Minimax Classifier with Box Constraint on the Priors

Cyprien Gilet, Susana Barbosa, Lionel Fillatre

Minimax Classifier with Box Constraint on the Priors

Cyprien Gilet
University of Côte d'Azur
CNRS, I3S laboratory
Sophia-Antipolis, France
gilet@i3s.unice.fr

Susana Barbosa
University of Côte d'Azur
CNRS, IPMC laboratory
Sophia-Antipolis, France
sudocarmo@gmail.com

Lionel Fillatre
University of Côte d'Azur
CNRS, I3S laboratory
Sophia-Antipolis, France
lionel.fillatre@i3s.unice.fr

Abstract

Learning a classifier in safety-critical applications like medicine raises several issues. Firstly, the class proportions, also called priors, are in general imbalanced or uncertain. Sometimes, experts are able to provide some bounds on the priors and taking into account this knowledge can improve the predictions. Secondly, it is also necessary to consider any arbitrary loss function given by experts to evaluate the classification decision. Finally, the dataset may contain both categorical and numeric features. In this paper, we propose a box-constrained minimax classifier which addresses all the mentioned issues. To deal with both categorical and numeric features, many works have shown that discretizing the numeric attributes can lead to interesting results. Here, we thus consider that numeric features are discretized. In order to address the class proportions issues, we compute the priors which maximize the empirical Bayes risk over a box-constrained probabilistic simplex. This constraint is defined as the intersection between the simplex and a box constraint provided by experts, which aims at bounding independently each class proportions. Our approach allows to find a compromise between the empirical Bayes classifier and the standard minimax classifier, which may appear too pessimistic. The standard minimax classifier, which has not been studied yet when considerring discrete features, is still accessible by our approach. When considering only discrete features, we show that, for any arbitrary loss function, the empirical Bayes risk, considered as a function of the priors, is a concave non-differentiable multivariate piecewise affine function. To compute the box-constrained least favorable priors, we derive a projected subgradient algorithm. The convergence of our algorithm is established. The performance of our algorithm is illustrated with experiments on the Framingham study database to predict the risk of Coronary Heart Disease (CHD).

1 Introduction

Context and problem statement The task of supervised classification is becoming increasingly promising in several real applications such as medical diagnosis, condition monitoring, or fraud detection. However, in such applications, we often have to face many difficulties. Firstly, the training set is generally imbalanced, i.e., the classes are not equally represented. In this case, minimizing the empirical risk leads the classifier to minimize the class-conditional risks of the classes with the largest number of samples. A minority class with just a small number of occurrences will tend to have a large class-conditional risk [10]. Furthermore, when some classes contain only a small number of samples, we can not claim that the class proportions of the training set correspond to the true state of nature. A classifier fitted on such a training set may have a poor performance on the test set [25]. Sometimes, experts in the application domain are generally able to provide us with some bounds on the class proportions. For example, in case of a medical disease, it is reasonable to bound the maximum frequency of a given disease. We can expect this bound to improve the performance of a classifier.

Preprint. Under review.

Secondly, the experts can require the use of a specific loss function for evaluating the classification decisions. For example, if the classifier confuses a throat infection with a cold, the consequences are not the same as confusing a throat infection with a lung cancer. Finally, we often have to deal with both numeric and categorical features. Many works have shown that the discretization of the numeric features can lead to results with better accuracy [7, 24, 32, 14, 22]. In this paper, we consider that the numeric attributes are discretized such that the classifier must only process discrete features. The goal of this paper is to build a classifier which addresses all these mentioned issues.

Related works A common approach to deal with imbalanced datasets is to balance the data by resampling the training set. But this approach may increase the misclassification risk when classifying some test samples which are imbalanced. Another common approach is the cost sensitive learning [3, 8] which aims at optimizing the cost of class misclassifications in order to counterbalance the number of occurrences of each class. However, this approach transforms the loss function provided by the experts, and these costs are generally difficult to tune. The task of learning the class-proportions which maximize the minimum empirical risk was already studied in past years. A pioneering work on the minimax criterion in the field of machine learning is [5]. This work studies the generalization error of a minimax classifier but it does not give any method to compute it. In [18], the authors proposed the Minimum Error Minimax Probability Machine for the task of binary classification only. The extension to multiple classes is difficult. This method is very close to [17]. The Support Vector Machine (SVM) classifier can also be tuned in order to minimize the maximum class-conditional risks. The study proposed in [6] is limited to the linear classifiers (using or not a feature mapping) and to the classification problems between only two classes. In [11], the authors proposed an approach which fits a decision rule by learning the probability distribution which minimizes the worst-case of misclassification over a set of distributions centered at the empirical distribution. When the class-conditional distributions of the training set belong to a known parametric family of probability distributions, the competitive minimax approach can be an interesting solution [12]. Finally, in [15], the authors proposed an interesting fixed-point algorithm based on generalized entropy and strict sense Bayesian loss functions. This approach alternates a resampling step of the learning set with an evaluation step of the class-conditional risk, and it leads to estimate the least-favorable priors. However, the fixed-point algorithm needs the minimax rule to be an equalizer rule. We can show that this assumption is in general not satisfied when considering discrete features. Moreover, when working with small datasets, some priors are not accessible when considering only the samples from the training set. Therefore, it is not always possible to re-sample the training set at each iteration.

Contributions In this paper, we propose a new method for computing the minimax classifier addressing all the previously mentioned issues. It is well known that the usual minimax classifier aims at finding the priors which maximize the minimum empirical risk over the probabilistic simplex [25]. These class proportions are called the least favorable priors. However, as discussed in [1], it appears that sometimes a minimax classifier can be too pessimistic since its associated least favorable priors might be too far from the state of nature, and the risk of misclassifications becomes too high. In this case, our approach is suitable to consider some box constraints on the priors in order to find an acceptable trade-off between addressing the priors issues and satisfying an acceptable risk. The resulting decision rule is the box-constrained minimax classifier. The contributions of the paper are the following. First, we calculate the optimal minimum empirical risk of the training set, also called the empirical Bayes risk. Second, we show that the empirical Bayes risk is a non-differentiable concave multivariate piecewise affine function with respect to the priors. The box-constrained minimax classifier is obtained by seeking at the maximum of the empirical Bayes risk over the box-constrained region. Third, we derive a projected subgradient algorithm which finds the least favorable proportions over the box-constrained simplex. In section 2, we present the box-constrained minimax classifier. In section 3, we study the empirical Bayes risk. Section 4 proposes an optimization algorithm to compute the box-constrained minimax classifier. Section 5 proposes some numerical experiments on the Framingham Heart study dataset [30]. Finally, Section 6 concludes the paper.

