Box-constrained optimization for minimax supervised learning

Cyprien Gilet, Susana Barbosa, Lionel Fillatre

ESAIM: PROCEEDINGS AND SURVEYS, August 2021, Vol. 71, p. 101-113
Didier Auroux, Jean-Baptiste Caillau, Régis Duvigneau, Abderrahmane Habbal, Olivier Pantz, Luc Pronzato, Ludovic Rifford

BOX-CONSTRAINED OPTIMIZATION FOR MINIMAX SUPERVISED LEARNING *, **

Cyprien Gilet[1], Susana Barbosa[2] and Lionel Fillatre[3]

Abstract. In this paper, we present the optimization procedure for computing the discrete box-constrained minimax classifier introduced in [1, 2]. Our approach processes discrete or beforehand discretized features. A box-constrained region defines some bounds for each class proportion independently. The box-constrained minimax classifier is obtained from the computation of the least favorable prior which maximizes the minimum empirical risk of error over the box-constrained region. After studying the discrete empirical Bayes risk over the probabilistic simplex, we consider a projected subgradient algorithm which computes the prior maximizing this concave multivariate piecewise affine function over a polyhedral domain. The convergence of our algorithm is established.

Résumé. Nous présentons dans cet article le problème d'optimisation lié au calcul d'un classifieur minimax à contrainte de boîte, ainsi que l'algorithme permettant de calibrer ce classifieur, que nous avons introduit dans [1, 2]. Notre approche considère des variables descriptives discrètes ou préalablement discrétisées. Les contraintes de boîte définissent des bornes sur chaque proporition par classe. Le classifieur est calibré en calculant la distribution a priori qui maximise le risque d'erreur minimum sur le simplexe contraint par boîte. Après avoir montré que ce risque d'erreur minimum est une fonction concave affine par morceaux avec un nombre fini de faces sur le simplexe, nous considérons un algorithme de sous-gradient projeté pour calculer la distribution a priori qui maximise ce risque de Bayes discret sur un domaine polyhédral. La convergence de l'algorithme est démontrée.

1. Introduction

Supervised classification is becoming essential in several real applications such as medical diagnosis, condition monitoring, or fraud detection. However, in such applications, we often have to face the following difficulties: imbalanced class proportions, prior probability shifts, presence of both numeric and categorical features (mixed attributes), and dependencies between some features.

Context and notation. Given $K \geq 2$ classes and a set $\mathcal{S} = \{(Y_i, X_i), i \in \mathcal{I}\}$ of m labeled training samples, the objective in fitting a supervised classifier [3, 4] is to learn a decision rule $\delta : \mathcal{X} \to \mathcal{Y} := \{1, \ldots, K\}$ which assigns each sample $i \in \mathcal{I}$ to a class $\delta(X_i) \in \mathcal{Y}$ from its feature vector $X_i := [X_{i1}, \ldots, X_{id}] \in \mathcal{X}$ composed of d

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

[1] University of Côte d'Azur, CNRS, I3S laboratory, Sophia-Antipolis, France.
[2] University of Côte d'Azur, CNRS, laboratory IPMC, Sophia-Antipolis, France.
[3] University of Côte d'Azur, CNRS, I3S laboratory, Sophia-Antipolis, France.

Article published online by EDP Sciences and available at https://www.esaim-proc.org or https://doi.org/10.1051/proc/202171109

observed attributes, and such that δ minimizes the empirical risk of classification errors

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

where $L : \mathcal{Y} \times \hat{\mathcal{Y}} \to [0, +\infty)$ is the loss function such that, for all $(k, l) \in \mathcal{Y} \times \hat{\mathcal{Y}}$ with $\hat{\mathcal{Y}} = \{1, \dots, K\}$ the set of predicted labels, $L(k, l) := L_{kl}$ corresponds to the loss, or the cost, of predicting the class l whereas the real class is k. Let $\Delta := \{\delta : \mathcal{X} \to \hat{\mathcal{Y}}\}$ be the set of all possible classifiers.

Dealing with imbalanced datasets. The risk of classification errors (1) can be written as (see [5])

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

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

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

In (3), $\hat{\mathbb{P}}_{\mathcal{S}}(\delta(X_i) = l \mid Y_i = k)$ corresponds to the empirical probability for the decision rule δ to predict the class l given that the true class is k on the set $\mathcal{S}$ of samples. When the class proportions $\hat{\pi}$ are imbalanced (that is when the classes are not equally represented), and as a consequence of (2), most classifiers essentially focus on the dominating classes containing the largest number of training samples, and underestimate the least represented ones [6–9]. Hence, the task of well classifying the instances from the smallest classes is difficult, which leads the minority classes to have a large conditional risk (3).

Dealing with prior probability shifts. Prior probability shift [10, 11] characterizes an evolution in the distribution of the priors between the training set and the test samples. We will use the notation δ_π for precising that the classifier δ was fitted with the prior distribution π in the K-dimensional probabilistic simplex $\mathbb{S} := \{\pi \in [0, 1]^K : \sum_{k=1}^K \pi_k = 1\}$. Let $\mathcal{S}' = \{(Y_i', X_i'), i \in \mathcal{I}'\}$ be a test dataset containing m' test instances and for which the class proportions $\pi' = [\pi_1', \dots, \pi_K']$ are unknown. The classifier $\delta_{\hat{\pi}}$ fitted on 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 attributes $X_i' \in \mathcal{X}$. To simplify our explanation, we assume that $\hat{\mathbb{P}}_{\mathcal{S}'}$ coincides with $\hat{\mathbb{P}}_{\mathcal{S}}$, i.e., there is no probability shift between the training set and the test set. Hence, our attention is focused on the prior shift, i.e., π' is different from $\hat{\pi}$. As explained in [5], the risk of classification errors, with respect to the fitted classifier $\delta_{\hat{\pi}}$, as a function of π', is

