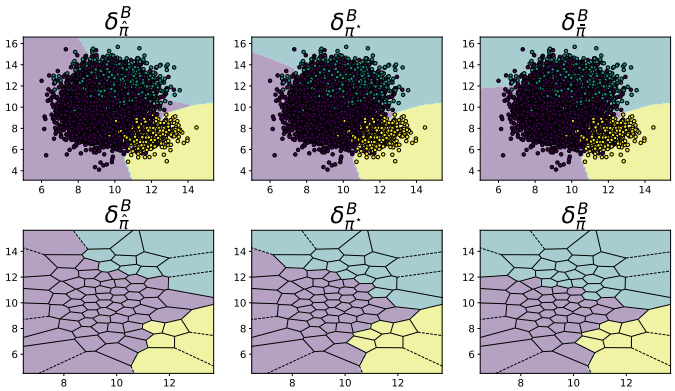


Fig. 6. Synthetic database: Impact of the priors on the class-label assignment to each discrete profile for the classifiers $\delta^B_{\hat{\pi}}$, $\delta^B_{\pi^\star}$ and $\delta^B_{\bar{\pi}}$.

the most adequate for equalizing the class conditional risks, are not balanced for this database. Under the hypothesis that the function V (20) calculated on the discretized database is close enough to the empirical Bayes risk associated to real features, 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.

We can observe in Fig. 7 that the minimax classifier $\delta^{\mathrm{B}}_{\bar{\pi}}$ perfectly balanced the class-conditional risks, and its associated global risks of errors are lower than the risks associated to the weighted Logistic Regression $\delta^{\mathrm{WLR}}_{\hat{\pi}}$ and the weighted Random Forest $\delta^{\mathrm{WRF}}_{\hat{\pi}}$. These two last approaches counterbalanced the class conditional risks contrary to the Logistic Regression $\delta^{\mathrm{LR}}_{\hat{\pi}}$, the discrete Bayes classifier $\delta^{\mathrm{B}}_{\hat{\pi}}$ and the K-Nearest-Neighbors $\delta^{\mathrm{NN}}_{\hat{\pi}}$, which leads to important global risks of errors. Finally, the Box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$ appears as a trade-off between the discrete Bayes classifier $\delta^{\mathrm{B}}_{\hat{\pi}}$ and the minimax classifier $\delta^{\mathrm{B}}_{\bar{\pi}}$. In other words, $\delta^{\mathrm{B}}_{\pi^\star}$ tends to balance the class-conditional risks while satisfying an acceptable global risk of errors with respect to the box-constraint.

Regarding the robustness of each classifier when dealing with prior probability shifts, we can observe in Fig. 8 that the two quantification approaches get low global risks when prior probability shifts occur over $\mathbb{U}$. Indeed, the task of estimating the class proportions $\pi^{(s)}$ of each test set $\mathcal{S}^{(s)}$ before applying the discrete Bayes classifier (19) allows to reach

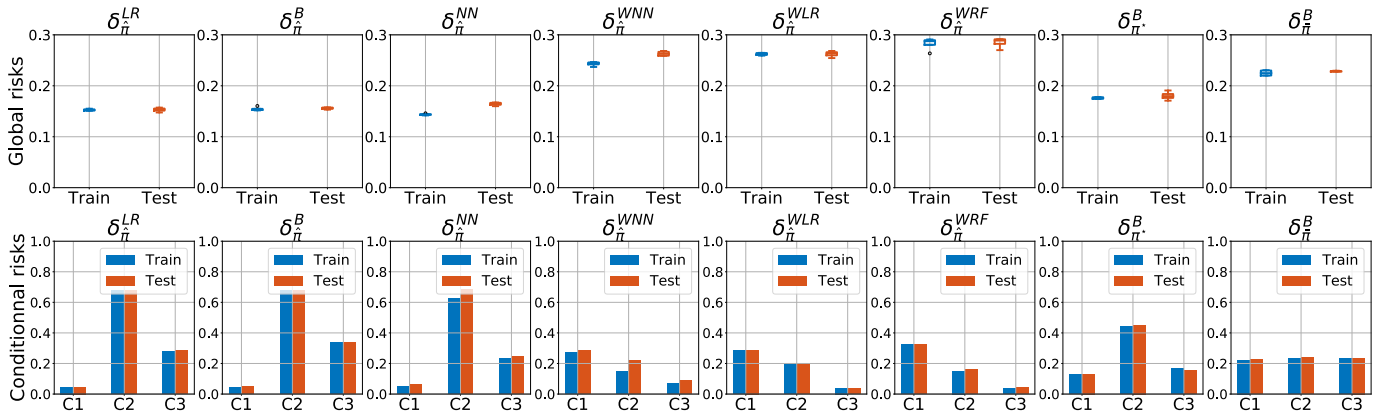


Fig. 7. Synthetic database: Comparison of the risks of misclassification after the 4-fold cross-validation procedure for which the class proportions of the test set were similar to the training set. On the top, the boxplots (training versus test) illustrate the dispersion of the global risks of misclassification. On the bottom, the barplots correspond to the average conditional risks associated to each class {C1, C2, C3} for each classifier.