2 Principle of box-constrained minimax classifier

Given $K \geq 2$ the number of classes, let $\mathcal{Y} = \{1, \ldots, K\}$ be the set of class labels and $\hat{\mathcal{Y}} = \mathcal{Y}$ the predicted labels. Let $\mathcal{X}$ be the space of all feature values. Let $L : \mathcal{Y} \times \hat{\mathcal{Y}} \to [0, +\infty)$ be the loss function such that, for all $(k, l) \in \mathcal{Y} \times \hat{\mathcal{Y}}$, $L(k, l) := L_{kl}$ corresponds to the loss, or the

cost, of predicting the class l whereas the real class is k. For example, the $L_{0\text{-}1}$ loss function is defined by $L_{kk} = 0$ and $L_{kl} = 1$ when $k \neq l$. Given a multiset $\{(Y_i, X_i), i \in \mathcal{I}\}$ containing a number m of labeled learning samples, the task of supervised classification is to learn a decision rule $\delta : \mathcal{X} \to \hat{\mathcal{Y}}$ which assigns each sample $i \in \mathcal{I}$ to a class $\hat{Y}_i \in \hat{\mathcal{Y}}$ from its feature vector $X_i := [X_{i1}, \ldots, X_{id}] \in \mathcal{X}$ composed of d observed features, and such that δ minimizes the empirical risk $\hat{r}(\delta) = \frac{1}{m} \sum_{i \in \mathcal{I}} L(Y_i, \delta(X_i))$ [31, 16, 9]. As explained in [25], this risk can be written as

$$\hat{r}\left(\delta_{\hat{\pi}}\right) \;=\; \sum_{k \in \mathcal{Y}} \hat{\pi}_k \, \hat{R}_k\left(\delta_{\hat{\pi}}\right), \tag{1}$$

where $\hat{\pi} = [\hat{\pi}_1, \ldots, \hat{\pi}_K]$ corresponds to the class proportions of the training set satisfying, for all $k \in \mathcal{Y}$, $\hat{\pi}_k = \frac{1}{m} \sum_{i \in \mathcal{I}} \mathbb{1}_{\{Y_i = k\}}$, [1] and where $\hat{R}_k\left(\delta_{\hat{\pi}}\right)$ corresponds to the empirical class-conditional risk associated to class k defined as

$$\hat{R}_k\left(\delta_{\hat{\pi}}\right) \;=\; \sum_{l \in \hat{\mathcal{Y}}} L_{kl} \, \hat{\mathbb{P}}(\delta_{\hat{\pi}}(X_i) = l \mid Y_i = k). \tag{2}$$

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

2.1 Minimax classifier principle

Let $\{(Y'_i, X'_i), i \in \mathcal{I}'\}$ be the multiset containing a number m' of test samples satisfying the unknown class proportions $\pi' = [\pi'_1, \ldots, \pi'_K]$. The classifier $\delta_{\hat{\pi}}$ fitted with the samples $\{(Y_i, X_i), i \in \mathcal{I}\}$ is then used to predict the classes Y'_i of the test samples $i \in \mathcal{I}'$ from their associated features $X'_i \in \mathcal{X}$. As described in [25], the risk of misclassification with respect to the classifier $\delta_{\hat{\pi}}$ and as a function of π' is defined by

$$\hat{r}\left(\pi', \delta_{\hat{\pi}}\right) = \sum_{k \in \mathcal{Y}} \pi'_k \hat{R}_k\left(\delta_{\hat{\pi}}\right). \tag{3}$$

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

$$\hat{r}\left(\pi', \delta_{\hat{\pi}}\right) \;=\; \pi'_1 \, \hat{R}_1\left(\delta_{\hat{\pi}}\right) + \pi'_2 \, \hat{R}_2\left(\delta_{\hat{\pi}}\right) = \pi'_1 \left(\hat{R}_1\left(\delta_{\hat{\pi}}\right) - \hat{R}_2\left(\delta_{\hat{\pi}}\right)\right) + \hat{R}_2\left(\delta_{\hat{\pi}}\right). \tag{4}$$

It is then clear that $\hat{r}\left(\pi', \delta_{\hat{\pi}}\right)$ is a linear function of π'_1. It is easy to verify that the maximum value of $\hat{r}\left(\pi', \delta_{\hat{\pi}}\right)$ is $M(\delta_{\hat{\pi}}) := \max\{\hat{R}_1\left(\delta_{\hat{\pi}}\right), \hat{R}_2\left(\delta_{\hat{\pi}}\right)\}$. Since $M(\delta_{\hat{\pi}})$ is larger than $\hat{r}\left(\pi', \delta_{\hat{\pi}}\right)$, it involves that the risk of the classifier can change significantly when π' differs from $\hat{\pi}$. More generally, for K classes, the maximum risk which can be attained by a classifier when π' is unknown is $M(\delta_{\hat{\pi}}) := \max\{\hat{R}_1\left(\delta_{\hat{\pi}}\right), \ldots, \hat{R}_K\left(\delta_{\hat{\pi}}\right)\}$. Hence, a solution to make a decision rule $\delta_{\hat{\pi}}$ robust with respect to the class proportions π' is to fit $\delta_{\hat{\pi}}$ by minimizing $M(\delta_{\hat{\pi}})$. As explained in [25], this minimax problem is equivalent to consider the following optimization problem:

$$\delta^B_{\hat{\pi}} \;=\; \operatorname*{argmin}_{\delta \in \Delta} \max_{\pi \in \mathbb{S}} \hat{r}(\pi, \delta_\pi) \;=\; \operatorname*{argmin}_{\delta \in \Delta} \max_{\pi \in \mathbb{S}} \hat{r}(\delta_\pi). \tag{5}$$

In [13], the famous Minimax Theorem establishes that

$$\min_{\delta \in \Delta} \max_{\pi \in \mathbb{S}} \hat{r}(\delta_\pi) \;=\; \max_{\pi \in \mathbb{S}} \min_{\delta \in \Delta} \hat{r}(\delta_\pi). \tag{6}$$

This theorem holds because our classification problem involves only discrete features. In the following, given $\pi \in \mathbb{S}$, we define $\delta^B_\pi := \operatorname{argmin}_{\delta \in \Delta} \hat{r}(\delta_\pi)$ as the optimal Bayes classifier for a given prior π. Hence, according to (6), provided that we can calculate δ^B_π for any $\pi \in \mathbb{S}$, the optimization problem (5) is equivalent to calculate the least favorable priors $\bar{\pi} := \operatorname{argmax}_{\pi \in \mathbb{S}} \hat{r}(\delta^B_\pi)$. The minimax classifier $\delta^B_{\bar{\pi}}$ is the Bayes classifier calculated with the prior $\bar{\pi}$.

[1] The indicator function of event E is denoted $\mathbb{1}_{\{E\}}$.