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

Since the $\hat{R}_k\left(\delta_{\hat{\pi}}\right)$'s do not depend on π', the risk (4) is clearly a linear function with respect to π'. Hence, when the distribution of the priors is uncertain and changes in time, the risk of classification errors is expected to evolve linearly. This is an important issue since we generally do not know when and why prior probability shifts may occur. Note that we have $\hat{r}\left(\delta_{\hat{\pi}}\right) = \hat{r}\left(\hat{\pi}, \delta_{\hat{\pi}}\right)$. The maximum value of $\hat{r}\left(\pi', \delta_{\hat{\pi}}\right)$ that can be reached is $M(\delta_{\hat{\pi}}) := \max_{k \in \mathcal{Y}} \hat{R}_k\left(\delta_{\hat{\pi}}\right)$. This issue is therefore especially highlighted when the class-conditional risks are imbalanced. An illustration of this issue is given in Figure 2 in Appendix A. The task of learning a robust classifier with respect to uncertain prior distributions is therefore necessary. This task is in the field of Bayesian Robustness [12] for Machine Learning.

Reminder on the minimax criterion. Training with imbalanced datasets and dealing with prior probability shifts share a common trait, namely the sensitivity to imbalanced class-conditional risks. In order to address this issue, a legitimate solution is to learn a minimax classifier [5, 12–14]. Instead of minimizing the global risk of classification errors (1), the objective of this approach is to minimize $\max_{k\in\mathcal{Y}} \hat{R}_k\left(\delta_{\hat{\pi}}\right)$. In other words, the minimax criterion tends to balance as more as possible the class-conditional risks (3), and according to equation (4), the resulted decision rule becomes robust when π' differs from $\hat{\pi}$. As explained in [5], learning a minimax classifier δ^M is equivalent to solve the following problem:

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

Introduction of the box-constrained minimax classifier. Although the minimax classifier is adequate for addressing the issues regarding the class proportions, such a decision rule can appear sometimes too pessimistic as discussed in [12, 15]. This drawback essentially occurs when prior probability shifts can append only over a subset of the simplex $\mathbb{S}$ and the global risk of classification errors associated with δ^M becomes too high. In order to alleviate this drawback, a solution is to shrink the priors constraint. In the literature, this task is called Γ-minimax classification [12], where Γ corresponds to a set containing only the acceptable prior distributions.

In this paper, we consider Γ to be a box-constraint $\mathbb{B}$ defined by

$$\mathbb{B} := \{\pi \in \mathbb{R}^K : \forall k \in \mathcal{Y},\; 0 \leq a_k \leq \pi_k \leq b_k \leq 1\}, \tag{6}$$

which allows to bound each class proportion π_k independently in the interval $[a_k, b_k]$ for all $k \in \mathcal{Y}$. The main asset of considering such a box-constraint stems from the fact that the experts of the application domain can easily and rationally build it, by providing some independent bounds $[a_k, b_k]$ on each class proportion[1].

When considering the new constraint (6) on the priors, we therefore set up the box-constrained simplex

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

Hence, to compute the box-constrained minimax classifier δ^C with respect to $\mathbb{U}$, the problem (5) becomes

$$\delta^C \;=\; \underset{\delta\in\Delta}{\operatorname{argmin}} \max_{\pi\in\mathbb{U}} \hat{r}(\pi,\delta). \tag{8}$$

Let us note that the minimax classifier (5) is a particular case of the box-constrained minimax classifier (8). Indeed, the minimax classifier is still accessible when considering $\mathbb{B} = [0,1]^K$, so that $\mathbb{U} = \mathbb{S}$.

Dealing with both numeric and categorical features. The task of dealing with both numeric and categorical attributes is difficult for reaching optimal results. 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 [13, 16]), this estimation is quite difficult. In such a case, a relevant solution is to discretize the numeric attributes in order to model the joint distribution of features with a probability mass function. Hence, since the number of values taken by the joint distribution is finite, we can estimate their probabilities of occurrence without making any assumptions of independence between the attributes. Many works have shown that the discretization of the numeric features generally leads to accurate results [17–21], with favorable statistical properties. For example, 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 [22–24]. In the following, we consider that all the features are discrete or beforehand discretized with a finite number of values.

[1]For example, in medical field, it may be reasonable to bound the maximum frequency of a given disease.

Contributions. In this paper, we provide a new algorithm to compute the box-constrained minimax classifier (8) in the context of discrete or beforehand discretized features. In section 2, we develop the procedure for solving the minimax optimization problem (8). This procedure is based on a projected subgradient algorithm, which computes the least favorable prior over the polyhedral constraint (7). The convergence of this algorithm is established. In section 3, we illustrate on a real public database the performance of the box-constrained minimax classifier with respect to the box-constraint bounds. Finally, section 4 concludes the paper.

2. Computation of the Discrete Box-Constrained Minimax Classifier

In the following, we consider that all the features are discrete or beforehand discretized. In this section, given a box-constraint $\mathbb{B}$, we present the optimization procedure for solving the minimax optimization problem (8).

2.1. Reasoning to compute our discrete Box-constrained minimax classifier

Dealing only with discrete or beforehand discretized features, it follows that each each attribute X_{ij} can take on a finite number of values t_j. Hence, 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 $\mathcal{T} = \{1, \ldots, T\}$ be the set of indices.