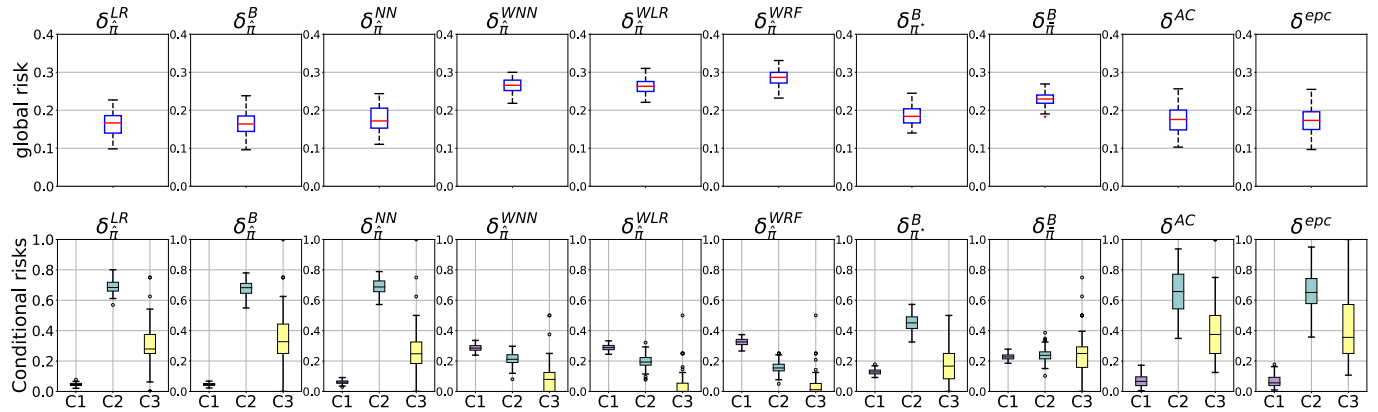


Fig. 8. Synthetic database: On the top, the boxplots illustrate the dispersion of the global risks of misclassification $[\hat{r}(\pi^{(1)}, \delta), \ldots, \hat{r}(\pi^{(100)}, \delta)]$ associated to each classifier $\delta \in \Delta^E$ when prior probability shifts occur over $\mathbb{U}$. On the bottom, the boxplots correspond to the dispersion of the conditional risks $[\hat{R}_k(\delta), \ldots, \hat{R}_k(\delta)]$ associated to each class {C1, C2, C3} of each classifier δ when dealing with these prior probability shifts.

the values of V (20) at these associated class proportions. However, as illustrated in Fig. 8, it follows that the values of the class-conditional risks become highly dispersed for these two quantification approaches. In other words, despite the fact that these quantification approaches allow to obtain satisfying global risks of errors when prior probability shifts occur, these methods do not guaranty robustness on the class-conditional risks. Contrary to these quantification approaches, the class-conditional risks associated to the other methods stay less dispersed when prior probability shifts occur. And regarding the dispersion of their associated global risks in Fig. 8, the minimax classifier $\delta^B_{\hat{\pi}}$, the weighted Logistic Regression $\delta^{\mathrm{WLR}}_{\hat{\pi}}$, the weighted Random Forest $\delta^{\mathrm{WRF}}_{\hat{\pi}}$, and the box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$ were the most robust when the class proportions of the 100 test sets differed from $\hat{\pi}$ since their associated risks $\hat{r}(\pi^{(s)}, \delta)$, $s \in \{1, \ldots, 100\}$ were the less dispersed. This means that these classifiers stay the more stable when prior probability shifts occur over $\mathbb{U}$. Between these four decision rules, the box-constrained minimax classifier gets the lowest global risks. In other words, it results that the box-constrained minimax classifier appears as the best classifier for ensuring an acceptable robustness in terms of both global and class-conditional risks, while respecting satisfactory global risks of error with an acceptable class-conditional risks balancing.

Let us finish the study of the synthetic dataset by showing the impact of the box-constraint over the decision regions of the classifier. Let us compare the Bayes classifier, the minimax classifier and the box-constrained classifier. Generally, the discretization of the features made by the Kmeans algorithm depends on the classifier. However, to show the impact of the priors, we set temporarily the same discrete features for all these classifiers. The partition of the input space is shown in Fig. 6. We can observe that the Bayes classifier favours the class 1. Our box-constrained minimax algorithm changes the class-label of certain regions (a region corresponds to a discrete profile) to give more importance to class 2 and class 3. These changes become more significant when we apply the minimax algorithm which clearly favours class 2 over class 1.

4.4.2 *Results associated with the real databases*

Regarding the six real databases, the results associated to each classifier for each criterion are presented in Table 2. In order to get a better overview of these results, we computed the average rank of each decision rule $\delta \in \Delta^E$ based on the six databases. Due to the computing time associated with the Weighted K-Nearest-Neighbors, we decided to not consider this classifier for these experiments. The priors $\pi^\star$ and $\bar{\pi}$ computed for each database using our algorithm

are summarized in Fig. 3. We can observe that the least favorable priors $\bar{\pi}$ for the databases Abalone, Scania Trucks and Satellite are not balanced and that they are different from the priors of the training set.

Concerning the global risks $\hat{r}(\hat{\pi}, \delta)$ and $\hat{r}(\pi', \delta)$, we can observe in Table 2 that the discrete Bayes classifier $\delta^{\mathrm{B}}_{\hat{\pi}}$ applied on the discretized databases can challenge the Logistic Regression $\delta^{\mathrm{LR}}_{\hat{\pi}}$ and the K-Nearest-Neighbors $\delta^{\mathrm{NN}}_{\hat{\pi}}$ applied both to the real features. Here again, this shows that the discretization impact is negligible. Furthermore, balancing the class-conditional risks implies that the classifiers $\delta^{\mathrm{WLR}}_{\hat{\pi}}$, $\delta^{\mathrm{WRF}}_{\hat{\pi}}$, $\delta^{\mathrm{B}}_{\bar{\pi}}$ get higher global risks than the decision rules $\delta^{\mathrm{B}}_{\hat{\pi}}$, $\delta^{\mathrm{LR}}_{\hat{\pi}}$, $\delta^{\mathrm{NN}}_{\hat{\pi}}$. Hence, the box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$ usually appears as a trade-off between the classifiers $\delta^{\mathrm{B}}_{\hat{\pi}}$, $\delta^{\mathrm{LR}}_{\hat{\pi}}$, $\delta^{\mathrm{NN}}_{\hat{\pi}}$ and the classifiers $\delta^{\mathrm{WLR}}_{\hat{\pi}}$, $\delta^{\mathrm{WRF}}_{\hat{\pi}}$, $\delta^{\mathrm{B}}_{\bar{\pi}}$.

Regarding the maximum of the class conditional risks, the minimax classifier $\delta^{\mathrm{B}}_{\bar{\pi}}$ can challenge the weighted Logistic Regression $\delta^{\mathrm{WLR}}_{\hat{\pi}}$ and the weighted Random Forest $\delta^{\mathrm{WRF}}_{\hat{\pi}}$ applied both to real features. Note that although $\delta^{\mathrm{WLR}}_{\hat{\pi}}$ and $\delta^{\mathrm{WRF}}_{\hat{\pi}}$ usually get convincing results for many of these real databases, it appears that these two classifiers suffer significantly on the two most difficult databases (Abalone and Scania trucks). Contrary to these two classifiers, our minimax algorithm clearly achieves the lowest value of the maximum class-conditional risks. This phenomena is also illustrated with the criterion ψ, since the minimax classifier $\delta^{\mathrm{B}}_{\bar{\pi}}$ appears as the most adequate for balancing the class-conditional risks. Here again, the box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$ generally appears as a trade-off between the classifiers $\delta^{\mathrm{B}}_{\hat{\pi}}$, $\delta^{\mathrm{LR}}_{\hat{\pi}}$, $\delta^{\mathrm{NN}}_{\hat{\pi}}$ and the classifiers $\delta^{\mathrm{WLR}}_{\hat{\pi}}$, $\delta^{\mathrm{WRF}}_{\hat{\pi}}$, $\delta^{\mathrm{B}}_{\bar{\pi}}$ for equalizing the class-conditional risks.

