Some contributions to multi-class learning

Christophe Denis

Some contributions to multi-class learning

Université Gustave Eiffel

Habilitation à Diriger des Recherches

Spécialité : Mathématiques Appliquées

présentée par

Christophe Denis

Soutenue publiquement le 21 janvier 2022 après avis des rapporteurs,

M.	**Gilles**	**Blanchard**	Université Paris-Saclay
Mme.	**Béatrice**	**Laurent-Bonneau**	INSA de Toulouse
Mme.	**Sara**	**van de Geer**	ETH Zürich

et devant le jury composé de :

M.	**Gilles**	**Blanchard**	Université Paris-Saclay
M.	**Antoine**	**Chambaz**	Université de Paris
Mme.	**Fabienne**	**Comte**	Université de Paris
M.	**Stéphane**	**Gaïffas**	Université de Paris
Mme.	**Béatrice**	**Laurent-Bonneau**	INSA de Toulouse
Mme.	**Adeline**	**Leclercq Samson**	Université Grenoble Alpes
Mme.	**Florence**	**Merlevède**	Université Gustave Eiffel

Contents

Avant-propos

Ce mémoire d'habilitation est une synthèse de mon travail de recherche réalisé depuis l'obtention de mon poste de Maître de conférences à l'Université Gustave Eiffel en 2013. Ma recherche s'inscrit dans le domaine de l'apprentissage statistique. Plus particulièrement, elle s'articule autour de deux thématiques: l'apprentissage supervisé sous contrainte et l'apprentissage pour des données à dépendance temporelle.

Ce manuscrit présente mes contributions à l'apprentissage multi-classes. Il se décompose en quatre chapitres. Un premier chapitre est consacré à une introduction à la classification multi-classes et introduit les notions importantes. Ce chapitre est aussi l'occasion d'exposer les enjeux et motivations des résultats présentés dans ce manuscrit. Les deux chapitres suivants sont consacrés à l'exposé de résultats obtenus en apprentissage sous contrainte, tandis que le dernier chapitre traite de la classification pour des données trajectorielles.

Une partie importante de mes travaux de recherche porte sur la classification sous contrainte. Au sein de cette thématique, j'ai principalement exploré deux sujets de recherche. D'une part la classification supervisée par ensemble (*set-valued classification*) sous contrainte de taille, dont je présente les résultats obtenus en collaboration avec E. Chzhen (LMO) et M. Hebiri (LAMA) au Chapitre 2. D'autre part, je m'intéresse à l'apprentissage sous contrainte d'équité (*fairness*) qui est un sujet d'importance croissante au sein de la communauté de l'apprentissage statistique. Mes travaux portant sur l'équité algorithmique sont issues en grande partie d'une collaboration avec E. Chzhen, M. Hebiri, L. Oneto (DIBRIS, University of Genoa), et M. Pontil (Istituto Italiano di Tecnologia et Université College London). Les résultats issus de cette collaboration s'inscrivent dans le cadre de la régression et de la classification binaire. Une extension au cadre de la classification multi-classes est présentée au Chapitre 3. Ce travail est le fruit d'une collaboration avec R. Elie (DeepMind), M. Hebiri et F. Hue (ENSAE). L'un des points communs aux procédures décrites dans le Chapitre 2 et le Chapitre 3 est que les méthodes d'apprentissage proposées peuvent tirer parti d'un cadre d'observation semi-supervisé.

En apprentissage multi-classes l'objectif est de prédire à partir de l'observation d'une variable X, dite covariable, sa classe d'appartenance Y (ou étiquette). L'apprentissage multi-classes est une thématique qui a été très étudiée au cours des deux dernières décénies, son potentiel applicatif irriguant toutes les branches de notre société. Due à

leur complexité croissante, de nombreux jeux de données multi-classes présentent une forte ambiguité; différentes étiquettes pouvant alors correspondre à la covariable X observée. Ce type de difficulté est notamment rencontré dans les problèmes d'annotation d'image comme pour la base de données `PlantNet` (voir Göeau, Joly, and Bonnet, 2015). Dans ces situations, l'approche classique qui consiste à ne prédire qu'au moyen d'une seule classe, n'est plus adaptée. En effet, prédire une seule étiquette parmi une grande liste de candidates peut s'avérer très peu informatif. Dans ce cas, une alternative est de considérer la classification par ensemble. C'est-à-dire que l'on va prédire une liste de classes candidates plutôt qu'une unique classe. Le Chapitre 2 présente mes contributions à cette thématique de recherche.

L'apprentissage sous contrainte d'équité (*fairness*) est un sujet de plus en plus prégnant dans les applications de l'apprentissage statistique. En effet, l'objectif de la *fairness* est de "corriger" le biais dans les données observées afin que les algorithmes d'apprentissage répondent à des contraintes éthiques. Une des idées sous-jacente est qu'un algorithme d'apprentissage ne doit pas être discriminant vis-vis d'une variable dite *sensible* tel que le genre ou l'ethnie. L'étude de la *fairness* s'inscrit ainsi au coeur des débats qui animent actuellement nos sociétés modernes. L'objectif du statisticien étant de comprendre les mécanismes pouvant induire ces biais. Le Chapitre 3 résume une contribution à ce problème dans le cadre de l'apprentissage multi-classes.

La classification de données à dépendance temporelle est un domaine important de recherche en Statistique, les récentes avancées technologiques (notament l'utilisation de capteurs) rendant très facile la collecte de ce type de données. Une partie de mon travail de recherche s'inscrit dans ce cadre. Durant ma thèse, j'ai travaillé sur la mise en place de méthodes de classification pour des données de maintien postural. Les troubles du maintien postural sont susceptibles d'entraîner une chute qui est l'une des premières causes de mortalité chez les personnes âgées. Les objectifs à plus long terme sont la mise au point de protocoles d'identification de troubles du maintien postural et l'adaptation au cas par cas des protocoles de rééducation fonctionnelle. L'une des spécificités de ce travail est l'exploitation d'une modélisation des données trajectorielles comme solutions d'équations différentielles stochastiques.

Une des perspectives soulevée par ces travaux de thèse a été l'étude d'un point de vue théorique et méthodologique de procédures de classification dans le cas où les variables explicatives sont solutions d'une équation différentielle stochastique. Au Chapitre 4, je présente les contributions obtenues en collaboration avec C. Dion-Blanc (LPSM) et M. Martinez (LAMA) et apportant une réponse à ce problème. Les résultats ont été obtenus pour des trajectoires unidimensionnelles et dans un cadre d'estimation paramétrique. Ce travail a ouvert de nombreuses perspectives de recherche. En particulier, la question de l'extension de ces résultats au cadre non-paramétrique et pour des trajectoires multi-dimensionnelles fait partie d'un projet de recherche actuellement en cours. Les premiers résultats obtenus dans cette direction sont également présentés au Chapitre 4 et la suite

du projet est l'objet du travail de thèse d'Eddy Ella-Mintsa (LAMA) que je co-encadre avec C. Dion-Blanc et V.C. Tran (LAMA).

Publications

Publications

[1] C. Denis, C. Dion, M. Martinez, *A ridge estimator of the drift from discrete repeated observations of the solution of a stochastic differential equation.* Bernoulli, 2021

[2] E. Chzhen, C. Denis, M. Hebiri, *Minimax semi-supervised confidence sets multi-class classification.* Bernoulli, 2021

[3] C. Denis, M. Hebiri, A.Zaoui, *Regression with reject option and application to kNN.* NeurIPS, 2020

[4] E. Chzhen, C. Denis, M. Hebiri, L.Oneto, M.Pontil, *Fair Regression with Wasserstein Barycenters.* NeurIPS, 2020

[5] E. Chzhen, C. Denis, M. Hebiri, L.Oneto, M.Pontil, *Fair Regression via Plug-in Estimator and Recalibration via Statistical Guarentees.* NeurIPS, 2020

[6] C. Denis, E. Lebarbier, C. Lévy-Leduc, O.Martin, L.Sansonnet, *A novel regularized approach for functional data clustering: An application to milking kinetics in dairy goats.* Journal of the Royal Statistical Society : Series C, 2020

[7] C. Denis, C. Dion, M. Martinez, *Consistent procedures for multiclass classification of discrete diffusion paths.* Scandinavian Journal of Statistics, 2020

[8] C. Denis, M. Hebiri, *Consistency of plug-in confidence sets for classification in semi-supervised learning.* Journal of Nonparametrics Statistics, 2020

[9] E. Chzhen, C. Denis, M. Hebiri, L.Oneto, M.Pontil, *Leveraging Labeled and Unlabeled Data for Consistent Fair Binary Classification.* NeurIPS, 2019

[10] C. Denis, M. Hebiri, *Confidence sets with expected sizes for Multiclass Classification.* Journal of Machine Learning Research , 2017

[11] Z. Ouni, C. Denis, C. Chauvel, A. Chambaz, *Contextual ranking by passive safety of generational classes of light vehicles.* Journal of the Royal Statistical Society : Series C , 2017

[12] C. Denis, M. Hebiri,*Confidence Sets for Classification.*
Statistical Learning and Data Sciences. SLDS , 2015

[13] C. Denis, *Classification in postural style based on stochastic process modeling.*
The International Journal of Biostatistics , 2014

[14] A. Chambaz , C. Denis, *Classification in Postural style.*
The Annals Of Applied Statistics , 2012

Preprints-Tecnical reports

[15] C. Denis, R. Elie, M. Hebiri, F. Hue, *Fairnes guarantee in multi-class classification*, 2021

[16] C. Denis, C. Dion, L. Sansonnet, *Multiclass classification for Hawkes processes*, 2021

[17] E. Chzhen, C. Denis, M. Hebiri, T.Lorieul, *Set-valued classification – overview via a unified framework*, 2021

[18] E. Chzhen, C. Denis, M. Hebiri, J. Salmon, *On the benefits of output sparsity for multi-label classification*, 2017

[19] C. Denis, *Top scoring pairs classifier: asymptotics and applications*, 2013

Chapter 1

Introduction to multi-class classification problem

Multi-class classification is one of the most studied statistical frameworks, arising in many fields which range from medical applications to social studies (*e.g.*, medical diagnosis, image recognition, text categorization to name a few). For a general introduction to this topic, we refer for instance to the well-established book by Devroye, Györfi, and Lugosi (1996). This chapter is dedicated to the presentation of main notions which are at the core of this manuscript. In particular, we introduce the general multi-class classification framework in Section 1.1. In Section 1.2, we sketch the main estimation techniques in multi-class classification. Section 1.3 introduces mixture models which are widely used to model multi-class classification problems. Finally, in Section 1.4 we present some modern challenges that are addressed in this manuscript.

1.1 General framework

In multi-class classification, a generic observation is a pair of random variables (X, Y) such that the feature vector X belongs to some space $\mathcal{X}$ and the label Y takes its values in $\mathcal{Y} = \{1, \dots, K\}$, with $K \geq 2$, and indicates the associated class to X. The distribution of (X, Y), denoted by $\mathbb{P}$, is assumed to be unknown to the statistician. In this setting, a classifier g is a measurable function which maps $\mathcal{X}$ onto $\mathcal{Y}$. Hence, a classifier is viewed as a prediction of the associated label Y. Throughout this manuscript, for a given classifier g, we consider the misclassification risk

$$R(g) = \mathbb{P}(g(X) \neq Y)$$

as the measure of the performance of a classifier g. The set of all classifiers is denoted by $\mathcal{G}$. Note that we often refer to multi-class (or multi-category) classification when $K \geq 3$. The specific case when $K = 2$ refers to the binary classification setting and is the most

studied classification problem in the machine learning community. We distinguish this case from the general multi-class classification framework where $K \geq 3$ even though these two problems share some similarities in nature. However, important technical differences also exist making the multi-class setting as a case study *per se*. We will discuss this point later.

1.1.1 Bayes classifier

At this step, a natural object to consider is a classifier g^* which achieves the minimum risk

$$g^* \in \arg\min_{g \in \mathcal{G}} R(g). \tag{1.1}$$

An elementary, yet important result in classification context is the characterization of the optimal rule g^*, namely the Bayes classifier , expressed for all $x \in \mathcal{X}$ as

$$g^*(x) \in \arg\max_{k \in \mathcal{Y}} p_k(x), \quad \text{with} \quad p_k(x) = \mathbb{P}(Y = k | X = x).$$

Clearly, since the distribution of (X, Y) is unknown, we do not have access to the Bayes classifier. Hence, an objective in multi-class classification is to build, based on a learning sample, an empirical classifier which mimics the Bayes classifier.

In view of the form of the optimal classifier, it is worth noting that the conditional probabilities p_k will play a central role in the estimation of g^*. If we consider the binary case ($K = 2$, and $\mathcal{Y} = \{0, 1\}$) the Bayes classifier can also be expressed simply as a thresholding of the conditional probability p_1

$$g^*(x) = \mathbb{1}_{\{p_1(x) \geq 1/2\}}.$$

This is in fact a specificity of the binary classification that eases the statistical analysis in this framework. This property does not hold in the multi-class setting and makes a major difference between the two settings.

As we have seen above, the optimal classifier achieves the minimum risk. Therefore, it is natural that the performance of a given classifier g is evaluated through its excess risk which is defined as

$$\mathcal{E}(g) = R(g) - R(g^*).$$

In particular, we can show a closed form of this excess risk.

Proposition 1.1. *For any $g \in \mathcal{G}$, the following holds*

$$\mathcal{E}(g) = \mathbb{E}_X \left[\sum_{k=1}^{K} \sum_{j \neq k} |p_k(X) - p_j(X)| \mathbb{1}_{\{g(X)=j,\, g^*(X)=k\}} \right].$$

Interestingly, the above expression of the excess risk in the multi-class setting is a straightforward but important generalization of the binary case which can be written as

$$\mathcal{E}(g) = \mathbb{E}_X \left[|2p_1(X) - 1| \mathbb{1}_{\{g^*(X) \neq g(X)\}} \right].$$

Next step consists in building an estimator of Bayes classifier and provides some theoretical insights, such as a control of its excess risk. This is the purpose of the following section. In particular, Section 1.2 can be viewed as a road-map for the study of classification procedures in multi-class framework.

1.2 Estimation strategy

In multi-class classification, the construction of a classification procedure relies on a learning sample $\mathcal{D}_n = \{(X_i, Y_i), \quad i = 1, \dots, n\}$ of i.i.d. copies of (X, Y). Hence, a classification procedure or an empirical classifier is a measurable (classifier-valued) function of the learning sample $\mathcal{D}_n$. Following the standard practice, we will often use the general notation $\hat{g}$ to denote an empirical classifier and drop the explicit dependency on the learning sample. The excess risk of a empirical classifier $\hat{g}$ is then a random variable (or, rather a random element, since $\hat{g}$ is classifier-valued, but we omit this benign distinction) that depends on $\mathcal{D}_n$. One basic and fundamental property that we expect from a good empirical classifier $\hat{g}$ is its point-wise consistency *w.r.t.* some class $\mathcal{P}$ of joint distributions P of (X, Y)

$$\lim_{n \to +\infty} \mathbb{E}\left[\mathcal{E}(\hat{g})\right] = 0, \quad \text{for} \quad P \in \mathcal{P},$$

which ensures that the classifier $\hat{g}$ is asymptotically as good as the Bayes classifier (note that at this stage we are not aiming at characterizing rates of convergence, since the above guarantee is not sufficient for this purpose). In principle, there are two main ways to build a consistent classifier, both of which are relying on the definition and the characterization of the Bayes classifier. The first strategy involves the estimation of the conditional probabilities and relies on the *plug-in* principle. The second one takes advantage of Equation (1.1) and relies on the empirical risk minimization principle.

1.2.1 Plug-in approach

Provided the characterization of the Bayes classifier, the most intuitive way to build consistent classification procedures relies on the construction of consistent estimators of the conditional probabilities p_k for all $k \in \{1, \dots, K\}$. More precisely, the first step consists in building estimators $\hat{p}_k$ of the p_k's based on the sample $\mathcal{D}_n$. We do not focus on this step which is rather well studied (van de Geer, 1990; Devroye, Györfi, and Lugosi, 1996; Wegkamp and van de Geer, 1996; Tsybakov, 2009a); let us, however, mention that

standard estimation methods can be used for this task such as the Kernel estimators or the k-Nearest Neighbor estimators (see Devroye, Györfi, and Lugosi, 1996, for instance).

Then, the empirical classifier is defined for all $x \in \mathcal{X}$ by

$$\hat{g}(x) \in \arg\max_{k \in \mathcal{Y}} \hat{p}_k(x).$$

This approach is motivated by the following result, which links the classification excess risk to its regression counterpart.

Proposition 1.2. *We have that*

$$\mathbb{E}\left[\mathcal{E}\left(\hat{g}\right)\right] \leq 2\mathbb{E}\left[\sum_{k=1}^{K} |\hat{p}_k(X) - p_k(X)|\right].$$

The above proposition ensures that the consistency of the $\hat{p}_k$'s implies the consistency of the plug-in classifier $\hat{g}$. Furthermore, one can also note that rates of convergence for the excess risk can trivially be deduced from the rates of convergence of the estimators $\hat{p}_k$ *w.r.t.* the L_1-norm. Importantly, the properties of the plug-in classifier $\hat{g}$ are inherited from the properties of the estimators $\hat{p}_k$. We refer to (Devroye, Györfi, and Lugosi, 1996; Yang, 1999; Audibert and Tsybakov, 2007; Tsybakov, 2009a) for more details.

1.2.2 Empirical risk minimization procedure

In this section, we describe another way to build consistent empirical classifiers. It relies on the empirical risk minimization principle which has been widely studied in the context of supervised learning.

Consider a set of classifiers $\mathcal{G}' \subset \mathcal{G}$, we define the empirical risk of a given classifier $g \in \mathcal{G}'$ by

$$\hat{R}(g) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{g(X_i) \neq Y_i\}},$$

and we naturally set the empirical risk minimizer (E.R.M.) $\hat{g}$ over the set $\mathcal{G}'$

$$\hat{g} \in \arg\min_{g \in \mathcal{G}'} \hat{R}(g).$$

From the definition of $\hat{g}$, we deduce that

$$\mathbb{E}\left[\mathcal{E}\left(\hat{g}\right)\right] = \mathbb{E}\left[R(\hat{g}) - R(\bar{g})\right] + \mathbb{E}\left[R(\bar{g}) - R(g^*)\right],$$

where $\bar{g} \in \arg\min_{\mathcal{G}'} R(g)$. Hence, the excess risk of the empirical risk minimizer is classically decomposed into two terms. The first one is the variance term which is usually studied using tools from the empirical process theory and depends on complexity assumptions imposed on the set $\mathcal{G}'$. In particular, assumptions on the entropy of the set $\mathcal{G}'$

are considered (van de Geer, 2000). The second one is the bias term and also depends on the complexity assumptions on $\mathcal{G}'$ and on the assumptions on the joint distribution of (X, Y). In this case, the properties of $\hat{g}$ are inherited from the properties of $\mathcal{G}'$. We refer to (Devroye, Györfi, and Lugosi, 1996; Vapnik, 1998; Massart and Nédélec, 2006) for the statistical property of the E.R.M. estimator.

Nevertheless, due to the non-convexity of the minimization problem defined by Equation (1.2.2), the estimator $\hat{g}$ is in general not computable. Although, the E.R.M. $\hat{g}$ offers appealing properties, it cannot be considered in this form for practical purposes. To avoid this issue, convex surrogates have been provided in the statistical literature. In particular, we refer to the work by Freund and Schapire (1997), Vapnik (1998), Friedman, Hastie, and R. Tibshirani (2000), Zhang (2004), Bartlett, Jordan, and McAuliffe (2006), Tewari and Bartlett (2007), and Yuan and Wegkamp (2010) for a complete study. Hereafter, we present the convexification of the problem (1.2.2) in the case of the square loss with the one-versus-all approach (Zhang, 2004). We consider this case for simplicity but also to highlight the link between classification and regression. Of course, other choices of loss functions can be considered and we refer to (Zhang, 2004; Bartlett, Jordan, and McAuliffe, 2006) for more details.

Convexification: from regression to classification. A particular feature of convexification is that it relies on a score function $\mathbf{f}(\cdot) = (f_1(\cdot), \ldots, f_K(\cdot))$, that is, a measurable mapping from $\mathcal{X}$ to $\mathbb{R}^K$. Each of these score functions results in an associated classifier $g_{\mathbf{f}}$ defined as

$$g_{\mathbf{f}}(x) \in \arg\max_{k \in \mathcal{Y}} f_k(x).$$

Even though, the score function is not well tailored for the classification task (it does not take values in a discrete set) we can, nevertheless, consider its L_2-risk as

$$R_2(\mathbf{f}) = \mathbb{E}\left[\sum_{k=1}^{K} (Z_k - f_k(X))^2\right], \quad \text{with} \quad Z_k = 2 \cdot \mathbb{1}_{\{Y=k\}} - 1.$$

Hence, the risk R_2 is simply the sum over all $k \in \mathcal{Y}$ of the L_2-risks associated to the regression problems of Z_k onto X. Interestingly, this formulation can be viewed as the generalization of the regression problems associated to K separate binary classification problems where for each $k \in \mathcal{Y}$, we focus on the classification problem $Y = k$ against $Y \neq k$. From this perspective, we can see this formulation as a one-versus-all problem. Here again, we can define the optimal score function $\mathbf{f}^*$ *w.r.t.* R_2

$$\mathbf{f}^* \in \arg\min_{\mathbf{f}} R_2(\mathbf{f}), \tag{1.2}$$

where the infimum is taken over all measurable functions. Hence, for each $k \in \mathcal{Y}$, we have that $f_k^*(X) = \mathbb{E}[Z_k|X] = 2p_k(X) - 1$. From the definition of $\mathbf{f}^*$, one can establish

Zhang's Lemma (see Zhang, 2004) which connects the regression problem defined by Equation (1.2) to the multi-class problem.

Lemma 1.1. *Let* $\mathbf{f}$ *a score function, then the following holds*

$$\mathbb{E}\left[R(g_{\mathbf{f}}) - R(g^*)\right] \leq \frac{1}{\sqrt{2}}\left(\mathbb{E}\left[R_2(\mathbf{f}) - R_2(\mathbf{f}^*)\right]\right)^{1/2}.$$

The immediate consequence of this result is that the properties of the classifier $g_{\mathbf{f}}$ can be studied through the properties of the score function $\mathbf{f}$. In view of this result, it is then natural to consider empirical risk minimizer estimator *w.r.t.* L_2-risk. More precisely, we define for a given score function $\mathbf{f}$ its empirical risk

$$\hat{R}_2(\mathbf{f}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}(f_k(X_i) - Z_k^i)^2, \quad \textit{where} \quad Z_k^i = 2 \cdot \mathbb{1}_{\{Y_i=k\}} - 1.$$

Therefore for a set $\mathcal{F}$ of score functions, the empirical risk minimizer $\hat{\mathbf{f}}$ is then defined as

$$\hat{\mathbf{f}} \in \arg\min_{\mathbf{f}\in\mathcal{F}} \hat{R}_2(\mathbf{f}).$$

Theoretical guarantees of $\hat{\mathbf{f}}$ are then obtained through complexity assumptions imposed on the set $\mathcal{F}$. Note that, the resulting estimator $\hat{\mathbf{f}}$ can be also interpreted as a plug-in classifier. Indeed, in view of the form the optimal score function f^*, estimators of conditional probabilities may be simply deduced by setting $\hat{p}_k = \frac{\hat{f}_k+1}{2}$.