Since $|\mathcal{X}| = T$ is finite, it follows that $|\Delta| = |\hat{\mathcal{Y}}|^{|\mathcal{X}|} = K^T$ is finite. When the set of classifiers Δ is finite, the famous Minimax Theorem [25] establishes that

$$\min_{\delta \in \Delta} \max_{\pi \in \mathbb{U}} \hat{r}(\pi, \delta) \;=\; \max_{\pi \in \mathbb{U}} \min_{\delta \in \Delta} \hat{r}(\pi, \delta). \tag{9}$$

Let us define δ_π^B the optimal Bayes classifier associated with the given priors $\pi \in \mathbb{S}$, such that

$$\delta_\pi^B \;:=\; \arg\min_{\delta \in \Delta} \hat{r}(\pi, \delta). \tag{10}$$

Let $V(\pi) = \hat{r}(\pi, \delta_\pi^B)$ denote the Bayes risk for a given π: it is the minimum risk for a given π. Hence, according to (9), and provided that we can calculate δ_π^B and its minimum Bayes risk $V(\pi)$ for any prior $\pi \in \mathbb{U}$, the optimization problem (8) is equivalent to compute the least favorable priors

$$\pi^\star \;:=\; \arg\max_{\pi \in \mathbb{U}} V(\pi), \tag{11}$$

so that the solution δ^C of (8) is the Bayes classifier given by (10) with the priors (11). The least favorable priors are generally difficult to compute as underlined in [12, 26–28].

Subsection 2.2 is devoted to the calculation of the minimum Bayes risk $V(\pi)$ over the simplex. Subsection 2.3 is devoted to compute the least favorable priors $\pi^\star$ solution of (11).

2.2. Calculation of the minimum empirical risk over the simplex

Dealing only with discrete or beforehand discretized features, we can estimate from the labeled learning instances $\mathcal{S} = \{(Y_i, X_i), i \in \mathcal{I}\}$ the probabilities $\hat{p}_{kt}$ of observing the feature profile $x_t \in \mathcal{X}$ given that the class label is k, for all $t \in \mathcal{T}$ and for all $k \in \mathcal{Y}$, such that

$$\hat{p}_{kt} \;:=\; \frac{1}{m_k} \sum_{i \in \mathcal{I}_k} \mathbb{1}_{\{X_i = x_t\}} \tag{12}$$

In (12), for all $k \in \mathcal{Y}$, $\mathcal{I}_k = \{i \in \mathcal{I} : Y_i = k\}$ denotes the set of learning samples from the class k, and $m_k = |\mathcal{I}_k|$ corresponds to the number of instances in $\mathcal{I}_k$. Since we can only consider the instances from the training set, the probabilities $\hat{p}_{kt}$ defined in (12) are assumed to be estimated once for all. Indeed, the statistical

estimation theory [29] 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, i.e., that the test samples follow the same theoretical class-conditional probabilities as the training samples.

The following theorem provides the analytic formula of the discrete Bayes classifier (10) associated with the training class proportions π, and its associated risk.

Theorem 1. *The empirical Bayes classifier δ_π^B, which minimizes the empirical risk (10) over Δ, is given by*

$$\delta_\pi^B : X_i \mapsto \arg\min_{l \in \hat{\mathcal{Y}}} \sum_{t \in \mathcal{T}} \sum_{k \in \mathcal{Y}} L_{kl}\, \pi_k\, \hat{p}_{kt}\, \mathbb{1}_{\{x_t = X_i\}}. \tag{13}$$

Its associated empirical risk is

$$V(\pi) = \hat{r}\left(\pi, \delta_\pi^B\right) = \sum_{k \in \mathcal{Y}} \pi_k \hat{R}_k\left(\delta_\pi^B\right), \tag{14}$$

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

$$\hat{R}_k\left(\delta_\pi^B\right) = \sum_{t \in \mathcal{T}} \sum_{l \in \hat{\mathcal{Y}}} L_{kl}\, \hat{p}_{kt}\, \mathbb{1}_{\{\lambda_{lt} = \min_{q \in \hat{\mathcal{Y}}} \lambda_{qt}\}}, \tag{15}$$

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

Proof. The proof is established in Theorem 1 in [1]. □

In other words, the function $V : \pi \in \mathbb{S} \mapsto V(\pi)$ gives the minimum value of the empirical risk when the class proportions are π and the class-conditional probabilities $\hat{p}_{kt}$ remain unchanged. The following proposition studies the function V over $\mathbb{S}$.

Proposition 1. *The empirical Bayes risk $V : \pi \mapsto V(\pi)$ is a concave multivariate piecewise affine function over the simplex $\mathbb{S}$ with a finite number of pieces. Moreover, if the following condition*

$$\exists\, (\pi, \pi', k) \in \mathbb{S} \times \mathbb{S} \times \mathcal{Y} : \hat{R}_k\left(\delta_\pi^B\right) \neq \hat{R}_k\left(\delta_{\pi'}^B\right) \tag{16}$$

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

Proof. The proof is established in Proposition 1, Proposition 2 and Corollary 1 in [1]. □

Note that the condition (16) is almost always satisfied. Otherwise, it would mean that each class conditional risk $\hat{R}_k\left(\delta_\pi^B\right)$ would remain equal whatever the prior $\pi \in \mathbb{S}$, even at the vertices of the simplex. The empirical Bayes risk V would be an affine function over $\mathbb{S}$.

2.3. Maximization of the minimum empirical risk V over $\mathbb{U}$

In order to compute our box-constrained minimax classifier, according to (11) and when considering (14), our objective is to solve the following optimization problem

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

Since $V : \pi \mapsto V(\pi)$ is in general non-differentiable provided that the condition (16) 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 [30] that follows 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 $n \geq 1$, $g^{(n)}$ denotes a subgradient of V at the point $\pi^{(n)}$, γ_n denotes the subgradient step, $\eta_n = \max\{1, \|g^{(n)}\|_2\}$, and $\mathrm{P}_{\mathbb{U}}$ denotes the exact projection onto the box-constrained simplex $\mathbb{U}$. Let us note that this algorithm remains applicable in the particular case where the condition (16) is not satisfied, i.e. when the function V is affine over $\mathbb{U}$. The following lemma gives a subgradient of the target function V.