Regarding the robustness of each classifier when dealing with prior probability shifts over $\mathbb{U}_\beta$, the results[1] are presented in Fig. 10. For these experiments, we observe similar results to those of the synthetic database. Although the classifiers $\delta^{\mathrm{B}}_{\bar{\pi}}$, $\delta^{\mathrm{WLR}}_{\hat{\pi}}$, $\delta^{\mathrm{WRF}}_{\hat{\pi}}$ generally get the highest global risks of errors, they were the most robust when the class proportions of the 1000 test sets differed from $\hat{\pi}$ since their associated risks $\hat{r}(\pi^{(s)}, \delta)$, $s \in \{1, \ldots, 100\}$ were the less dispersed. Our box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$ appears here again as a trade-off between the classifiers $\delta^{\mathrm{B}}_{\hat{\pi}}$, $\delta^{\mathrm{LR}}_{\hat{\pi}}$, $\delta^{\mathrm{NN}}_{\hat{\pi}}$ and the classifiers $\delta^{\mathrm{WLR}}_{\hat{\pi}}$, $\delta^{\mathrm{WRF}}_{\hat{\pi}}$, $\delta^{\mathrm{B}}_{\bar{\pi}}$ for equalizing the class-conditional risks and for satisfying acceptable global risks of errors. Finally, all these classifiers ensure a better stability of the class-conditional risks than the two quantification approaches δ^{AC} and δ^{epc}.

Finally, if we look at the processing training times, we have to note that the task of discretizing the features as described in subsection 4.1 induces a higher processing training time for the classifiers $\delta^{B}_{\hat{\pi}}$, $\delta^{B}_{\pi^\star}$, $\delta^{B}_{\bar{\pi}}$. Excluding this preprocessing time of discretizing the features, we can 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 [55], whereas concerning $\delta^{B}_{\pi^\star}$, the procedure for projected onto $\mathbb{U}$ is more complex.

1. Since the results associated with the Diabetes, Scania Trucks, Abalone, NASA pc3 and Satellite databases are similar to those associated to the Synthetic and the Framingham databases, we just present here the results associated to the Framingham database.

4.4.3 *Impact of the Box-constraint radius*

We have previously seen that the box-constrained minimax classifier $\delta^{B}_{\pi^\star}$ allows to find a trade-off between achieving an acceptable global risk and equalizing the class-conditional risks. 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 (28). 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. 9, the more ρ_β increases, the more equalizer $\delta^{B}_{\pi^\star}$ becomes. It follows that $\delta^{B}_{\pi^\star}$ becomes more accurate 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 $V(\bar{\pi})$. Therefore, the experts can easily tighten or spread the box-constraint bounds in order to find an acceptable trade-off.

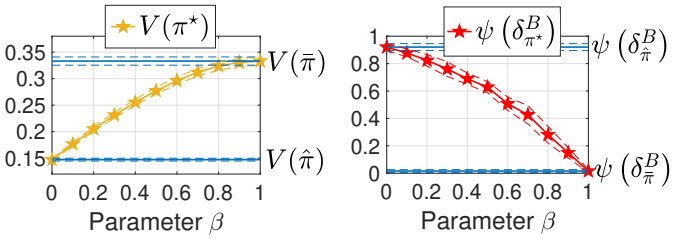


Fig. 9. Framingham database: Impact of the box-constraint radius on $\delta^{\mathrm{B}}_{\pi^\star}$ when β increases from 0 to 1 in (28), after a 4-fold cross-validation procedure. As β increases, the box-constraint radius increases which changes the values of $\pi^\star$, and therefore the values of $V(\pi^\star)$ and $\psi(\delta^{\mathrm{B}}_{\pi^\star})$. Results are presented as $\mathrm{mean} \pm \mathrm{std}$.

4.5 Application to a large scale database

The previous results illustrated that sometimes the weighted approaches like the Weighted Logistic Regression or the Weighted Random Forest can perform well too in the task of equalizing the class-conditional risks. However, these weighted approaches can suffer when dealing with a large number of classes K. Furthermore, when the class proportions of the training set are balanced, the class-conditional risks can even be highly unequal due to the complexity of the classification problem. In such a case, these weighted classifiers become unable to balance the class-conditional risks when considering weights inversely proportional to the class proportions. Unlike these weighted methods, our minimax classifier $\delta^{\mathrm{B}}_{\bar{\pi}}$ is able to face these difficulties for minimizing the maximum of the class-conditional risks.

Let us consider the CIFAR-100 database [43] that contains 60,000 images with $K = 100$ classes for which the class proportions are perfectly balanced. For this experiment, we considered the features extracted from the last hidden layer of the convolutional neural networks EfficientNet-B0 [75]. We then discretized the features using the Kmeans procedure and we compared on the same database the efficiency of the Weighted Logistic Regression $\delta^{\mathrm{WLR}}_{\hat{\pi}}$, the Weighted Random Forest $\delta^{\mathrm{WRF}}_{\hat{\pi}}$, and the Discrete Minimax Classifier $\delta^{\mathrm{B}}_{\bar{\pi}}$, for minimizing the maximum of the class-conditional risks. We do not apply the box-constrained minimax classifier on this dataset; we prefer focusing our attention to the equalization of the class-conditional risks on such a large scale dataset. The training set, respectively the test

TABLE 2
Results associated with each classifier $\delta \in \Delta^E$ and each database after the 4-fold cross-validation procedure. The notation δ (R) means that the classifier δ was applied on the real features. The notation δ (K) means that the classifier δ was performed on the discretized version of each database using the Kmeans algorithm. 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 with the 6 databases. Furthermore, the computing time criterion associated with the classifiers $\{\delta^B_{\hat{\pi}}, \delta^B_{\pi^\star}, \delta^B_{\bar{\pi}}\}$ takes into account the preprocessing task of discretizing the data as described in subsection 4.1.