2.2 Benefits of Box-constrained minimax classifier

Sometimes, the minimax classifier may appear too pessimistic since the least favorable priors $\bar{\pi}$ may be too far from the priors $\hat{\pi}$ of the training set, and experts may consider that the class proportions $\bar{\pi}$ is unrealistic. For example in Figure 1, right, let suppose that the proportions of class 1 are 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 for 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 $\delta^B_{\bar{\pi}}$ is more robust and the box-constrained minimax classifier $\delta^B_{\pi^\star}$ has no loss. If we look now at the point a_1, the minimax classifier is disappointing but the loss of the box-constrained minimax classifier is still acceptable. In other words, the box-constrained minimax classifier seems to provide us with a reasonable trade-off between the loss of performance and the robustness to the prior change. To our knowledge, the concept of box-constrained minimax classifier has not been studied yet. More generally, in the case where we bound independently each class proportion, we therefore consider the box-constrained simplex

$$\mathbb{U} := \mathbb{S} \cap \mathbb{B}, \tag{7}$$

where $\mathbb{B} := \{\pi \in \mathbb{R}^K : \forall k = 1, \ldots, K,\ 0 \leq a_k \leq \pi_k \leq b_k \leq 1\}$ is the box constraint which delimits independently each class proportion. Hence, to compute the box-constrained minimax classifier with respect to $\mathbb{B}$, we consider the minimax problem $\delta^B_{\pi^\star} = \operatorname{argmin}_{\delta \in \Delta} \max_{\pi \in \mathbb{U}} \hat{r}(\delta_\pi)$, and according to (6), provided that we can calculate δ^B_π for any $\pi \in \mathbb{U}$, this problem leads to the optimization problem

$$\pi^\star \ = \ \arg\max_{\pi \in \mathbb{U}} \hat{r}(\delta^B_\pi). \tag{8}$$

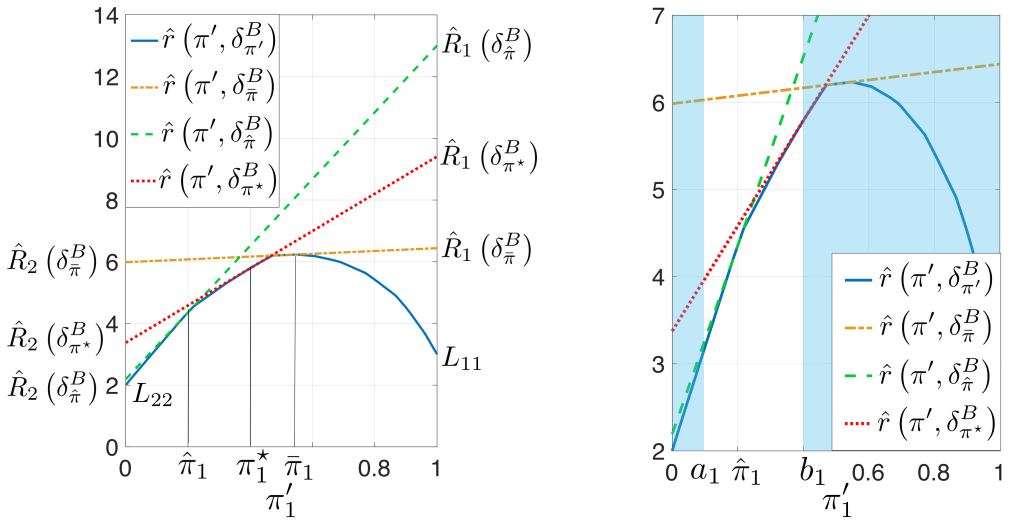

Figure 1: Comparison between the empirical Bayes classifier $\delta^B_{\hat{\pi}}$, the minimax classifier $\delta^B_{\bar{\pi}}$ and the box-constrained minimax classifier $\delta^B_{\pi^\star}$. Let us note that these results come from a synthetic dataset for which $K = 2$ classes. The generation of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

3 Discrete empirical Bayes risk

This section defines the empirical Bayes risk and studies its behavior as a function of the priors.

3.1 Empirical Bayes risk for the training set prior

For all $k \in \mathcal{Y}$, let $\mathcal{I}_k = \{i \in \mathcal{I} : Y_i = k\}$ be the set of learning samples from the class k, and $m_k = |\mathcal{I}_k|$ the number of samples in $\mathcal{I}_k$. Thus with these notations and in link with (2), we can write

$$\hat{\mathbb{P}}(\delta_{\hat{\pi}}(X_i) = l \mid Y_i = k) \ = \ \frac{1}{m_k} \sum_{i \in \mathcal{I}_k} \mathbb{1}_{\{\delta_{\hat{\pi}}(X_i) = l\}}. \tag{9}$$

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

$$\hat{p}_{kt} = \frac{1}{m_k} \sum_{i \in \mathcal{I}_k} \mathbb{1}_{\{X_i = x_t\}} \tag{10}$$

the probability estimate of observing the features profile $x_t \in \mathcal{X}$ with the class label k. In the context of statistical hypothesis testing theory, [29] 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 $\delta_{\hat{\pi}} \in \Delta$ with discrete features in the context of machine learning.

Lemma 1. *The empirical risk $\hat{r}\left(\delta_{\hat{\pi}}\right)$ of a decision rule $\delta_{\hat{\pi}} \in \Delta$ fitted on the train dataset is*

$$\hat{r}\left(\delta_{\hat{\pi}}\right) \;=\; \sum_{t \in \mathcal{T}} \sum_{k \in \mathcal{Y}} \sum_{l \in \hat{\mathcal{Y}}} L_{kl}\, \hat{\pi}_k\, \hat{p}_{kt}\, \mathbb{1}_{\{\delta_{\hat{\pi}}(x_t) = l\}}. \tag{11}$$

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

Let us note that the performance of any classifier δ trained on the learning dataset depends only on the probabilities $\hat{p}_{kt}$ and the priors $\hat{\pi}_k$. In this sense the set of values $\{\hat{p}_{kt}, \hat{\pi}_k\}$ can be viewed as an exhaustive statistics of the training dataset. The empirical Bayes classifier which minimizes $\hat{r}\left(\delta_{\hat{\pi}}\right)$ on the train dataset is given in the following Theorem.

Theorem 1. *The empirical Bayes classifier $\delta^B_{\hat{\pi}}$, fitted on the training set satisfying the class proportions $\hat{\pi} \in \mathbb{S}$, which minimizes over Δ the empirical risk $\hat{r}\left(\delta_{\hat{\pi}}\right)$, is*

$$\delta^B_{\hat{\pi}} : X_i \mapsto \underset{l \in \hat{\mathcal{Y}}}{\arg\min} \sum_{t \in \mathcal{T}} \sum_{k \in \mathcal{Y}} L_{kl}\, \hat{\pi}_k\, \hat{p}_{kt}\, \mathbb{1}_{\{X_i = x_t\}}. \tag{12}$$

Its associated empirical Bayes 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 $k \in \mathcal{Y}$, the empirical class-conditional risk associated to class k is

$$\hat{R}_k\left(\delta^B_{\hat{\pi}}\right) \;=\; \sum_{t \in \mathcal{T}} \sum_{l \in \hat{\mathcal{Y}}} L_{kl}\, \hat{p}_{kt}\, \mathbb{1}_{\{\lambda_{lt} = \min_{q \in \hat{\mathcal{Y}}} \lambda_{qt}\}}, \tag{13}$$

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 D.2. □

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

3.2 Empirical Bayes risk extended to any prior over the simplex

Since we can only consider the samples from the training set, the probabilities $\hat{p}_{kt}$ defined in (10) are assumed to be estimated once for all. Indeed, the statistical estimation theory [26] has established that the estimates $\hat{p}_{kt}$ correspond to the maximum likelihood estimates of the true probabilities p_{kt} for all couples $(k, t) \in \mathcal{Y} \times \mathcal{T}$. By estimating these probabilities with the full training set, we get the best unbiased estimate with the smallest variance. This paper assumes that these class-conditional probabilities are representative of the test set. However, as explained in Section 2, we can not be confident in the class proportions estimate $\hat{\pi}_k$. They are probably biased by the data collection. For this reason, the empirical Bayes risk must be viewed as a function of the class proportions.