Lemma 1. *Given $\pi \in \mathbb{U}$, the vector composed of 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. 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 $v \in \mathbb{R}^K$. In our case, given $\pi \in \mathbb{U}$, let consider $\pi' \in \mathbb{U}$. Denoting $\hat{R}\left(\delta_\pi^B\right)$ the vector $\hat{R}\left(\delta_\pi^B\right) := \left[\hat{R}_1\left(\delta_\pi^B\right), \ldots, \hat{R}_K\left(\delta_\pi^B\right)\right]$ of all class-conditional risks, we get:

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

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

In the following, we choose $g^{(n)} = \hat{R}\left(\delta^B_{\pi^{(n)}}\right)$ at each iteration $n \geq 1$ in (18). The following theorem establishes the convergence of the iterates (18) to $\pi^\star$.

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

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

the sequence of iterates (18) converges strongly 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 [30]. Here we have the strong convergence since $\pi^{(n)}$ belongs to a finite dimensional space. □

Remark 1. *In the general case where the empirical Bayes risk V is not constantly equal to zero over $\mathbb{S}$, the subgradient $\hat{R}\left(\delta^B_{\pi^\star}\right)$ at the box-constrained minimax optimum is not null. Otherwise, the associated risk $V(\pi^\star)$ would vanish due to (14). This would contradict the fact that $\pi^\star$ is solution of (17).*

According to Remark 1, in the general case where the empirical Bayes risk V is not constantly equal to zero over $\mathbb{S}$, the sequence of iterates (18) is infinite, and we need to consider a stopping criterion. To this aim, we propose to follow the reasoning in [31] which leads to the following corollary.

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

$$\left|\max_{n\leq N}\left\{V\left(\pi^{(n)}\right)\right\} - V(\pi^\star)\right| \leq \max\left\{1, \sqrt{\sum_{k=1}^{K}\left[\sum_{l=1}^{K} L_{kl}\right]^2}\right\} \frac{\rho^2 + \sum_{n=1}^{N}\gamma_n^2}{2\sum_{n=1}^{N}\gamma_n} \tag{20}$$

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

Proof. The proof is detailed in Appendix B. □

In practice we can 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.

2.4. **Exact projection onto the box constrained region**

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

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

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

$$\mathbb{U} \ = \ \bigcap_{i=1}^{2K+2} U_i. \tag{21}$$

In other words, our box-constrained simplex $\mathbb{U}$ is a polyhedral set. Thus, in order to compute the exact projection onto $\mathbb{U}$, we propose to use the algorithm provided in [32] which computes the exact projection onto polyhedral sets in Hilbert spaces. Let us note that in the case where we are interested in computing the minimax classifier (5), we have $\mathbb{U} = \mathbb{S}$, and we can perform the projection onto $\mathbb{S}$ using the algorithm provided in [33] for which the complexity is lower.

2.5. **Box-constrained minimax classifier Algorithm**

The procedure for computing the box-constrained minimax classifier $\delta^B_{\pi^\star}$ is summarized in Algorithm 1. In practice, we choose the sequence of steps $(\gamma_n)_{n \geq 1} = 1/n$ which satisfies (19).

Algorithm 1 Box-constrained Minimax Classifier

Input: $(Y_i, X_i)_{i \in \mathcal{I}}$, K, N.
Compute $\pi^{(1)} = \hat{\pi}$
Compute the $\hat{p}_{kt}$'s as described in (12).
$r^\star \leftarrow 0, \quad \pi^\star \leftarrow \pi^{(1)}$
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 (15)
  end for
  $r^{(n)} = \sum_{k=1}^K \pi_k^{(n)} g_k^{(n)}$ see (14)
  if $r^{(n)} > r^\star$ **then**
    $r^\star \leftarrow r^{(n)}, \quad \pi^\star \leftarrow \pi^{(n)}$
  end if
  $\gamma_n \leftarrow 1/n, \quad \eta_n \leftarrow \max\{1, \|g^{(n)}\|_2\}, \quad \pi^{(n+1)} \leftarrow \mathrm{P}_{\mathbb{U}}\left(\pi^{(n)} + \gamma_n\, g^{(n)}/\eta_n\right)$
end for
Output: $r^\star$, $\pi^\star$ and $\delta^B_{\pi^\star}$ provided by (13) with $\pi = \pi^\star$.

3. Numerical experiments

Database description. For illustrating the interest of our box-constrained minimax classifier, we applied our algorithm on the real public Framingham database [34]. The objective of the Framingham study is to predict

the development of a Coronary Heart Disease (CHD) within 10 years based on $d = 15$ clinical and biological features (7 categorical and 8 numeric). In this paper, we do not study the effects of the discretization of continuous features, and we consider the discretized attributes as built in [1, 2]. This database contains $\mathrm{K} = 2$ classes, with class 2 corresponding to individuals who have developed a CHD, and class 1 corresponding to the others. For this database, 3,658 patients have been followed for 10 years. Among these patients, 85% did not developped a CHD, while 15% developped a CHD within 10 years. In other words, the dataset is imbalanced: $\hat{\pi} = [0.85, 0.15]$, which complicates the task of well predicting a CHD based on the labeled learning observations. For this experiment, let us consider the $L_{0\text{-}1}$ loss function, such that $L_{11} = L_{22} = 0$, and $L_{12} = L_{21} = 1$.