Criteria	Databases	Classifiers						
		$\delta^{LR}_{\hat{\pi}}$ (R)	$\delta^{B}_{\hat{\pi}}$ (K)	$\delta^{NN}_{\hat{\pi}}$ (R)	$\delta^{WLR}_{\hat{\pi}}$ (R)	$\delta^{WRF}_{\hat{\pi}}$ (R)	$\delta^{B}_{\pi^\star}$ (K)	$\delta^{B}_{\bar{\pi}}$ (K)
Training $\hat{r}(\hat{\pi}, \delta)$	Framingham	0.15 ± 0.00	0.15 ± 0.00	$\mathbf{0.14 \pm 0.00}$	0.34 ± 0.01	0.28 ± 0.01	0.19 ± 0.01	0.34 ± 0.01
	Diabetes	0.29 ± 0.01	0.25 ± 0.03	$\mathbf{0.22 \pm 0.00}$	0.31 ± 0.01	0.23 ± 0.02	0.25 ± 0.02	0.27 ± 0.02
	Abalone	0.34 ± 0.01	0.33 ± 0.01	$\mathbf{0.32 \pm 0.01}$	1.06 ± 0.77	0.97 ± 0.16	0.43 ± 0.02	0.65 ± 0.04
	Scania Trucks	3.13 ± 0.03	0.96 ± 0.02	3.51 ± 0.10	0.71 ± 0.02	$\mathbf{0.61 \pm 0.03}$	2.58 ± 0.31	4.24 ± 0.93
	NASA pc3	0.10 ± 0.00	$\mathbf{0.09 \pm 0.00}$	0.10 ± 0.00	0.89 ± 0.01	0.17 ± 0.01	0.16 ± 0.02	0.30 ± 0.01
	Satellite	$\mathbf{0.003 \pm 0.0}$	0.007 ± 0.0	0.008 ± 0.0	0.019 ± 0.0	0.024 ± 0.0	0.023 ± 0.0	0.045 ± 0.01
Classifier Average Rank		3.00	**2.17**	2.33	5.00	3.83	3.67	5.50
Test $\hat{r}(\pi', \delta)$	Framingham	$\mathbf{0.15 \pm 0.00}$	$\mathbf{0.15 \pm 0.00}$	$\mathbf{0.15 \pm 0.01}$	0.35 ± 0.01	0.30 ± 0.02	0.22 ± 0.02	0.37 ± 0.01
	Diabetes	0.30 ± 0.03	0.29 ± 0.02	0.28 ± 0.03	0.31 ± 0.04	$\mathbf{0.25 \pm 0.03}$	0.29 ± 0.02	0.31 ± 0.01
	Abalone	$\mathbf{0.34 \pm 0.02}$	0.36 ± 0.01	0.37 ± 0.02	1.07 ± 0.78	1.05 ± 0.20	0.49 ± 0.02	0.67 ± 0.06
	Scania Trucks	3.20 ± 0.25	0.99 ± 0.06	3.68 ± 0.24	0.85 ± 0.14	$\mathbf{0.73 \pm 0.01}$	2.67 ± 0.42	4.32 ± 0.96
	NASA pc3	$\mathbf{0.10 \pm 0.01}$	0.11 ± 0.01	$\mathbf{0.10 \pm 0.01}$	0.89 ± 0.01	0.20 ± 0.01	0.20 ± 0.02	0.32 ± 0.01
	Satellite	$\mathbf{0.007 \pm 0.00}$	0.008 ± 0.0	0.008 ± 0.0	0.022 ± 0.0	0.026 ± 0.0	0.025 ± 0.0	0.050 ± 0.01
Classifier Average Rank		**2.17**	**2.17**	2.50	4.33	3.17	3.33	5.33
Training $\max_{k \in \hat{\mathcal{Y}}} \hat{R}_k(\delta)$	Framingham	0.95 ± 0.01	0.94 ± 0.01	0.89 ± 0.01	0.34 ± 0.00	$\mathbf{0.33 \pm 0.02}$	0.67 ± 0.03	0.35 ± 0.01
	Diabetes	0.60 ± 0.05	0.47 ± 0.06	0.46 ± 0.02	0.34 ± 0.01	$\mathbf{0.25 \pm 0.02}$	0.38 ± 0.06	0.29 ± 0.03
	Abalone	3.25 ± 0.49	3.06 ± 0.19	4.05 ± 0.46	1.35 ± 0.79	1.07 ± 0.14	0.93 ± 0.26	$\mathbf{0.83 \pm 0.17}$
	Scania Trucks	271 ± 10	39.8 ± 4.5	298 ± 13	38.5 ± 2.7	19.8 ± 1.9	10.8 ± 2.7	$\mathbf{6.5 \pm 1.2}$
	NASA pc3	1.00 ± 0.00	0.92 ± 0.02	0.96 ± 0.01	0.99 ± 0.01	$\mathbf{0.18 \pm 0.01}$	0.57 ± 0.02	0.30 ± 0.01
	Satellite	0.20 ± 0.03	0.44 ± 0.04	0.56 ± 0.02	$\mathbf{0.02 \pm 0.01}$	0.13 ± 0.01	0.23 ± 0.06	0.05 ± 0.01
Classifier Average Rank		6.16	5.33	6.00	3.33	2.00	3.33	**1.83**
Test $\max_{k \in \hat{\mathcal{Y}}} \hat{R}_k(\delta)$	Framingham	0.96 ± 0.02	0.97 ± 0.02	0.93 ± 0.02	$\mathbf{0.37 \pm 0.01}$	0.41 ± 0.02	0.73 ± 0.03	0.45 ± 0.04
	Diabetes	0.62 ± 0.05	0.51 ± 0.05	0.55 ± 0.04	0.34 ± 0.04	$\mathbf{0.29 \pm 0.01}$	0.44 ± 0.07	0.34 ± 0.01
	Abalone	3.29 ± 1.06	3.68 ± 0.38	3.79 ± 0.71	$\mathbf{1.61 \pm 0.64}$	1.64 ± 0.69	1.99 ± 0.22	1.92 ± 0.55
	Scania Trucks	272 ± 18	43.2 ± 5.2	312 ± 15	51.9 ± 17.9	30.5 ± 4.8	17.9 ± 5.2	$\mathbf{11.6 \pm 3.1}$
	NASA pc3	1.00 ± 0.00	0.98 ± 0.02	0.99 ± 0.02	0.99 ± 0.01	$\mathbf{0.30 \pm 0.05}$	0.74 ± 0.07	0.40 ± 0.05