Let us denote δ^B_{π} the empirical Bayes classifier fitted on a training set with the class proportions $\pi \in \mathbb{S}$, keeping unchanged the class-conditional probabilities $\hat{p}_{kt}$:

$$\delta^B_{\pi} : X_i \mapsto \underset{l \in \hat{\mathcal{Y}}}{\arg\min} \sum_{t \in \mathcal{T}} \sum_{k \in \mathcal{Y}} L_{kl}\, \pi_k\, \hat{p}_{kt}\, \mathbb{1}_{\{X_i = x_t\}}. \tag{14}$$

From Theorem 1, it follows that the minimum empirical Bayes risk for any prior π is given by the function $V : \mathbb{S} \mapsto [0, 1]$ defined by

$$V(\pi) \;:=\; \hat{r}\left(\delta_\pi^B\right) \;=\; \sum_{k\in\mathcal{Y}} \pi_k \hat{R}_k\left(\delta_\pi^B\right) \tag{15}$$

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

$$\hat{R}_k\left(\delta_\pi^B\right) = \sum_{t\in\mathcal{T}} \sum_{l\in\hat{\mathcal{Y}}} L_{kl}\, \hat{p}_{kt}\, \mathbb{1}_{\left\{\sum_{k\in\mathcal{Y}} L_{kl}\, \pi_k\, \hat{p}_{kt} = \min_{q\in\hat{\mathcal{Y}}} \sum_{k\in\mathcal{Y}} L_{kq}\, \pi_k\, \hat{p}_{kt}\right\}}. \tag{16}$$

The function $V : \pi \mapsto V(\pi)$ gives the minimum value of the empirical Bayes risk when the class proportions are π and the class-conditional probabilities $\hat{p}_{kt}$ remain unchanged. In other words, a classifier can be said robust to the priors if its risk remains very close to $V(\pi)$ whatever the value of $\pi \in \mathbb{S}$.

It is well known in the literature [25, 9] that the Bayes risk, as a function of the priors, is concave over the probabilistic simplex $\mathbb{S}$. The following proposition shows that this result holds when considering the empirical Bayes risk (15). Let us note that all the results are given for $\pi \in \mathbb{S}$, but they also hold over the box-constrained probabilistic simplex $\mathbb{U}$ since $\mathbb{U} \subset \mathbb{S}$.

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

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

Then, the following proposition and its corollary show that the minimum empirical risk V is not differentiable provided that exist $\pi, \pi' \in \mathbb{S}$ and $k \in \mathcal{Y}$ such that $\hat{R}_k\left(\delta_\pi^B\right) \neq \hat{R}_k\left(\delta_{\pi'}^B\right)$. When this condition is not satisfied, it means that all the class conditional risks are equal whatever the prior, and in other words, that the empirical Bayes risk is an affine function over the simplex.

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 D.4. □

Corollary 1. *If there exist $\pi, \pi' \in \mathbb{S}$ and $k \in \mathcal{Y}$ such that $\hat{R}_k\left(\delta_\pi^B\right) \neq \hat{R}_k\left(\delta_{\pi'}^B\right)$, then V is non-differentiable over the simplex $\mathbb{S}$.*

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

According to (15), the optimization problem (8) is equivalent to the optimization problem

$$\pi^\star \;=\; \arg\max_{\pi\in\mathbb{U}} V(\pi). \tag{17}$$

Since $V : \pi \mapsto V(\pi)$ is non-differentiable over $\mathbb{U}$ provided that there exist $\pi, \pi' \in \mathbb{U}$ and $k \in \mathcal{Y}$ such that $\hat{R}_k\left(\delta_\pi^B\right) \neq \hat{R}_k\left(\delta_{\pi'}^B\right)$, it is necessary to develop an optimization algorithm adapted to both the non-differentiability of V and the domain $\mathbb{U}$.

4 Maximization over the box-constrained probabilistic simplex

We are interested in solving the optimization problem (17). In order to compute the least favorable priors $\pi^\star$ which maximize V over the box-constrained simplex $\mathbb{U}$ in the general case where V is non-differentiable, we propose to use a projected subgradient algorithm based on [2] and following the scheme

$$\pi^{(n+1)} = \mathrm{P}_{\mathbb{U}}\left(\pi^{(n)} + \frac{\gamma_n}{\eta_n}\, g^{(n)}\right). \tag{18}$$

In (18), at each iteration n, $g^{(n)}$ denotes a subgradient of V at $\pi^{(n)}$, γ_n denotes the sub-gradient step, $\eta_n = \max\{1, \|g^{(n)}\|_2\}$, and $\mathrm{P}_{\mathbb{U}}$ denotes the projection onto the box-constrained simplex $\mathbb{U}$. Let us note that this algorithm also holds in the particular case where the hypothesis "for all $(\pi, \pi', k) \in \mathbb{U} \times \mathbb{U} \times \mathcal{Y}, \hat{R}_k\left(\delta_\pi^B\right) = \hat{R}_k\left(\delta_{\pi'}^B\right)$" is satisfied, i.e. 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 $\hat{R}\left(\delta_\pi^B\right) := \left[\hat{R}_1\left(\delta_\pi^B\right), \ldots, \hat{R}_K\left(\delta_\pi^B\right)\right] \in \mathbb{R}^K$ composed by all the class-conditional risks is a subgradient of the empirical Bayes risk $V : \pi \mapsto V(\pi)$ at the point π.*

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

In the following, we choose $\hat{R}\left(\delta_\pi^B\right) = \left[\hat{R}_1\left(\delta_\pi^B\right), \ldots, \hat{R}_K\left(\delta_\pi^B\right)\right]$ as subgradient in (18). The following theorem establishes the convergence of the iterates (18) 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*

$$\inf_{n\geq 1} \gamma_n > 0, \quad \sum_{n=1}^{+\infty} \gamma_n^2 < +\infty, \quad \sum_{n=1}^{+\infty} \gamma_n = +\infty, \tag{19}$$

the sequence of iterates following the scheme (18) converges to a solution $\pi^\star$ of (17), whatever the initialization $\pi^{(1)} \in \mathbb{S}$.