1.3 The case of the mixture model

In Section 1.2, we provide a high-level description of general methods to estimate the Bayes classifier. Of course, the theoretical properties of these procedures depend on the assumption on the joint distribution of (X, Y). In this section, we consider the particular case where the distribution of (X, Y) comes from a mixture model. For simplicity, we assume that $\mathcal{X} = \mathbb{R}^d$ and present the mixture model for multi-class classification in the parametric setting.

Let $\Theta \subset \mathbb{R}^p$. We assume that Y is distributed according to $(\pi_k)_{k\in\mathcal{Y}}$. Conditional on $Y = k$, we assume that X admits a density $f_{\theta_k}(\cdot) = f(\theta_k, \cdot)$ *w.r.t.* the Lebesgue measure. Note that the function f is supposed to be known while for each k the parameter θ_k is unknown. In this case, the marginal density of X is given for all $x \in \mathbb{R}^d$ by

$$f_X(x) = \sum_{k=1}^{K} \pi_k f_{\theta_k}(x).$$

Hence, under this parametric assumption, the classes are simply discriminated by the parameters $(\theta_k)_{k\in\mathcal{Y}}$. Thanks to the Bayes formula, we easily deduced that

$$p_k(x) = \frac{\pi_k f_{\theta_k}(x)}{f_X(x)}\mathbb{1}_{\{f_X(x)\neq 0\}}.$$

In other words, in the considered parametric setting, the Bayes classifier is known up to the parameter $(\theta_k)_k$ and the marginal distribution of the label. Even though restrictive, such a modelling can be appealing from the computational and statistical perspectives, leading to faster algorithms and better statistical rates. Indeed, the distribution of Y can be estimated by the empirical frequencies $\hat{\pi}_k$ while estimator $\hat{\theta}_k$ of parameter θ_k can be obtained by standard parametric methods, which are typically reduced to convex optimization problems. Naturally, given estimators $\hat{\pi}_k$ and $\hat{\theta}_k$, we consider the plug-in classifier $\hat{g}$ defined by

$$\hat{p}_k(x) = \frac{\hat{\pi}_k f_{\hat{\theta}_k}(x)}{\hat{f}_X(x)} \mathbb{1}_{\{\hat{f}_X(x) \neq 0\}}.$$

Apart from their appealing computational and statistical properties, mixture models and the described methodology results in classifiers which are easily interpretable, which should be contrasted with general "black box" algorithms (*e.g.* neural networks). It goes without saying that the numerical performance of this estimation procedure strongly depends on the considered model and requires an important modeling effort.

1.4 Overview of the results

In the last decade, several challenges have emerged in multi-class classification. In particular, modern multi-class datasets are often heterogeneous, complex, and also large scale, leading to multi-class datasets which may involve a large number of observations, of variables, and/or of classes. In the present manuscript, we will try to address some of these characteristics providing a specific answer adapted to the problem in hand.

Due to the high ambiguity between classes inherent in large scale multi-class datasets, single-output algorithms often exhibit poor performance. As an illustration, the error rate of state-of-the art methods is around 20% on the `ImageNet` dataset (Xie et al., 2017). In this context, we may question the relevance and value of single-output predictions. It is precisely within these high ambiguity problems that we benefit from the set-valued classification approach. Indeed, set-valued predictors, which allow to output a set of possible class candidates rather than an single class, are dedicated to handle the ambiguity in multi-class classification. In Chapter 2, we present a statistical analysis of plug-in and empirical risk minimization methods in the framework of set-valued classification with controlled expected size.

Mitigating bias in data is an active research field in the machine learning community. This is premised on the fact that learning algorithms may inherit bias in the data during the training process, leading to undesired knock-on effect on future decisions. In particular, severe conflicts may arise with ethical criteria of the modern society using algorithms that only have a purpose of prediction accuracy. *Algorithmic fairness*, which has been emerging in the last few years, try to give a solution to the problem of mitigating the bias

in data. In Chapter 3, we provide results obtained in fairness in the context of multi-class classification.

The recent advance of modern technologies has generated a large number of datasets which can be modeled as functional data. In this context, a major challenge is to provide learning algorithms that are designed to handle temporal data. Many methods have been developed to treat such type of data (see (Ramsay and Silverman, 2007) for an overview). In Chapter 4, we present a contribution to this research area by considering the specific case where the feature is modeled as a diffusion sample path. More precisely, results provided in Chapter 4 focus on the multi-class problem where the data come from a solution of a *mixture of stochastic differential equations*, and can be viewed as an extension of the mixture model presented in Section 1.3 dedicated to functional data classification.

Chapter 2

Set-valued classification

In this chapter, I present results obtained for set-valued classification. This setting is motivated by the fact that modern multi-class datasets can be extremely ambiguous due to the large variety of label candidates. In this kind of situation, single-output classifier can lead to wrong and unsatisfactory predictions. In contrast, set-valued classifiers allow to predict multiple candidate labels and provide an alternative that may improve the prediction accuracy. In the set-valued classification framework, a predictor is allowed to predict not only a single label, but a set of candidates labels. Hence, it offers a natural way to work with the ambiguity.

Based on previously obtained results in (Denis and Hebiri, 2020) in the context of classification with reject option, we propose in (Denis and Hebiri, 2017) the set-valued approach, where the size of the output (*e.g.* the number of the predicted labels) can be controlled in expectation by the practitioner. Up to our knowledge, that was the first work dealing with this framework at that time. We refer to this new setting as *set-valued classification with control expected size*. This approach shares some similarities with the top-k procedure which *always* outputs k candidate labels (Lapin, Hein, and Schiele, 2015). In particular, by controlling the size of the output, both approaches ensures the interpretability of the output. However, in contrast to the top-k procedure, controlling the size in expectation of the predictor offers the opportunity of handling the heterogeneity of the feature space. That is to say, the size of the output is adaptive *w.r.t.* the marginal distribution of X.

In Section 2.3, we present a general procedure which is the result of several works. In particular, this procedure can leverage unlabeled data to satisfy the size constraint in expectation. We then show that consistent set-valued predictors under expected size constraint can be obtained in a semi-supervised way. Interestingly, the proposed methodology can be generalized to other learning problems under constraint. As an important development, we extend it to algorithmic fairness (Chzhen et al., 2020a,b). In (Denis and Hebiri, 2017), we propose an empirical risk minimization procedure which is presented in Section 2.4. In light of this work, we propose in (Chzhen, Denis, and Hebiri, 2021)

a minimax analysis of the set-valued classification with controlled expected size setting. Importantly, this work highlights the relevance of the semi-supervised approach in the set-valued classification framework. A part of these results are presented in Section 2.5.

2.1 General framework

In this section, we introduce main definitions and notations. A set-valued classifier Γ is a (measurable) function which maps $\mathcal{X}$ onto $2^{\mathcal{Y}}$. The set of all set-valued classifiers is denoted by $\mathbf{\Gamma}$. Naturally, two parameters arise in this framework: the *error rate* and the *(expected) size* of a set-valued which are defined as

$$\underbrace{P(\Gamma) = \mathbb{P}(Y \notin \Gamma(\boldsymbol{X}))}_{\text{error}}, \qquad \underbrace{\mathrm{S}(\Gamma) = \mathbb{E}\,|\Gamma(\boldsymbol{X})|}_{\text{size}} .$$

These two notions appear as fundamental and they have different names depending on the community and the field (for instance, the error rate is often called coverage, recall, or risk). The balance between the set size and the error rate is a common denominator between all set-valued classifiers, and depending on the application they should be considered in a different way. Hence different framework of set-valued classification are studied in the literature. In the following section we present the most considered set-valued settings and try to give some of their characteristics.

2.1.1 Set-valued approaches

Arguably the most natural set-valued classifier is the one that outputs a fixed amount of candidate labels for each instance. This type of set-valued classification strategies is called top-k prediction (Lapin, Hein, and Schiele, 2015; Oh, 2017), where k is the amount of candidate labels predicted. For an observation x, the optimal way to output k candidates is to select those that correspond to k highest conditional probabilities $p_1(x), \dots, p_K(x)$. For instance, top-5 prediction is the one chosen for the `ImageNet` dataset (Russakovsky et al., 2015). One of the advantage of this approach is that the interpretability of the output is controlled through the parameter k. Indeed, The top-k procedure is then defined as

$$\textbf{Top-}k\textbf{:} \qquad \Gamma^*_{\mathrm{top(s)}} \in \arg\min \left\{ \mathbb{P}(Y \notin \Gamma(\boldsymbol{X})) \,:\, \forall \boldsymbol{x} \in \mathbb{R}^d \;\; |\Gamma(\boldsymbol{x})| = \mathrm{s} \right\} ,$$

with some $\mathrm{s} \in \mathcal{Y}$. One of the drawback of this approach is that it does not take into account inhomogeneous structure of the problem. Roughly speaking, there is no reason to always predict the same amount of labels all the time. As a remedy, we propose to replace the *hard constraint* on the size of the output by a constraint on the *expected size*. More formally, for a given $\mathrm{s} \in (0, K)$, we consider the following constraint problem

$$\mathrm{s}\textbf{-Oracle:} \qquad \Gamma^*_{\mathrm{s}} \in \arg\min \{ \mathbb{P}(Y \in \Gamma(X)) : \;\; \mathbb{E}_X |\Gamma(X)| \leq \mathrm{s} \}. \tag{2.1}$$

Interestingly, while this formulation generalizes the top-k approach, it also preserve its main characteristic. Indeed, since the parameter $s \in (0, K)$ is fixed before hand, by considering this approach we also control the interpretability and stability of the outcome.

Yet another way to define a set-valued classification framework is proposed by Vovk (2002a,b) and Vovk, Gammerman, and Shafer (2005) and statistically addressed by Sadinle, Lei, and Wasserman (2018), where for a fixed $\alpha \in (0,1)$, they define Γ^*_α as

$$\textbf{Controlled error-rate:} \qquad \Gamma^*_\alpha \in \arg\min \{\mathbb{E}|\Gamma(\boldsymbol{X})| \,:\, \mathbb{P}(Y \notin \Gamma(\boldsymbol{X})) \leq \alpha\} \quad ,$$

that is, Γ^*_α is the "smallest" set-valued classifier with controlled probability of error. Even though this framework is intuitive, this apporach does not allow to control the size of the output. In certain situations it suffers from the lack of interpretability and the lack of stability *w.r.t.* the parameter α. In (Chzhen, Denis, and Hebiri, 2021; Chzhen et al., 2021), we illustrate this phenomena. Other approach based on the control of the point-wise error, $\mathbb{P}(Y \notin \Gamma(X)|X = x)$ are recently investigated in the literature (Györfi and Walk, 2020; Romano, Sesia, and Candès, 2020).

In the rest of this chapter, we present the results obtained in (Denis and Hebiri, 2017; Chzhen, Denis, and Hebiri, 2021) for the statistical problem induced by Equation (2.1).

2.2 Set-valued classification with controlled expected size

Let $s \in \mathcal{Y}$. This section establishes the important properties of the optimal set-valued predictor with controlled expected size Γ^*_s. We recall its definition here.

$$\textbf{s-Oracle:} \qquad \Gamma^*_s \in \arg\min \{P(\Gamma) \,:\, \Gamma \in \boldsymbol{\Gamma} \text{ s.t. } S(\Gamma) \leq s\} \quad ,$$

2.2.1 Properties of s-Oracle.

Let us start by the following mild continuity assumption which is assumed throughout this chapter.

Assumption 1 (Continuity of CDF). *For all $k \in \mathcal{Y}$ the cumulative distribution function (CDF) $F_{p_k}(\cdot) := \mathbb{P}_{\boldsymbol{X}}(p_k(\boldsymbol{X}) \leq \cdot)$ of $p_k(\boldsymbol{X})$ is continuous on $(0,1)$.*

Continuity Assumption 1 is central. It allows to express the s-Oracle set-valued classifier Γ^*_s in the form of thresholding and to highlight the relation of the constrained minimization with its unconstrained counterpart.

Proposition 2.1. *Let the function $G : [0,1] \to [0,K]$ be defined for all $t \in [0,1]$ as*

$$G(t) := \sum_{k=1}^{K} \left(1 - F_{p_k}(t)\right) = \sum_{k=1}^{K} \mathbb{P}_X(p_k(X) > t) \quad ,$$

then under Assumption 1, for $\mathrm{s} \in (0, K)$*, the* s*-Oracle set-valued classifier* Γ^*_{s} *can be obtained for all* x *as*

$$\Gamma^*_{\mathrm{s}}(x) = \left\{ k \in \mathcal{Y} \,:\, p_k(x) \geq G^{-1}(\mathrm{s}) \right\}, \tag{2.2}$$

where we denote by G^{-1} *the generalized inverse of* G *defined for all* $\mathrm{s} \in (0, K)$ *as* $G^{-1}(\mathrm{s}) := \inf\{t \in [0,1] \,:\, G(t) \leq \mathrm{s}\}$.

Note that the threshold $G^{-1}(\mathrm{s})$ depends on the joint distribution $\mathbb{P}$ and thus, is unknown beforehand. Furthermore, under Assumption 1, the considered framework is well posed in the sense that the s-Oracle set-valued classifier Γ^*_{s} is unique up to changes on sets of $\mathbb{P}_X$ zero measure.

Theorem 2.1. *For every* $\mathrm{s} \in (0, K)$*, under Assumption 1 the* s*-Oracle set-valued classifier* Γ^*_{s} *defined in Proposition 2.1 is unique up to changes on* $\mathbb{P}_X$ *zero measure. That is, for all set-valued classifier such that* $\mathrm{S}(\Gamma) \leq \mathrm{s}$ *either of the following conditions hold*

- $P(\Gamma) > P(\Gamma^*_{\mathrm{s}})$,
- $\Gamma(\boldsymbol{x}) = \Gamma^*_{\mathrm{s}}(\boldsymbol{x})$ *for almost all* $\boldsymbol{x} \in \mathbb{R}^d$ w.r.t. $\mathbb{P}_X$.

Under Assumption 1, we can also provide another characterization of the s-Oracle set-valued classifier. The next proposition establishes that s-Oracle can be obtained via an unconstrained minimization, which trades-off the error and the size.

Proposition 2.2. *Let* $\mathrm{s} \in (0, K)$*, and assume that Assumption 1 is fulfilled, then the* s*-Oracle defined in Equation (2.2) is a minimizer over* $\boldsymbol{\Gamma}$ *of the following risk*

$$R_{\mathrm{s}}(\Gamma) = P(\Gamma) + G^{-1}(\mathrm{s})\,\mathrm{S}(\Gamma) \ .$$

Consequently, the accuracy of a set-valued classifier Γ can be quantified according to its excess risk

$$R_{\mathrm{s}}(\Gamma) - R_{\mathrm{s}}(\Gamma^*_{\mathrm{s}}) = \sum_{k=1}^{K} \mathbb{E}_{\mathbb{P}_X}\left[|p_k(\boldsymbol{X}) - G^{-1}(\mathrm{s})| \mathbf{1}_{\{k \in \Gamma(X) \triangle \Gamma^*_{\mathrm{s}}(X)\}} \right] \ , \tag{2.3}$$

One can already observe that the above excess risk of any set-valued classifier Γ relies on the behavior of the conditional probabilities p_k around the threshold $G^{-1}(\mathrm{s})$.

2.2.2 Measures of performance.

Let us conclude this section by introducing performance measures that we will study in the context of set-valued classification with controlled expected size. Of course, the objective is to measure the "distance" *w.r.t.* the oracle set-valued classifier. Let $s \in (0, K)$

and Γ a set-valued classifier. The most intuitive measure of performance is the Hamming distance, which is a suitable way to quantify the distance between two sets. The Hamming risk of Γ is defined as

$$H(\Gamma) = \mathbb{E}_X \left[|\Gamma(X) \triangle \Gamma_s^*(X)| \right].$$

Hence, since $\triangle$ stands for the symmetric difference, we then have that $H(\Gamma) \geq 0$ and that $H(\Gamma) = 0$ implies that $\Gamma(x) = \Gamma_s^*(x)$ almost surely *w.r.t.* $\mathbb{P}_X$. In view of Proposition 2.2, another natural measure of performance is the excess risk $R_s(\Gamma) - R_s(\Gamma_s^*)$. Note that the excess risk is bounded by the Hamming risk, and consistency *w.r.t.* to the Hamming risk implies consistency *w.r.t.* the risk R_s.

2.3 Plug-in set-valued classifier

In this section, we present a general data-driven procedure which relies on the plug-in principle. Let us point out that the feasible set $\{\Gamma \in \mathbf{\Gamma} : \mathrm{S}(\Gamma) \leq \mathrm{s}\}$ of the above problem is distribution dependent. It implies that *a priori* we cannot decide whether a given set-valued classifier is feasible. However, this set only depends on the marginal distribution $\mathbb{P}_X$ of the features, which motivates us to introduce unlabeled sample in the observational model. Hence, we are interested in the semi-supervised setup of this problem. That is, in what follows it is assumed that two independent samples are provided, a labeled one $\mathcal{D}_n^{\mathrm{L}} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ and an unlabeled sample $\mathcal{D}_N^{\mathrm{U}} = \{X_{n+1}, \ldots, X_{n+N}\} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_X$ both being independent from (X, Y).

Based on the learning samples, the statistical goal is then to build an empirical set-valued classifier $\hat{\Gamma} : (\mathbb{R}^d \times \mathcal{Y})^n \times (\mathbb{R}^d)^N \to \mathbf{\Gamma}$, which mimics the behavior of Γ_s^*. In particular one of the desired property of $\hat{\Gamma}$ is that the expected size of the empirical set-valued classifier satisfy the size constraint. However, as noticed above, since the constraint on the size depends on the marginal distribution $\mathbb{P}_X$, we relax this constraint and focus on empirical set-valued for which $\mathbb{E}\left[\mathrm{S}(\hat{\Gamma})\right] \to s$.

2.3.1 Construction of the estimator

The expression of the optimal set-valued classifier provided Equation (2.2) naturally suggests a plug-in approach. The proposed procedure is two-step procedure. In a first step, based on the labeled sample, we build preliminary estimators $\hat{p}_k$ of the posterior probabilities p_k. In order to build the plug-in estimator, we consider for all $t \in (0, 1)$

$$\widetilde{G}(t) = \sum_{k=1}^{K} \mathbb{P}_X \left(\hat{p}_k(X) \geq t | \mathcal{D}_n \right),$$

and then consider the pseudo-oracle set valued classifier defined for all $x \in \mathcal{X}$ as

$$\widetilde{\Gamma}(x) = \{k \in \mathcal{Y}, \;\; \hat{p}_k(x) \geq \tilde{G}^{-1}(s)\}.$$

Before to go further, it is important to note that, as for the s-Oracle classifier, the continuity of $\widetilde{G}$ is required.

Assumption 2. *For all $k \in \mathcal{Y}$, conditionally on the data, the cumulative distribution function $F_{\hat{p}_k}(t) := \mathbb{P}_X(\hat{p}_k(X) \leq t)$ of $\hat{p}_k(\boldsymbol{X})$ is continuous on $(0,1)$.*

The above assumption is important since it ensures that

$$\mathbb{E}\left[\mathcal{S}(\tilde{\Gamma})\right] = s.$$

A key point of our study is that it is always possible to ensure that this assumption is satisfied by introducing a random perturbation. Let $(\zeta_1, \ldots, \zeta_K)$ be K i.i.d. random variables distributed according to a uniform distribution on $[0, u]$ ($u > 0$) and independent from all other variables. For each $k \in \mathcal{Y}$, we introduce the randomized estimator

$$\bar{p}_k(X, \zeta_k) = \hat{p}_k(X) + \zeta_k, \quad \text{and} \quad \tilde{G}_u(t) = \sum_{k=1}^{K} \mathbb{P}_{X,\zeta_k}(\bar{p}_k(X, \zeta_k) \geq t | \mathcal{D}_n).$$

Then, the function $\tilde{G}_u$ is continuous and the resulting randomized set-valued classifier satisfies the size constraint. Since the distribution of $\tilde{G}_u$ depends on the distribution of X and $(\zeta_k)_{k\in\mathcal{Y}}$, in a second step, based on the unlabeled dataset $\mathcal{D}_N^{\mathrm{U}}$ and (ζ_k^i) i.i.d. from a Uniform distribution on $[0, u]$ and independent from all other random variable, we define its empirical counterpart for all $t \in (0,1)$

$$\hat{G}_u(t) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{1}_{\{\bar{p}_k(X_{n+i}, \zeta_k^i) \geq t\}}.$$

The resulting set-valued classifier is then defined as

$$\hat{\Gamma}_u(X, \zeta) = \{k \in \mathcal{Y}, \ \ \bar{p}_k(X, \zeta_k) \geq \tilde{G}_u^{-1}(s)\}.$$

Note that this approach is not explicitly presented in (Denis and Hebiri, 2017; Chzhen, Denis, and Hebiri, 2021), but rigorous presentation of this construction can be found in (Denis, Hebiri, and Zaoui, 2020) in the context of regression with reject option.

Of course, if the size of the unlabeled $N = 0$, we simply split the labeled sample into two independent samples. Then, one sample is used to estimate the posterior probabilities while the second one is used to calibrate the threshold. Interestingly, the proposed procedure is general and we successfully extend this methodology to other learning problem under constraint (Chzhen et al., 2020b; Denis, Hebiri, and Zaoui, 2020; Denis et al., 2021). In particular, we show an application of this methodology in Chapter 3.

2.3.2 Theoretical properties

In this section, we establish the first properties of the plug-in set-valued classifier.

Universal expected size guarantee. The first result is a distribution-free result which shows that the set-valued classifier $\hat{\Gamma}_u$ meets, asymptotically, the expected size constraint.

Theorem 2.2. *Let $s \in \mathcal{Y}$, and $u > 0$, there exists a universal constant such that for any distribution of (X, Y) and any estimators $\hat{p}_k$ of p_k, it holds that*

$$\mathbb{E}\left[\left|\mathbb{E}_\zeta\left[\mathcal{S}(\hat{\Gamma}_u)\right] - s\right|\right] \leq \frac{C}{\sqrt{N}}.$$

Interestingly, this result highlights that the proposed post-processing procedure satisfies the expected size constraint up to a reminder term of order $N^{-1/2}$ and an rely on any off-the-shell estimator of the posterior probabilities p_k. Importantly, only unlabeled data are required to calibrate the threshold.

Consistency. We now establish, under mild assumptions, the consistency of $\hat{\Gamma}_u$ and then show that the proposed set-valued classifier mimics the oracle one both in term of error and expected size.