	Satellite	0.31 ± 0.07	0.52 ± 0.10	0.57 ± 0.09	$\mathbf{0.11 \pm 0.03}$	0.20 ± 0.06	0.37 ± 0.03	0.20 ± 0.10
Classifier Average Rank		5.50	5.33	5.83	2.33	**1.83**	3.33	2.17
Training $\psi(\delta)$	Framingham	0.94 ± 0.01	0.93 ± 0.01	0.88 ± 0.01	0.02 ± 0.02	0.07 ± 0.03	0.56 ± 0.04	$\mathbf{0.01 \pm 0.01}$
	Diabetes	0.48 ± 0.05	0.33 ± 0.12	0.38 ± 0.03	0.07 ± 0.03	0.07 ± 0.02	0.20 ± 0.09	$\mathbf{0.02 \pm 0.02}$
	Abalone	3.14 ± 0.49	2.88 ± 0.18	3.96 ± 0.46	1.33 ± 0.80	1.05 ± 0.15	0.59 ± 0.26	$\mathbf{0.27 \pm 0.11}$
	Scania Trucks	267 ± 11	39.3 ± 4.5	297 ± 13.1	38.2 ± 2.7	19.5 ± 1.9	8.3 ± 3.1	$\mathbf{2.5 \pm 1.5}$
	NASA pc3	1.00 ± 0.00	0.92 ± 0.02	0.96 ± 0.01	0.99 ± 0.01	0.04 ± 0.03	0.45 ± 0.04	$\mathbf{0.02 \pm 0.01}$
	Satellite	0.20 ± 0.03	0.44 ± 0.04	0.56 ± 0.02	0.02 ± 0.01	0.11 ± 0.01	0.21 ± 0.06	$\mathbf{0.002 \pm 0.0}$
Classifier Average Rank		5.83	5.17	6.00	3.33	2.67	3.17	**1.00**
Test $\psi(\delta)$	Framingham	0.95 ± 0.02	0.96 ± 0.02	0.91 ± 0.02	$\mathbf{0.04 \pm 0.03}$	0.13 ± 0.03	0.60 ± 0.04	0.09 ± 0.03
	Diabetes	0.49 ± 0.03	0.34 ± 0.10	0.42 ± 0.03	0.09 ± 0.02	0.08 ± 0.03	0.24 ± 0.12	$\mathbf{0.05 \pm 0.01}$
	Abalone	3.18 ± 1.05	3.48 ± 0.38	3.67 ± 0.71	1.60 ± 0.66	1.53 ± 0.72	1.63 ± 0.20	$\mathbf{1.40 \pm 0.61}$
	Scania Trucks	272 ± 18	42.7 ± 5.1	312 ± 15	51.6 ± 18	30.1 ± 4.9	15.5 ± 5.1	$\mathbf{7.3 \pm 3.0}$
	NASA pc3	1.00 ± 0.00	0.96 ± 0.02	0.99 ± 0.02	0.99 ± 0.00	0.10 ± 0.06	0.60 ± 0.10	$\mathbf{0.09 \pm 0.05}$
	Satellite	0.31 ± 0.07	0.52 ± 0.10	0.57 ± 0.09	$\mathbf{0.09 \pm 0.04}$	0.18 ± 0.07	0.35 ± 0.03	0.15 ± 0.09
Classifier Average Rank		5.83	5.33	6.17	3.00	2.50	3.50	**1.33**
Training Time (s)	Framingham	2.65 ± 0.34	34.8 ± 5.1	$\mathbf{0.003 \pm 0.00}$	2.05 ± 0.41	0.23 ± 0.05	34.2 ± 0.57	34.8 ± 3.36
	Diabetes	0.67 ± 0.35	10.5 ± 0.1	$\mathbf{0.01 \pm 0.00}$	0.73 ± 0.23	0.13 ± 0.00	12.7 ± 0.3	10.9 ± 0.03
	Abalone	5.39 ± 1.09	37.8 ± 0.59	$\mathbf{0.003 \pm 0.00}$	10.58 ± 0.04	0.22 ± 0.00	158.5 ± 1.6	41.2 ± 0.1
	Scania Trucks	408 ± 33	799 ± 18	$\mathbf{4.37 \pm 1.50}$	392 ± 34	5.72 ± 4.61	847 ± 13	812 ± 8
	NASA pc3	2.19 ± 2.57	13.5 ± 0.1	$\mathbf{0.003 \pm 0.00}$	5.56 ± 0.50	0.19 ± 0.00	15.9 ± 0.2	14.3 ± 0.2
	Satellite	6.60 ± 0.36	43.8 ± 06	$\mathbf{0.005 \pm 0.0}$	20.4 ± 0.1	0.22 ± 0.00	45.0 ± 0.5	45.5 ± 1.6
Classifier Average Rank		3.33	4.83	**1.00**	3.67	2.00	6.33	5.83
Predictions Time (s)	Framingham	$\mathbf{6.8 \times 10^{-4}}$	1.0×10^{-2}	4.8×10^{-2}	1.0×10^{-3}	1.5×10^{-2}	9.9×10^{-3}	1.2×10^{-2}
	Diabetes	5.8×10^{-4}	2.7×10^{-3}	6.9×10^{-3}	$\mathbf{5.5 \times 10^{-4}}$	9.6×10^{-3}	2.7×10^{-3}	2.7×10^{-3}
	Abalone	6.4×10^{-4}	4.0×10^{-2}	3.6×10^{-2}	$\mathbf{6.2 \times 10^{-4}}$	1.5×10^{-2}	3.8×10^{-2}	3.8×10^{-2}
	Scania Trucks	$\mathbf{3.3 \times 10^{-3}}$	2.1×10^{-1}	17.9 ± 1.2	3.6×10^{-3}	5.2×10^{-2}	2.2×10^{-1}	2.3×10^{-1}
	NASA pc3	7.8×10^{-4}	5.4×10^{-3}	1.4×10^{-2}	$\mathbf{5.8 \times 10^{-4}}$	1.2×10^{-2}	4.9×10^{-3}	4.8×10^{-3}
	Satellite	7.1×10^{-4}	1.4×10^{-2}	1.4×10^{-2}	$\mathbf{6.8 \times 10^{-4}}$	1.3×10^{-2}	1.4×10^{-2}	1.4×10^{-2}
Classifier Average Rank		1.67	4.33	5.50	**1.33**	4.33	4.00	4.33