Proof. The proof is a consequence of Theorem 1 in [2]. Note that here we have the strong convergence since $\pi^{(n)}$ belongs to a finite dimensional space as mentioned in the introduction of [2]. □

Remark 1. *It is worth noting that, when the empirical Bayes risk V is not zero everywhere, the subgradient $\hat{R}\left(\delta_{\pi^\star}^B\right)$ at the box-constrained minimax optimum does not vanish, otherwise the associated risk $V(\pi^\star)$ would be null too. This would be a contradiction with the fact that $\pi^\star$ is a solution of (17). Hence, the sequence $(\pi^{(n)})_{n\geq 1}$ generated by (18) when $g^{(n)} = \hat{R}\left(\delta_{\pi^{(n)}}^B\right)$ at each step is infinite.*

When considering the general case where V is not uniformly null over $\mathbb{S}$, according to Remark 1, we need a stopping criterion since the sequence $(\pi^{(n)})_{n\geq 1}$ generated by (18) is infinite when $g^{(n)} = \hat{R}\left(\delta_{\pi^{(n)}}^B\right)$ at each step. We propose to follow the reasoning in [4] which shows that the difference between the box-constrained minimax risk and the worst empirical Bayes risk computed until the iteration N is bounded by:

$$\left| \max_{n\leq N} \left\{ V\left(\pi^{(n)}\right) \right\} - V(\pi^\star) \right| \leq \frac{\rho^2 + \sum_{n=1}^N \gamma_n^2}{2\sum_{n=1}^N \gamma_n}, \tag{20}$$

where ρ is a constant satisfying $\|\pi^{(1)} - \pi^\star\|_2 \leq \rho$. We propose to choose $\rho^2 = K$ since all the proportions belong to the probabilistic simplex. Since (20) converges to 0 as $N \to \infty$, we can choose a small tolerance $\varepsilon > 0$ as a stopping criterion: we fix ε and, then, we compute $N = N_\varepsilon$ such that the bound in (20) is smaller than ε.

When considering the sequence of iterates (18), we need to compute the exact projection onto the box-constrained probabilistic simplex $\mathbb{U}$ at each iteration n. To perform this projection, we propose to consider the algorithm provided by [28], which computes the exact projection onto polyhedral sets in Hilbert spaces. In Appendix B, we show how to apply this exact projection to our box-constrained simplex $\mathbb{U}$. Finally, the procedure for computing the box-constrained minimax classifier is summarized in the step by step Algorithm 1 in Appendix C.

5 Numerical experiments

Dataset description For illustrating the interest of our box-constrained minimax classifier in medicine, we applied our algorithm to the Framingham Heart database [30]. This database contains the clinical observations of 3,658 individuals (after removing individuals with missing values) who have been followed for 10 years. The objective of the Framingham study was to predict the development of a Coronary Heart Disease (CHD) within 10 years based on $d = 15$ observed features measured at inclusion. We therefore have K = 2 classes, with class 2 corresponding to individuals who have developed a CHD, and class 1 corresponding to the others. Among the 15 features, 7 are categorical (*sex, education, smoking status, previous history of stroke, diabetes, hypertension, antihypertensive treatment*) and 8 are numeric (*age, number of cigarettes per day, cholesterol levels, systolic blood pressure, diastolic blood pressure, heart rate, body mass index (BMI), glycemia*). The dataset is imbalanced: $\hat{\pi} = [0.85, 0.15]$, which means that 15% of the individuals have developed a CHD within 10 years. For this experiment, we considered the $L_{0\text{-}1}$ loss function.

Features discretization In order to apply our algorithm, we need to discretize the numerical features. To this aim, many methods can be applied as explained in [7, 24]. We can use supervised discretization methods such as [19, 21, 20], or unsupervised methods such as the Kmeans algorithm [23]. Here we decided to quantize the features using the Kmeans 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 is quantized with the index of the centroid closest to it, i.e., $Q(X_i) = j$ where $Q : \mathbb{R}^d \mapsto \{1, \ldots, T\}$ denotes the k-means quantizer and j is the index of the centroid of the cluster in which X_i belongs to. The choice of T is important since it has an impact on the generalization error. It was established from a 10-sub-fold cross-validation over the main training set, and such that the generalization error computed over the validation set, as a function of T, should not exceed the training error by more than 1%. An example of this procedure is given in Figure 3, left.

Box-constraint generation In order to illustrate the benefits of the box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$ compared to the minimax classifier $\delta^{\mathrm{B}}_{\bar{\pi}}$ 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 $\beta \in [0, 1]$,

$$\mathbb{B}_\beta = \left\{\pi \in \mathbb{R}^K : \forall k \in \mathcal{Y}, \quad \hat{\pi}_k - \rho_\beta \leq \pi_k \leq \hat{\pi}_k + \rho_\beta\right\}, \quad \rho_\beta := \beta \left\|\hat{\pi} - \bar{\pi}\right\|_\infty. \tag{21}$$

Our box-constrained probabilistic simplex is therefore $\mathbb{U}_\beta = \mathbb{S} \cap \mathbb{B}_\beta$. Thus, when $\beta = 0$, $\mathbb{B}_0 = \{\hat{\pi}\}$, $\mathbb{U}_0 = \{\hat{\pi}\}$ and $\pi^\star = \hat{\pi}$. When $\beta = 1$, $\bar{\pi} \in \mathbb{B}_1$, hence $\bar{\pi} \in \mathbb{U}_1$ and $\pi^\star = \bar{\pi}$. For the next experiment, after having estimated the proportions $\hat{\pi}$ and $\bar{\pi}$ over the main dataset, we chose $\beta = 0.5$ which results that $\mathbb{B}_{0.5} = \{\pi \in \mathbb{R}^2 : 0.68 \leq \pi_1 \leq 1,\ 0 \leq \pi_2 \leq 0.32\}$. In other words, we consider that the proportion of individuals who tend to develop a CHD should not exceed 32%. Let us note that here and in the following, the least favorable priors $\bar{\pi}$ were estimated using our box-constrained minimax algorithm when considering $\mathbb{B} = [0, 1] \times [0, 1]$, so that $\mathbb{U} = \mathbb{S}$. The minimax classifier is a particular case of the box-constraint minimax classifier.

Results We performed a 10-fold cross-validation and we applied our box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$ when considering the box $\mathbb{B}_{0.5}$ described above. We compared $\delta^{\mathrm{B}}_{\pi^\star}$ to the Logistic Regression $\delta^{\mathrm{LR}}_{\hat{\pi}}$, the Random Forest $\delta^{\mathrm{RF}}_{\hat{\pi}}$, the discrete Bayes classifier $\delta^{\mathrm{B}}_{\hat{\pi}}$ (12), and the minimax classifier $\delta^{\mathrm{B}}_{\bar{\pi}}$. We applied $\delta^{\mathrm{LR}}_{\hat{\pi}}$ and $\delta^{\mathrm{RF}}_{\hat{\pi}}$ to both the original dataset and the discretized dataset, in order to evaluate the impact of the discretization. We can observe in Figure 2 that the performances associated to $\delta^{\mathrm{LR}}_{\hat{\pi}}$ and $\delta^{\mathrm{RF}}_{\hat{\pi}}$ are similar when considering real or discretized features. And these performances are moreover similar to the discrete Bayes classifier $\delta^{\mathrm{B}}_{\hat{\pi}}$. However, when regarding the class conditional-risks, the classifiers $\delta^{\mathrm{LR}}_{\hat{\pi}}$, $\delta^{\mathrm{RF}}_{\hat{\pi}}$ and $\delta^{\mathrm{B}}_{\hat{\pi}}$ are not satisfying when predicting accurately the patients who tend to develop a CHD. To do so, it is rather preferable to consider our minimax classifier $\delta^{\mathrm{B}}_{\bar{\pi}}$, even if it appears globally too pessimistic. In the case where the global risk of $\delta^{\mathrm{B}}_{\bar{\pi}}$ is not acceptable, it is therefore preferable to reduce the box-constraint area and consider the box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$, which is a trade-off between $\delta^{\mathrm{B}}_{\hat{\pi}}$ and $\delta^{\mathrm{B}}_{\bar{\pi}}$. The box-constraint area has an impact on the results and this aspect is discussed in the next paragraph. Let us note that, for the training steps of this procedure, our algorithm computed $\bar{\pi} = [0.52 \pm 0.01, 0.58 \pm 0.01]$ and $\pi^\star = [0.68 \pm \mathrm{e}^{-3}, 0.32 \pm \mathrm{e}^{-3}]$ such as $V(\bar{\pi}) = 0.33 \pm 0.01$ and $V(\pi^\star) = 0.28 \pm 0.01$. Finally, the results associated to the test steps presented in Figure 2 were computed when considering each whole fold test set satisfying the class proportions $\pi' = \hat{\pi}$.