Procedure of the experiment and results. In the following, let $\bar{\pi} := \operatorname{argmax}_{\pi \in \mathbb{S}} V(\pi)$ be the least favorable priors over the simplex $\mathbb{S}$, and thus let $\delta^{\mathrm{B}}_{\bar{\pi}}$ be the minimax classifier δ^{M} solution of (5). The box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$ solution of (8) aims to find a trade-off between achieving an acceptable global risk and balancing the class-conditional risks with respect to the box-constraint (6). In other words, the box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$ is designed to find a trade-off between the discrete Bayes classifier $\delta^{\mathrm{B}}_{\hat{\pi}}$ (13) associated with the class proportions of the training set $\hat{\pi}$, and the minimax classifier $\delta^{\mathrm{B}}_{\bar{\pi}}$. These results depend on the box-constraint bounds. In practice, the box-constraint can be established by the experts of the application field by bounding some or all the prior probabilities independently. If the results are not enough satisfying, the experts can easily tighten or spread the box-constraint bounds in order to find an acceptable trade-off between balancing the class-conditional risks and achieving an acceptable global risk of error.

For this experiment, in order to illustrate this property, we compared $\delta^{\mathrm{B}}_{\hat{\pi}}$, $\delta^{\mathrm{B}}_{\bar{\pi}}$ and $\delta^{\mathrm{B}}_{\pi^\star}$ for different box-constraint bounds. To this aim, we consider the box-constraints $\mathbb{B}_\beta$ centered in $\hat{\pi}$, and such that, given $\beta \in [0, 1]$,

$$\mathbb{B}_\beta = \left\{ \pi \in \mathbb{R}^K : \forall k \in \mathcal{Y},\ \hat{\pi}_k - \rho_\beta \leq \pi_k \leq \hat{\pi}_k + \rho_\beta \right\}, \tag{22}$$

with $\rho_\beta := \beta \left\| \hat{\pi} - \bar{\pi} \right\|_\infty = \beta \max_{k \in \mathcal{Y}} |\hat{\pi}_k - \bar{\pi}_k|$. Our box-constrained probabilistic simplex is therefore $\mathbb{U}_\beta = \mathbb{S} \cap \mathbb{B}_\beta$. Thus, when $\beta = 0$, $\mathbb{B}_0 = \{\hat{\pi}\}$, hence $\mathbb{U}_0 = \{\hat{\pi}\}$ and $\pi^\star = \hat{\pi}$. When $\beta = 1$, $\bar{\pi} \in \mathbb{B}_1$, hence $\bar{\pi} \in \mathbb{U}_1$ and $\pi^\star = \bar{\pi}$. Note that the minimax classifier $\delta^{\mathrm{B}}_{\bar{\pi}}$ was trained using our Algorithm 1 when considering $\mathbb{U} = \mathbb{S}$, and for this particular case the projection onto the simplex $\mathbb{S}$ was performed using the algorithm provided in [33].

The procedure of our experiment is the following: we performed a 10-fold cross-validation, that is we randomly split the main dataset such that 90% of the instances composed training set and the 10% staying instances belong to the test set. We repeated ten times this splitting, and at each repetition of this cross-validation, we ranged β from 0 to 1 so that we increased the box-constraint radius, and we measured $V(\pi^\star)$ and $\psi\left(\delta^{\mathrm{B}}_{\pi^\star}\right)$, where $\psi : \Delta \to \mathbb{R}^+$ such that, for all $\delta \in \Delta$,

$$\psi(\delta) \ := \ \max_{k \in \mathcal{Y}} \hat{R}_k(\delta) - \min_{k \in \mathcal{Y}} \hat{R}_k(\delta), \tag{23}$$

In other words, the criterion ψ aims to measure how a given classifier $\delta \in \Delta$ performs for balancing the class-conditional risks.

The results of the experiment are presented in Figure 1. We can observe that as β increases, and thus as the radius ρ_β increases, then the better $\delta^{B}_{\pi^\star}$ performs for balancing the class-conditional risks, and thus the better $\delta^{B}_{\pi^\star}$ performs for well predicting the patients who tend to develop a CHD. However, as β increases, and thus as the radius ρ_β increases, the more pessimistic $\delta^{B}_{\pi^\star}$ becomes since $V(\pi^\star)$ converges to $V(\bar{\pi})$ which is the maximum value of V. Concerning the computing time, at each iteration of the cross-validation procedure, the time to train our Discrete Box-constrained Minimax Classifier $\delta^{B}_{\pi^\star}$ was around $0.67s$ for each parameter β, against $0.01s$ for the Discrete Bayes Classifier $\delta^{\mathrm{B}}_{\hat{\pi}}$ (13). Note that the computing time associated with the Discrete Minimax Classifier $\delta^{\mathrm{B}}_{\bar{\pi}}$ was around $0.35s$ using the algorithm provided in [33] to project onto the simplex $\mathbb{S}$, which is faster than the training time associated with $\delta^{B}_{\pi^\star}$.

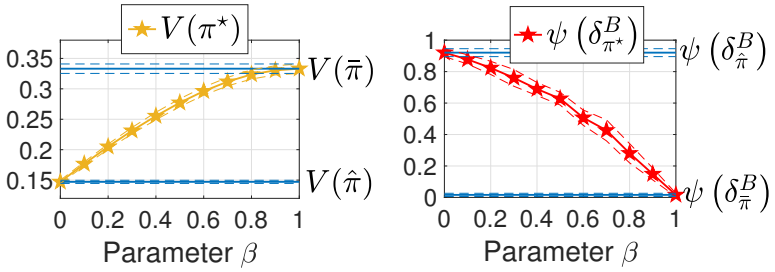

FIGURE 1. Impact of the box-constraint radius on $\delta^{\mathrm{B}}_{\pi^\star}$ when β increases from 0 to 1 in (22), after a 10-fold cross-validation procedure. The results are presented as [mean $\pm$ std].