set, is composed of 40,000 instances, resp. 20,000 instances. Both the training and test sets satisfied the balanced class proportions $\hat{\pi} = [1/100, \ldots, 1/100]$.

The Weighted Logistic Regression $\delta^{\text{WLR}}_{\hat{\pi}}$ and the Weighted Random Forest $\delta^{\text{WRF}}_{\hat{\pi}}$ have similar results than the Discrete Bayes Classifier $\delta^{\text{B}}_{\hat{\pi}}$. Their global risks of errors reach 0.161 on the training set and 0.167 on the test set. But we can observe in Fig. 11 that, even though the class pro-

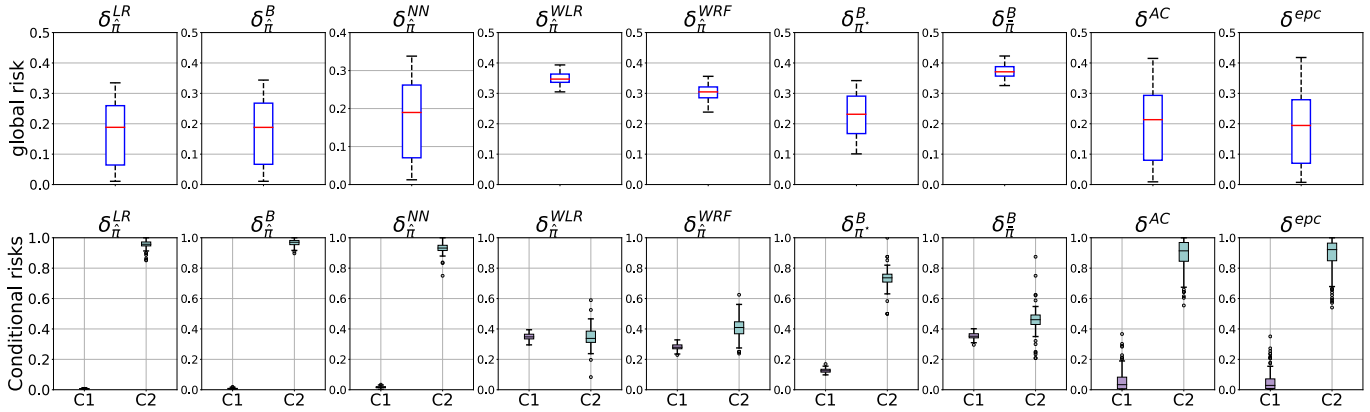


Fig. 10. Framingham database: On the top, the boxplots illustrate the dispersion of the global risks of misclassification $[\hat{r}(\pi^{(1)}, \delta), \ldots, \hat{r}(\pi^{(100)}, \delta)]$ associated to each classifier $\delta \in \Delta^E$ when prior probability shifts occur over $\mathbb{U}$. On the bottom, the boxplots correspond to the dispersion of the conditional risks $[\hat{R}_k(\delta), \ldots, \hat{R}_k(\delta)]$ associated to each class $\{\text{C1}, \text{C2}\}$ of each classifier δ when dealing with these prior probability shifts.

Fig. 11. CIFAR-100 database: Class-conditional risks associated with the Dicrete Bayes Classifier $\delta^{\text{B}}_{\hat{\pi}}$, the Weighted Logistic Regression $\delta^{\text{WLR}}_{\hat{\pi}}$, the Weighted Random Forest $\delta^{\text{WRF}}_{\hat{\pi}}$, and the Discrete Minimax Classifier $\delta^{\text{B}}_{\hat{\pi}}$ on both the training and test sets.

portions are perfectly balanced, their class-conditional risks are highly unequal, achieving $\psi(\delta^{\text{B}}_{\hat{\pi}}) = 0.508$, $\psi(\delta^{\text{WLR}}_{\hat{\pi}}) = 0.545$, $\psi(\delta^{\text{WRF}}_{\hat{\pi}}) = 0.528$ on the training set, and $\psi(\delta^{\text{B}}_{\hat{\pi}}) = 0.560$, $\psi(\delta^{\text{WLR}}_{\hat{\pi}}) = 0.630$, $\psi(\delta^{\text{WRF}}_{\hat{\pi}}) = 0.625$ on the test set. Since the class proportions are perfectly balanced, the Weighted Logistic Regression and the Weighted Random Forest were not able to balance the class-conditional risks when considering their class-weights inversely proportional to the class proportions. Because of the large number of classes, it is too difficult to manually optimize these class-weights. Despite these difficulties, we can observe that our minimax classifier $\delta^{\text{B}}_{\hat{\pi}}$ performed well to minimize the maximum of the class-conditional risks and to balance these risks per class, achieving $\hat{r}(\hat{\pi}, \delta^{\text{B}}_{\hat{\pi}}) = 0.283$ and $\psi(\delta^{\text{B}}_{\hat{\pi}}) = 0.23$ on the training set, and $\psi(\delta^{\text{B}}_{\hat{\pi}}) = 0.35$ and $\hat{r}(\hat{\pi}, \delta^{\text{B}}_{\hat{\pi}}) = 0.294$ on the test set.

5 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 aims to address 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 us to find a trade-off between minimizing the maximum of the class-conditional risks and achieving an acceptable global risk of errors. Our approach also allows to easily consider the classic minimax criterion which remains generally challenging to compute in many application fields.

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

Future work will be devoted to adapt our algorithm for training a minimax regret classifier [19], [31], studying the generalization error of our minimax classifier, and improving 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 sincerely thank Marie Guyomard and Nicolas Glaichenhaus for their contributions and their 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] W. Liu and S. Chawla, "Class confidence weighted knn algorithms for imbalanced data sets," *PAKKD*, 2011.

[9] 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.

[10] 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.

[11] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognition*, 2012.

[12] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*. MIT Press, 2008.

[13] P. González, A. Castaño, C. Nitesh, and J. J. Del Coz, "A review on quantification learning," *ACM Computing Surveys*, 2017.