Theorem 2.3. *Let $s \in \mathcal{Y}$. Assume that $u = u_n \to 0$ and that for each $k \in \mathcal{Y}$, we have $\mathbb{E}\left[|\hat{p}_k(X) - p_k(X)\right] \to 0$. Under Assumption 1, it holds that*

$$\lim_{n,N\to+\infty} \mathbb{E}\left[H\left(\hat{\Gamma}_{u_n}\right)\right] = 0.$$

This result ensures that asymptotically $\hat{\Gamma}_u = \Gamma_s^*$ provided that the perturbation u is sufficiently small and that estimators $\hat{p}_k$ are consistent *w.r.t.* to L_1-risk. As an immediate consequence of the above result, we also have that $\hat{\Gamma}_u$ is consistent with respect to the risk R_s. Note that we only establish the consistency of the plug-in set-valued classifier. In Section 2.5, we also show that plug-in classifier can achieve optimal rates of convergence.

2.3.3 Numerical evaluation

In this section, we illustrate the numerical properties of the plug-in set-valued classifier. One of our primary objectives is to provide experimental evidences that highlight the importance of the assumptions involved in our framework. In particular, we focus on Assumption 2 that is required from $\hat{p}_1, \ldots, \hat{p}_K$ and moreover, in case $N = 0$, we highlight the importance of data splitting for construction of $\hat{p}_1, \ldots, \hat{p}_K$ and estimation of G. To this end, we compare the performance of the set-valued classifier $\hat{\Gamma}_u$ against several natural alternatives:

a) classifier that *violates the continuity Assumption 2*: exploits an estimator $\hat{p}_k$ of the regression functions p_k so that Assumption 2 is violated;

b) classifier that *violates the independence between samples $\mathcal{D}_n^{\mathrm{L}}$ and $\mathcal{D}_N^{\mathrm{U}}$*: uses the X_i's used for training the regression functions $\hat{p}_k$ to also estimate G function.

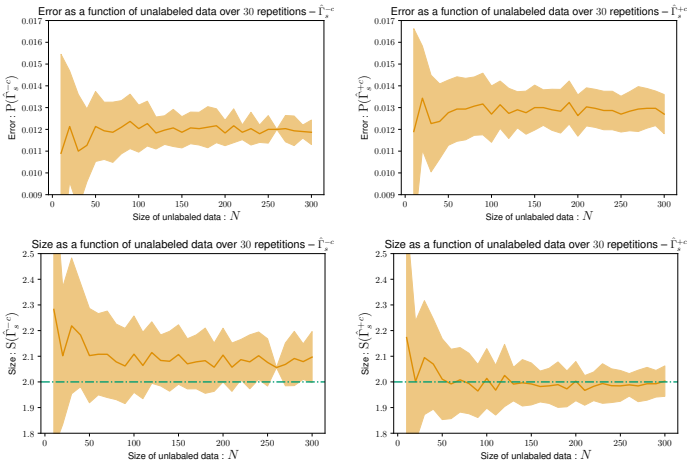

Figure 2.1: Importance of Assumption 2. Set-valued classifier $\hat{\Gamma}_s^{-c}$ does not satisfy Assumption 2. Set-valued classifier $\hat{\Gamma}_s^{+c}$ modifies the output of the random forest to force continuity.

We consider the `MNIST` dataset (LeCun et al., 1998) which is composed of images of handwritten digits from 0 to 9. The goal is to predict which digit is present on the image. For the base estimator $\hat{p}_k$ we select the `random forest` method (Breiman, 2001) which *does not* satisfy the continuity Assumption 2.

We fix $u = 10^{-6}$ for the construction of $\hat{\Gamma}_u$ (in the sequel, the dependency *w.r.t.* u is dropped) while we do not add the random perturbation when we evaluate the plug-in set-valued classifier in the case where Assumption 2 is not satisfied. Note that in the last case, the construction of the set-valued classifier is the same as $\hat{\Gamma}_u$. We simply replace $\bar{p}_k$ by $\hat{p}_k$.

Consequently, we compare the performance of the following set-valued classifier $\hat{\Gamma}_s^{+c}$ (builded with the random perturbation), $\hat{\Gamma}_s^{-c}$ (builded without the random perturbation), $\hat{\Gamma}_s^{+sp}$ (builded by using the splitting of the sample), and $\hat{\Gamma}_s^{-sp}$ (without splitting). Note that the randomization is used to build $\hat{\Gamma}_s^{+sp}$ and $\hat{\Gamma}_s^{-sp}$. For each set-valued classifier, the empirical error rate $\mathcal{P}(\hat{\Gamma}_s)$ and empirical expected size $S(\hat{\Gamma}_s)$ are evaluated by cross-validation.

General observations. Let us focus on the set-valued $\hat{\Gamma}_s^{+c}$ (*Continuous*) which is the semi-supervised method presented in Section 2.3.1. It satisfies all the required assumptions and hence Theorem 2.2 is applicable to this classifier. Figure 2.1-right-bottom displays the expected size of $\hat{\Gamma}_s^{+c}$ for $s = 2$. It highlights that even with a moderate unlabeled sample size N, the set-valued classifier has the prescribed expected size.
In addition, when we compare the values of the errors $P(\hat{\Gamma}_s^{+c})$ in all the boxes in Figure 2.3, we observe that this error is indeed decreasing *w.r.t.* s. It is important to notice

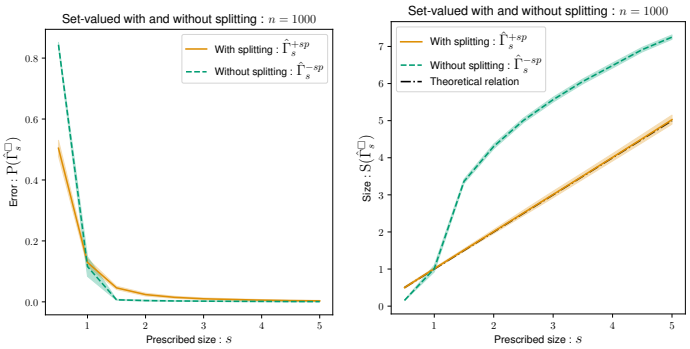

Figure 2.2: Importance of splitting in case $\mathcal{D}_N^{\mathrm{U}} = \emptyset$. Set-valued classifier $\hat{\Gamma}_s^{-sp}$ uses the same labeled data twice. Set-valued classifier $\hat{\Gamma}_s^{+sp}$ *splits* labeled data to force independence.

that the error of the corresponding single-output classifier (random forest without the calibration step) is 0.140 and then the use of the set-valued classifier is relevant already for moderate values of the size as s $= 2$. In this case, $P(\hat{\Gamma}_2^{+c})$ equals 0.014 – the error is 10 times lower.

Continuity of the estimator. As it was mentioned in the beginning of this section, the random forest classifier does not satisfy Assumption 2. Our goal here is to understand the importance of this assumption. Figure 2.1-left-bottom demonstrates that in the absence of the continuity Assumption 2, the set-valued classifier $\hat{\Gamma}_{\mathrm{s}}^{-c}$ (*Not continuous*) which does not modify $\hat{p}_1, \ldots, \hat{p}_K$ has a systematic bias in terms of the size across a wide range of N. Meanwhile, $\hat{\Gamma}_{\mathrm{s}}^{+c}$ (*Continuous*) successfully captures the prescribed size in average, see Figure 2.1-right. We also note that in both cases the variance of the outcome reduces with the growth of the unlabeled data N. Finally, we highlight that the error of $\hat{\Gamma}_{\mathrm{s}}^{-c}$ is slightly lower than that of $\hat{\Gamma}_{\mathrm{s}}^{+c}$. However, this is attributed to the *larger* output size and not its superior performance.

Data splitting. The second important conclusion we report deals with the relevance of the independence condition between the dataset used to estimate the regression functions and the dataset used to estimate the function G. Figure 2.2-right displays that the set-valued classifier $\hat{\Gamma}_{\mathrm{s}}^{-sp}$ (*Without splitting*) consistently over-estimates the size. By time, the expected size can even be twice as large as the desired size. In contrast, the size of the set-valued classifier $\hat{\Gamma}_{\mathrm{s}}^{+sp}$ (*With splitting*) follows the diagonal line illustrating again that the proposed construction succeeds to satisfy the size constraint. According to the error, we can see from Figure 2.2-left that the set-valued classifier without splitting outperforms slightly the set-valued classifier with splitting. However, this should be tempered by the fact that the size of the former is larger. In addition this could be related to the fact that we used more data to estimate the p_k's for the method without splitting.

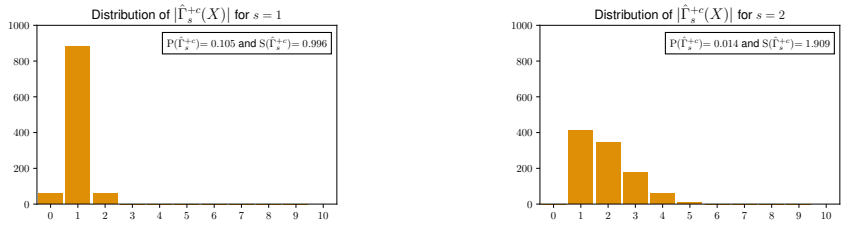

Figure 2.3: Distribution of the size of $\hat{\Gamma}_s^{+c}$ in a single outcome of the experiments across various values of s.

Which sizes do we get? We end this section by a thiner description of the size of the set-valued classifier $\hat{\Gamma}_s^{+c}$ (*Continuous*). For a single outcome of the experiment we report on Figure 2.3 the distribution of the size of the set-valued classifier for $s = 1, 2$. One draw the following conclusion. First of all we note that setting $s = 1$ is not equivalent to the set-up of the single-output classification. Indeed, the plot on top-left of Figure 2.3 shows that even though most of the times the corresponding set-valued classifier outputs only one candidates, there are situations where no labels or two predicted label candidates are provided. Moreover, the error of $\hat{\Gamma}_1^{+c}$ (with $s = 1$) is 0.105, while the error of the corresponding single-output classifier (pure random forest without the second step) is 0.14. Hence, set-valued classifiers can improve the performance even in the case of $s = 1$. Besides, note that for values of $s = 2$ the corresponding set-valued classifier *significantly improves* the error, while having small size in average. Finally, we highlight how the set-valued classifier with the controlled expected size is different from the top-s procedure.

2.4 Empirical risk minimization for set-valued classification

In the previous section, we provide a general procedure to build set-valued classifier based on any estimators of the posterior probabilities. Furthermore, we have shown that from the numerical point of view, this method , which is easily implementable, exhibits good performance. Now, the question of the aggregation of set-valued classifier is then legitimate. This is one of the motivation of Denis and Hebiri (2017). In this work, we study set-valued classifier based on empirical risk minimization technique. Similarly to Section 1.2.2 in Chapter 1, due to the non convexity of the problem, we focus on convex surrogate of the initial problem that is tailored for the set-valued classification problem. In view of Equation (2.2), we propose a two-step procedure. In a first step we build, based on a proper convex loss, a vector $\mathbf{f}$ of score functions. In a second step, as for the plug-in procedure, we calibrate a threshold to satisfy the expected size constraint. Hence, the main question is which convex loss suits for set-valued classifier. Note that throughout this chapter, we assume that Assumption 1 is fulfilled.

2.4.1 Convexification of the initial problem

Let $\mathbf{f} = (f_1, \ldots, f_K) : \mathcal{X} \to \mathbb{R}^K$ be a score function and $G_{\mathbf{f}}(\cdot) = \sum_{k=1}^{K}(1 - F_{f_k}(\cdot))$, where F_{f_k} is the CDF of $f_k(\boldsymbol{X})$. Analogously to Assumption 1, let us assume the continuity of $G_{\mathbf{f}}$. This allows us to write that for any $s \in (0, K)$, there exists $\delta \in \mathbb{R}$, such that $G_{\mathbf{f}}(-\delta) = s$. The set-valued classifier $\Gamma_{\mathbf{f},\delta}$ associated to f and δ is defined by

$$\Gamma_{\mathbf{f},\delta}(X) = \{k \in \mathcal{Y} \ : \ f_k(\boldsymbol{X}) \geq -\delta\}.$$

The definition of parameter δ, implies that $S(\Gamma_{\mathbf{f},\delta}) = s$. The excess risk formula provided in Equation (2.3) suggests the one-versus-all approach (see Section 1.2.2 in Chapter 1). For simplicity, we present the proposed method for the L_2-risk. We refer to (Denis and Hebiri, 2017) for a study of general loss functions.

We recall that for a score function $\mathbf{f}$, its associated L_2-risk is defined as

$$R_2(\mathbf{f}) = \mathbb{E}\left[\sum_{k=1}^{K}(Z_k - f_k(X))^2\right], \ \text{ with } \ Z_k = 2\,\mathbb{1}_{\{Y=k\}} - 1, k = 1, \ldots, K.$$

Let $\mathcal{F}$ a convex set of score function , we aim at solving

$$\underbrace{\bar{\mathbf{f}} \in \arg\min_{\mathbf{f}\in\mathcal{F}} R_2(\mathbf{f}),}_{\text{optimal over } \mathcal{F}} \qquad \underbrace{\mathbf{f}^* \in \arg\min_{\mathbf{f}} R_2(\mathbf{f}),}_{\text{overall optimal}}$$

for the purpose of building the optimal set-valued classifiers $\Gamma_{\bar{\mathbf{f}},\delta}$ and $\Gamma_{\mathbf{f}^*,\delta}$ respectively. The following result show that the square loss is set-valued calibrated (Zhang, 2004; Bartlett, Jordan, and McAuliffe, 2006; Yuan and Wegkamp, 2010).

Proposition 2.1. *Let* $s \in (0, K)$. *There exists* $\delta^* \in \mathbb{R}$ *such that*

$$\Gamma_{\mathbf{f}^*,\delta^*} = \Gamma_s^* \ ,$$

The property of calibration means that the set-valued classifier based on $\mathbf{f}^*$, the minimizer of the L_2-risk, is the s-Oracle Γ_s^*. Roughly speaking, minimizing the Ł$_2$-risk is the same as minimizing the risk R_s.

2.4.2 Data-driven procedure

Recall that we have in hand a labeled $\mathcal{D}_n^{\mathrm{L}}$ and an unlabeled $\mathcal{D}_N^{\mathrm{U}}$ datasets. Similarly to the construction of plug-in set-valued classifier, the labeled dataset is used to build $\hat{\mathbf{f}}$ the minimizer of the empirical L_2-risk.

$$\hat{\mathbf{f}} \in \arg\min_{\mathbf{f}\in\mathcal{F}} \hat{R}_2(\mathbf{f}) \ \text{ with } \ \hat{R}_2(\mathbf{f}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}(Z_k^i - f_k(\boldsymbol{X}_i))^2,$$

where $Z_k^i = 2\,\mathbf{1}_{\{Y_i=k\}} - 1$ for all $k = 1, \dots, K$ and $\mathcal{F}$ is a convex set of score functions. At this stage, we need to specify the $\delta > 0$ such that $\mathcal{S}(\Gamma_{\hat{\mathbf{f}},\delta}) = \mathrm{s}$. For simplicity, we assume that the conditional distribution function of the score function $\hat{f}_k$ are continuous. As in Section 2.3 this condition on the estimator can be satisfied using randomization of the functions $\hat{f}_k$ without changing the statistical performance of the resulting set-valued predictor. Then, based on $\mathcal{D}_N^{\mathrm{U}}$, we define the empirical set-valued classifier

$$\hat{\Gamma}(x) = \left\{k \in \mathcal{Y} : \hat{f}_k(x)) \geq \hat{G}^{-1}(s)\right\} \quad \text{with} \quad \hat{G}(\cdot) = \frac{1}{N} \sum_{\boldsymbol{X} \in \mathcal{D}_N^{\mathrm{U}}} \sum_{k=1}^{K} \mathbf{1}_{\{\hat{f}_k(\boldsymbol{X}) \geq \cdot\}}. \tag{2.4}$$

2.4.3 Theoretical guarantees

Now we investigate rates of convergence for the empirical set-valued classifiers $\hat{\Gamma}$ *w.r.t.* the risk R_s. Note that the distribution-free result provided in Section 2.3 holds for $\hat{\Gamma}$. The rate of convergence of the set-valued classifier Γ is obtained under the classical margin assumption (Mammen and Tsybakov, 1999).

Assumption 3 (α-margin assumption)**.** *We say that the distribution $\mathbb{P}$ of the pair $(\boldsymbol{X}, Y) \in \mathbb{R}^d \times \mathcal{Y}$ satisfies α-margin assumption if there exists $C_1 > 0$ and $t_0 \in (0, 1)$ such that for every positive $t \leq t_0$*

$$\mathbb{P}_X\left(0 < \left|p_k(\boldsymbol{X}) - G^{-1}(\mathrm{s})\right| \leq t\right) \leq C_1 t^{\alpha} \ .$$

The exponent α will directly specify the rates of convergence and the classification problem gets easier with the growth of this parameter. It is important to note that since we assume that the distribution functions of $p_k(\boldsymbol{X})$ are continuous for each k, we have $\mathbb{P}_X\left(0 < |p_k(\boldsymbol{X}) - G^{-1}(\mathrm{s})| \leq t\right) \to 0$ with $t \to 0$. Therefore, the margin condition only specifies the rate of this decay to 0. We can now state our error control for the empirical set-valued classifiers defined by (2.4).

Theorem 2.4. *Assume that $\|f\|_\infty \leq B$ for all $f \in \mathcal{F}$. Let $M_n = \mathcal{N}(1/n, L_\infty, \mathcal{F})$ be the covering number of $\mathcal{F}$* w.r.t. *L_∞-norm with closed balls with radius $1/n$. Grants Assumptions 1 and 3*

$$\mathbb{E}\left[R_s(\hat{\Gamma}) - R_s(\Gamma_s^*)\right] \lesssim \left\{\inf_{\mathbf{f} \in \mathcal{F}} (R_2(\mathbf{f}) - R_2(f^*)) + \frac{\log(M_n)}{n}\right\}^{\alpha/(\alpha+2)} + \frac{1}{\sqrt{N}} \ ,$$

where the leading constant depends only on L, B, and α.

The rate of convergence for the excess error is of order $\left(\frac{\log(M_n)}{n}\right)^{\alpha/(\alpha+2)} + \frac{1}{\sqrt{N}}$. In particular, it establishes the consistency of the procedure provided that $\alpha > 0$. Besides, same results holds for the Hamming distance. At the time it was proved, this result is the first, up to our knowledge, that provides a bound on the excess risk for set-valued

classifiers in multi-class setting. Compared to the literature, the exponent $\alpha/(\alpha+2)$ is not classical and it is not clear whether it is improvable. Indeed, in the standard multi-class framework, this exponent is of order $(\alpha+1)/(\alpha+2)$ (see Zhang (2004)). The second part of the rates which is of order $N^{-1/2}$ relies on the estimation of the function $\widetilde{G}(t) = \sum_{k=1}^{K}(1 - F_{\hat{\mathbf{f}}_k}(t))$, which serves as a pseudo-oracle CDF that knows the marginal distribution $\mathbb{P}_X$.

2.4.4 Numerical experiments

According to the developed methodology, we propose an aggregation procedure based on the cross-validation principle. Formally, the procedure is applied in the case where $\mathcal{F}$ is the convex hull of a finite family of score functions (we refer to (Denis and Hebiri, 2017) for more details). Here, we consider a family of 4 score functions based respectively on the random forest, the softmax regression, the support vector machines, and the k nearest neighbors (with $k = 11$) procedures. Note that the numerical results are presented for the boosting loss for which Theorem 2.4 apply.

We evaluate the performance of the procedure on two real datasets: the *Forest type mapping* dataset and the *one-hundred plant species leaves* dataset coming from the UCI database. We refer to these two datasets as `Forest` and `Plant` respectively. The `Forest` dataset consists of $K = 4$ classes and 523 labeled observations In the `Plant` dataset, there are $K = 100$ classes and 1600 labeled observations.

To get an indication of the statistical significance of the aggregated procedure (referred as `CV`) we compare it to the set-valued classifiers that result from each component of the library in terms of empirical error and empirical expected size. Without going in deep details, we mention that we split the dataset in three: the labeled $\mathcal{D}_n^{\mathrm{L}}$ and an unlabeled $\mathcal{D}_N^{\mathrm{U}}$ datasets of size n and N respectively to train the set-valued classifiers. A third labeled dataset of size M is used to compute error and size. We set the sizes of the samples as $n = 200$, $N = 100$ and $M = 223$ for the `Forest` dataset, and $n = 1000$, $N = 200$ and $M = 400$ for the `Plant` one. The empirical error and expected size is computed by cross-validation.

As a benchmark, we notify that the misclassification errors of the best classifier from the library for the `Forest` dataset is evaluated at 0.15 , whereas in the `Plant` dataset, it is evaluated at 0.40. Note that The performance of the classical methods is rather weak in the latter dataset.

The results are reported in Table 2.1, and confirm our expectations. A general observation is that the size constraint is quite well satisfied for all of the methods. Also, moderate level of constraint s leads to drastic improvement in terms of errors as compared to our benchmarks. Moreover, we observe that the risk gets drastically better with moderate s as compared to the *best* misclassification risk. For instance, for the `Plant`, the error rate of the set-valued classifier with s = 2 based on random forests is 0.18 whereas

`Forest` ($K = 4$)

		s-set				
s		`rforest`	`softmax reg`	`svm`	`kknn`	`CV`
2	P	0.02 (0.02)	0.06 (0.02)	0.02 (0.01)	0.05 (0.03)	0.02 (0.01)
	S	2.00 (0.09)	2.00 (0.08)	2.00 (0.09)	2.00 (0.08)	2.00 (0.08)

`Plant` ($K = 100$)

		s-set				
s		`rforest`	`softmax reg`	`svm`	`kknn`	`CV`
2	P	0.18 (0.03)	0.77 (0.02)	0.32 (0.04)	0.20 (0.03)	0.17 (0.03)
	S	2.00 (0.09)	2.02 (0.18)	1.99 (0.10)	2.00 (0.08)	2.00 (0.08)
10	P	0.02 (0.01)	0.42 (0.04)	0.03 (0.02)	0.08 (0.03)	0.02 (0.01)
	S	9.95 (0.38)	10.06 (0.58)	9.98 (0.22)	9.98 (0.23)	9.96 (0.37)

Table 2.1: For each dataset, we derive the estimated error P and size S of the different set-valued classifiers *w.r.t.* s. We compute the means and standard deviations (between parentheses) over the repetitions. For each s, the set-valued classifiers are based on—from left to right—`rforest`, `softmax reg` and `svm`, `kknn` and `CV` which are respectively the random forest, the softmax regression, support vector machines, k nearest neighbors and the aggregation procedure. Top: the dataset is the `Forest` – the dataset is the `Plant`.

the misclassification error rate of the best component in the library is 0.40. Interestingly the aggregated set-valued classifier (`CV`) outperforms all components of the library in all of the experiments. A last observation that motivates the use of aggregation procedure.

2.5 Optimal rates of convergence

In this section, we focus on the study of optimal rates of convergence for set-valued classifiers with controlled expected size. We exploit classical non-parametric theory tools to derive upper and lower bounds on the excess-risk (Mammen and Tsybakov, 1999; Yang, 1999; Györfi et al., 2002; Massart and Nédélec, 2006; Audibert and Tsybakov, 2007). From the technical point of view, our work is close in spirit to the one by Audibert and Tsybakov (2007) who study the statistical performance of plug-in classification rules under assumptions which involve the smoothness of the regression function and the margin condition 3. They derive fast rates of convergence (faster than $n^{-1/2}$) for plug-in classifiers based on local polynomial estimators (Stone, 1977; Tsybakov, 1986; Audibert and Tsybakov, 2007) and show their optimality in the minimax sense.