4. Conclusion

This paper presents the optimization procedure for computing a box-constrained minimax classifier in the context of discrete or discretized features with multiple classes, a positive loss function, and some dependencies between the features. This minimax classifier aims to address the issues of imbalanced datasets and prior probability shifts. Our method is in the field of Γ-minimaxity and Bayesian Robustness for Machine Learning. Our approach is designed for considering independent bounds on the class proportions, which can be easily and rationally provided by the experts from the application domain, and which allow us to find a trade-off between minimizing the maximum of the class conditional risks, and achieving an acceptable global risk of errors, based on the interest or the knowledge of the experts.

The computation of the box-constrained minimax classifier results from the computation of the least favorable prior which maximizes the minimum empirical risk of classification errors over the box-constrained probabilistic simplex, using a projected subgradient algorithm. The convergence of our algorithm is established.

An important work would be to improve the computation time of the exact projection onto the box-constrained simplex, which would be essential for dealing with databases containing a large number of classes.

References

[1] C. Gilet, S. Barbosa, and L. Fillatre, "Discrete box-constrained minimax classifier for uncertain and imbalanced class proportions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[2] 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.

[3] V. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10 5, pp. 988–99, 1999.

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

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

[6] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1263–1284, 2009.

[7] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, pp. 429–449, 2002.

[8] 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.

[9] Q. Dong, S. Gong, and X. Zhu, "Imbalanced deep learning by minority class incremental rectification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[10] 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.

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

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

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

[14] 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.
[15] 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.
[16] G. V. Trunk, "A problem of dimensionality: A simple example," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1979.
[17] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," *International Conference on Machine Learning*, 1995.
[18] 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.
[19] 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.
[20] 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.
[21] 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.
[22] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, 2nd ed. Springer-Verlag New York, 1996.
[23] 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.
[24] 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.
[25] T. Ferguson, *Mathematical Statistics : A Decision Theoretic Approach*. Academic Press, 1967.
[26] L. Fillatre, "Constrained epsilon-minimax test for simultaneous detection and classification," *IEEE Transactions on Information Theory*, vol. 57, no. 12, pp. 8055–8071, 2011.
[27] 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.
[28] L. Fillatre, "Constructive minimax classification of discrete observations with arbitrary loss function," *Signal Processing*, vol. 141, pp. 322–330, 2017.
[29] C. R. Rao, *Linear Statistical Inference and its Applications*. Wiley, 1973.
[30] 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.
[31] 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.
[32] K. E. Rutkowski, "Closed-form expressions for projectors onto polyhedral sets in Hilbert spaces," *SIAM Journal on Optimization*, vol. 27, pp. 1758–1771, 2017.
[33] L. Condat, "Fast projection onto the simplex and the ℓ_1 ball," *Mathematical Programming*, vol. 158, no. 1, pp. 575–585, 2016.
[34] A. J. Wawrzyniak, *Framingham Heart Study*. New York, NY: Springer New York, 2013, pp. 811–814.
[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