Changes in the priors of the test set In order to study the robustness of each classifier when the class proportions π' of the test set are uncertain, we uniformly generated 1,000 random priors $\pi^{(s)}$, $s \in \{1, \ldots, 1000\}$, over the box-constrained simplex $\mathbb{U}_{0.5}$ using the procedure [27]. For each repetition of the previous cross-validation, we generated 1000 test subsets satisfying one of the random priors $\pi^{(s)}$. Each fitted classifier was then tested when considering all the 1000 random priors uniformly dispersed over $\mathbb{U}_{0.5}$. In Figure 3, right, we observe that when the class proportions of the test set changed uniformly over $\mathbb{U}_{0.5}$, the minimax classifier $\delta^{\mathrm{B}}_{\bar{\pi}}$ was the most robust since the most stable, but it was also the most pessimistic contrary to the other classifiers. The box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$ appears here again as a trade-off between $\delta^{\mathrm{B}}_{\bar{\pi}}$ and $\delta^{\mathrm{B}}_{\hat{\pi}}$.

Impact of the Box-constraint area In order to measure the impact of the box-constraint area on $\delta^{B}_{\pi^\star}$, we resized the radius ρ_β of $\mathbb{B}_\beta$ in (21) by changing the value of β from 0 to 1. Let consider the function $\psi : \Delta \to \mathbb{R}^+$ such that

$$\psi(\delta) = \max_{k \in \mathcal{Y}} \hat{R}_k(\delta) - \min_{k \in \mathcal{Y}} \hat{R}_k(\delta), \tag{22}$$

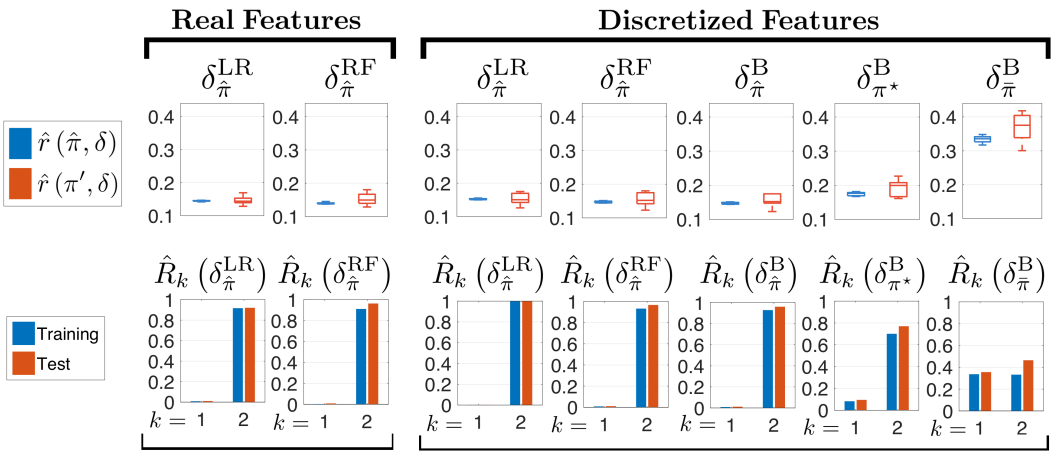

Figure 2: The boxplots (training versus test) illustrate the dispersion of the global risks of misclassification. The barplots correspond to the average class-conditional risk associated to each classifier.

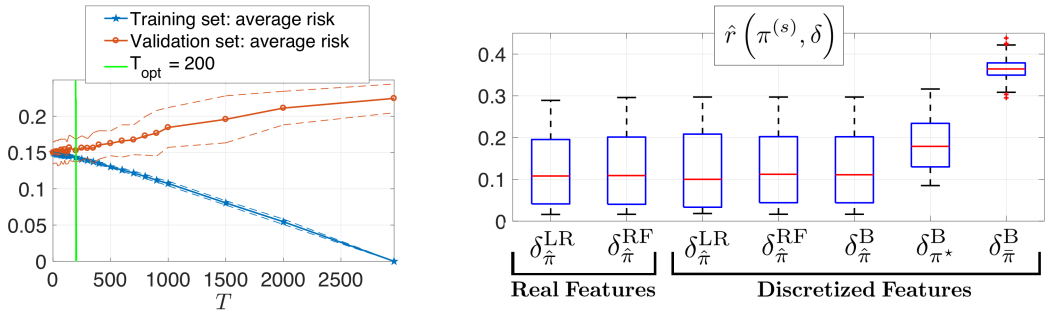

Figure 3: **Left.** Risks $\hat{r}\left(\hat{\pi}, \delta^B_{\hat{\pi}}\right)$ as a function of the number of centroids T. The dashed curves show the standard-deviation around the mean. **Right.** Evaluation of the robustness of each classifier when $\pi' = \pi^{(s)}$ changes over $\mathbb{U}_{0.5}$. Here, $\hat{r}(\pi^{(s)}, \delta)$ corresponds to the 10-fold cross-validation average risk associated to the test set satisfying the priors $\pi^{(s)} \in \mathbb{U}_{0.5}$, $s \in \{1, \ldots, 1000\}$.

which aims at measuring how equalizer a given classifier $\delta \in \Delta$ is. In Figure 4, left, we observe that when β increases from 0 to 1, $V(\pi^\star)$ increases from $V(\hat{\pi})$ to $V(\bar{\pi})$. At the same time, in Figure 4, right, when β increases from 0 to 1, $\psi\left(\delta^B_{\pi^\star}\right)$ decreases from $\psi\left(\delta^B_{\hat{\pi}}\right)$ to $\psi\left(\delta^B_{\bar{\pi}}\right)$. Hence, the larger the box-constraint area is, the more equalizer the classifier $\delta^B_{\pi^\star}$ is, but the more pessimist $\delta^B_{\pi^\star}$ becomes, since $V(\pi^\star)$ becomes much bigger than $V(\hat{\pi})$. In the case where $\delta^B_{\pi^\star}$ appears globally too pessimistic, it would be rather interesting to reduce the box-constraint area in order to find a trade-off between equalizing the class conditional risks and decreasing the empirical risk $V(\pi^\star)$ close enough to $V(\hat{\pi})$. In other words, the box-constrained minimax classifier $\delta^B_{\pi^\star}$ allows to find a compromise between satisfying an acceptable global risk and minimizing the risk of missing the individuals who tend to develop a CHD.