The motivation of this section is two folds

Q1. What is a minimax setup in this problem and what are the minimax rates of convergence?

Q2. Can we statistically justify the introduction of the unlabeled data $\mathcal{D}_N^{\mathrm{U}}$ from the

minimax perspective? To be more precise, we would like to understand whether the rates of convergence are affected by N – the size of the unlabeled sample.

Neither of these natural questions have been considered and answered in the previous literature. For simplicity, we only present the result obtained in (Chzhen, Denis, and Hebiri, 2021) *w.r.t.* the measure of risk R_s.

The central notion we manipulate in this section is the minimax rate of convergence in the semi-supervised setting. Let us denote by $\mathbb{P}$ a family of joint distribution of (X, Y), and let $\hat{\Gamma}$ be an estimator based on $\mathcal{D}_n^{\mathrm{L}}$ and $\mathcal{D}_N^{\mathrm{U}}$. We introduce the maximal risk

$$\Psi_{n,N}\left(\hat{\Gamma}, \mathcal{P}\right) = \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}}}\left[R_s(\hat{\Gamma}) - R_s(\Gamma_s^*)\right].$$

The minimax rate of convergence is then defined as follows.

Definition 2.1 (Minimax rate of convergence)**.** *For a given family $\mathcal{P}$ of joint distributions on $\mathbb{R}^d \times \mathcal{Y}$ the minimax rates are defined as*

$$\mathcal{E}_{n,N}(\mathcal{P}) := \inf_{\hat{\Gamma}} \Psi_{n,N}\left(\hat{\Gamma}, \mathcal{P}\right),$$

where the infimum is taken over all set-valued classifiers based on $\mathcal{D}_n^{\mathrm{L}}$ and $\mathcal{D}_N^{\mathrm{U}}$. The sequence $\mathcal{E}_{n,0}(\mathcal{P})$ corresponds to the supervised *regime, while the sequence $\mathcal{E}_{n,N}(\mathcal{P})$ for $N \geq 1$ corresponds to the* semi-supervised *regime.*

2.5.1 Assumptions

In this part we state all the assumptions used in this work and state the family of distributions $\mathcal{P}$ which drives the minimax rates. The first assumption is the margin assumption Assumption 3 that we already introduced in the previous Section 2.4.

The second assumption restricts the set of possible marginal distributions of the feature vectors. Following Audibert and Tsybakov (2007), we first introduce the notion of regular set. Let c_0 and r_0 be two positive constants. We say that a Borel set $A \subset \mathbb{R}^d$ is a (c_0, r_0)-regular set if

$$\mathrm{Leb}\left(A \cap \mathcal{B}(\boldsymbol{x}, r)\right) \geq c_0 \,\mathrm{Leb}\left(\mathcal{B}(\boldsymbol{x}, r)\right), \qquad \forall r \in (0, r_0],\ \forall \boldsymbol{x} \in A\ .$$

Definition 2.2 (Strong density)**.** *We say that the probability measure $\mathbb{P}_X$ on $\mathbb{R}^d$ satisfies the $(\mu_{\min}, \mu_{\max}, c_0, r_0)$-strong density assumption if it is supported on a compact (c_0, r_0)-regular set $A \subset \mathbb{R}^d$ and has a density μ w.r.t. the Lebesgue measure such that $\mu(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathbb{R}^d \setminus A$ and*

$$0 < \mu_{\min} \leq \mu(\boldsymbol{x}) \leq \mu_{\max} < \infty, \quad \forall \boldsymbol{x} \in A\ .$$

Let us mention, that there are various ways to relax this assumption. For instance, it is possible to get rid of the lower bound on the density (Audibert and Tsybakov, 2007; Kpotufe and Martinet, 2018). Besides, the compactness of the support can also be relaxed and replaced by a proper tail condition (Gadat, Klein, and Marteau, 2016). This type of relaxations are not altering our conclusions about the effect of unlabeled data and thus, for simplicity, we provide the analysis under the strong density assumption.

The next assumption is standard in non-parametric statistics, and states that the conditional distribution of Y is smooth.

Definition 2.3 (Hölder class, Tsybakov, 2009a)**.** *We say that a function $h : \mathbb{R}^d \to \mathbb{R}$ is (β, L)-Hölder for $\beta > 0$ and $L > 0$ if h is $\lfloor \beta \rfloor$ times continuously differentiable and $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ we have*

$$\left| h(\boldsymbol{x}') - h_{\boldsymbol{x}}(\boldsymbol{x}') \right| \leq L \left\| \boldsymbol{x} - \boldsymbol{x}' \right\|^{\beta} \quad ,$$

where $h_{\boldsymbol{x}}(\cdot)$ is the Taylor polynomial of degree $\lfloor \beta \rfloor$ of $h(\cdot)$ at the point $\boldsymbol{x} \in \mathbb{R}^d$. The set of all functions from $\mathbb{R}^d$ to $\mathbb{R}$ satisfying the above conditions is called $(\beta, L, \mathbb{R}^d)$-Hölder and is denoted by $\mathcal{H}(\beta, L, \mathbb{R}^d)$.

Finally, we are in position to define the family of distributions $\mathcal{P}$ that governs the rates of convergence.

Definition 2.4. *We denote by $\mathcal{P}(L, \beta, \alpha)$ the set of joint distributions on $\mathbb{R}^d \times \mathcal{Y}$ which satisfy the following conditions*

- *the marginal $\mathbb{P}_{\boldsymbol{X}}$ satisfies the $(\mu_{\min}, \mu_{\max}, c_0, r_0)$-strong density,*
- *for all $k \in \mathcal{Y}$ the k^{th} regression function $p_k(\cdot) = \mathbb{P}(Y = k \mid \boldsymbol{X} = \cdot)$ belongs to the $(\beta, L, \mathbb{R}^d)$-Hölder class, that is, $p_k \in \mathcal{H}(\beta, L, \mathbb{R}^d)$ for all $k \in \mathcal{Y}$,*
- *for all $k \in \mathcal{Y}$ the regression function p_k satisfy the (C_1, t_0, α)-Margin assumption,*
- *for all $k \in \mathcal{Y}$, the cumulative distribution function F_{p_k} of $p_k(\boldsymbol{X})$ is continuous.*

The family of distributions $\mathcal{P}(L, \beta, \alpha)$ resembles the one considered in (Audibert and Tsybakov, 2007) in the context of binary classification. The major difference is the continuity Assumption 1, which poses certain restrictions and does not allow to re-use in a straightforward way their construction for lower bounds.

2.5.2 Lower bound

In this section we establish minimax lower bounds on the introduced risk measures. Our rates highlight the benefit of the semi-supervised approaches in the context of the set-valued classification with controlled expected size.

Theorem 2.5. *Let $K \geq 3$, $s \leq \lfloor K/2 \rfloor - 1$. If $2\alpha\lceil \frac{\beta}{2} \rceil \leq d$, then for all $n, N \in \mathbb{N}$ it holds that*

$$\mathcal{E}_{n,N}(\mathcal{P}(L,\beta,\alpha)) \gtrsim n^{-\frac{(1+\alpha)\beta}{2\beta+d}} \bigvee (n+N)^{-1/2}.$$

The above lower bound imply that the best rate in the supervised regime is $n^{-1/2}$ across all the risk. Therefore, even if the margin assumption is very strong ($\alpha \gg 1$) supervised methods ($N = 0$) *cannot* achieve fast rates. This fact is the major difference with classical setups where the value of threshold is known (such as classification and level set estimation). Indeed, under the same assumptions on the family of distributions, without the continuity Assumption 1, the minimax rate in those frameworks is $n^{-(1+\alpha)\beta/(2\beta+d)}$ as proved for instance in (Audibert and Tsybakov, 2007; Rigollet and Vert, 2009) and unlabeled data *cannot* improve it. In contrast, this limitation can be neglected in the semi-supervised regime. Indeed, for sufficiently large unlabeled sample, the dominant term in the lower bound is of order $n^{-(1+\alpha)\beta/(2\beta+d)}$, which can be faster than $n^{-1/2}$. More precisely, the following relations are *necessary* to get fast rates of convergence

$$(n+N)^{-1/2} = o\left(n^{-(1+\alpha)\beta/(2\beta+d)}\right), \quad n^{-(1+\alpha)\beta/(2\beta+d)} = o(n^{-1/2}) \ .$$

The condition on the left hand side ensures that we have enough unlabeled data to eliminate the impact of not knowing threshold $G^{-1}(s)$ in Equation (2.2). Whereas, the condition on the right hand side ensures that the classification problem with "known" threshold admits fast rates. The above discussion suggests that the lack of knowledge of the threshold $G^{-1}(s)$ is significant, and the considered framework is more difficult from the statistical perspective than its more classical counterparts.

Finally, Note that the second part of the rate in all three cases is $(n+N)^{-1/2}$ instead of $N^{-1/2}$. Actually, if $n \gg N$, from purely minimax perspective it is impossible to obtain a lower bound with $N^{-1/2}$ instead of $(n+N)^{-1/2}$. Indeed, one can always split the labeled sample erasing labels from one of them. Such splitting artificially augments the size of unlabeled sample N by some fraction of n. Of course, the regime $n \gg N$ is not particularly interesting, since, despite the fact that $N \neq 0$, still corresponds to the essentially supervised setup.

2.5.3 Upper bound

In this section, we build a set-valued classifier that achieves (up to a log factor) the lower bound stated in Theorem 2.5. We first establish the obtained upper bound and then present the construction of the estimator.

Theorem 2.6. *Let $K \in \mathbb{N}$, $s \in (0, K)$, then there exists an estimator $\hat{\Gamma}$ such that for all $n, N \in \mathbb{N}$ we have*

$$\Psi_{n,N}\left(\hat{\Gamma}, \mathcal{P}(L,\beta,\alpha)\right) \lesssim \left(\frac{n}{\log n}\right)^{-\frac{(1+\alpha)\beta}{2\beta+d}} \bigvee (n+N)^{-1/2}.$$

As an immediate consequence, from the above result is that the lower bound of Theorem 2.5 are achievable.

Construction of the estimator. The construction of the set-valued classifier is similar to the one described in Section 2.3 with technical modifications for matching the obtained lower bound. In particular, in order to get the rate $(n+N)^{1/2}$, we do use the whole labeled sample to build estimators of the posterior probability. These estimators $\hat{p}_k$ are constructed using an arbitrary half $\mathcal{D}_{\lfloor n/2 \rfloor}$ of the labeled dataset $\mathcal{D}_n^{\mathrm{L}}$ and the following assumption is required.

Assumption 4 (Exponential concentration)**.** *There exist estimators $\hat{p}_k$ for all $k \in \mathcal{Y}$ based on $\mathcal{D}_{\lfloor n/2 \rfloor}$ and positive constants C_1, C_2 and $\delta_0 \geq 0$ such that for all $k \in \mathcal{Y}$ and all $n \geq 2$ we have for all $\delta > \delta_0 n^{-\beta/(2\beta+d)}$*

$$\sup_{\mathbb{P} \in \mathcal{P}(L,\beta,\alpha)} \mathbf{P}\left(|\hat{p}_k(\boldsymbol{x}) - p_k(\boldsymbol{x})| \geq \delta\right) \leq C_1 \exp\left(-C_2 n^{\frac{2\beta}{2\beta+d}} \delta^2\right) \quad ,$$

for almost all $\boldsymbol{x} \in \mathbb{R}^d$ w.r.t. $\mathbb{P}_{\boldsymbol{X}}$.

Assumption 5 (Continuity of CDF)**.** *For all $k \in \mathcal{Y}$, conditionally on the data, the cumulative distribution function $F_{\hat{p}_k}(t) := \mathbb{P}_{\boldsymbol{X}}(\hat{p}_k(\boldsymbol{X}) \leq t)$ of $\hat{p}_k(\boldsymbol{X})$ is almost surely $\mathbb{P}^{\otimes \lfloor n/2 \rfloor}$ continuous on $(0,1)$.*

First let us point out that Assumption 4 induces that for all $n \geq 2$ and all $\alpha > 0$

$$\sup_{\mathbb{P} \in \mathcal{P}(L,\beta,\alpha)} \mathbf{E}\, \|\boldsymbol{p} - \hat{\boldsymbol{p}}\|_{\infty,\mathbb{P}_X}^{1+\alpha} \lesssim \left(\frac{n}{\log n}\right)^{-\frac{(1+\alpha)\beta}{2\beta+d}} \quad ,$$

where $\boldsymbol{p}(\cdot) = (p_1(\cdot), \ldots, p_K(\cdot))^\top$ and $\hat{\boldsymbol{p}}(\cdot) = (\hat{p}_1(\cdot), \ldots, \hat{p}_K(\cdot))^\top$. Assumption 4 is commonly used in the statistical community when we deal with rates of convergence in the classification settings (Audibert and Tsybakov, 2007; Lei, 2014; Sadinle, Lei, and Wasserman, 2018). It is for instance satisfied by the locally polynomial estimator (Stone, 1977; Tsybakov, 1986; Audibert and Tsybakov, 2007) with $\delta_0 = 0$. As already seen, Assumption 5 can always be satisfied by slightly processing any estimator $\hat{p}$. In (Chzhen, Denis, and Hebiri, 2021), we propose to perturb the estimator with a *deterministic* perturbation to fit with the minimax framework.

Finally, To make our presentation mathematically correct we introduce the following notation $\mathcal{D}_n^{\mathrm{L}} = \mathcal{D}_{\lfloor n/2 \rfloor} \bigcup \mathcal{D}_{\lceil n/2 \rceil}$, where $\mathcal{D}_{\lfloor n/2 \rfloor}$ is the dataset used to build the estimators $\hat{p}_k$ for $k \in \mathcal{Y}$. Now, all the labels are removed from $\mathcal{D}_{\lceil n/2 \rceil}$. That is, $\mathcal{D}_{\lceil n/2 \rceil}$ consists of $\lceil n/2 \rceil$ *i.i.d.* samples from $\mathbb{P}_X$. Consequently, the semi-supervised estimator of $G(\cdot)$ is defined as

$$\hat{G}(\cdot) = \frac{1}{\lceil n/2 \rceil + N} \sum_{\boldsymbol{X} \in \mathcal{D}_N^{\mathrm{U}} \bigcup \mathcal{D}_{\lceil n/2 \rceil}} \sum_{k=1}^{K} \mathbf{1}_{\{\hat{p}_k(X) > \cdot\}}.$$

Finally, the set-valued classification procedure $\hat{\Gamma}$ is defined point-wise as

$$\hat{\Gamma}(\boldsymbol{x}) = \left\{k \in \mathcal{Y} \,:\, \hat{p}_k(\boldsymbol{x}) \geq \hat{G}^{-1}(\mathrm{s})\right\}.$$

Sketch of proof. Now, we briefly discuss about the log factor obtained in Theorem 2.6. We introduce an intermediate quantity $\tilde{G}(\cdot) = \sum_{k=1}^{K} \mathbb{P}_X\left(\hat{p}_k(X) > \cdot\right)$, and the associated set-valued classifier, which we refer to as the pseudo Oracle classifier given for all $x \in \mathbb{R}^d$ by

$$\tilde{\Gamma}(x) = \{k \in \mathcal{Y}, \;\; \hat{p}_k(x) \geq \tilde{G}^{-1}(\mathrm{s})\}.$$

The set-valued $\tilde{\Gamma}$ assumes knowledge of the marginal distribution $\mathbb{P}_X$, it is seen as an idealized version of $\hat{\Gamma}$, and formally corresponds to the case $N = +\infty$. Besides, thanks to Assumption 2, this pseudo Oracle satisfies the size constraints, that is $\mathrm{S}(\tilde{\Gamma}) = \mathrm{s}$ almost surely. Now, we start with the following decomposition

$$R_s(\hat{\Gamma}) - R_s(\Gamma_s^*) \;\; = \;\; \underbrace{R_s\left(\hat{\Gamma}\right) - R_s\left(\tilde{\Gamma}_s\right)}_{(1)} + \underbrace{R_s(\tilde{\Gamma}) - R_s(\Gamma_s^*)}_{(2)}$$

The first term (1) in the r.h.s. of the above inequality is relies on a C.D.F. estimation problem. For the second term (2), we observe that

$$k \in (\Gamma_s^*(X) \,\Delta\, \tilde{\Gamma}_s(X)) \Rightarrow \left|p_k(X) - G^{-1}(s)\right| \leq |G^{-1}(s) - \tilde{G}^{-1}(s)| + |\hat{p}_k(X) - p_k(X)| \quad (2.5)$$

Now, the crucial step of our analysis is the following lemma, that bounds the difference between $\tilde{G}^{-1}(\mathrm{s})$ and $G^{-1}(\mathrm{s})$ in terms of the difference between $\hat{p}_k$'s and p_k's.

Lemma 2.1 (Upper bound on the thresholds)**.** *Let Assumption 1 be satisfied, then for all* $\mathrm{s} \in (0, K)$

$$\left|G^{-1}(\mathrm{s}) - \tilde{G}^{-1}(\mathrm{s})\right| \leq \|\boldsymbol{p} - \hat{\boldsymbol{p}}\|_{\infty, \mathbb{P}_X}, \quad \textit{almost surely } \mathbb{P}^{\otimes n} \otimes \mathbb{P}_{\boldsymbol{X}}^{\otimes N} \;.$$

The difference $|G^{-1}(\mathrm{s}) - \tilde{G}^{-1}(\mathrm{s})|$ resembles the Wasserstein infinity distance which gives an alternative approach to prove Lemma 2.1, see (Bobkov and Ledoux, 2016). Lemma 2.1 explains the extra $\log n$ factor that appears in the upper bound, as the minimax estimation in sup norm contains the $\log n$ factor, see for instance (Stone, 1982; Tsybakov, 2008). From Equation (2.5) and Assumption 3,

$$\begin{aligned} R_s(\hat{\Gamma}) - R_s(\Gamma_s^*) &\leq 2 \max_k \|\hat{p}_k - p_k\|_\infty \sum_{k=1}^{K} P_X\left(\left|p_k(X) - G^{-1}(s)\right| \leq 2 \max_k \|\hat{p}_k - p_k\|_\infty\right) \\ &\leq C \max_k \|\hat{p}_k - p_k\|_\infty^{1+\alpha}. \end{aligned}$$

It is intuitively clear that if, on top of Lemma 2.1, we manage to control the difference $|\tilde{G}^{-1}(\mathrm{s}) - \hat{G}^{-1}(\mathrm{s})|$ then the proof of the upper bound would simply follow the

arguments of Audibert and Tsybakov (2007). Yet, such a control is not feasible under our assumptions. To see this, notice that conditionally on $\mathcal{D}_n$ the quantity $|\tilde{G}^{-1}(s) - \hat{G}^{-1}(s)|$ resembles the deviation of quantile from its empirical version. However, classical result[1] on asymptotic normality of sample quantiles (Ma and Robinson, 1998, Theorem 2) tells that in order to have a central limit theorem with $(n+N)^{-1/2}$ rate it is necessary and sufficient to require $\tilde{G}'(\tilde{G}^{-1}(s)) > 0$. From the minimax perspective, this condition cannot be satisfied since we do not require any lower bound on the derivative of $G(\cdot)$.

In (Chzhen, Denis, and Hebiri, 2021) we demonstrate that the upper bound can be improved if we assume that the derivative of $G(\cdot)$ is uniformly lower bounded, that is, $G^{-1}(\cdot)$ has some regularity.

2.6 Conclusion

In this chapter, we studied the set-valued classification problem with a controlled expected size. The theoretical analysis started with a setting that allowed us to get distribution free results and then we added some assumptions to build a minimax analysis where we emphasized the unlabeled data as a key feature to get fast rate of convergence in some situations. From this perspective, our analysis wants to be a promotion for semi-supervised methods that can be exploited in several fields such as the *fairness* problem that we described in Chapter 3. We paid a particular attention to understand the meaning and the implications of the continuity assumption we imposed on the conditional probability CDF. One of our main challenges was the study and the control of differences of quantiles, the function G^{-1} being in the center of our attention, through tools from empirical processes, rank statistics and non parametric theory.

In (Denis and Hebiri, 2017; Chzhen, Denis, and Hebiri, 2021), the dependency in the number of labels K has not been considered. Yet, set-valued classification can face extreme classification scenarios. The prediction ability of set-valued classifiers should be investigated in a context where the number of labels is large, as well as the number of observations and features. By step, the first consideration should be to tackle the question of large number of labels. Techniques from high dimensional statistic should be used, in particular, one could think of a new notion of sparsity assumption adapted to the set-valued classification. A possible direction is to introduce a margin assumption that involves a number $s^* \ll K$ which reflects, in the set-valued framework, the optimal size s. In this case, one may expect only logarithmic deflation of the rate in terms of K, and maybe linear (or surperlinear) in terms of the sparsity level s^*. This line of research might also be related to multi-label classification where this kind of behaviors have already been obtained for instance in (Jain, Prabhu, and Varma, 2016) and in a previous work (Chzhen et al., 2017).

[1] We can arrive to a similar conclusion from (Bobkov and Ledoux, 2016, Theorem 5.11)

Chapter 3

Multi-class classification under demographic parity constraint

In this chapter, I present a contribution for multi-class classification under fairness constraint. Algorithmic fairness has become very popular during the last decade (Calders, Kamiran, and Pechenizkiy, 2009; Zemel et al., 2013; Lum and Johndrow, 2016; Zafar et al., 2017; Agarwal et al., 2018; Donini et al., 2018; Agarwal, Dudik, and Wu, 2019; Barocas, Hardt, and Narayanan, 2019; Chzhen et al., 2019; Chiappa et al., 2020). It helps addressing an important social problem: mitigating historical bias contained in the data. This is a crucial issue in many applications such as loan assessment, health care, or even criminal sentencing. The common objective in algorithmic fairness is to reduce the influence of a sensitive attribute on a prediction.

In recent years, various authors and communities have been proposing different formal definition of the notion of fairness, equality, and justice. An attractive formalism relies on the idea that a prediction must not discriminate based on some characteristics of an instance (sensitive feature) such as the gender or the ethnicity. We refer the interested reader to (Barocas, Hardt, and Narayanan, 2019; Mehrabi et al., 2019) for a general introduction to the problem of fair prediction and to (Barrio, Gordaliza, and Loubes, 2020; Oneto and Chiappa, 2020) for a review of the most recent theoretical advances.

In (Chzhen et al., 2019), we investigated the binary classification and regression settings under fairness constraints. One of the major contribution of these works is to provide a closed form of the oracle predictors and derive post-processing estimation procedures which are able to leverage unlabeled data and exhibit good numerical performance. In the same spirit, we investigate in (Denis et al., 2021) the fair multi-class classification framework. Imposing fairness constraint in the multi-class problem has only been briefly discussed in (Ye and Xie, 2020) by considering Support Vector Machine (SVM) fair prediction. Our contribution provides a deeper analysis of the problem.