[14] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. John Wiley and Sons, 2000.

[15] G. Forman, "Counting positives accurately despite inaccurate classification," *Proceedings of ECML'05*, 2005.

[16] ——, "Quantifying counts and costs via classification," *Data Mining and Knowledge Discovery*, 2008.

[17] L. Milli, A. Monreale, G. Rossetti, F. Giannotti, D. Pedreschi, F. Sebastiani, and ISTI-CNR, "Quantification trees," *IEEE 13th International Conference on Data Mining*, 2013.

[18] P. Kar, S. Li, H. Narasimhan, and S. Chawla, "Online optimization methods for the quantification problem," *arXiv:1605.04135v3*, 2016.

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

[20] 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.

[21] A. A. Borovkov, *Mathematical Statistics*. Gordon and Breach Sciences Publishers, Amsterdam, 1998.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

[29] 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.

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

[31] 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.

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

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

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

[39] 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.

[40] 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.

[41] L. Mena and J. A. Gonzalez, "Machine learning for imbalanced datasets: Application in medical diagnostic," in *FLAIRS Conference*, 2006.

[42] M. Rastgoo, G. Lemaître, J. Massich, O. Morel, F. Marzani, R. Garcia, and F. Meriaudeau, "Tackling the problem of data imbalancing for melanoma classification," in *BIOSTEC - 3rd International Conference on Bioimaging*, 2016.

[43] A. Krizhevsky, "Learning multiple layers of features from tiny images," 2009, https://www.cs.toronto.edu/ kriz/cifar.html.

[44] R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review," *Statist. Sci.*, vol. 17, no. 3, pp. 235–255, 08 2002. [Online]. Available: https://doi.org/10.1214/ss/1042727940

[45] 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.

[46] 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.

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

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

[49] L. A. Dalton and M. R. Yousefi, "On optimal bayesian classification and risk estimation under multiple classes," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2015, no. 1, p. 8, 2015.

[50] L. A. Dalton and E. R. Dougherty, *Optimal Bayesian Classification*. SPIE Press Book, 2020.

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

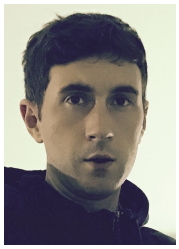
[52] Y. I. Alber, A. N. Iusem, and M. V. Solodov, "On the projected subgradient method for nonsmooth convex optimization in a hilbert space," *Mathematical Programming*, vol. 81, pp. 23–35, 1998.

[53] 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.

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

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

[56] 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.
[57] "Discrete box-constrained minimax classifier algorithm," https://github.com/cypgilet/.
[58] 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.
[59] 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.
[60] 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.
[61] S. C. AB, "Aps failure at scania trucks data set," 2016, https://www.kaggle.com/uciml/aps-failure-at-scania-trucks-data-set.
[62] 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.
[63] "Satellite database," https://www.openml.org/d/40900.
[64] R. Kerber, "Chimerge: Discretization of numeric attributes," *AAAI-92 Proceedings*, pp. 123–127, 1992.
[65] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," *IEEE, International Conference on tools with Artificial Intelligence*, 1995.
[66] A. L. Kurgan and K. J. Cios, "Caim discretization algorithm," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, pp. 145–153, 2004.
[67] 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.
[68] C. Gentile, S. Li, and G. Zappella, "Online clustering of bandits," in *Proceedings of Machine Learning Research*, E. P. Xing and T. Jebara, Eds., vol. 32, no. 2. Bejing, China: PMLR, 22–24 Jun 2014, pp. 757–765.
[69] S. Li, C. Gentile, and A. Karatzoglou, "Graph clustering bandits for recommendation," *arXiv:1605.00596*, 2016.
[70] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd ed. Chapman & Hall, 1990.
[71] F. Pedregosa, G. Varoquaux, A. Gramfort, and et al, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[72] J. Barranquero, P. González, J. Díez, and J. José Del Coz, "On the study of nearest neighbor algorithms for prevalence estimation in binary problems," *Pattern Recognition*, 2013.
[73] H. Kawakubo, M. Plessis, and M. Sugiyama, "Computationally efficient class-prior estimation under class balance change using energy distance," *IEICE Transactions on Information and Systems*, pp. 176–186, 01 2016.
[74] W. J. Reed, "Random points in a simplex," *Pacific J. Math.*, vol. 54, no. 2, pp. 183–198, 1974.
[75] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *Proceedings of the 36th International Conference on Machine Learning*, 2019.
[76] N. I. of Diabetes, Digestive, and K. Diseases, "Pima indians diabetes database," 1988, https://www.kaggle.com/uciml/pima-indians-diabetes-database.
[77] T. Menzies, J. Greenwald, and A. Frank, "Data mining static code attributes to learn defect predictors," *IEEE Transactions on Software Engineering*, vol. 32, 2007.
[78] 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.
[79] T. McCabe, "A complexity measure," *IEEE Transactions on Software Engineering*, vol. 32, 1976.
[80] L. Halstead, "Elements of software science," *Elsevier*, 1977.



Cyprien Gilet received an MSc degree in applied mathematics (Master MIGS, Dijon, France) in 2017, and he is currently a PhD student in Machine Learning at Université 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 physicians with the diagnosis of patients. 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 qualification in applied mathematics for biological sciences from Nova University in Lisbon (2006) and a PhD degree in tropical medicine from the University of Liverpool (2012). From 2013 to 2016 she worked as a postdoctoral researcher at the University of São Paulo in Brazil. Since 2017 she has been a postdoctoral 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 an MSc degree in decision and information engineering and a PhD 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 has been a full Professor at Université 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.

APPENDIX A
SYNTHETIC DATABASES GENERATION

A.1 Synthetic database for Figure 1

The results presented in Fig. 1 come from a synthetic dataset. This dataset was generated as follows: We considered $K = 2$ classes and $d = 3$ features. We generated $m = 20,000$ instances such that for each instance $i \in \mathcal{I}$, $Y_i \sim \text{Cat}(K, \hat{\pi})$ 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 $L_{11} = 3$, $L_{12} = 15$, $L_{21} = 25$, $L_{22} = 2$.