6 Conclusion

This paper proposes a box-constrained minimax classifier which i) is robust to the imbalanced or uncertain class proportions, ii) includes some bounds on the class proportions, iii) can take into account any given loss function, and iv) is suitable for working on discrete/discretized features. In future work, we propose to investigate the robustness of the classifier with respect to the class-conditional probabilities $\hat{p}_{kt}$.

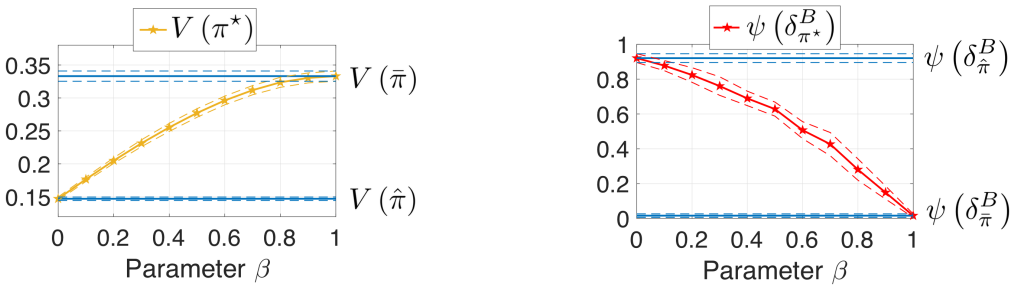

Figure 4: Impact of the box-constraint area on $\delta^{\mathrm{B}}_{\pi^\star}$ when β increases from 0 to 1 in (21), after a 10-fold cross-validation procedure. Results are presented as $\mathrm{mean} \pm \mathrm{std}$.

Acknowledgements

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

References

[1] Rocío Alaiz-Rodríguez, Alicia Guerrero-Curieses, and Jesús Cid-Sueiro. Minimax regret classifier for imprecise class distributions. *Journal of Machine Learning Research*, 8:103–130, Jan 2007.

[2] Ya. I. Alber, A. N. Iusem, and M. V. Solodov. On the projected subgradient method for nonsmooth convex optimization in a hilbert space. *Mathematical Programming*, 81(1):23–35, Mar 1998.

[3] Bernardo Ávila Pires, Csaba Szepesvari, and Mohammad Ghavamzadeh. Cost-sensitive multiclass classification risk bounds. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1391–1399, Atlanta, Georgia, USA, 17–19 Jun 2013.

[4] Stephen Boyd, Lin Xiao, and Almir Mutapcic. Lecture notes: Subgradient methods, stanford university, 2003. URL: `http://web.mit.edu/6.976/www/notes/subgrad_method.pdf`.

[5] Adam Cannon, James Howse, Don Hush, and Clint Scovel. Learning with the Neyman-Pearson and min-max criteria. *Los Alamos National Laboratory, Tech. Rep. LA-UR*, pages 02–2951, 2002.

[6] Mark A Davenport, Richard G Baraniuk, and Clayton D Scott. Tuning support vector machines for minimax and Neyman-Pearson classification. *IEEE transactions on pattern analysis and machine intelligence*, 32(10):1888–1898, 2010.

[7] James Dougherty, Ron Kohavi, and Mehran Sahami. Supervised and unsupervised discretization of continuous features. *International Conference on Machine Learning*, 1995.

[8] Chris Drummond and Robert 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.

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

[10] Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'01, pages 973–978, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

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

[12] Meir Feder and Neri Merhav. Universal composite hypothesis testing: A competitive minimax approach. *IEEE Transactions on information theory*, 48(6):1504–1517, 2002.

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

[14] Salvador García, Julián Luengo, and Francisco Herrera. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems*, 98:1–29, 2016.

[15] 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*, 34(4):383–392, Nov 2004.

[16] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag New York, 2nd edition, 2009.

[17] Huang Kaizhu, Yang Haiqin, King Irwin, and R. Lyu Michael. Imbalanced learning with a biased minimax probability machine. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(4):913–923, Aug 2006.

[18] Huang Kaizhu, Yang Haiqin, King Irwin, R. Lyu Michael, and Laiwan Chan. The minimum error minimax probability machine. *Journal of Machine Learning Research*, page 1253–1286, 2004.

[19] Randy Kerber. Chimerge: Discretization of numeric attributes. *AAAI-92 Proceedings*, pages 123–127, 1992.

[20] A. Lukasz Kurgan and Krysztof J. Cios. Caim discretization algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 16:145–153, 2004.

[21] Huan Liu and Rudy Setiono. Chi2: Feature selection and discretization of numeric attributes. *IEEE, International Conference on tools with Artificial Intelligence*, 1995.

[22] Jonathan L. Lustgarten, Vanathi Gopalakrishnan, Himanshu Grover, and Shyam Visweswaran. Improving classification performance with discretization on biomedical datsets. *AMIA 2008 Symposium Proceedings*, pages 445–449, 2008.

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

[24] Liu Peng, Wang Qing, and Gu Yujia. Study on comparison of discretization methods. *IEEE, International Conference on Artificial Intelligence and Computational Intelligence*, pages 380–384, 2009.

[25] H. V. Poor. *An Introduction to Signal Detection and Estimation*. Springer-Verlag New York, 2nd edition, 1994.

[26] C. Radhakrishna Rao. *Linear Statistical Inference and its Applications*. Wiley, 1973.

[27] W. J. Reed. Random points in a simplex. *Pacific J. Math.*, 54(2):183–198, 1974.

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

[29] M.I. Schlesinger and Václav Hlavác. *Ten Lectures on Statistical and Structural Pattern Recognition*. Springer Netherlands, 1st edition, 2002.

[30] Boston University, the National Heart Lung, and Blood Institute. The framingham heart study, From 1948. Downloaded data: `https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset`.

[31] Vladimir Vapnik. An overview of statistical learning theory. *IEEE transactions on Neural Networks*, 10 5:988–99, 1999.

[32] Ying Yang and Geoffrey I. Webb. Discretization for naive-bayes learning: managing discretization bias and variance. *Machine Learning*, 74(1):39–74, Jan 2009.