3.1 Fair multi-class classification

In this section, we introduce the general framework and also define and discuss the notion of fairness that we consider. In addition, we exhibit a closed form of the optimal fair predictor in Section 3.1.3.

3.1.1 Statistical setting

In fair multi-class classification, compared to the usual multi-class setting, we assume that the input feature is a couple (X, S) where X corresponds to the vector of covariates and S is the sensitive feature (*e.g.* gender, ethnicity, qualification, birth place, ...). For simplicity, we assume that $S \in \{-1, 1\}$. The distribution of the sensitive feature S is denoted by $(\pi_s)_{s\in\mathcal{S}}$, and we assume that $\min_{s\in\mathcal{S}} \pi_s > 0$. A complete observation is of the form (X, S, Y) where $Y \in \mathcal{Y}$ is the label associated to (X, S). In this context, a classification rule g is a function mapping $\mathcal{X} \times \{-1, 1\}$ onto $\mathcal{Y}$, whose performance is again evaluated through the misclassification risk

$$\mathcal{R}(g) = \mathbb{P}\left(g(X, S) \neq Y\right).$$

For $k \in \mathcal{Y}$, $p_k(X, S)$ denotes the conditional probability $\mathbb{P}\left(Y = k|X, S\right)$. Recall that a Bayes classifier minimizes the misclassification risk and is defined as

$$g^*(x, s) \in \arg\max_{k\in\mathcal{Y}} p_k(x, s)\ , \quad \text{for all } (x, s) \in \mathcal{X} \times \mathcal{S}\ .$$

3.1.2 Multi-class classification with demographic parity

Several notion of fairness have been studied in the binary classification framework, such as *Equalized-Odds*, *Equality of Opportunity*, or *Demographic Parity*. We refer for instance to (Hardt, Price, and Srebro, 2016; Barocas, Hardt, and Narayanan, 2019) for an overview of these notions. In this chapter, we consider muti-class classification problems under Demographic Parity (DP) fairness constraint (Calders, Kamiran, and Pechenizkiy, 2009), that requires the independence of the prediction function from the sensitive feature S. The DP constraint is perhaps the most intuitive notion of fairness which is common to the classification and regression setting. DP constraint has a recognized interest in various applications; this constraint could be typically imposed in loan agreement without gender attributes or in the context of crime prediction without ethnicity discrimination. Formally, we define DP constraint as follows

Definition 3.1 (Demographic parity)**.** *We say that a classifier $g \in \mathcal{G}$ (and write $g \in \mathcal{G}_{\text{fair}}$) with respect to the distribution $\mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$ if*

$$\mathbb{P}\left(g(X, S) = k|S = 1\right) = \mathbb{P}\left(g(X, S) = k|S = -1\right), \quad \forall k \in \mathcal{Y}\ .$$

The above definition naturally extends to the multi-class setting the DP considered in binary classification (Agarwal, Dudik, and Wu, 2019; Gordaliza et al., 2019; Jiang et al., 2019; Oneto, Donini, and Pontil, 2019; Chiappa et al., 2020) .

Intuitively, when fairness is required, two important aspects of a classifier g need to be controlled: its misclassification risk $\mathcal{R}(g)$ and its unfairness, which we define below.

Definition 3.2 (Unfairness measure)**.** *The unfairness of a classifier $g \in \mathcal{G}$ is quantified by*

$$\mathcal{U}(g) := \sum_{k=1}^{K} \left| \mathbb{P}\left(g(X,S) = k | S = 1\right) - \mathbb{P}\left(g(X,S) = k | S = -1\right) \right| \quad .$$

Naturally, taking into account the definition above, a classifier g is fair if and only if $\mathcal{U}(g) = 0$.

Alternative measures of unfairness could be considered. For instance, once can replace the summation by a maximum over k in the above definition. However, summing over all possible labels is more informative and appears more naturally when controlling the prediction risk (see Theorem 3.2).

3.1.3 Optimal fair classifier

In this section, we provide an explicit formulation of the optimal fair classifiers *w.r.t.* the misclassification risk under DP constraint. An optimal fair classifier is a solution of

$$\min_{g \in \mathcal{G}_{\text{fair}}} \mathcal{R}(g).$$

The difficulty of obtaining an optimal fair classifier consists in properly balancing the misclassification risk together with the fairness criterion. Let g be a classifier and $\lambda = (\lambda_1, \ldots, \lambda_K) \in \mathcal{R}^K$ the Lagrange multiplier of the above minimization problem. We introduce a risk measure, referred as *fair-risk*, which provides a trade-off between the accuracy of g and its unfairness

$$\mathcal{R}_\lambda(g) := \mathcal{R}(g) + \sum_{k=1}^{K} \lambda_k \left[\mathbb{P}\left(g(X,S) = k | S = 1\right) - \mathbb{P}\left(g(X,S) = k | S = -1\right) \right].$$

In order to be able to derive a characterization of the optimal fair classifier, we require the following assumption on the random variables $p_k(X,s)$.

Assumption 6. *The mapping $t \mapsto \mathbb{P}\left(p_k(X,S) - p_j(X,S) \leq t | S = s\right)$ is assumed continuous, for any $k, j \in \mathcal{Y}$ and $s \in \mathcal{S}$.*

This assumption requires that the distribution of the differences $p_k(X,S) - p_j(X,S)$ has no atoms. Note that this assumption is similar to Assumption 1. In fact, it plays the same role and is specific to the multi-class framework. In particular, we observe that in the binary case ($K = 2$), the above assumption simply boils down to the one considered

in (Chzhen et al., 2019) that requires the continuity of $t \mapsto \mathbb{P}\left(p_k(X,S) \leq t | S = s\right)$. It is, however, clear that in the general case $K \geq 3$ these two conditions describe different sets of distributions.

From Assumption 6, we deduce a characterization of the optimal fair classifier.

Proposition 3.1. *Let Assumption 6 be satisfied and define $\lambda^* \in \mathbb{R}^K$ by*

$$\lambda^* \in \arg\min_{\lambda \in \mathbb{R}^K} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s}\left[\max_k \left(\pi_s p_k(X,s) - s\lambda_k\right)\right].$$

Then, $g^*_{\text{fair}} \in \arg\min_{g \in \mathcal{G}_{\text{fair}}} \mathcal{R}(g)$ if and only if $g^*_{\text{fair}} \in \arg\min_{g \in \mathcal{G}} \mathcal{R}_{\lambda^*}(g)$.

In other words, the optimum of the risk $\mathcal{R}(g)$ over the class of fair classifiers is also maximizing the fair-risk $\mathcal{R}_{\lambda^*}$. By construction, $\mathcal{R}_{\lambda^*}$ is a risk measure which optimally balances both classification accuracy and unfairness. Proposition 3.1 directly implies that $\mathcal{R}_{\lambda^*}(g) \geq \mathcal{R}_{\lambda^*}(g^*_{\text{fair}}) = \mathcal{R}(g^*_{\text{fair}}) \geq 0$, for $g \in \mathcal{G}$. We now quantify the performance of any classifier $g \in \mathcal{G}$ through its *excess fair-risk*

$$\mathcal{E}_{\text{fair}}(g) := \mathcal{R}_{\lambda^*}(g) - \mathcal{R}_{\lambda^*}(g^*_{\text{fair}}).$$

Furthermore, Proposition 3.1 directly implies a closed form expression of optimal fair classifiers, which is the bedrock of our procedure. Any optimal fair classifier is simply maximizing scores, which are obtained by shifting the original conditional probabilities in a optimal manner.

Corollary 3.1. *Under Assumption 6, an optimal fair classifier is characterized by*

$$g^*_{\text{fair}}(x,s) \in \arg\max_k \left(\pi_s p_k(x,s) - s\lambda^*_k\right) \quad , \quad (x,s) \in \mathcal{X} \times \mathcal{S}\,.$$

Note that a similar characterization of a optimal fair predictor is obtained in (Chzhen et al., 2020a)

3.2 General estimation procedure

In this section, we provide now a plug-in estimator for the optimal fair classifier g^*_{fair}. The proposed procedure is inspired from the methodology described in Section 2.3 of Chapter 2. In particular, we propose an algorithm that enjoys strong theoretical guarantees both in terms of fairness and risk. In particular, we exhibit in Section 3.2.2 distribution-free fairness guarantee.

3.2.1 Plug-in estimator

We are given two datasets. The first *labeled* one $\mathcal{D}_n = \{(X_i, S_i, Y_i),\ i = 1, \dots, n\}$ consists of *i.i.d.* samples from the distribution $\mathbb{P}$. This is the classical dataset used for training estimators $(\hat{p}_k)_k$ of the conditional probabilities $(p_k)_k$, *e.g.*, Random Forest, SVM, *etc.* The second *unlabeled* dataset $\mathcal{D}'_N$ consists of N *i.i.d.* copies of (X, S). the sample $\mathcal{D}'_N$ is collected and split in the following way: we set $(S_1, \dots, S_N)$ the *i.i.d.* sample of sensitive features used to compute empirical frequencies $(\hat{\pi}_s)_{s\in\mathcal{S}}$ for estimating $(\pi_s))_{s\in\mathcal{S}}$. The number of observations corresponding to $S = s$ is denoted N_s, for $s \in \mathcal{S}$. Of course $N_{-1} + N_1 = N$. For $s \in \mathcal{S}$, the feature vector in $\mathcal{D}'_N$ denoted $X^s_1, \dots, X^s_{N_s}$ is composed by *i.i.d.* data from $\mathbb{P}_{X^s}$, the distribution of $X|S = s$. All samples are assumed independent.

In order to derive consistency results on the excess fair-risk and the unfairness of our plug-in rule, we require continuity conditions on the random variables $\hat{p}_k(X, S)$, in the spirit of Assumption 2 (conditional on the learning sample). However, such property is automatically satisfied whenever perturbing the $(\hat{p}_k)_k$ with a small random noise. To this end, we introduce $\bar{p}_k(X, S, \zeta_k) := \hat{p}_k(X, S) + \zeta_k$, for a given uniform perturbation ζ_k on $[0, u]$.

Given $(\zeta_k)_{k\in\mathcal{Y}}$ and $(\zeta^s_{k,i})$ independent copies of a Uniform distribution on $[0, u]$, we define the randomized fair classifier $\hat{g}$ as

$$\hat{g}(x, s) = \arg\max_{k\in\mathcal{Y}} \left(\hat{\pi}_s \bar{p}_k(x, s, \zeta_k) - s\hat{\lambda}_k\right) \ , \quad \text{for all } (x, s) \in \mathcal{X} \times \mathcal{S} \ ,$$

with $\hat{\lambda} \in \mathbb{R}^K$ given as

$$\hat{\lambda} \in \arg\min_{\lambda} \sum_{s\in\mathcal{S}} \frac{1}{N_s} \sum_{i=1}^{N_s} \left[\max_{k\in\mathcal{Y}} \left(\hat{\pi}_s \bar{p}_k(X^s_i, s, \zeta^s_{k,i}) - s\lambda_k\right)\right] \ . \tag{3.1}$$

Note that the construction of the plug-in rule $\hat{g}$ relies on (x, s) but also on the perturbations ζ and $\zeta^s_{k,i}$ for $k \in \mathcal{Y}$, $i \in N_s$ and $s \in \mathcal{S}$. Note that the objective function in Equation (3.1) is convex but non smooth due to the evaluation of the function (hard) *max*. In (Denis et al., 2021), we provide an alternative algorithm based on soft-max regularization. Finally, as for the procedure provided in Section 2.3 of Chapter 2, the proposed estimator relies on a post-processing approach. In a first step, preliminary estimators of p_k are builded while in a second step these estimators are re-calibrated to meet fairness criterion.

3.2.2 Statistical guarantees

We are now in position to derive fairness and consistency guarantees of the plug-in procedure.

Universal fairness guarantee. We first focus on fairness assessment and prove that the plug-in estimator $\hat{g}$ is asymptotically fair. The convergence rate on the unfairness to zero is parametric with the number of unlabeled data N. Notably, as in set-valued classification, the fairness guarantee is distribution-free and holds for any estimators of the conditional probabilities.

Theorem 3.1. *For any distribution $\mathbb{P}$, there exists a constant $C > 0$ which only depends on K and $\min_{s\in\mathcal{S}} \pi_s$, such that for any estimators $\hat{p}_k$ we have,*

$$\mathbb{E}\left[\mathcal{U}(\hat{g})\right] \leq \frac{C}{\sqrt{N}}.$$

Consistency of the excess fair-risk. We now consider the excess risk of the estimator $\hat{g}$. We define the L_1-norm in $\mathbb{R}^K$ between the estimator $\hat{\mathbf{p}} := (\hat{p}_1, \ldots, \hat{p}_K)$ and the vector of the conditional probabilities $\mathbf{p} := (p_1, \ldots, p_K)$ as $\|\hat{\mathbf{p}} - \mathbf{p}\|_1 = \sum_{k\in\mathcal{Y}} |\hat{p}_k(X,S) - p_k(X,S)|$.

Theorem 3.2. *Let Assumption 6 be satisfied, then the following holds*

$$\mathbb{E}\left[\mathcal{E}_{\text{fair}}(\hat{g})\right] \leq C\left(\mathbb{E}\left[\|\hat{\mathbf{p}} - \mathbf{p}\|_1\right] + \sum_{s\in\mathcal{S}} \mathbb{E}\left[|\hat{\pi}_s - \pi_s|\right] + \mathbb{E}\left[\mathcal{U}(\hat{g})\right] + u\right) \quad .$$

The above result highlights that the excess fair-risk of $\hat{g}$ depends on 1) the quality of the estimators of the conditional probabilities through its L_1-risk; 2) the quality of the estimators of $(\pi_s)_{s\in\mathcal{S}}$; 3) the unfairness of the classifier; and 4) the upper-bound u on the regularizing perturbations. Consequently, $\hat{g}$ is consistent *w.r.t.* the excess-fair risk as soon as the estimator $\hat{\mathbf{p}}$ is consistent in L_1-norm.

Corollary 3.2. *If $\mathbb{E}\left[\|\hat{\mathbf{p}} - \mathbf{p}\|_1\right] \to 0$ and $u = u_n \to 0$ when $n \to \infty$, we have*

$$\mathbb{E}\left[\mathcal{E}_{\text{fair}}(\hat{g})\right] \to 0, \qquad \text{as} \quad n, N \to \infty \ .$$

We emphasize that Theorem 3.1 and Corollary 3.2 directly imply that $\hat{g}$ performs asymptotically as well as g^*_{fair} in terms of both fairness and accuracy.

3.3 Numerical experiments

In this section, we discuss several numerical aspects of the proposed algorithm. As a benchmark, we introduce an alternative approach that enforces fairness on each individual score in Section 3.3.1. Then, we illustrate efficiency of our procedure on real datasets in Section 3.3.2.

3.3.1 Alternative strategy for fair multi-class classification

The procedure developed in Section 3.2 enforces the score maximizer to be fair. An alternative approach suggested in (Ye and Xie, 2020) consists in imposing fairness at the level of each posterior probability instead of their maximizer. That is to say, we require that the posterior probabilities are DP fair. However, Ye and Xie, 2020 does not provide any theoretical analysis of this approach. In (Chzhen et al., 2020b), we provide a solution of the regression problem under DP constraint. Following this idea, we consider the L_2-risk defined in Chapter 1 and using the same notations, we consider a score function $\mathbf{f}$ for which we define the DP constraint

Definition 3.3. *We say that* $\mathbf{f} : \mathcal{X} \times \{-1, 1\} \mapsto \mathbb{R}^K$ *is* score-fair *in demographic parity if each coordinate of* $\mathbf{f}$ *is fair* w.r.t. *the demographic parity notion of fairness (Chzhen et al., 2020b).*

Consequently, a possible way to tackle fair multi-class classification is to consider the following minimization problem

$$\mathbf{f}^*_{\text{score}-\text{fair}} \in \arg\min \left\{ R_2(\boldsymbol{f}) : \mathbf{f} \text{ is } \textit{score-fair} \right\} \quad .$$

While this approach seems to be rather natural, let us emphasize that *score-fair* DP does not imply DP for the score maximizer, since the maximum, unlike thresholding, operation does not preserve the DP property. Optimal *score-fair* functions rely on the L_2-risk and be easily characterized following the approach in (Chzhen et al., 2020b; Gouic, Loubes, and Rigollet, 2020). In particular, Theorem 2.3 in (Chzhen et al., 2020b) identifies the distribution of score-fair classifier $\mathbf{f}^*_{\text{score}-\text{fair}}$ as solutions of a Wasserstein barycenter problem. We refer to (Denis et al., 2021) for the details of the estimation procedure of $\mathbf{f}^*_{\text{score}-\text{fair}}$.

3.3.2 Application to real datasets

In this section, we illustrate the performance of our method *argmax-fair*, the alternative approach *score-fair*, and the classifier builded without fairness constraint (*unfair*), for both linear and non-linear multi-class classification. For linear models, we consider the one-versus-all logistic regression (reglog) and the SVM with linear kernel (linearSVC); for non-linear models: SVM model with Gaussian kernel (GaussSVC) and Random Forests algorithm (RF).

Datasets. The performance of our method is evaluated on four benchmark datasets for which we provide a short description.

Communities&Crime (CRIME) dataset contains socio-economic, law enforcement, and crime data about communities in the US with 1994 examples. The task is to predict the number of violent crimes per 10^5 population which, we divide into $K = 7$ balanced classes based on equidistant quantiles. Following Calders et al., 2013 and Chzhen et al., 2020c the binary sensitive feature is the percentage of black population.

METHOD \ DATA	CRIME, K = 7		LAW, K = 4		WINE, K = 5		CMC, K = 3	
	Accuracy	Unfairness	Accuracy	Unfairness	Accuracy	Unfairness	Accuracy	Unfairness
reglog + unfair	0.34 ± 0.02	1.12 ± 0.07	0.43 ± 0.01	0.89 ± 0.05	0.54 ± 0.01	0.47 ± 0.05	0.52 ± 0.02	0.78 ± 0.16
reglog + score-fair	0.33 ± 0.01	0.78 ± 0.09	0.42 ± 0.01	0.09 ± 0.02	0.54 ± 0.01	0.08 ± 0.03	0.51 ± 0.02	0.25 ± 0.1
reglog + argmax-fair	0.28 ± 0.01	0.26 ± 0.07	0.42 ± 0.01	0.05 ± 0.02	0.54 ± 0.02	0.04 ± 0.01	0.52 ± 0.02	0.19 ± 0.1
linearSVC + unfair	0.36 ± 0.02	1.12 ± 0.07	0.43 ± 0.01	0.97 ± 0.07	0.53 ± 0.01	0.27 ± 0.05	0.51 ± 0.02	0.63 ± 0.22
linearSVC + score-fair	0.31 ± 0.02	0.88 ± 0.05	0.42 ± 0.01	0.1 ± 0.03	0.53 ± 0.01	0.1 ± 0.07	0.53 ± 0.02	0.26 ± 0.16
linearSVC + argmax-fair	0.29 ± 0.02	0.25 ± 0.08	0.42 ± 0.01	0.04 ± 0.02	0.53 ± 0.01	0.06 ± 0.04	0.52 ± 0.02	0.2 ± 0.12
GaussSVC + unfair	0.36 ± 0.02	1.4 ± 0.13	0.43 ± 0.01	1.04 ± 0.04	0.53 ± 0.01	0.28 ± 0.06	0.51 ± 0.02	1.0 ± 0.17
GaussSVC + score-fair	0.35 ± 0.02	1.02 ± 0.07	0.42 ± 0.01	0.16 ± 0.04	0.55 ± 0.01	0.12 ± 0.04	0.51 ± 0.02	0.16 ± 0.09
GaussSVC + argmax-fair	0.3 ± 0.02	0.22 ± 0.05	0.42 ± 0.01	0.10 ± 0.03	0.55 ± 0.01	0.06 ± 0.03	0.5 ± 0.03	0.2 ± 0.08
RF + unfair	0.37 ± 0.02	1.02 ± 0.04	0.40 ± 0.01	0.65 ± 0.04	0.66 ± 0.01	0.31 ± 0.05	0.55 ± 0.02	0.35 ± 0.18
RF + score-fair	0.34 ± 0.02	0.67 ± 0.06	0.39 ± 0.01	0.11 ± 0.05	0.66 ± 0.01	0.09 ± 0.03	0.52 ± 0.03	0.21 ± 0.08
RF + argmax-fair	0.3 ± 0.02	0.33 ± 0.11	0.39 ± 0.01	0.07 ± 0.02	0.66 ± 0.01	0.08 ± 0.02	0.55 ± 0.02	0.22 ± 0.13

Table 3.1: Performance (accuracy & unfairness) of the methods for all datasets and classifiers. We report the means and standard deviations over the 30 repetitions. Colored values highlight fairness.

Law School Admissions (LAW) dataset (Wightman and Ramsey, 1998) presents national longitudinal bar passage data and has 20649 examples. The task is to predict a students GPA divided into $K = 4$ classes based on equidistant quantiles. The sensitive attribute is the race (white versus non-white).

Wine Quality (WINE) dataset (Cortez et al., 2009) reports the description of 6497 wines and the task is to predict the quality graded by the experts. The quality is between 3 (bad) and 9 (good) but we consider only $K = 5$ classes (4 to 8) due to a too low frequency of the class 3 and 9 (resp. 5 and 30 examples). The sensitive attribute is the color (red versus white).

Contraceptive Method Choice (CMC) dataset is about 1987 National Indonesia Contraceptive Prevalence Survey. The problem is to predict the contraceptive method choice of a woman (no use, long-term or short-term methods) based on her demographic and socio-economic characteristics. The sensitive feature is the religion (Islam versus Non-Islam).

Performance. Results are presented in Table 3.1 and highlight the effectiveness of our method. As an example, for the LAW dataset and the GaussSVC with *argmax-fair*, the unfairness is divided by almost 25 (0.97 to 0.04). Furthermore, the *argmax-fair* procedure outperforms the *unfair* and the *score-fair* algorithms for the datasets CRIME, LAW and WINE in terms of unfairness: However, we observe a small decrease of the models accuracy (relatively small compared to the gain in fairness). Note that for the dataset CMC, *score-fair* and *argmax-fair* achieve similar performance.

3.4 Conclusion

In the multi-class classification framework, we provide an optimal fair classification rule under DP constraint and derive misclassification and fairness guarantees of the as-

sociated plug-in fair classifier. Our approach achieves distribution-free fairness and can be applied on top of any probabilistic base estimator. We illustrate the proficiency of our procedure on various synthetic and real datasets, notably in comparison to the *score-fair* approach suggested in (Ye and Xie, 2020). The efficiency of our algorithm in terms of fairness is particularly salient for datasets with large historical bias.

However, our numerical study also outlines the downside of fairness proficiency in terms of classification accuracy. One should hereby be very cautious when using classifiers with strong fairness guarantee, as it possibly degrades the classification quality. This calls for an analysis of classification problems with fairness constraints from a multi-objective perspective and paves the way for characterizing the Pareto front between fairness and accuracy objectives. An interesting approach to handle both fairness and accuracy in the multi-class setting will be to consider set-valued classification under fairness constraint.