A.2 Synthetic database for Section 4

The synthetic database considered in Section 4 was generated as follows: We considered $K = 3$ classes {C1, C3, C3} and $d = 2$ features. We generated $m = 10,000$ instances such that for each instance $i \in \mathcal{I}$, $Y_i \sim \text{Cat}(K, \hat{\pi})$ with $\hat{\pi} = [0.8, 0.15, 0.05]$. For $j \in \{1, 2\}$, we generated the features X_{ij} as follows:

$$X_{ij} \;=\; \mathbb{1}_{\{Y_i=1\}} U_i \;+\; \mathbb{1}_{\{Y_i=2\}} V_i \;+\; \mathbb{1}_{\{Y_i=3\}} W_i \,,$$

with $U_i \sim \mathcal{N}(\mu_{1j}, \sigma_{1j})$, $V_i \sim \mathcal{N}(\mu_{2j}, \sigma_{2j})$, $W_i \sim \mathcal{N}(\mu_{3j}, \sigma_{3j})$ where

$$\mu = \begin{bmatrix} 9.5 & 10 \\ 10 & 12 \\ 11.7 & 7.6 \end{bmatrix}, \quad \sigma = \begin{bmatrix} 1 & 1.5 \\ 1.1 & 1.2 \\ 0.8 & 0.8 \end{bmatrix}.$$

With this aim, we used the algorithm *datasets.make_blobs* provided by Scikit-Learn [71]. The scatter plot of the database generated is provided in Fig. 5. For this database, we finally considered the $L_{0\text{-}1}$ loss function.

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

Let us recall 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 can therefore write $\mathbb{U}$ as

$$\mathbb{U} \;=\; \bigcap_{i=1}^{2K+2} U_i. \tag{30}$$

In [54], 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 (30).

APPENDIX C
PROOFS OF THE PAPER

C.1 Proof of Lemma 1

From (4), (5), (3) and (6) 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:

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

□

C.2 Proof of Theorem 1

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

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

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 (7), and for all $\delta \in \Delta$, we get

$$\hat{r}(\delta_{\hat{\pi}}) \;\ge\; \sum_{t \in \mathcal{T}} \sum_{l \in \hat{\mathcal{Y}}} \sum_{k \in \mathcal{Y}} L_{kl}\, \hat{\pi}_k\, \hat{p}_{kt}\, \mathbb{1}_{\{\lambda_{lt} = \min_{q \in \hat{\mathcal{Y}}} \lambda_{qt}\}}. \tag{31}$$

It follows that (31) is a lower bound of the empirical Bayes risk. It is straightforward to verify that the decision rule (8) achieves the lower bound (31). Hence, the classifier (8) minimizes (7), and its associated empirical Bayes risk is:

$$\hat{r}\left(\delta^B_{\hat{\pi}}\right) \;=\; \sum_{t\in\mathcal{T}}\sum_{l\in\hat{\mathcal{Y}}}\sum_{k\in\mathcal{Y}} L_{kl}\,\hat{\pi}_k\,\hat{p}_{kt}\,\mathbb{1}_{\{\lambda_{lt}=\min_{q\in\hat{\mathcal{Y}}}\lambda_{qt}\}}. \quad (32)$$

Finally, from (4) and (32), we identify the empirical class-conditional risk of class $k \in \mathcal{Y}$ as (9). □

C.3 Proof of Proposition 1

Let $\alpha \in [0,1]$ and let consider the priors $\pi, \pi', \pi'' \in \mathbb{S}$ 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 (21), 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 (20) 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 recall 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 us 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 [53] when considering the subgradient definition associated with a concave function, we can show that at iteration $N \geq 1$

$$\begin{aligned}
&V(\pi^\star) - \max_{n\leq N}\left\{V\left(\pi^{(n)}\right)\right\}\\
&\qquad\leq \frac{\left\|\pi^{(1)} - \pi^\star\right\|_2^2 + \sum_{n=1}^N \frac{\gamma_n^2}{\eta_n^2}\left\|g^{(n)}\right\|_2^2}{2\sum_{n=1}^N \frac{\gamma_n}{\eta_n}}. \quad (33)
\end{aligned}$$

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

$$\sum_{n=1}^N \frac{\gamma_n^2}{\eta_n^2}\left\|g^{(n)}\right\|_2^2 \;\leq\; \sum_{n=1}^N \gamma_n^2. \quad (34)$$

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,

$$\sum_{n=1}^N \frac{\gamma_n}{\eta_n} \;\geq\; \frac{1}{\max\{1, h(L)\}}\sum_{n=1}^N \gamma_n. \quad (35)$$

Finally, from (33), (34) and (35), we get (26). □

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

Supplementary Material

Database descriptions

Framingham Heart database: This database comes from the Framingham Heart study [58], and contains the clinical observations of 3,658 individuals (after removing individuals with missing values) who were followed for 10 years. The objective of the Framingham study was to predict the development of 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 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 developed CHD within 10 years. For this database, we considered the $L_{0\text{-}1}$ loss function.

Diabetes prediction database: Another example of the application of machine learning in the field of medicine is to predict the onset of diabetes based on diagnostic measurements. We consider here the database studied in [59] which was originally provided by the National Institute of Diabetes and Digestive and Kidney Diseases, and available at [76]. 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 class 2 corresponds to the patients who 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 [60]. This dataset is composed of 8 features (1 categorical and 7 numerical) from which the objective is to predict the age of each abalone. The initial ages to be predicted ranged from 1 to 29. For this experiment, we decided to consider $K = 5$ classes $\{A_1, A_2, A_3, A_4, A_5\}$ associated with the age groups $\{[\leq 4], [5, 10], [11, 15], [16, 20], [\geq 21]\}$ and with 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 farther the predicted class is from the true class, the more important this error is.

APS Failure Trucks database: This real condition monitoring database [61] focuses on the 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 use. 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:

$$L = \begin{bmatrix} 0 & 10 \\ 500 & 0 \end{bmatrix}, \tag{36}$$

so that the cost of predicting a nonexistent 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, of which 68,494 do not present any failure and 815 do 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 [62], [77]. More details on this database and on this task are given in [77] and [78]. For our experiments, we downloaded the data from `https://www.openml.org/d/1050`. This database is composed of 1,563 samples and 37 attributes measured with McCabe [79] and Halstead [80] "module"-based metrics. We have $K = 2$ classes, where class 2 corresponds to the defect programs. The class proportions $\hat{\pi} = [0.8976, 0.1024]$ are imbalanced, which complicates the task of detecting defective programs. For this database, we considered the $L_{0\text{-}1}$ loss function.

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