Appendix

Minimax Classifier with Box Constraint on the Priors

Cyprien Gilet, Susana Barbosa, Lionel Fillatre

A Generation of the synthetic dataset for Figure 1

The results presented in Figure 1 come from a synthetic dataset. This dataset was generated as follow. We considered $K = 2$ classes and $d = 3$ features. We generated $m = 20,000$ samples such that for each sample $i \in \mathcal{I}$, $Y_i \sim \text{Cat}(K, \hat{\pi})$ with $\hat{\pi} = [0.2, 0.8]$. The categorical distribution, which is denoted as $\text{Cat}(K, \pi)$, 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 matrix containing the values L_{kl}:

$$L = \begin{bmatrix} 3 & 15 \\ 25 & 2 \end{bmatrix}.$$

B Projection onto the box-constrained simplex

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

$$\begin{aligned}
U_1 &= \left\{\pi \in \mathbb{R}^K : \langle \pi, e_1 \rangle \le b_1\right\} \\
&\vdots \\
U_K &= \left\{\pi \in \mathbb{R}^K : \langle \pi, e_K \rangle \le b_K\right\} \\
U_{K+1} &= \left\{\pi \in \mathbb{R}^K : \langle \pi, -e_1 \rangle \le -a_1\right\} \\
&\vdots \\
U_{2K} &= \left\{\pi \in \mathbb{R}^K : \langle \pi, -e_K \rangle \le -a_K\right\} \\
U_{2K+1} &= \left\{\pi \in \mathbb{R}^K : \langle \pi, \mathbf{1}_K \rangle \le 1\right\} \\
U_{2K+2} &= \left\{\pi \in \mathbb{R}^K : \langle \pi, -\mathbf{1}_K \rangle \le -1\right\}
\end{aligned} \tag{23}$$

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 a vector composed of ones. We therefore can write $\mathbb{U}$ as

$$\mathbb{U} \;=\; \bigcap_{i=1}^{2K+2} U_i. \tag{24}$$

In [28], 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 (24). Let us define $\{\eta_1, \dots, \eta_{2K+2}\}$ and $\{u_1, \dots, u_{2K+2}\}$ such that, for $i \in \{1, \dots, 2K+2\}$,

$$\eta_i = \begin{cases} b_i & \text{if} \quad i \in \{1, \dots, K\} \\ -a_{(i-K)} & \text{if} \quad i \in \{K+1, \dots, 2K\} \\ 1 & \text{if} \quad i = 2K+1 \\ -1 & \text{if} \quad i = 2K+2 \end{cases}, \quad u_i = \begin{cases} e_i & \text{if} \quad i \in \{1, \dots, K\} \\ -e_{(i-K)} & \text{if} \quad i \in \{K+1, \dots, 2K\} \\ \mathbf{1}_K & \text{if} \quad i = 2K+1 \\ -\mathbf{1}_K & \text{if} \quad i = 2K+2. \end{cases}$$

According to Theorem 1 in [28] and when considering (24), given $x \in \mathbb{R}^K$, the projection over the box-constrained region is

$$\text{P}_{\mathbb{U}}(x) \;=\; x - \sum_{i=1}^{2K+2} \nu_i u_i,$$

where the ν_i's satisfy the three conditions:

1. For all $i \in \{1, \ldots, 2K+2\}$, $\nu_i \geq 0$;
2. For all $i \in \{1, \ldots, 2K+2\}$,
$$\langle x, u_i \rangle - \eta_i - \sum_{j=1}^{2K+2} \nu_i \langle u_i, u_j \rangle \leq 0;$$
3. For all $i \in \{1, \ldots, 2K+2\}$,
$$\nu_i \left(\langle x, u_i \rangle - \eta_i - \sum_{j=1}^{2K+2} \nu_i \langle u_i, u_j \rangle \right) = 0.$$

In [28], the authors proposes an algorithm which allows to compute the ν_i's satisfying the three previous conditions in order to compute the exact projection of $x \in \mathbb{R}^K$ onto $\mathbb{U}$.

C Box-constrained minimax classifier Algorithm

The procedure for computing the box-constrained minimax classifier is summarized in the step by step algorithm 1. We choose the sequence of steps $(\gamma_n)_{n\geq 1} = (1/n)_{n\geq 1}$ which satisfies (19). Let us remind that, in the inputs, K is the number of classes, N is the number of iterations for performing (18). Finally, in Algorithm 1, $\mathrm{P}_{\mathbb{U}}$ denotes the procedure which allows to project onto $\mathbb{U}$ (see [28]).

Algorithm 1 Computation of the Box-constrained minimax classifier

Input: $(Y_i, X_i)_{i\in\mathcal{I}}$, K, N.
Compute $\pi^{(1)} = \hat{\pi}$
Compute the $\hat{p}_{kt}$'s as described in (10).
$r^\star \leftarrow 0$
$\pi^\star \leftarrow \pi^{(1)}$
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 (16)
  end for
  $r^{(n)} = \sum_{k=1}^K \pi_k^{(n)} g_k^{(n)}$ see (1)
  if $r^{(n)} > r^\star$ **then**
    $r^\star \leftarrow r^{(n)}$
    $\pi^\star \leftarrow \pi^{(n)}$
  end if
  $\gamma_n \leftarrow 1/n$
  $\eta_n \leftarrow \max\{1, \|g^{(n)}\|_2\}$
  $z^{(n)} \leftarrow \pi^{(n)} + \gamma_n\, g^{(n)}/\eta_n$
  $\pi^{(n+1)} \leftarrow \mathrm{P}_{\mathbb{U}}\left(z^{(n)}\right)$
end for
Output: $r^\star$,$\pi^\star$, and $\delta^B_{\pi^\star}$ provided by (14) with $\pi = \pi^\star$.

D Proofs of the paper

D.1 Proof of Lemma 1

From (1), (2), (9) and (10) it follows that:

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

The indicator function in the last equation can be rewritten as

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

Hence, we finally get:

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

□

D.2 Proof of Theorem 1

From Lemma 1, we get

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

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

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

The last inequality can be rewritten as

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

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

$$\hat{r}(\delta_{\hat{\pi}}) \geq \sum_{t\in\mathcal{T}}\sum_{l\in\hat{\mathcal{Y}}}\sum_{k\in\mathcal{Y}} L_{kl}\,\hat{\pi}_k\,\hat{p}_{kt}\,\mathbb{1}_{\{\lambda_{lt}=\min_{q\in\hat{\mathcal{Y}}}\lambda_{qt}\}}. \tag{25}$$

It follows that (25) is a lower bound of the empirical Bayes risk. It is straightforward to verify that the decision rule

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

achieves the lower bound (25). Hence, it minimizes (11). Its associated empirical Bayes risk is:

$$\hat{r}\left(\delta^B_{\hat{\pi}}\right) = \sum_{t\in\mathcal{T}}\sum_{l\in\hat{\mathcal{Y}}}\sum_{k\in\mathcal{Y}} L_{kl}\,\hat{\pi}_k\,\hat{p}_{kt}\,\mathbb{1}_{\{\lambda_{lt}=\min_{q\in\hat{\mathcal{Y}}}\lambda_{qt}\}}. \tag{26}$$

From (1) and (26), we identify the empirical class-conditional risk of class $k\in\mathcal{Y}$ as

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

□

D.3 Proof of Proposition 1

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

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

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

D.4 Proof of Proposition 2

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

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

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 (16), 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 (15) 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, for all $t \in \mathcal{T}$, we can show that $\pi' \in [\pi]$ if and only if $\delta^B_{\pi'}(x_t) = \delta^B_{\pi}(x_t)$. 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. □

D.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 by two affine pieces since it is impossible to have a single equivalence class. Hence, V is non-differentiable over the intersections of these pieces. □

D.6 Proof of Lemma 2

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

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

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