Chapter 4

Multi-class classification for diffusion paths

In this chapter, I present a generalization of the mixture model described in Chapter 1 to the case where the feature consists of observations of a diffusion process, that is solution of a stochastic differential equation (s.d.e.), observed over a fixed time interval $[0, T]$. As a consequence, the drift function of the s.d.e. in this model depends on the label of the observation. While mixture models are widely studied from a statistical learning perspective, the presented results are among the first that deal with the s.d.e. framework and constitute a new contribution to the functional data learning field.

Usually, in the context of parameters estimation of a diffusion process, the only available data consists of a single observation (continuous or discrete) of a diffusion path. Within this context, $T \to +\infty$ and ergodicity properties of the diffusion are used in force to handle this problem. One of the specificity of the model we study in this chapter is that the horizon time T is fixed and the ergodicity of the diffusion process is not required. However, in the multi-class setting, this framework can be considered since a learning sample of independent copies of (X, Y) is available. The asymptotic in T is then replaced by the asymptotic *w.r.t.* the sample size. Throughout this chapter, we assume that data are collected at discrete times.

Up to our knowledge, the work of Cadre (2013) is the first one that tackles the problem of classification in the s.d.e. framework. There, the authors focus on binary classification for continuous observations and provide an estimation procedure relying on the empirical risk minimization strategy. However, this method cannot be implemented for practical purpose since the proposed procedure does not consider E.R.M. estimator based on convex surrogate of the misclassification risk (see Section 1.2.2 in Chapter 1). More recently, the work of Gadat, Gerchinovitz, and Marteau (2020) provides a minimax analysis of a much simpler mixture model where the input feature is modeled as a solution of a white noise equation.

In (Denis, Dion, and Martinez, 2020), we extend the results obtained by Cadre (2013)

to discrete time observations in the multi-class framework. In the parametric setting, we provide theoretical guarantees for different estimation procedures which are easily implementable. A part of these results are presented in Section 4.3.

In addition, we also investigate in (Denis, Dion, and Martinez, 2020) the parametric estimation of the drift function in our specific framework. Motivated by the extension of our results to the nonparametric case, a second line of results presented in this chapter focuses on the estimation of the drift function in the context of i.i.d. repeated observations. This part is developed in (Denis, Dion, and Martinez, 2021) where we propose a novel constrained estimator relying on a ridge type constraint. This procedure offers appealing properties and can be viewed as an alternative strategy to the cutoff procedure developed in (Comte and Genon-Catalot, 2020) studied in the context of continuous observations. The construction of the procedure and its main properties are provided in Section 4.4.

4.1 Model and assumptions

In this chapter, the feature X is a mixture of Brownian motion with drift. More precisely, X takes its value in $\mathcal{X} := (C([0,T]), \mathcal{C})$ the set of real valued continuous functions with its corresponding σ-algebra endowed by the uniform topology. Given a starting point $x_0 \in \mathbb{R}$, the process $X = (X_t)_{t\in[0,T]}$ is assumed to come from the following diffusion model

$$\begin{cases} X_0 &= x_0 \\ dX_t &= b^*_Y(X_t)dt + \sigma(X_t)dW_t, \end{cases} \tag{4.1}$$

where $(W_t)_{t\geq 0}$ denotes a standard Brownian motion on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and such that the label Y is independent of $(W_t)_{t\geq 0}$ with known distribution under $\mathbb{P}$ given by $(\pi_k)_{k\in\mathcal{Y}}$. The function $b^* = (b^*_1, \ldots, b^*_K)$ is a vector of K unknown Borel real functions. The real-valued function σ is assumed to be known. In the sequel, $(\mathcal{F}^X_t)_{t\geq 0} := \{\sigma(X_s : s \leq t) ; t \geq 0\}$ denotes the natural filtration of the process X.

4.1.1 Assumptions

Throughout this chapter, we make the following assumptions.

Assumption 7 (Ellipticity and regularity)**.** *There exist strictly positive constants σ_0, σ_1 such that*

$$0 < \sigma_0 \leq \sigma(x) \leq \sigma_1, \quad \forall x \in \mathbb{R}.$$

There exists a positive constant L_0 such that

$$\sup_{i\in\mathcal{Y}} |b^*_i(x) - b^*_i(y)| + |\sigma(x) - \sigma(y)| \leq L_0|x-y|, \quad \forall (x,y) \in \mathbb{R}^2.$$

Assumption 7 ensures the existence and uniqueness of a strong solution for Equation (4.1) and that $\mathbb{E}[\sup_{t\in[0,T]} |X_t|^q] < \infty$ for any integer $q \geq 1$. Furthermore, it implies that

$$\sup_{i\in\mathcal{Y}} |b_i^*(x)| \leq C_0(1 + |x|). \tag{4.2}$$

Finally, b^* satisfies the following condition.

Assumption 8 (Novikov condition)**.**

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^T \frac{b_i^{*2}}{\sigma^2}(X_s)ds\right)\right] < +\infty, \quad \forall i \in \mathcal{Y}.$$

Here Novikov's condition (Assumption 8) is sufficient to apply Girsanov's theorem (Revuz and Yor, 2013)

4.1.2 Bayes Classifier

The Bayes classifier g^* is then defined as

$$g^*(X) \in \arg\max_{k\in\mathcal{Y}} p_k(X), \quad p_k(X) := \mathbb{P}\left(Y = k|\mathcal{F}_T^X\right).$$

The following proposition, which is an extension of the one obtained in (Cadre, 2013) in the context of binary classification, provides a closed form of the Bayes classifier. This result relies on the application of the Girsanov's theorem.

Proposition 4.1. *For all $t \in (0, T)$ and each $i \in \mathcal{Y}$ we define*

$$F^k := \int_0^T \frac{b_k^*}{\sigma^2}(X_s)dX_s - \frac{1}{2}\int_0^T \frac{(b_k^*)^2}{\sigma^2}(X_s)ds.$$

The sequence of conditional probabilities satisfies

$$p_k(X) = \mathbb{P}\left(Y = k|\mathcal{F}_T^X\right) = \varphi_k(F) \quad \mathbb{P} - a.s$$

where $F = (F^1, \ldots, F^K)$, and $\varphi_k : (x_1, \ldots, x_K) \mapsto \dfrac{p_k e^{x_k}}{\sum_{j=1}^K p_j e^{x_j}}$ are the softmax functions.

Proposition 4.1 is a key result and is the bedrock of the classification procedures presented in this chapter. Indeed, it highlights the dependency of the Bayes classifier *w.r.t.* the unknown function b^*. It naturally suggests that consistent classification rule can be derived from estimator of b^*.

4.2 Estimation strategy

Let $(X_t)_{t\in[0,T]}$ be the solution of (4.1). We assume that the observation consists of a single discretized sample path $\overline{X}(\omega) := (X_{k\Delta}(\omega))_{k\in\{0,\dots,n\}}$ with $T = n\Delta$. Note that while we assume that T is fixed, we consider the asymptotic $n \to +\infty$. As we assume that observations are collected at discrete times, the learning sample $\mathcal{D}_N$ consists in i.i.d. observations $(\overline{X}^{(j)}, Y^{(j)})_{j=1,\dots,N}$ of $(\overline{X}, Y)$.

4.2.1 Discrete observations classifier

Hereafter, we define a set of classifiers which are based on the discrete time observations $(X_{k\Delta})_{k\in\{0,\dots,n\}}$. We refer these classifiers as *discrete observations classifiers*. In particular, we provide a control of the excess risk of these classifiers.

For a trajectory X and $b = (b_1, \dots, b_K)$ a vector of K Borel real functions, we define for $i \in \mathcal{Y}$ the discrete version of F based on $(X_{k\Delta})_{k\in\{0,\dots,n\}}$ and b:

$$\overline{F}_b^i := \sum_{k=0}^{n-1} \left(\frac{b_i}{\sigma^2}(X_{k\Delta})(X_{(k+1)\Delta} - X_{k\Delta}) - \frac{\Delta}{2}\frac{b_i^2}{\sigma^2}(X_{k\Delta}) \right), \ \overline{F}_b := (\overline{F}_b^1, \dots, \overline{F}_b^K).$$

Then we set $\overline{p}_b^k(X) := \varphi_k(\overline{F}_b)$, $k = 1, \dots, K$.

Finally, for any function b, we then naturally define the discrete observations classifier $\overline{g}_b$ by

$$\overline{g}_b(X) := \arg\max_{k\in\mathcal{Y}} \overline{p}_b^k(X).$$

4.2.2 Comparison inequality

The following proposition establish a bound for the excess risk of some discrete observations classifier $\overline{g}_b$, and highlights its link with a suitable distances between b and b^*. We introduce the norm $\|.\|_T$ defined for a real valued function f and a process X from model (4.1):

$$\|f\|_T^2 := \sup_{t\in[0,T]} \mathbb{E}[|f(X_t)|^2].$$

Moreover, for a function $b = (b_1, \dots, b_K)$, we set $\|b\|_T := \max_{i\in\mathcal{Y}} \|b_i\|_T$.

Proposition 4.2. *Let $b = (b_1, \dots, b_K)$. Assume that there exists $C_b > 0$ and $\gamma \geq 1$ such that $\sup_{i\in\mathcal{Y}} |b_i(x)| \leq C_b(1 + |x|)^\gamma$. Then, the following holds*

$$\begin{aligned}\sum_{i=1}^K \mathbb{E}_X\left[\left|\overline{p}_b^i(X) - p_i(X)\right|\right] &\leq C\left(\sqrt{\Delta} + \sum_{i=1}^K \left(\mathbb{E}_X\left[\frac{1}{n}\sum_{k=0}^{n-1}(b_i - b_i^*)^2(X_{k\Delta})\right]\right)^{1/2}\right) \\ &\leq C_1\left(\sqrt{\Delta} + \|b - b^*\|_T\right).\end{aligned}$$

where C, C_1 are positive constants which depend on T, K, C_b and on the constants in the Assumptions 7, 8.

Therefore, if we manage to build consistent estimator $\hat{b}$ of b^* *w.r.t.* $\|.\|_T$, the resulting plug-in classifier $\bar{g}_{\hat{b}}$ is also consistent. In (Denis, Dion, and Martinez, 2020), we investigate the parametric setting and provide theoretical guarantees for estimator of the drift defined as the minimizer of a contrast function based on the Gaussian log-likelihood approximation. We will discuss the extension to the nonparametric case later (see Section 4.3.3). In the next section, we present results obtained for classifiers defined as empirical risk minimizers which take advantage of Proposition 4.2.

4.3 Classification procedure based on emprirical risk minimization

In (Cadre, 2013), the empirical risk minimization procedure is used in the context of binary classification where the features come from continuous diffusion sample paths and are discriminated by their drift. Following the same idea, we investigate the case where the estimator of b^* is defined as an empirical risk minimizer. To this end, we assume that b^* belongs to a set of functions $\mathcal{B}$. Besides, for $b \in \mathcal{B}$, we introduce the empirical risk of a discrete observations classifier $\overline{g}_b$ by

$$\widehat{R}(\overline{g}_b) = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}_{\{\overline{g}_b(\overline{X}^{(j)}) \neq Y^{(j)}\}}.$$

Now, assume that there exists an ε-net $\mathcal{B}_\varepsilon \subseteq \mathcal{B}$ with respect to the norm $\|.\|_T$. We define the estimator $\widehat{b}^\varepsilon = (\widehat{b}_1^\varepsilon, \ldots, \widehat{b}_K^\varepsilon)$ as

$$\widehat{b}^\varepsilon \in \arg\min_{b \in \mathcal{B}_\varepsilon} \widehat{R}(\overline{g}_b). \tag{4.3}$$

As mentioned in Chapter 1, this estimator can not be considered in practice. One of the contribution provided in (Denis, Dion, and Martinez, 2020) is to bypass this issue by proposing a procedure which involves a convex surrogate of the minimization problem defined in Equation (4.3). The procedure is described in Section 4.3.1. However, from a theoretical perspective it is always interesting to study the properties of $\widehat{b}^\varepsilon$, at least as a benchmark. The following theorem establishes the rates of convergence of the classification procedure $\overline{g}_{\widehat{b}^\varepsilon}$ *w.r.t.* its excess risk.

Theorem 4.1. *Assume that there exists $u > 0, C > 0$ such that $\log(\mathrm{Card}(\mathcal{B}_\varepsilon)) \leq C\varepsilon^{-u}$. Let $\Delta = O\left(N^{-2/(2+u)}\right)$ and $\varepsilon \propto N^{-1/(2+u)}$. The empirical risk minimizer satisfies*

$$\mathbb{E}\left[\mathcal{E}(\overline{g}_{\widehat{b}^\varepsilon})\right] \leq \frac{C}{N^{1/(2+u)}}.$$

where C is a positive constant which depends on T, K, and on the constants in the Assumptions 7 and 8.

The proof of this result relies on Proposition 4.2 and classical tools of empirical process. It shows that, provided that the time step Δ is sufficiently small, the obtained rates of convergence is the same to the one obtained when the feature X belongs to $\mathbb{R}$. This rate of convergence is obtained in (Cadre, 2013) in the context of binary classification with continuous time observations.

Hereafter, we provide an example of set $\mathcal{B}$ for which Theorem 4.1 applies. Define $\psi(x) := C_1(1+|x|)$ and $C_1 \geq C_0$ (with C_0 given in Equation (4.2)). Let $\beta \geq 1$ and consider

$$\mathcal{F}_\beta = \left\{ b \in \mathcal{C}^\beta,\ \forall j \in \mathbb{Z},\ i = 0, \dots, \beta\ \sup_{[j,j+1[} \left| \frac{d^i b}{dx^i} \right| \leq \psi(|j|),\ |b(x)| \leq \psi(x) \right\}.$$

Application of the result given in (Van Der Vaart and Wellner, 1996) (see the proof of Theorem 2.7.1) shows that Theorem 4.1 can be applied with $\mathcal{B} = (\mathcal{F}_\beta)^K$ with $u = 1/\beta$. In this case, we obtain that the rate of convergence of the empirical risk minimizer is of order of $N^{-\beta/2\beta+1}$ where β is the smoothness of drift functions b_i^*. Hence, we obtain similar rate as in the classical classification setting (Yang, 1999).

4.3.1 One-versus-All approach

In this section, we focus on the case where the set $\mathcal{B}$ is a parametric family of drift functions. The set $\mathcal{B}$ is defined as follows

$$\mathcal{B} = \{(b(\theta_i, .))_{i \in \mathcal{Y}},\ \forall i \in \mathcal{Y},\ \theta_i \in \Theta\},$$

where $\Theta \subset \mathbb{R}^d$ is compact and for each $\theta \in \Theta$, $x \mapsto b(\theta, x)$ is a real valued function which satisfies Assumptions 7, 8. Moreover, we assume that the function b is known. For each $i \in \mathcal{Y}$, we denote the drift functions by $b_i^*(x) := b(\theta_i^*, x)$, $\theta_i^* \in \Theta$ (and $\pi^* = \pi_{b_{\theta^*}}$). Furthermore, for $\theta = (\theta_1, \dots, \theta_K) \in \Theta^K$, we denote the vector $(b(\theta_i, .))_{i \in \mathcal{Y}}$ by $b_\theta = (b_{\theta_1}, \dots, b_{\theta_K})$. Finally, for $\theta \in \Theta^K$, we also define $\|\theta\| = \max_{i \in \mathcal{Y}} \|\theta_i\|_\infty$. We consider the following assumption

Assumption 9. *Function b is Lipschitz-continuous with respect to $\theta \in \Theta$:*

$$|b(\theta, x) - b(\theta', x)| \leq C\,(1 + |x|)\,\|\theta - \theta'\|_\infty.$$

This assumption implies that for $\theta, \theta^{'} \in \Theta$, we have

$$\|b(\theta, \cdot) - b(\theta', \cdot)\|_T \leq C\|\theta - \theta^{'}\|_\infty,$$

for a constant C depending on T.

Classification procedure. We derive a classification procedure based on the one-versus-all approach. It involves a convex surrogate of the minimization problem defined in Equation (4.3). Hence, we consider the square loss and apply the methodology described in Chapter 1. Let $\mathbf{f}(\cdot) = (f^1(\cdot), \dots, f^k(\cdot)$ a score function, its associated L_2 risk is

$$R_2(\mathbf{f}) = \left[\sum_{k=1}^{K} \mathbb{E}\left(Z_k - f^k(X)\right)^2\right], \; Z_k = 2\,\mathbb{1}_{\{Y=k\}} - 1.$$

In this case, we recall that

$$f^{*k}(X) = 2p_k(X) - 1, \;\; k \in \mathcal{Y}.$$

In view of the form of the optimal score function f^*, we naturally define the estimator $\widehat{\theta}$ of the true parameter θ^*

$$\widehat{\theta} \in \arg\min_{\theta \in \Theta^K} \widehat{R}_2(\overline{f}_\theta), \;\; \overline{f}^k_\theta = 2\overline{p}^k_{b_\theta} - 1, \;\; k \in \mathcal{Y}, \tag{4.4}$$

with $\widehat{R}_2$ the empirical counterpart of R_2. The following theorem establishes the consistency of the proposed procedure.

Theorem 4.2. *Assume that $\Theta = [0,1]^d$ and that there exists $\alpha \geq 2$ such that $\Delta \propto O(N^{-\alpha})$. Under Assumption 9, the classification procedure $\overline{g}_{b_{\widehat{\theta}}}$ given by (4.4) satisfies,*

$$\mathbb{E}\left[R(\overline{g}_{b_{\widehat{\theta}}}) - R(g^*)\right] \leq O\left(\sqrt{\frac{d\log(N)}{N}}\right).$$

We can note that up to the logarithmic factor, we obtain the usual parametric rate of convergence which is of order of $N^{-1/2}$. Interestingly, if we consider $\widehat{\theta} \in \arg\min_{\theta \in \Theta_N} \widehat{R}(\overline{g}_{b_\theta})$ with Θ_N a $1/N$-net of Θ^K, one can show that the rate of convergence is also of order $N^{-1/2}$. Hence, from a theoretical point of view, the use of convex surrogate does not degrade the performance of the classification procedure when $\alpha \geq 2$.

4.3.2 Numerical evaluation

In this section, we provide a short simulation study to evaluate the performance of the procedure described in Section 4.3.1. Let us describe the models under consideration for our numerical experiments. We fix $K = 3$, $\pi_k = 1/K$ and $\sigma = 1$. We consider the following examples

- model 1 Additive OU $b(\theta, x) = -(x - \theta)$, $x_0 = 4$
- model 2 Multiplicative OU $b(\theta, x) = -\theta x$, $x_0 = 4$

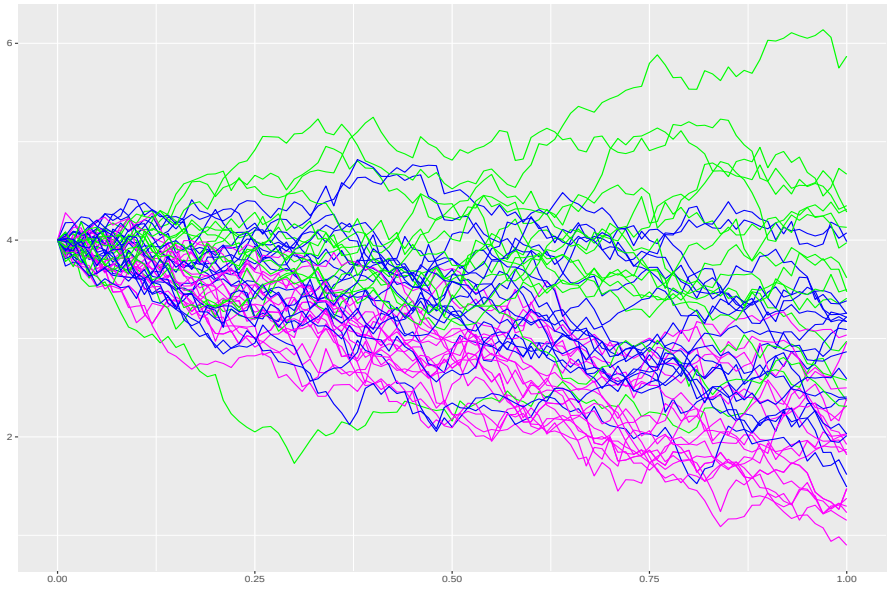

Figure 4.1: Illustration of the classification problem with 3 classes (blue, green and magenta) for the three different values of θ^*) for model *Additive* with $n = 100,\ N = 100$.

- model 3 Polynomial $b(\theta, x) = -(x-\theta)^3 - (x+\theta)^3$, $x_0 = 4$
- model 4 Hyperbolic $b(\theta, x) = -\theta x/\sqrt{1+x^2}$, $x_0 = 4$

We compare the results on the design: $\theta^* = \{1, 2, 4\}$ for model 1, 2, 4, and $\theta^* = \{1/4, 1/2, 1\}$ for model 3. The models 1 and 2 are widely used in practical applications, and they satisfy all the assumptions required for our theoretical results, while the model 3 does not fulfill the Assumption 7, illustrating the robustness of the classification procedure. The model 4 is widely used in mathematical finance to model log-returns of assets prices in stock markets.

Figure 4.1 displays some trajectories generated according to the model 1 (Additive). At first sight, without the knowledge of the labels, it seems to be difficult to assign a class to each trajectory. It illustrates the difficulty of this classification problem (see Table 4.1). As a benchmark, we also evaluate the procedure provided in (Denis, Dion, and Martinez, 2020) based on the contrast estimation which is referred as `MLE`. This procedure relies on the parametric estimation of parameter θ_k. The procedure based on the one-versus-all strategy is referred as `OVA` $\overline{g}_{\widehat{\theta}}$ with $\widehat{\theta}$ given in Equation (4.4).

We fix $n \in \{50, 250\}$, $\Delta = 1/n$, $N = 500$ For each model we provide an evaluation of the misclassification risk of the two classification procedures by using the Monte-Carlo method over 100 repetitions. Furthermore, the risk of the Bayes rule is evaluated independently with a sample of size 10000. The results are summarized in Table 4.1 .

	Oracle	MLE	OVA
Model 1	0.31 (.002)	0.31 (0.01)	0.31 (0.01)
Model 2	0.12 (.003)	0.12 (0.01)	0.12 (0.01)
Model 3	0.22 (.003)	0.23 (0.01)	0.22 (0.01)
Model 4	0.33(.004)	0.33 (0.02)	0.33 (0.01)

Table 4.1: Average and standard deviation of the misclassification error rate of the two procedures with $n = 250$ and $N = 500$.

First of all, note that model 1 and model 4 seem to be more tricky for the misclassification risk. This is due to the fact that the classes generated by θ_1^* and θ_2^* are much overlapped. On the contrary, the classification problem involved by model 2 is more easier. Second, all the classification procedures perform well. Indeed, the evaluation of the misclassification risk are closed to the Bayes risk with small variances. Furthermore the two procedures have similar performance.

4.3.3 A first conclusion

This section is dedicated to a preliminary discussion regarding the presented results. The contribution of (Denis, Dion, and Martinez, 2020) should be viewed as a first step of the study of the problem defined by Equation (4.1). Notably, we show that standard techniques of statistical learning can be successfully apply to the mixture diffusion model. We establish, provided that the time step Δ is sufficiently small, that the classification procedure presented in Section 4.3 achieves standard rates of convergence. However, results provided in (Denis, Dion, and Martinez, 2020) mainly focus on the parametric setting and require that the weights of the mixture as well as the diffusion coefficient are known.