A. Illustration of prior probability shifts issue

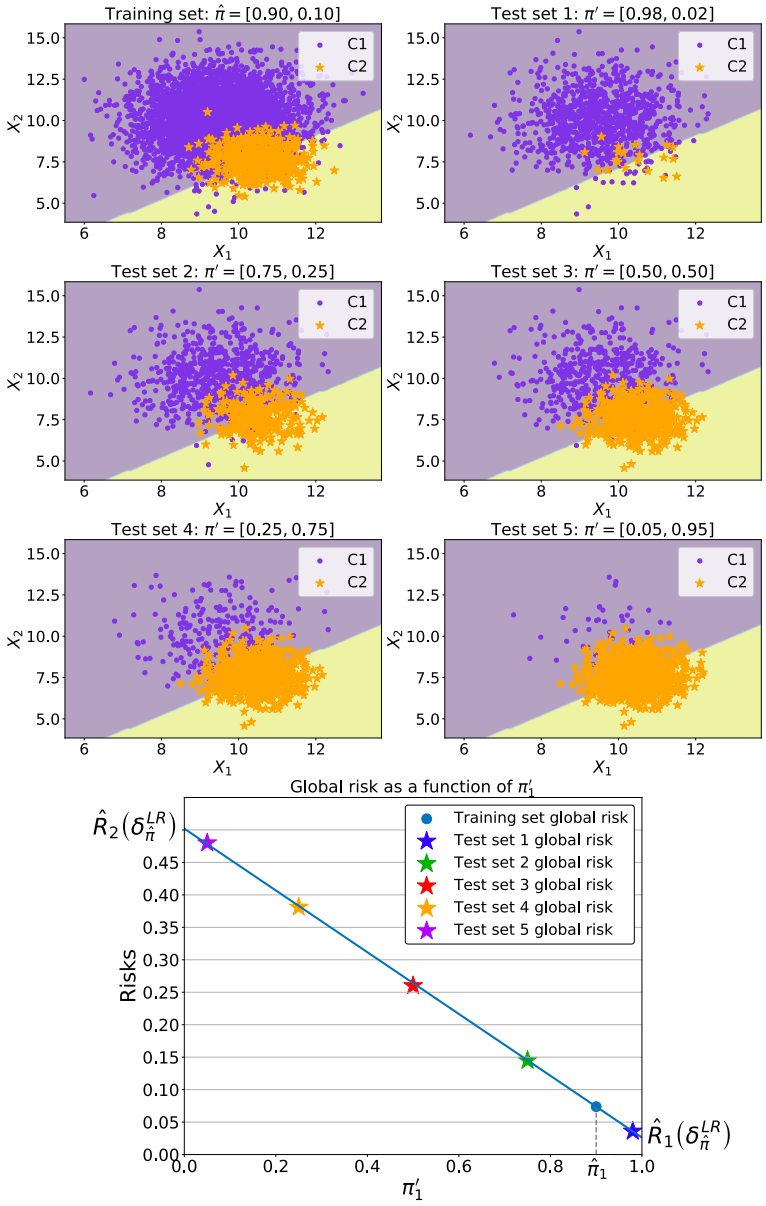

FIGURE 2. For this experiment we generated a training dataset (Up-Left) containing $m = 5,000$ instances described by $d = 2$ features and clustered into $K = 2$ classes which satisfies the class proportions $\hat{\pi} = [0.90, 0.10]$. We then trained the Logistic Regression $\delta^{LR}_{\hat{\pi}}$ on this training set and we applied it on 5 different test sets containing $m' = 1,000$ instances. Each dataset was generated using the `make_blobs` function provided by Scikit-Learn [35] from the same features distributions in each class, but the test sets differ according to the class proportions π' ranging over the simplex $\mathbb{S}$. The last subfigure describes the global risk associated with each dataset. Since we have $K = 2$ classes, the global risk (4) associated with $\delta^{LR}_{\hat{\pi}}$ can be written as $\hat{r}\left(\pi', \delta^{LR}_{\hat{\pi}}\right) = \pi'_1[\hat{R}_1\left(\delta^{LR}_{\hat{\pi}}\right) - \hat{R}_2\left(\delta^{LR}_{\hat{\pi}}\right)] + \hat{R}_2\left(\delta^{LR}_{\hat{\pi}}\right)$.

B. Proof of Corollary 1

Let $n \geq 1$ and let $z^{(n+1)} := \pi^{(n)} + \frac{\gamma_n}{\eta_n} g^{(n)}$. We have

$$\begin{aligned}\left\|z^{(n+1)} - \pi^\star\right\|_2^2 &= \left\|\pi^{(n)} + \frac{\gamma_n}{\eta_n} g^{(n)} - \pi^\star\right\|_2^2 \\ &= \left\|\pi^{(n)} - \pi^\star + \frac{\gamma_n}{\eta_n} g^{(n)}\right\|_2^2 \\ &= \left\|\pi^{(n)} - \pi^\star\right\|_2^2 + 2\frac{\gamma_n}{\eta_n}\left\langle g^{(n)}, \pi^{(n)} - \pi^\star\right\rangle + \frac{\gamma_n^2}{\eta_n^2}\left\|g^{(n)}\right\|_2^2.\end{aligned}$$

Since $g^{(n)}$ is a subgradient of V at the point $\pi^{(n)}$, it follows that

$$\left\langle g^{(n)}, \pi^\star - \pi^{(n)}\right\rangle + V\left(\pi^{(n)}\right) \geq V(\pi^\star) \quad \Leftrightarrow \quad \left\langle g^{(n)}, \pi^{(n)} - \pi^\star\right\rangle \leq V\left(\pi^{(n)}\right) - V(\pi^\star).$$

Thus,

$$\left\|z^{(n+1)} - \pi^\star\right\|_2^2 \leq \left\|\pi^{(n)} - \pi^\star\right\|_2^2 + 2\frac{\gamma_n}{\eta_n}\left(V\left(\pi^{(n)}\right) - V(\pi^\star)\right) + \frac{\gamma_n^2}{\eta_n^2}\left\|g^{(n)}\right\|_2^2.$$

As explained in [31],

$$\left\|\pi^{(n+1)} - \pi^\star\right\|_2^2 = \left\|\mathrm{P}_{\mathbb{U}}\left(z^{(n+1)}\right) - \pi^\star\right\|_2^2 \leq \left\|z^{(n+1)} - \pi^\star\right\|_2^2.$$

Thus,

$$\left\|\pi^{(n+1)} - \pi^\star\right\|_2^2 \leq \left\|\pi^{(n)} - \pi^\star\right\|_2^2 + 2\frac{\gamma_n}{\eta_n}\left(V\left(\pi^{(n)}\right) - V(\pi^\star)\right) + \frac{\gamma_n^2}{\eta_n^2}\left\|g^{(n)}\right\|_2^2.$$

Applying the last inequality recursively, it follows that

$$\left\|\pi^{(n+1)} - \pi^\star\right\|_2^2 \leq \left\|\pi^{(1)} - \pi^\star\right\|_2^2 + 2\sum_{i=1}^n \frac{\gamma_i}{\eta_i}\left(V\left(\pi^{(i)}\right) - V(\pi^\star)\right) + \sum_{i=1}^n \frac{\gamma_i^2}{\eta_i^2}\left\|g^{(i)}\right\|_2^2.$$

Since $\left\|\pi^{(n+1)} - \pi^\star\right\|_2^2 \geq 0$,

$$\begin{aligned}0 &\leq \left\|\pi^{(1)} - \pi^\star\right\|_2^2 + 2\sum_{i=1}^n \frac{\gamma_i}{\eta_i}\left(V\left(\pi^{(i)}\right) - V(\pi^\star)\right) + \sum_{i=1}^n \frac{\gamma_i^2}{\eta_i^2}\left\|g^{(i)}\right\|_2^2 \\ \Leftrightarrow \quad 2\sum_{i=1}^n \frac{\gamma_i}{\eta_i}\left(V(\pi^\star) - V\left(\pi^{(i)}\right)\right) &\leq \left\|\pi^{(1)} - \pi^\star\right\|_2^2 + \sum_{i=1}^n \frac{\gamma_i^2}{\eta_i^2}\left\|g^{(i)}\right\|_2^2.\end{aligned}$$