A second step of our research on this topic has been to focus on the study of plug-in procedures in the nonparametric setting including the estimation of the weights and the diffusion coefficient in the procedure. In light of this, the objective is to study a procedure where drift functions b_k are separately estimated by a nonparametric method. Naturally, the first challenge is to focus on the estimation of the drift function within the i.i.d. framework. Indeed, this problem is usually tackled in the setting where the observation consists in a single path. In this context, an estimator of the drift function is evaluated when the horizon time T tends to infinity and under the assumption that the underlying process is ergodic. On the contrary, in our framework the horizon time is fixed and we do not require ergodicity property. The asymptotic *w.r.t.* T is then replaced by the asymptotic with respect to the learning sample. Clearly, these two settings are different and classical approaches can not be considered for the estimation of the drift function under the i.i.d. framework.

4.4 Estimation of the drift function under i.i.d. framework: a first step toward the plug-in procedure

In this section, we focus on the following model

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \quad X_0 = x_0 \tag{4.5}$$

where $x_0 \in \mathbb{R}$ is known, and $(W_t)_{t\geq 0}$ denotes a standard Brownian motion.

The estimation of the drift function of a diffusion process from a single path is a well known problem (see for instance Kutoyants, 2004). More precisely, one can cite, (Yoshida, 1992; Gobet, 2002) for the case of continuous ergodic diffusions, (Bibby and Sørensen, 1995; Kessler, Sørensen, et al., 1999) for martingale estimation functions, (Gobet, Hoffmann, Reiß, et al., 2004) in the low frequency context, and (Hoffmann, 1999; Dalalyan et al., 2005; Comte, Genon-Catalot, and Rozenholc, 2007; Schmisser, 2013) in the nonparametric context. In the Bayesian literature, the asymptotic properties of minimum contrast estimators are studied for example in (Meulen and Van Zanten, 2013; Gugushvili and Spreij, 2014; Koskela, Spano, and Jenkins, 2019).

Nevertheless, it seems that very few works investigate the estimation of the drift function from a sample of i.i.d. observations (diffusion paths) when the horizon time T is assumed to be fixed. In fact, up to our knowledge, only Comte and Genon-Catalot (2020) and Della Maestra and Hoffmann (2021) deals with this framework. However, the work of Della Maestra and Hoffmann (2021) focuses on the more general setting of stochastic system of N interacting particles and does not directly handle the i.i.d. framework. The closest contribution to ours is provided in (Comte and Genon-Catalot, 2020) where the authors consider a least squares contrast estimator (based on continuous observations). In order to ensure the stability of the estimator, the authors propose to insert a cutoff function. More precisely, the estimator is set to the zero function according to some threshold which depends on the dimension of the considered space of approximation. This procedure may reduce the dimension of the spaces of approximation on which the resulting estimator is non trivial and can lead to some limitations in practice. As an alternative strategy, we propose in (Denis, Dion, and Martinez, 2021) to build a regularized estimator based on a ridge constraint.

4.4.1 Assumptions and notations

We present results obtained in (Denis, Dion, and Martinez, 2021) for the estimation of the drift function on a compact interval that we assume to be $[0, 1]$ for simplicity. That is to say, we focus on the estimation of $\tilde{b}(.) = b(.)\mathbb{1}_{[0,1]}(.)$.

We consider the assumptions detailed in Section 4.1. We also assume that

Assumption 10. *σ belongs $\mathcal{C}_b^2(\mathbb{R})$.*

Note that under the considered assumptions, the process $(X_t)_{t\in[0,T]}$ admits a transition density $(t,x) \mapsto p(t,x_0,x)$.

Finally, we assume that $N \in \mathbb{N}^*$ independent discrete observations $\left(\overline{X}^{(1)},\dots,\overline{X}^{(N)}\right)$ coming from independent solutions $\left(X^{(1)},\dots,X^{(N)}\right)$ of (4.1) are available. We refer to the vector of observations $\left(\overline{X}^{(1)},\dots,\overline{X}^{(N)}\right)$ as the learning sample.

We introduce some additional notations and then give key result on the transition density p. For a real valued function h defined on $\mathbb{R}$, we denote $\|h\|_{n,b}$ the empirical integrated norm defined as:

$$\|h\|_{n,b}^2 := \mathbb{E}_X\left[\frac{1}{n}\sum_{k=0}^{n-1} h^2(X_{k\Delta})\right],$$

where $\mathbb{E}_X$ is the expectation with respect to the law $\mathbb{P}_X$ of the discrete path $\overline{X}$ defined by (4.5). Its standard L^2-norm is denoted by $\|h\|$. Let us also introduce the following empirical norm

$$\|h\|_{N,n}^2 := \frac{1}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1} h^2\left(X_{k\Delta}^{(j)}\right).$$

The following lemma highlights the connection between the norms $\|.\|_{n,b}$ and $\|.\|$

Lemma 4.1. *Under assumptions 7, and 10, there exists $\pi_1 > \pi_0 > 0$, such that for all $y \in [0,1]$ and $n \geq 4$, we have*

$$\pi_0 \leq \frac{1}{n}\sum_{k=1}^{n-1} p(k\Delta, x_0, y) \leq \pi_1.$$

This result is one the main tool for the study of the rates of convergence. Notably, for a function h such that $\mathrm{supp}(h) \subseteq [0,1]$, we have $\|h\|^2 \leq \frac{1}{\pi_0}\|h\|_{n,b}^2$.

4.4.2 Ridge estimator for drift function

In this section, we describe our regularized procedure which relies on a projection estimator based on the B-spline basis.

***B*-spline basis.** Let $K_N \in \mathbb{N}^*$, $A_N, B_N \in \mathbb{R}$, $A_N < B_N$, and $M \in \mathbb{N}^*$. Let us introduce the sequence of knots $\mathbf{u} = (u_{-M},\dots,u_{K_N+M})$ such that for $i = 0,\dots,K_N$

$$u_i = A_N + i\frac{(B_N - A_N)}{K_N},$$

$u_{-M} = \dots = u_{-1} = u_0 = A_N$, and $u_{K_N} = u_{K_N+1} = \dots = u_{K_N+M} = B_N$. We consider the B-splines functions $(B_{i,M,\mathbf{u}})_{i=-M,\dots,K_N-1}$ of degree M associated to the knot vector $\mathbf{u}$. The B-splines functions are defined as follows (see for instance Györfi et al., 2006, and references therein).

Definition 4.1. *the B-spline function of degree ℓ with knots vector $\mathbf{u}$ is recursively defined for all $x \in \mathbb{R}$ by,*

$$B_{i,\ell,\mathbf{u}}(x) = \mathbb{1}_{[u_i,u_{i+1})}(x),$$

for $\ell = 0$, and $i = -M, \dots, K_N + M - 1$, and

$$B_{i,\ell+1,\mathbf{u}}(x) = \frac{x - u_i}{u_{i+\ell+1} - u_i} B_{i,\ell,\mathbf{u}}(x) + \frac{u_{i+\ell+2} - x}{u_{i+\ell+2} - u_{i+1}} B_{i+1,\ell,\mathbf{u}}(x),$$

for $\ell = 0, \dots, M-1$, and $i = -M, \dots, K_N + M - l - 2$. We use the convention $0/0 = 0$.

Note that the B-spline functions are positive functions. According to the choice of the knot vector $\mathbf{u}$, the B-spline functions are zero outside $[A_N, B_N]$. Besides, these functions are linearly independent even though their supports are not disjoint. The main advantage of these piecewise polynomial functions is that they satisfy some global smoothness conditions. This kind of attractive property is particularly interesting when we want to build smooth estimates. Finally, the B-spline space $\mathcal{S}_{K_N,M,\mathbf{u}}$ is defined as

$$\mathcal{S}_{K_N,M,\mathbf{u}} = \text{span}\{(B_{i,M,\mathbf{u}}) \ : \ i = -M, \dots, K_N - 1\}.$$

Hence, the linear space $\mathcal{S}_{K_N,M,\mathbf{u}}$ has dimension $\dim(\mathcal{S}_{K_N,M,\mathbf{u}}) = K_N + M$. We also recall that if $h \in \mathcal{S}_{K_N,M,\mathbf{u}}$, then h is $M-1$ continuously differentiable on $[A_N, B_N)$ and zero outside of $[A_N, B_N)$. Another appealing property of the B-spline is that for all $x \in [A_N, B_N)$, $\sum_{i=-M}^{K_N-1} B_{i,M,\mathbf{u}}(x) = 1$.

Constrained estimation based on the B-spline basis. Since we focus on the estimation of b on $[0,1]$, we set $A_N = 0$, and $B_N = 1$. For $L_N > 0$, we define the constrained subspace

$$\mathcal{S}_{K_N,L_N,M} := \left\{ h = \sum_{i=-M}^{K_N-1} a_i B_{i,M,\mathbf{u}} \in \mathcal{S}_{K_N,M,\mathbf{u}} \ : \ \|\mathbf{a}\|_2^2 \leq (K_N + M) L_N \right\}. \tag{4.6}$$

The subspace $\mathcal{S}_{K_N,L_N,M}$ is composed of functions $h = \sum_{i=-M}^{K_N-1} a_i B_{i,M,\mathbf{u}}$ for which we ensure uniform boundedness on the coefficients a_i. Then, in view of the properties of the B-spline, functions of $\mathcal{S}_{K_N,L_N,M}$ are bounded w.r.t. $\|\cdot\|_\infty$ and $\|\cdot\|$. Note that the choice of the tuning parameter in the constraint ensures a control of the bias term. Indeed, assume that $\|\tilde{b}\|_\infty \leq \sqrt{L_N}$, then there exists $\tilde{h} \in \ \mathcal{S}_{K_N,L_N,M}$ such that

$$\left|\tilde{h}(x) - \tilde{b}(x)\right| \leq \frac{C}{K_N}, \ \ \forall x \in (0,1).$$

We consider the estimator $\widehat{b}_{N,n}$ defined as the minimizer of a least square contrast

$$\widehat{b}_{N,n} \in \arg \min_{h \in \mathcal{S}_{K_N,L_N,M}} \gamma_{N,n}(h), \tag{4.7}$$

where for $h \in \mathcal{S}_{K_N,L_N,M}$,

$$\gamma_{N,n}(h) := \frac{1}{N\,n}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\left(Z_{k\Delta}^{(j)} - h(X_{k\Delta}^{(j)})\right)^2, \quad Z_{k\Delta}^{(j)} := \frac{X_{(k+1)\Delta}^{(j)} - X_{k\Delta}^{(j)}}{\Delta}.$$

Note that the resulting estimator is then defined as $\widehat{b}_{N,n}(.) = \sum_{i=-M}^{K_N-1} \widehat{a}_i B_{i,M,\mathbf{u}}(.)$ where the vector $\widehat{\mathbf{a}} = {}^t(\widehat{a}_{-M},\dots,\widehat{a}_{K_N-1}) \in \mathbb{R}^{K_N+M}$ is the ridge estimator (see Hastie, Tibshirani, and Friedman, 2001):

$$\widehat{\mathbf{a}} = \arg\min_{\|\mathbf{a}\|_2^2 \le (K_N+M)L_N} \|\mathbf{Z} - \mathbf{B}\mathbf{a}\|_2^2,$$

where the vector $\mathbf{Z} = {}^t(Z_{\Delta}^{(j)},\dots,Z_{n\Delta}^{(j)},\ j = 1,\dots n)$ belongs to $\mathbb{R}^{Nn}$ and the matrix $\mathbf{B} = (B_{i,M,\mathbf{u}}(\mathbf{X}_j))_{j,i} \in \mathbb{R}^{(Nn)\times(K_N+M)}$, with $\mathbf{X}_j = {}^t(X_{\Delta}^{(j)},\dots,X_{n\Delta}^{(j)})$. This problem has a unique solution which ensures that the resulting estimator $\widehat{b}_{N,n}$ is always well defined. Moreover, this procedure offers attractive numerical properties.

4.4.3 Optimal rates of convergence

This section is dedicated to the study of the rates of convergence of the estimator $\widehat{b}_{N,n}$. We assume that $x_0 \in (0,1)$. The rate of convergence of the estimation procedure is studied over the class of Hölder functions.

Assumption 11. *For $\beta \in [1, M+1]$, and $R > 0$, the restriction $\tilde{b} := b_{|[0,1]}$ of b to $[0,1]$ belongs to the Hölder ball $\Sigma(\beta, R)$: the function $\tilde{b}$ is $l = \lfloor\beta\rfloor$ times differentiable on $(0,1)$ and its derivative $\tilde{b}^{(l)}$ satisfies*

$$\forall x, y \in (0,1), \quad \left|\tilde{b}^{(l)}(x) - \tilde{b}^{(l)}(y)\right| \le R\,|x-y|^{\beta-l}.$$

Upper bound. In order to derive optimal rate of convergence, we consider a slightly modified version of the estimator defined in Equation (4.7). The truncated estimator is defined as follow

$$\widehat{b}_{N,n}^{L_N}(x) := \begin{cases} \widehat{b}_{N,n}(x) & \text{if } |\widehat{b}_{N,n}(x)| \le \sqrt{L_N}, \\ \operatorname{sgn}(\widehat{b}_{N,n}(x))\sqrt{L_N} & \text{if } |\widehat{b}_{N,n}(x)| > \sqrt{L_N}. \end{cases}$$

We remind the reader that L_N is the multiplicative factor that controls the bound on the Euclidean norms of the parameter coefficients $\mathbf{a}$ for all functions belonging to $\mathcal{S}_{K_N,L_N,M}$ (see Equation (4.6)). First, for N large enough, since $\widetilde{b}$ is bounded, $\|\widetilde{b}\|_\infty \le \sqrt{L_N}$ which implies $\|\widetilde{b} - \widehat{b}_{N,n}^{L_N}\|_b \le \|\widetilde{b} - \widehat{b}_{N,n}\|_b$. Therefore, the consistency of $\widehat{b}_{N,n}$ implies the consistency of $\widehat{b}_{N,n}^{L_N}$. Moreover, let us notice that $\|\widehat{b}_{N,n}\|_\infty < \sqrt{(K_N+M)L_N}$ while the truncated estimator $\widehat{b}_{N,n}^{L_N}$ satisfies $\|\widehat{b}_{N,n}^{L_N}\|_\infty < \sqrt{L_N}$. This property is particularly important in Theorem 4.3 to reduce the order of the variance term with respect to K_N.

Proposition 4.3. *Grant Assumptions 7, and 11. For N large enough, the following holds*

$$\mathbb{E}\left[\left\|\widehat{b}_{N,n}^{L_N} - \tilde{b}\right\|_{N,n}^2\right] \leq C\left(\left(\frac{M+1}{K_N}\right)^{2\beta} + \frac{K_N + L_N}{N} + \Delta\right),$$

where $C > 0$ is a constant depending only on σ_1, π_0, T, M, and R.

The obtained bound is composed of three terms. The first one, which relies on the spline approximation properties, gives the order of the bias under the assumption that the function $\tilde{b}$ is Hölder. The second is the variance term which is of order $(K_N + L_N)/N$. Note that the bound on the variance term relies on the equivalence between the empirical norm and the L_2-norm over $[0,1]$. In (Denis, Dion, and Martinez, 2021), it is shown that without the ellipticity assumption the control of the variance term is of order $((K_N + L_N)/N)^{1/2}$. Finally, the last term is the error due to the discretization. Combining Proposition 4.3 with concentration arguments and lemma 4.1, we obtain the following result. Let us introduce $\mathcal{K}_N = \{1, \ldots, K_N^*\}$ with $K_N^* = \sqrt{N/\log^2(N)}$.

Theorem 4.3. *Grant Assumptions 7, 10, and 11. Let $K_N \in \mathcal{K}_N$. Assume that $L_N = \log(N)$ and $\Delta = O(1/N)$, then for N large enough the following holds*

$$\mathbb{E}\left[\|\widehat{b}_{N,n}^{L_N} - \widetilde{b}\|^2\right] \leq C\left(\left(\frac{M+1}{K_N}\right)^{2\beta} + \frac{\log^2(N) K_N}{N}\right),$$

where $C > 0$ is a constant depending only on σ_1, T, M and R.

From this result, one can see that the rate of convergence is, up to a logarithmic factor, the optimal nonparametric rate in the regression setting (Tsybakov, 2009b). Indeed, since $\beta \geq 1$, for $K_N = \left\lfloor \left(N/\log^2(N)\right)^{1/(2\beta+1)} \right\rfloor \in \mathcal{K}_N$, we obtain

$$\mathbb{E}\left[\|\widehat{b}_{N,n}^{L_N} - \widetilde{b}\|^2\right] \lesssim \left(\frac{\log^2(N)}{N}\right)^{\frac{2\beta}{2\beta+1}}. \tag{4.8}$$

This inequality shows that, regarding to the L_2-risk, the problem of estimating the drift function on a compact set based on repeated observations is equivalent to the estimation of a function in the regression setting (provided that the time step Δ is small enough). Let us comment the logarithm factors. The first $\log(N)$ is due to the fact that there is no prior knowledge on the bound of $\|\widetilde{b}\|_\infty$. The second one is due to the control of the supremum of an empirical process over the subset $\mathcal{S}_{K_N, L_N, M}$.

Lower bound. We establish a lower bound on the L_2-risk for the Hölder class of functions $\Sigma(\beta, R)$ with regularity parameter β, defined in Assumption 11.

Theorem 4.4. *Grant Assumptions 7, an 10, and 11. There exists two constants $c_1, c_0 > 0$ such that for N large enough and $\widehat{b}$ constructed from $(X^1, \dots, X^N)$,*

$$\sup_{b\,:\,\widetilde{b}\in\Sigma(\beta,R)} \mathbb{E}\left[\left\|\widehat{b}-\widetilde{b}\right\|^2\right] \geq c_1 N^{-2\beta/(2\beta+1)},$$

The proof of the Theorem follows the same lines of Theorem 2.8 in (Tsybakov, 2009b) except for the control of the Kullback-Leibler divergence. In Theorem 4.4, this control relies on the Girsanov formula. Hence, Theorem 4.4 and Equation (4.8) establish that the estimator $\widehat{b}^{L_N}_{N,n}$ is optimal in the minimax sense. However, the choice of K_N depends on the regularity β of the function $\widetilde{b}$ which is unknown in practice.

Adaptive estimator. We propose an adaptive estimator based on a penalized contrast. To alleviate the notations, the parametes K_N is denoted by K. Besides, in order to highlight the dependency on K, the estimator $\widehat{b}^{L_N}_{N,n}$ is denoted $\widehat{b}_K$ (and we choose $L_N = \log(N)$). Our adaptive procedure relies on the dyadic B-splines. That is to say, we assume that K belongs to $\mathcal{K} = \{2^p,\ p = 0, \dots, p_{\max}\}$ with $2^{p_{\max}} \leq \sqrt{N/\log^2(N)}$. Hence, this particular choice ensures that the spaces $S_{K,M,\mathbf{u}}$ are nested (for $K < K'$, $S_{K,M,\mathbf{u}} \subset S_{K',M,\mathbf{u}}$) which is an important property in light of the proof of Theorem 4.5. We define the following estimator

$$\widehat{K} = \arg\min_{K\in\mathcal{K}} \left\{\gamma_{N,n}(\widehat{b}_K) + \mathrm{pen}(K)\right\}, \tag{4.9}$$

and then consider the estimator $\widehat{b}_{\widehat{K}}$ defined as the minimizer of a penalized contrast. To penalize the complexity of $S_{K,L,M}$, we choose a penalty term $\mathrm{pen}(K) \geq 44\frac{\log^2(N)(K+M)}{N}$ for N large enough. Now, we state the following result

Theorem 4.5. *Grant Assumptions 7, 10, and 11. Assume that $L_N = \log(N)$ and $\Delta = O(1/N)$. The estimator $\widehat{b}_{\widehat{K}}$ of $\widetilde{b}$ satisfies*

$$\mathbb{E}\left[\|\widehat{b}_{\widehat{K}}-\widetilde{b}\|^2\right] \leq 2\inf_{K\in\mathcal{K}}\left\{\inf_{h\in S_{K,L,M}}\left\|h-\widetilde{b}\right\|^2_{n,b} + \mathrm{pen}(K)\right\} + \frac{C}{N},$$

where C is a positive constant depending on σ_1, π_0, T, M and R.

This result shows that the estimator $\widehat{b}_{\widehat{K}}$ achieves the bias-variance compromise over the model collection $(S_{K,L,M})_{K\in\mathcal{K}}$. In particular, whenever $\widetilde{b} \in \Sigma(\beta, R)$, with $\beta \leq M+1$, the estimator $\widehat{b}_{\widehat{K}}$ reaches the optimal rate up to a logarithmic factor. Note that the penalty term can be chosen equal to $44\frac{\log^2(N)(K+M)}{N}$. However, in practice it is better to consider $\mathrm{pen}(K) = c\frac{\log(N)^2(K+M)}{N}$ where the constant c is calibrated through numerical experiments.

4.4.4 Numerical experiments

In this Section, we briefly illustrate the performance of the proposed estimator. We choose $n\Delta = T = 1$ with $n = 100$. The sample size N is fixed to 1000. Our estimators are based on the cubic ($M = 3$) B-spline basis. We restrict our investigation to the set $\mathcal{K} = \{2^p, p = 0, 1, 2, 3, 4, 5\}$ (thus $\dim(\mathcal{S}_{K,M,\mathbf{u}}) = 2^p + 3,\ p = 0, 1, 2, 3, 4, 5$). Finally, according to our theoretical results, the constant coefficient L_N is chosen equal to $\log(N)$. After numerical investigations which are detailed in (Denis, Dion, and Martinez, 2021), we fix $c = 0.01$.

We consider the three following models to illustrate the accuracy of the estimator.

- model 1 $b(x) = 1 - x$, $\sigma(x) = 1$
- model 2 $b(x) = (1 - x^2)(-2\text{atanh}(x) - x)$, $\sigma(x) = 1 - x^2$
- model 3 $b(x) = 0.1(-\sin(2\pi x) + \cos(2\pi x) + 16\sin(3\pi x) - 5\cos(3\pi x))$, $\sigma(x) = 1$

The first model is widely used diffusion models. The model 2 possess a non constant diffusion coefficient and do not satisfy the ellipticity assumption 7. Finally, The model 3 has a multimodal drift function. It requires to explore more possible values of K (larger dimension).

We focus on the estimation of $\widetilde{b} = b\mathbb{1}_{[-1,1]}$. Note that for model 2 we have $\widetilde{b} = b$. Figure 4.2 displays ten realizations of the estimators $\widehat{b}_{\widehat{K}}$ on the three models. We can see that these estimates perform quite well. Regarding the chosen dimension, for models 1,2 the value $\widehat{K} = 1$ is mostly chosen while $\widehat{K} = 8$ is mostly selected for model 3. This is not surprising since the drift functions of model 1,2 are quite simple whereas the multimodal aspect of the drift function of model 3 requires to select larger $\widehat{K}$. Hence, for model 3 the estimation of $\tilde{b}$ is more challenging.

Note that, we also investigate in (Denis, Dion, and Martinez, 2021) the estimation of b without restriction on the estimation interval. In this case, it is natural to build our procedure on the random interval defined as $\left[\min\left((\bar{X}^1, \ldots, \bar{X}^N\right), \max\left((\bar{X}^1, \ldots, \bar{X}^N\right)\right]$. In this case, we also show that the estimation procedure has good performance.

4.5 Discussion and perspectives

In Section 4.4, we provide a new procedure to estimate the drift function of homogeneous diffusion process in the i.i.d. framework. It is the starting point of the study of a plug-in classification procedure for the mixture model defined by Equation (4.1) in the nonparametric setting. This work is a part of the Ph.D. thesis of E. Ella Mintsa whom I co-supervise with C. Dion-Blanc and V.C. Tran.

Let me briefly present the main idea of the strategy followed by E. Ella-Mintsa. Considering the result obtained in Proposition 4.2, a first step consists in the construction of

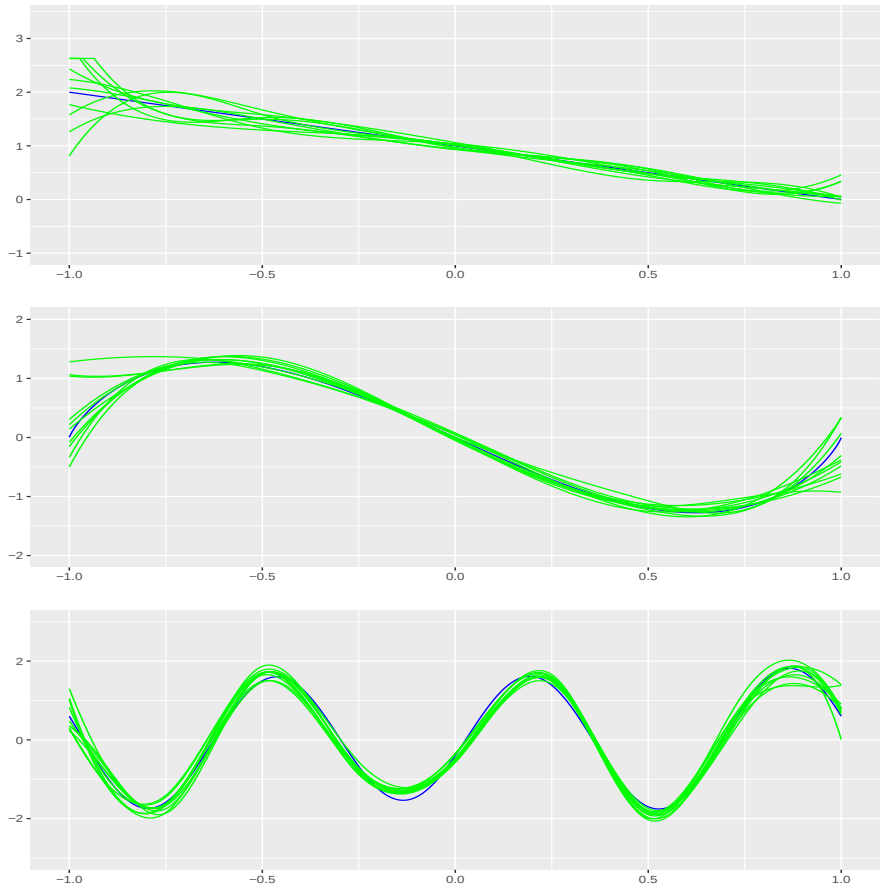

Figure 4.2: The three graphs show the three models 1-2-3 (top to bottom) and on each of them the true drift function in blue (dark) and 10 estimates $\widehat{b}_{\widehat{K}}$ in green (light grey), on the compact interval $[-1, 1]$

the ridge estimator $\hat{b}_k$ of b_k^* for each k, based on the observations of the learning sample for which the associated label is k. A first difficulty is that the obtained excess risk of the resulting plug-in classifier satisfies

$$\mathbb{E}\left[\mathcal{E}(\bar{g}_{\hat{b}})\right] \leq C\left(\sqrt{\Delta} + \sum_{k=1}^{K} \mathbb{E}_Y\left[\|\hat{b}_k - b_k^*\|_{n,b_Y^*}\right]\right).$$

Therefore, the theoretical properties of the ridge estimator presented in Section 4.4 are not sufficiently sharp to ensure that for each k

$$\sum_{k=1}^{K} \mathbb{E}\left[\|\hat{b}_k - b_k^*\|_{n,b_Y^*}\right] \to 0.$$

Indeed, when the s.d.e. is driven by b_j^* with $j \neq k$, the estimator $\hat{b}_k$ may be not consistent. Furthermore, our result holds only for the estimation of the drift function over a compact interval. To bypass these difficulties, the strategy is to obtained a similar result as Lemma 4.1 for a compact interval $[-A_N, A_N]$, with $A_N \to +\infty$. Importantly, the dependency of the constants π_0 and π_1 *w.r.t.* A_N should be carefully evaluated. In this case, for each $t \in [0,T]$, $\mathbb{P}\left(|X_t| \geq A_N\right)$ is easily controlled by the Markov Inequality, and then one can obtain the following bound on the excess risk

$$\mathbb{E}\left[\mathcal{E}(\bar{g}_{\hat{b}})\right] \leq C\sqrt{\Delta} + R_{A_N} + C_{A_N} \sum_{k=1}^{K} \mathbb{E}\left[\|\hat{b}_k - b_k^*\|_{[-A_N,A_N]}\right],$$

where $R_{A_N} \to 0$ and C_{A_N} is a positive constant depending on π_0 and π_1. Hence, the control of C_{A_N} will allow to derive the consistency of the plug-in classifier.

The other line of the extension of the results provided in this chapter is to consider the case where the distribution of Y and the coefficient diffusion σ are unknown. While the estimation of the weights of the mixture is not really a difficult task as it can be easily handled by considering the empirical distribution of Y, the estimation of σ is more intricate. To tackle this problem, E. Ella Mintsa proposes a version of the estimator described in Section 4.4 dedicated to the estimation of σ.

Another extension is to investigate a classification procedure based on the empirical risk minimization principle. In this case, we can consider a similar procedure as the one described in Section 4.3.1 where the minimization is performed over the set $\mathcal{S}_{K_N,L_N,M}$ defined in Equation (4.6).

An important part of the study of these procedures will be also to evaluate their numerical performance. In particular, a comparison with other classification procedures for functional data such as depth classification or recurrent neural networks will highlight the relevance of the proposed approach.

Lastly, the generalization of the initial model is an important guideline for further research. The extension to the case of inhomogeneous diffusions as well as considering the case of multidimensional diffusions will cover a broader class of possible applications.

References

Stone, C. (1977). "Consistent nonparametric regression". *Ann. Statist.*, pp. 595–620 (pp. 34, 38).

— (1982). "Optimal global rates of convergence for nonparametric regression". *Ann. Statist.*, pp. 1040–1053 (p. 39).

Tsybakov, A. (1986). "Robust reconstruction of functions by the local-approximation method". *Problemy Peredachi Informatsii* 22.2, pp. 69–84 (pp. 34, 38).

van de Geer, S. (1990). "Estimating a Regression Function". *The Annals of Statistics* 18.2, pp. 907 –924 (p. 15).

Yoshida, Nakahiro (1992). "Estimation for diffusion processes from discrete observation". *Journal of Multivariate Analysis* 41.2, pp. 220 –242 (p. 60).

Bibby, Bo Martin and Sørensen, Michael (1995). "Martingale estimation functions for discretely observed diffusion processes". *Bernoulli*, pp. 17–39 (p. 60).

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. Vol. 31. Applications of Mathematics (New York). New York: Springer-Verlag (pp. 13, 15–17).

Van Der Vaart, Aad W and Wellner, Jon A (1996). "Weak convergence". *Weak convergence and empirical processes*. Springer, pp. 16–28 (p. 56).

Wegkamp, M. and van de Geer, S. (1996). "Consistency for the least squares estimator in nonparametric regression". *The Annals of Statistics* 24.6, pp. 2513 –2523 (p. 15).

Freund, Y. and Schapire, R. E. (1997). "A decision-theoretic generalization of on-line learning and an application to boosting". *Journal of computer and system sciences* 55.1, pp. 119–139 (p. 17).

LeCun, Yann, Bottou, Léon, Bengio, Yoshua, and Haffner, Patrick (1998). "Gradient-based learning applied to document recognition". *Proceedings of the IEEE* 86.11, pp. 2278–2324 (p. 28).

Ma, C. and Robinson, J. (1998). "17 Approximations to distributions of sample quantiles". *Handbook of Statistics* 16, pp. 463–484 (p. 40).

Vapnik, V. N. (1998). *Statistical learning theory*. Wiley (p. 17).

Wightman, L. F. and Ramsey, H. (1998). *LSAC national longitudinal bar passage study*. Law School Admission Council (p. 48).

Hoffmann, M (1999). "Adaptive estimation in diffusion processes". *Stochastic processes and their Applications* 79.1, pp. 135–163 (p. 60).

Kessler, Mathieu, Sørensen, Michael, et al. (1999). "Estimating equations based on eigenfunctions for a discretely observed diffusion process". *Bernoulli* 5.2, pp. 299–314 (p. 60).

Mammen, E. and Tsybakov, A. B. (1999). "Smooth discrimination analysis". *Ann. Statist.* 27.6, pp. 1808–1829 (pp. 32, 34).

Yang, Y. (1999). "Minimax nonparametric classification: Rates of convergence". *IEEE Transactions on Information Theory* 45.7, pp. 2271–2284 (pp. 16, 34, 56).

Friedman, J., Hastie, T. J., and R. Tibshirani, Robert (2000). "Additive logistic regression: a statistical view of boosting". *Ann. Statist.* 28.2, pp. 337–407 (p. 17).

van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press (p. 17).

Breiman, L. (2001). "Random Forests". *Mach. Learn.* 45.1, pp. 5–32 (p. 28).

Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc. (p. 63).

Gobet, Emmanuel (2002). "LAN property for ergodic diffusions with discrete observations". *Annales de l'Institut Henri Poincare (B) Probability and Statistics* 38.5, pp. 711–737 (p. 60).

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer Ser. Statist. New York: Springer-Verlag (p. 34).

Vovk, V. (2002a). "Asymptotic optimality of transductive confidence machine". *Algorithmic learning theory*. Vol. 2533. Lecture Notes in Comput. Sci. Berlin: Springer, pp. 336–350 (p. 23).

— (2002b). "On-line confidence machines are well-calibrated". *Proceedings of the Forty-Third Annual Symposium on Foundations of Computer Science*. Los Alamitos: CA. IEEE Computer Society, pp. 187–196 (p. 23).

Gobet, Emmanuel, Hoffmann, Marc, Reiß, Markus, et al. (2004). "Nonparametric estimation of scalar diffusions based on low frequency data". *The Annals of Statistics* 32.5, pp. 2223–2253 (p. 60).

Kutoyants, Y (2004). *Statistical Inference for Ergodic Diffusion Processes*. Springer, London (p. 60).

Zhang, T. (Feb. 2004). "Statistical behavior and consistency of classification methods based on convex risk minimization". *Ann. Statist.* 32.1, pp. 56–85 (pp. 17, 18, 31, 33).

Dalalyan, Arnak et al. (2005). "Sharp adaptive estimation of the drift function for ergodic diffusions". *The Annals of Statistics* 33.6, pp. 2507–2528 (p. 60).

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. New York: Springer (p. 23).

Bartlett, P., Jordan, M., and McAuliffe, J. (2006). "Convexity, classification, and risk bounds". *J. Amer. Statist. Assoc.* 101.473, pp. 138–156 (pp. 17, 31).

Györfi, László, Kohler, Michael, Krzyzak, Adam, and Walk, Harro (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media (p. 61).

Massart, P. and Nédélec, É (Oct. 2006). "Risk bounds for statistical learning". *Ann. Statist.* 34.5, pp. 2326–2366 (pp. 17, 34).

Audibert, J-Y. and Tsybakov, A. B. (2007). "Fast learning rates for plug-in classifiers". *Ann. Statist.* 35.2, pp. 608–633 (pp. 16, 34–38, 40).

Comte, F., Genon-Catalot, V., and Rozenholc, Y (May 2007). "Penalized nonparametric mean square estimation of the coefficients of diffusion processes". *Bernoulli* 13.2, pp. 514–543 (p. 60).

Ramsay, James O and Silverman, Bernard W (2007). *Applied functional data analysis: methods and case studies*. Springer (p. 20).

Tewari, A. and Bartlett, P. (2007). "On the consistency of multiclass classification methods". *Journal of Machine Learning Research* 8, pp. 1007–1025 (p. 17).

Tsybakov, A. (2008). *Introduction to Nonparametric Estimation*. Springer Ser. Statist. Springer New York (p. 39).

Calders, T., Kamiran, F., and Pechenizkiy, M. (2009). "Building classifiers with independency constraints". *IEEE international conference on Data mining* (pp. 41, 42).

Cortez, Paulo, Cerdeira, António, Almeida, Fernando, Matos, Telmo, and Reis, José (2009). "Modeling wine preferences by data mining from physicochemical properties". *Decision Support Systems* 47.4, pp. 547–553 (p. 48).

Rigollet, P. and Vert, R (Nov. 2009). "Optimal rates for plug-in estimators of density level sets". *Bernoulli* 4, pp. 1154–1178 (p. 37).

Tsybakov, A. B. (2009a). *Introduction to nonparametric estimation*. Springer Series in Statistics. New York: Springer (pp. 15, 16, 36).

Tsybakov, A.B. (2009b). *Introduction to nonparametric estimation*. Springer Series in Satistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats (pp. 64, 65).

Yuan, M. and Wegkamp, M. (2010). "Classification methods with reject option based on convex risk minimization". *J. Mach. Learn. Res.* 11, pp. 111–130 (pp. 17, 31).

Cadre, B. (2013). "Supervised classification of diffusion paths". *Mathematical Methods of Statistics* 22.3, pp. 213–225 (pp. 51, 53, 55, 56).

Calders, T., Karim, A., Kamiran, F., Ali, W., and Zhang, X. (2013). "Controlling attribute effect in linear regression". *IEEE International Conference on Data Mining* (p. 47).

Meulen, F. van der and Van Zanten, H. (2013). "Consistent nonparametric Bayesian inference for discretely observed scalar diffusions." *Bernoulli* 19, pp. 44–63 (p. 60).

Revuz, Daniel and Yor, Marc (2013). *Continuous martingales and Brownian motion*. Vol. 293. Springer Science & Business Media (p. 53).

Schmisser, E. (2013). "Penalized nonparametric drift estimation for a multidimensional diffusion process". *Statistics* 47.1, pp. 61–84 (p. 60).

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). "Learning fair representations". *International Conference on Machine Learning* (p. 41).

Gugushvili, S and Spreij, P. (2014). "Consistent non-parametric Bayesian estimation for a time-inhomogeneous Brownian motion." *ESAIM: Probability and Statistics* 18, pp. 332–341 (p. 60).

Lei, J. (2014). "Classification with confidence". *Biometrika* 101.4, pp. 755–769 (p. 38).

Göeau, H., Joly, A., and Bonnet, P. (2015). "LifeCLEF plant identification task 2015". *CLEF working notes* (p. 8).

Lapin, Maksim, Hein, Matthias, and Schiele, Bernt (2015). "Top-k multiclass SVM". *Advances in Neural Information Processing Systems*, pp. 325–333 (pp. 21, 22).

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, Andrej, Khosla, Aditya, Bernstein, Michael, et al. (2015). "Imagenet large scale visual recognition challenge". *International journal of computer vision* 115.3, pp. 211–252 (p. 22).

Bobkov, S. and Ledoux, M. (2016). "One-dimensional empirical measures, order statistics and Kantorovich transport distances". *Memoirs of the American Mathematical Society* (pp. 39, 40).

Gadat, S., Klein, T., and Marteau, C. (2016). "Classification in general finite dimensional spaces with the k-nearest neighbor rule". *The Annals of Statistics* 44.3, pp. 982–1009 (p. 36).

Hardt, M., Price, E., and Srebro, N. (2016). "Equality of opportunity in supervised learning". *Neural Information Processing Systems* (p. 42).

Jain, H., Prabhu, Y., and Varma, M. (2016). "Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications". *KDD*, pp. 935–944 (p. 40).

Lum, K. and Johndrow, J. (2016). "A statistical framework for fair predictive algorithms". *arXiv preprint arXiv:1610.08077* (p. 41).

Chzhen, E., Denis, C., Hebiri, M., and Salmon, J (2017). "On the benefits of output sparsity for multi-label classification". preprint (p. 40).

Denis, C. and Hebiri, M. (2017). "Confidence sets with expected sizes for Multiclass Classification". *Journal of Machine Learning Research* 18.1, pp. 3571–3598 (pp. 21, 23, 26, 30, 31, 33, 40).

Oh, S. (2017). "Top-k hierarchical classification". *AAAI Conference on Artificial Intelligence* (p. 22).

Xie, Saining, Girshick, Ross, Dollár, Piotr, Tu, Zhuowen, and He, Kaiming (2017). "Aggregated residual transformations for deep neural networks". *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500 (p. 19).

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017). "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment". *International Conference on World Wide Web* (p. 41).

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). "A reductions approach to fair classification". *arXiv preprint arXiv:1803.02453* (p. 41).

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. (2018). "Empirical risk minimization under fairness constraints". *Neural Information Processing Systems* (p. 41).

Kpotufe, S. and Martinet, G. (2018). "Marginal Singularity, and the Benefits of Labels in Covariate-Shift". *Conference On Learning Theory*, pp. 1882–1886 (p. 36).

Sadinle, M., Lei, J., and Wasserman, L. (2018). "Least ambiguous set-valued classifiers with bounded error levels". *Journal of the American Statistical Association*, pp. 1–12 (pp. 23, 38).

Agarwal, A., Dudik, M., and Wu, Z. S. (2019). "Fair Regression: Quantitative Definitions and Reduction-Based Algorithms". *International Conference on Machine Learning* (pp. 41, 43).

Barocas, Solon, Hardt, Moritz, and Narayanan, Arvind (2019). *Fairness and Machine Learning*. http://www.fairmlbook.org. fairmlbook.org (pp. 41, 42).

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. (2019). "Leveraging Labeled and Unlabeled Data for Consistent Fair Binary Classification". *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., pp. 12739–12750 (pp. 41, 44).

Gordaliza, P., Del Barrio, E., Fabrice, G., and Loubes, J. M. (2019). "Obtaining fairness using optimal transport theory". *International Conference on Machine Learning* (p. 43).

Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. (2019). "Wasserstein fair classification". *arXiv preprint arXiv:1907.12059* (p. 43).

Koskela, J., Spano, D., and Jenkins, P.A. (2019). "Consistency of Bayesian nonparametric inference for discretely observed jump diffusions." *Bernoulli* 25, pp. 2183–2205 (p. 60).

Mehrabi, Ninareh, Morstatter, Fred, Saxena, Nripsuta, Lerman, Kristina, and Galstyan, Aram (2019). "A survey on bias and fairness in machine learning". *arXiv preprint arXiv:1908.09635* (p. 41).

Oneto, L., Donini, M., and Pontil, M. (2019). "General fair empirical risk minimization". *arXiv preprint arXiv:1901.10080* (p. 43).

Barrio, Eustasio del, Gordaliza, Paula, and Loubes, Jean-Michel (2020). "Review of Mathematical frameworks for Fairness in Machine Learning". *arXiv preprint arXiv:2005.13755* (p. 41).

Chiappa, S., Jiang, R., Stepleton, T., Pacchiano, A., Jiang, H., and Aslanides, J. (2020). "A general approach to fairness with optimal transport". *AAAI* (pp. 41, 43).

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. (2020a). "Fair Regression via Plug-in Estimator and RecalibrationWith Statistical Guarantees". *Advances in Neural Information Processing Systems* (pp. 21, 44).

— (2020b). "Fair Regression with Wasserstein Barycenters". *Advances in Neural Information Processing Systems* (pp. 21, 26, 47).

— (2020c). "Fair Regression with Wasserstein Barycenters". *arXiv preprint arXiv:2006.07286* (p. 47).

Comte, F. and Genon-Catalot, V. (2020). "Nonparametric drift estimation for i.i.d. paths of stochastic differential equations". *The Annals of Statistics* 48.6, pp. 3336 –3365 (pp. 52, 60).

Denis, C. and Hebiri, M. (2020). "Consistency of plug-in confidence sets for classification in semi-supervised learning". *Journal of Nonparametric Statistics* 32.1, pp. 42–72 (p. 21).

Denis, Christophe, Dion, Charlotte, and Martinez, Miguel (2020). "Consistent procedures for multiclass classification of discrete diffusion paths". *Scandinavian Journal of Statistics* 47, pp. 516–554 (pp. 51, 52, 55, 58, 59).

Denis, Christophe, Hebiri, Mohamed, and Zaoui, Ahmed (2020). "Regression with reject option and application to kNN". *NeurIPS 2020 - 34th Conference on Neural Information Processing Systems* (p. 26).

Gadat, Sébastien, Gerchinovitz, Sébastien, and Marteau, Clément (2020). "Optimal functional supervised classification with separation condition". *Bernoulli* (p. 51).

Gouic, Thibaut Le, Loubes, Jean-Michel, and Rigollet, Philippe (2020). "Projection to Fairness in Statistical Learning". *arXiv preprint arXiv:2005.11720* (p. 47).

Gyöfi, L. and Walk, H. (2020). "Nearest neighbor based conformal prediction". *Submitted* (p. 23).

Oneto, Luca and Chiappa, Silvia (2020). "Fairness in Machine Learning". *Recent Trends in Learning From Data*. Springer, pp. 155–196 (p. 41).

Romano, Y., Sesia, M, and Candès, E.J (2020). "Classification with Valid and Adaptive Coverage". *arXiv preprint arXiv:2006.02544* (p. 23).

Ye, Q. and Xie, W. (2020). "Unbiased Subdata Selection for Fair Classification: A Unified Framework and Scalable Algorithms". *arXiv preprint arXiv:2012.12356* (pp. 41, 47, 49).

Chzhen, E., Denis, C., and Hebiri, M. (2021). "Minimax semi-supervised set-valued approach to multi-class classification". *Bernoulli* 27.4, pp. 2389 –2412 (pp. 21, 23, 26, 35, 38, 40).

Chzhen, E., Denis, C., Hebiri, M., and Lorieul, T. (2021). "Set-valued classification – overview via a unified framework". preprint (p. 23).

Della Maestra, L. and Hoffmann, M. (2021). "Nonparametric estimation for interacting particle systems : McKean-Vlasov models". *Probability Theory and Related Fields* (p. 60).

Denis, C., Elie, R., Hebiri, M., and Hue, F. (2021). "Fairness guarantee in multi-class classification". preprint (pp. 26, 41, 45, 47).

Denis, Christophe, Dion, Charlotte, and Martinez, Miguel (2021). "A ridge estimator of the drift from discrete repeated observations of the solution of a stochastic differential equation." *Bernoulli* 27, pp. 2675–2713 (pp. 52, 60, 64, 66).