By definition of $\pi^\star$, $V(\pi^\star) \geq \max_{i=1\ldots,n} V\left(\pi^{(i)}\right)$, thus

$$2\sum_{i=1}^n \frac{\gamma_i}{\eta_i}\left(V(\pi^\star) - V\left(\pi^{(i)}\right)\right) \geq 2\sum_{i=1}^n \frac{\gamma_i}{\eta_i}\left(V(\pi^\star) - \max_{i\leq n}\left\{V\left(\pi^{(i)}\right)\right\}\right).$$

It follows that,

$$V(\pi^\star) - \max_{i\leq n}\left\{V\left(\pi^{(i)}\right)\right\} \leq \frac{\left\|\pi^{(1)} - \pi^\star\right\|_2^2 + \sum_{i=1}^n \frac{\gamma_i^2}{\eta_i^2}\left\|g^{(i)}\right\|_2^2}{2\sum_{i=1}^n \frac{\gamma_i}{\eta_i}}. \tag{24}$$

Let us remind that for all $i \in \{1, \ldots, n\}$, $\eta_i = \max\left\{1, \left\|g^{(i)}\right\|_2\right\}$. We can therefore distinguish two cases :

- **Case 1.** If it exists $i \in \{1, \dots, n\}$ such that $\left\|g^{(i)}\right\|_2 < 1$, then $\eta_i^2 = 1$, and

$$\frac{\gamma_i^2}{\eta_i^2} \left\|g^{(i)}\right\|_2^2 = \gamma_i^2 \left\|g^{(i)}\right\|_2^2 \leq \gamma_i^2.$$

- **Case 2.** If $i \in \{1, \dots, n\}$ such that $\left\|g^{(i)}\right\|_2 \geq 1$, then $\eta_i^2 = \left\|g^{(i)}\right\|_2^2$, and

$$\frac{\gamma_i^2}{\eta_i^2} \left\|g^{(i)}\right\|_2^2 = \gamma_i^2.$$

Hence,

$$\sum_{i=1}^{n} \frac{\gamma_i^2}{\eta_i^2} \left\|g^{(i)}\right\|_2^2 \leq \sum_{i=1}^{n} \gamma_i^2.$$

Applying the last inequality in (24), it follows that

$$V(\pi^\star) - \max_{i \leq n} \left\{V\left(\pi^{(i)}\right)\right\} \leq \frac{\left\|\pi^{(1)} - \pi^\star\right\|_2^2 + \sum_{i=1}^{n} \gamma_i^2}{2 \sum_{i=1}^{n} \frac{\gamma_i}{\eta_i}}. \tag{25}$$

Furthermore, since for all $i \in \{1, \dots, n\}$, $\eta_i = \max\left\{1, \left\|g^{(i)}\right\|_2\right\}$, and since we consider the subgradient

$$g^{(i)} = \hat{R}\left(\delta^B_{\pi^{(i)}}\right) := \left[\hat{R}_1\left(\delta^B_{\pi^{(i)}}\right), \dots, \hat{R}_K\left(\delta^B_{\pi^{(i)}}\right)\right],$$

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

$$\hat{R}_k\left(\delta^B_{\pi^{(i)}}\right) = \sum_{l \in \hat{\mathcal{Y}}} L_{kl}\, \hat{\mathbb{P}}_{\mathcal{S}}\left(\delta^B_{\pi^{(i)}}(X_s) = l \mid Y_s = k\right),$$

it follows that, for all $i \in \{1, \dots, n\}$,

$$\left\|g^{(i)}\right\|_2 = \sqrt{\sum_{k=1}^{K} \left[\hat{R}_k\left(\delta^B_{\pi^{(i)}}\right)\right]^2} = \sqrt{\sum_{k=1}^{K} \left[\sum_{l=1}^{K} L_{kl}\, \hat{\mathbb{P}}_{\mathcal{S}}\left(\delta^B_{\pi^{(i)}}(X_s) = l \mid Y_s = k\right)\right]^2} \leq \sqrt{\sum_{k=1}^{K} \left[\sum_{l=1}^{K} L_{kl}\right]^2}.$$

Let us define $h(L) := \sqrt{\sum_{k=1}^{K} \left[\sum_{l=1}^{K} L_{kl}\right]^2}$. Thus, for all $i \in \{1, \dots, n\}$, $\eta_i \leq \max\{1, h(L)\}$, and then

$$\begin{aligned} \forall i \in \{1, \dots, n\}, \quad \frac{\gamma_i}{\eta_i} \geq \frac{\gamma_i}{\max\{1, h(L)\}} \quad &\Rightarrow \quad \sum_{i=1}^{n} \frac{\gamma_i}{\eta_i} \geq \frac{1}{\max\{1, h(L)\}} \sum_{i=1}^{n} \gamma_i \\ &\Rightarrow \quad \frac{1}{\sum_{i=1}^{n} \frac{\gamma_i}{\eta_i}} \leq \frac{1}{\frac{1}{\max\{1, h(L)\}} \sum_{i=1}^{n} \gamma_i}. \end{aligned}$$

Finally, coming back to equation (25), and since $\left\|\pi^{(1)} - \pi^\star\right\|_2^2 \leq K$, it follows that

$$V(\pi^\star) - \max_{i \leq n} \left\{V\left(\pi^{(i)}\right)\right\} \leq \max\{1, h(L)\} \frac{K + \sum_{i=1}^{n} \gamma_i^2}{2 \sum_{i=1}^{n} \gamma_i}.$$

□